UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Identificação de discursos de ódio em Redes Sociais

Caroline Porfirio Rodrigues

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

Caroline Porfirio Rodrigues

Identificação de discursos de ódio em Redes Sociais

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Diego Raphael Amancio

Versão original

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Rodrigues, Caroline Porfirio
S856m Identificação de discursos de ódio em Redes Sociais / Caroline Porfirio Rodrigues ; orientador Diego Raphael Amancio. – São Carlos, 2023.
61 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Amancio, Diego Raphael , orient. II. Título.

Caroline Porfirio Rodrigues

Identificação de discursos de ódio em Redes Sociais

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Diego Raphael Amancio

Original version

São Carlos

2023

AGRADECIMENTOS

Neste momento significativo, gostaria de expressar minha mais profunda gratidão a Deus, minha família e ao corpo docente do MBA em Inteligencia Artificial e Big Data.

Primeiramente, agradeço a Deus por me abençoar e capacitar ao longo deste desafiador caminho acadêmico. Sua graça tem sido minha âncora nos momentos difíceis e minha luz nos momentos de dúvida.

À minha família, meu alicerce e maior fonte de apoio, quero expressar minha sincera gratidão. Seu amor, incentivo constante e compreensão foram os pilares que sustentaram minha jornada.

Aos dedicados membros do corpo docente do MBA, estendo minha gratidão pela excelência em educação e orientação. Seus ensinamentos e conhecimentos enriqueceram meu aprendizado e continuarão a influenciar minha carreira.

Este trabalho é uma conquista compartilhada e uma representação de nossos esforços coletivos. Agradeço profundamente a todos que desempenharam um papel importante em minha jornada acadêmica e pessoal.

Com gratidão,

RESUMO

Rodrigues, C. P. **Identificação de discursos de ódio em Redes Sociais**. 2023. 61p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

A crescente ubiquidade das redes sociais transformou-as em um meio poderoso de expressão, comunicação e compartilhamento de informações em todo o mundo. No entanto, essa disseminação maciça de conteúdo digital também trouxe consigo um desafio crítico: a proliferação de discursos de ódio e conteúdo prejudicial nas plataformas de mídia social. Esses discursos não apenas corroem o ambiente online, mas também têm o potencial de incitar violência, promover a discriminação e prejudicar comunidades inteiras. Portanto, a detecção eficaz de discursos de ódio em redes sociais tornou-se uma questão imperativa. Este projeto de pesquisa se dedica à detecção de discursos de ódio em tweets em inglês coletados da plataforma Twitter. Foram abordadas técnicas de Processamento de Linguagem Natural (PLN), que inclui normalização de texto, tokenização e vetorização. Também foi implementado o balanceamento de dados usando a técnica Synthetic Minority Over-sampling Technique (SMOTE), gerando amostras sintéticas da classe alvo. A metodologia inclui a aplicação e análise comparativa de três algoritmos de aprendizado de máquina: Support Vector Machine (SVM), Regressão Logística (LR) e Naive Bayes (NB). Essa análise foi conduzida em dois cenários distintos: um com amostras sintéticas (Grupo A) e outro apenas com amostras originais (Grupo B). Dado que o objetivo principal deste estudo é a identificação precisa de discursos de ódio, o recall é uma métrica de avaliação de relevância significativa para os modelos. Os resultados indicaram que o balanceamento de dados no Grupo A resultou em melhorias substanciais no recall para todos os algoritmos, tornando-os mais eficazes na identificação de discursos de ódio. Por outro lado, os modelos no Grupo B demonstraram altas taxas de precisão, mas com recall notavelmente inferior, sugerindo uma tendência a classificar erroneamente muitos exemplos de discursos de ódio como conteúdo não ofensivo. O algoritmo Naive Bayes registrou resultados superiores em termos de recall, com uma pontuação de 0.88 e uma precisão de 0.4, mas destaca-se também o modelo de Regressão Logística com um recall de 0.6, precisão de 0.68 e F1-score de 0.64, apresentando uma performance mais equilibrada na tarefa de classificação. Este estudo oferece insights para a compreensão da eficácia das abordagens de classificação na detecção de discursos de ódio em mídias sociais e ressalta a importância do balanceamento de dados para aprimorar o desempenho desses modelos.

Palavras-chave: Detecção de discurso de ódio, Processamento de Linguagem Natural, Balanceamento de dados, Aprendizado de Máquina, Regressão Logística, SVM, Naive-Bayes.

ABSTRACT

Rodrigues, C. P. **Identification of hate speech on Social Media**. 2023. 61p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

The increasing ubiquity of social media has transformed them into a powerful means of expression, communication, and information sharing worldwide. However, this massive dissemination of digital content has also brought a critical challenge: the proliferation of hate speech and harmful content on social media platforms. These speeches not only erode the online environment but also have the potential to incite violence, promote discrimination, and harm entire communities. Therefore, the effective detection of hate speech on social networks has become an imperative issue. This research project is dedicated to the detection of hate speech in English tweets collected from the Twitter platform. Natural Language Processing (NLP) techniques were employed, including text normalization, tokenization, and vectorization. Data balancing was also implemented using the Synthetic Minority Over-sampling Technique (SMOTE), generating synthetic samples of the target class. The methodology involves the application and comparative analysis of three machine learning algorithms: Support Vector Machine (SVM), Logistic Regression (LR), and Naive Bayes (NB). This analysis was conducted in two distinct scenarios: one with synthetic samples (Group A) and another with only original samples (Group B). Given that the main objective of this study is the accurate identification of hate speech, recall is a metric of significant relevance for the models. The results indicated that data balancing in Group A resulted in substantial improvements in recall for all algorithms, making them more effective in identifying hate speech. On the other hand, models in Group B demonstrated high accuracy rates but with notably lower recall, suggesting a tendency to misclassify many instances of hate speech as non-offensive content. The Naive Bayes algorithm recorded superior results in terms of recall, with a score of 0.88 and a precision of 0.4. Additionally, the Logistic Regression model stood out with a recall of 0.6, precision of 0.68, and an F1-score of 0.64, presenting a more balanced performance in the classification task. This study provides insights into the effectiveness of classification approaches in detecting hate speech on social media and underscores the importance of data balancing to enhance the performance of these models.

Keywords: Hate speech detection, Natural Language Processing, Data balancing, Machine Learning, Logistic Regression, SVM, Naive-Bayes.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

USP	Universidade de São Paulo
USPSC	Campus USP de São Carlos
AM	Aprendizado de Máquinas
NPL	Natural Procesing language
PLN	Processamento de linguagem natural
SVM	Support Vector Machines
LR	Logistic Regression
NB	Naive Bayes
SMOTE	Synthetic Minority Over-sampling Technique
NLTK	Natural Language Toolkit
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
AUC	Area Under the Receiver Operating Characteristic Curve
ELI5	Explain Like I'm 5
LIME	Local Interpretable Model-Agnostic Explanations

SUMÁRIO

1 INTRODUÇÃO

As redes sociais vêm desempenhando um papel importante na vida pessoal e profissional das pessoas, ganhando um espaço cada vez maior no dia a dia da população. Entre a variedade de facilidades que as redes sociais proporcionam estão a conexão com outras pessoas ao redor do mundo, o compartilhamento de informações, o entretenimento e a comercialização de marcas e produtos.

A facilidade de acesso a essas informações compartilhadas e o contato com grupos e pessoas com interesses comuns, ainda que possa ser benéfica, pode também facilitar e instigar a propagação de notícias falsas e discursos de ódio contra indivíduos. A Organização das Nações Unidas (ONU) define o discurso de ódio como "qualquer tipo de comunicação, seja oral ou escrita, —ou também comportamental—, que ataque ou use linguagem pejorativa ou discriminatória em referência a uma pessoa ou grupo com base no que são, ou seja, com base em sua religião, etnia, nacionalidade origem, raça, cor, ascendência, gênero ou outras formas de identidade".

A propagação de discursos de ódio nas redes sociais pode ocorrer de diversas maneiras, que incluem postagens pessoais, compartilhamento direto de imagens e conteúdos contendo manifestações de ódio, comentários e respostas em postagens de outros usuários, o que pode levar também a organizações de grupos de ódio, que se coordenam para promover discursos de preconceito e intolerância. Algoritmos de recomendação também podem contribuir para a disseminação de discursos de ódio, sugerindo conteúdos similares a usuários que já demonstram interesse por esse tema. A intensificação da disseminação de discursos de ódio pode desencadear consequências ainda mais graves, levando a situações em que indivíduos cometem crimes de ódio, que são atos criminosos motivados por preconceito ou intolerância direcionados a grupos específicos de pessoas. Os crimes de ódio são caracterizados por dois elementos essenciais: a prática de uma infração criminal e a presença de uma motivação preconceituosa. Segundo a *Organization for Security and Cooperation in Europe (OSCE)*, um crime de ódio ocorre quando um perpetrador intencionalmente direciona suas ações contra um indivíduo ou propriedade devido a um ou mais traços identitários, expressando hostilidade em relação a esses aspectos identitários durante a execução do crime.

Discursos de ódio também podem ser propagados de forma velada por meio de mensagens que usam palavras ou frases que aparentemente são inocentes, mas que carregam conotações negativas dirigidas a uma pessoa ou grupo específico. Alguns usuários, com objetivo de driblar algoritmos usados para identificação de este tipo de conteúdo nas redes sociais, substituem letras por símbolos em palavras usadas para discursos de ódio, ou substituem palavras que fazem referência a grupos específicos para postar conteúdos de

preconceito dirigidos a esses grupos em forma de códigos. Magu, Joshi and Luo (2017) apresentaram alguns dos códigos mais usados identificados na rede social twiter para fazer referência a comunidades em postagens de ódio, no idioma inglês.

Code word	Actual word
Google	Black
Yahoo	Mexican
Skype	Jew
Bing	Chinese
Skittle	Muslim
Butterfly	Gay

Figura 1 – Palavras-códigos comumente usadas em referência a comunidades em Redes Sociais

A propagação deste tipo de conteúdo pode ser facilitada por algumas características das redes sociais, como por exemplo a possibilidade de realizar postagens anônimas ou com usuários falsos, dificultando o rastreio e a identificação direta de usuários que propagam conteúdos de ódio. O uso de trolls e bots também pode contribuir para disseminar discursos de ódio e desinformação nas redes sociais, criado um ambiente hostil para os usuários e vítimas desse preconceito. A propagação de discursos de ódio pode ocorrer em qualquer rede social. O maior ou menor uso de uma rede em específico para propagar esse tipo de conteúdo pode ser influenciada por diversos fatores, como por exemplo cultura, quantidade de usuários e as normas e políticas de moderação de conteúdo de cada rede social.

Um estudo feito por Silva and Roman (2021) mostrou que, entre as principais redes sociais usadas atualmente, estudos de identificação de discursos de ódio se concentram principalmente em conteúdos extraídos das redes Twitter e Facebook, já que nessas redes as postagens são realizadas principalmente por meio de textos, enquanto que em outras redes como Instagram e Youtube, as postagens são acompanhadas de imagens e/ou vídeos, o que poderia dificultar e complicar as análises de conteúdo, uma vez que seria necessário combinar e processar esses dados com os textos que os acompanham.

1.1 Justificativa e Motivação

Os discursos de ódio são discursos que expressam intolerância, preconceito e hostilidade a certos grupos de pessoas. Tais discursos podem levar a discriminação, violência e exclusão desses grupos, podendo ter um impacto significativo na saúde mental e na segurança de usuários online e offline, bem como prejudicar a coesão social e ferir os direitos humanos.

Por essa razão, a detecção e o combate aos discursos de ódio é uma tarefa funda-

mental para garantir a segurança e proteger a dignidade e os direitos humanos de todos os indivíduos, independentemente de suas características e crenças pessoais, em ambientes virtuais.

A utilização de técnicas de Processamento de Linguagem Natural (PLN) é uma ferramenta apropriada para a detecção de discursos de ódio em larga escala, uma vez que permite a análise de grandes volumes de textos e a identificação de padrões que podem indicar a presença de conteúdos de preconceito e intolerância, aplicando algoritmos de aprendizado de máquinas e técnicas de mineração de dados para a identificação de palavras e frases que são frequentemente associadas a discursos de ódio. A identificação de discursos de ódio usando PLN também poderá contribuir para a prevenção da propagação de violência e preconceitos contra grupos específicos, permitindo o reconhecimento de conteúdos ofensivos e a aplicação de medidas corretivas adequadas, que podem incluir a remoção do conteúdo, o bloqueio de usuários e a identificação de grupos e comunidades online que propagam esse tipo de discurso.

Dessa forma, este projeto poderá impactar significativamente na identificação e remoção de conteúdos com teor negativo, contribuindo para a promoção de um ambiente virtual mais sano e respeitoso.

1.2 Objetivos

Este projeto de pesquisa tem como objetivo explorar as técnicas de PLN mais recentes para detectar discursos de ódio em postagens do tipo tweets, extraídos da rede social Twitter, desenvolvendo um sistema automatizado que possa analisar grandes volumes de textos em ambientes virtuais e identificar conteúdos que são classificados como discursos de ódio, ou seja, que promovam algum tipo de intolerância, preconceito ou hostilidade em relação a determinado grupo social com base em sua raça, gênero, crença, orientação sexual ou outro.

Para alcançar estes objetivos, serão usados técnicas de PLN, como:

1. Pré-processamento de texto: responsável por preparar os textos para análise, incluindo a remoção de pontuação, a lematização e a tokenização.

2. Algoritmos de classificação: responsáveis por classificar o texto em categorias, como "discurso de ódio"ou "conteúdo seguro". Alguns exemplos de algoritmos de classificação incluem SVM (Support Vector Machines), Regressão Logística e Naive Bayes.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Mineração de dados

A quantidade de dados gerados atualmente é imensa, o que incluiu os dados gerados por redes sociais, como por exemplo o Twitter e Facebook. Em média 350.000 *tweets* são feitos por minuto, correspondendo a aproximadamente 500 milhões de tweets por dia. Toda essa quantidade de dados gerados deve ser analisada e processada para a extração de padrões, transformando-se assim em informações que podem ser usadas para a toma de decisões.

Um dos processos para manipular e analisar essa grande quantidade de dados é a chamada mineração de dados. Segundo Kaufman and Rousseeuw (2009) a mineração de dados "(...) é definido como o processo de descobrimento de padrões nos dados. O processo deve ser automático ou (mais comumente) semiautomático. Os padrões descobertos devem ser significantes de modo que levem a alguma vantagem, geralmente uma vantagem econômica". Dessa forma, a aplicação de técnicas de mineração de dados pode auxiliar na transformação dessa grande quantidade de dados em informações e conhecimentos úteis para a resolução de problemas e tomada de decisões.

Os dados usados nos processos de mineração de dados podem ser classificados como estruturados, semi-estruturados ou não estruturados. Dados estruturados são aqueles que seguem formatos e esquemas específicos e definidos, podendo ser armazenados em tabelas ou banco de dados relacionais. Dados semi-estruturados possuem alguma estrutura, porém não seguem um padrão definido. Dados não estruturados são dados que não possuem uma estrutura ou um padrão bem definido, podendo ser compostos por diferentes elementos dentro de um todo. Exemplos de dados não estruturados incluem e-mails, áudios, imagens, vídeos, posts de redes sociais, entre outros.

Devido à sua estrutura definida, os dados estruturados são mais facilmente analisados através da aplicação de processos de mineração de dados. Por outro lado, a análise de dados não estruturados torna-se mais complexa devido à falta de padronização em seu conteúdo. A mineração de texto pode ser vista como uma subárea da mineração de dados, diferenciando-se principalmente pelo tipo de dados manipulados (não estruturados versus estruturados) e pelas etapas de pré-processamento. Estas etapas visam aplicar métodos para obter uma representação estruturada dos textos, facilitando, assim, a análise e extração de informações significativas a partir de dados textuais (Sinoara, Marcacini and Rezende (2021)). Por esse motivo, para a análise de dados não estruturados se faz necessário a aplicação de técnicas especificas de mineração de texto , como por exemplo as tecnologias de Processamento de Linguagem Natural (PLN) e técnicas de Aprendizado de

Maquinas (AM).

2.2 Mineração de textos

Para Feldman (2013), a mineração de textos pode ser definida como um método de extração de informações a partir de banco de dados não estruturados ou semi-estruturados. Dessa forma, a aplicação de técnicas de mineração de textos para se analisar grandes volumes de dados provenientes de redes sociais pode ser indicado.

Ainda segundo Feldman (2013), a análise de textos pode ocorrer em três diferentes níveis, sendo a nível de documento, que consiste em classificar um documento como um todo de acordo ao tipo de sentimento que ele expressa (por exemplo uma classificação em sentimentos positivos, negativos ou neutros), a nível de sentença, analisando uma sentença específica de um documento e a nível entidade ou aspecto, que analisa todas as expressões presentes em um documento.

Morais and Ambrósio (2007) descrevem as etapas de mineração de texto como sendo: seleção do documento, definição do tipo de abordagem dos dados, preparação dos dados, indexação e normalização, cálculo da relevância dos termos, seleção dos termos e pós-processamento. Logo, se deve definir o tipo de abordagem que será utilizado para a análise dos dados textuais, podendo esta ser do tipo semântica, baseada na funcionalidade dos termos encontrados no texto ou do tipo estatística, baseada na frequência dos termos. Ambas as técnicas podem ser usadas separadamente ou combinadas.

2.3 Pré-processamento

Para Rezende, Marcacini and Moura (2011), conjuntos de documentos de textos são representados por grandes números de atributos, o que pode dificultar a caracterização do texto em análise. Os textos podem conter um ou mais atributos irrelevantes para o processo de classificação.

O pré-processamento de textos é uma etapa importante no PLN que tem como objetivo preparar o texto para análise, removendo informações irrelevantes e transformando o dado a um formato mais estruturado para seu posterior processamento. Alguns passos do pré-processamento incluem:

1. Remoção de pontuações e caracteres especiais, como por exemplo “%”,”&”,”$”.

2. Remoção das chamadas *stopwords*, que são palavras que conectam as demais palavras de uma sentença e não aportam com informação relevante para a análise.

3. Normalização de palavras, convertendo palavras de diferentes formas ou ortografias em uma forma padronizada, por exemplo transformando palavras no plural em

seu formato singular (como exemplo, as palavras árvore e árvores serão todas padronizadas para o formato árvore).

4. *Stemming*, processo de reduzir as palavras as suas raízes, removendo prefixos e sufixos para redução da variação. Um exemplo de stemming aplicado as variações das palavras “amigo”, “amiga”, “amigos”, “amigável” reduzirá todas ao formato base de “amig”.

5. Lematização, processo de redução das palavras a sua forma base, ou lemma, considerando sua morfologia e classificação vocabular. Por exemplo, se considerarmos as palavras “aprendendo”, “aprendiz”, “aprendido” e “aprendizado”, a palavra base é sempre “aprender”. O processo de lematização irá aplicar este mesmo conceito as palavras identificadas no texto a ser analisado.

6. Tokenização, que consiste em dividir o texto em unidades menores de significado, como palavras ou frases, ajudando a estruturar o texto que alimentará os modelos de NPL aplicados.

7. Vetorização, processo de transformar o texto em uma representação numérica, que será mais bem compreendida por algoritmos de NPL, criando vetores de palavras ou frases.

Desta forma, o objetivo de realizar um pré-processamento de dados textuais não-estruturados é extrair uma representação estruturada e manipulável que preserve as principais características dos dados para sua análise por algoritmos de NPL (Rezende et al, 2011).

2.4 Processamento de Linguagem Natural (PLN)

Processamento de Linguagem Natural (PLN) é um campo de estudo que tem como objetivo o desenvolvimento de programas e algoritmos que sejam capazes de entender e interpretar a linguagem humana. Trata-se de uma área multidisciplinar, com aplicações nas áreas de Computação, Linguística, Estatística, Psicologia, entre outras.

As aplicações de técnicas de NPL (do inglês, Natural Procesing Language) são complexas devido a que essa área envolve a compreensão de uma das mais complexas habilidades do ser humano: a linguagem. A linguagem possui uma imensa variedade de idiomas e dialetos, que são falados em todo o mundo, tendo cada uma delas suas particularidades de estrutura gramatical, vocabulário e regras de sintaxe.

Por outro lado, linguagens também são ricas em ambiguidades, ironias, sarcasmos e metáforas que, ainda que fazem parte da comunicação entre seres humanos, pode ser muito difícil para a compreensão por parte de uma máquina. Não obstante, palavras também podem ter seu significado alterado de acordo ao contexto ao qual são inseridas.

No caso de identificação de discursos de ódio em textos extraídos de redes sociais, esta tarefa também pode ser complexa levando em consideração que palavras podem ser usadas de forma negativa com conotação preconceituosa se usadas fora de contexto.

Na imagem abaixo é possível observar um tweet, no idioma português, postado por um usuário em sua rede social e que pode ser considerado como conteúdo ofensivo e discurso de ódio contra pessoas pretas. Ainda que para humanos seja fácil identificar que as palavras como "preto" e "macaquinho" foram usadas em um contexto com objetivo ofensivo e discriminatório, a mesma interpretação para uma máquina pode ser mais complexa.

Figura 2 – Conteúdo racista postado por usuario na plataforma Twitter

Mesmo com toda essa complexidade, muitos algoritmos de NPL são capazes de identificar palavras ou conteúdos de ódio em redes sociais, e são usados para eliminar este tipo de conteúdo de suas plataformas. A própria rede Twitter afirma ter um sistema de identificação de conteúdo de ódio, além de prover a seus usuários a possibilidade de denunciar possíveis violações as regras da plataforma, as quais serão aplicadas medidas tais como a remoção do conteúdo e o banimento do usuário.

2.5 Aprendizado de Máquinas

O aprendizado de máquina concentra-se em capacitar os computadores a aprender a partir de dados, de modo que possam melhorar suas performances em tarefas específicas ao longo do tempo. Em vez de serem explicitamente programados para realizar uma tarefa, os sistemas de aprendizado de máquina são projetados para aprender padrões e regras a partir dos dados disponíveis. Isso é alcançado por meio da construção de algoritmos e modelos que podem generalizar a partir dos exemplos observados, permitindo que esses sistemas tomem decisões ou façam previsões a novos dados não vistos anteriormente. O aprendizado de máquina é amplamente aplicado em uma variedade de campos, incluindo reconhecimento de padrões, processamento de linguagem natural, visão computacional, análise de dados, otimização, entre outros, desempenhando um papel fundamental na resolução de problemas complexos e na extração de insights valiosos a partir de grandes volumes de informações. Os métodos de aprendizado de máquina têm assumido um espaço considerável na aplicação de tarefas relacionadas ao agrupamento e classificação de texto. Algoritmos como o Naive Bayes, Support Vector Machines (SVM), bem como as Redes Neurais Profundas, que

incluem arquiteturas como LSTM, GRU e RCNN, além de modelos fundamentados em árvores de decisão, estão se destacando nesse contexto (KOWSARI; HEIDARYSAFA, 2019). Essa ampla variedade de técnicas oferece um conjunto diversificado de soluções para lidar com a complexidade e a natureza não estruturada dos dados textuais.

Os algoritmos de *machine learning* podem ser categorizados com base na necessidade ou não de treinamentos com supervisão humana, sendo eles:

2.5.1 Aprendizado Supervisionado

No campo do aprendizado supervisionado, um modelo é treinado com um conjunto de dados rotulados, ou seja, dados de entrada e saída conhecidos, e em seguida é capaz de fazer previsões sobre novos dados apresentados com base no que aprendeu na fase de treinamento com os dados fornecidos. Destacam-se dois tipos primários de tarefas, com base na natureza das saídas desejadas: a primeira categoria é a classificação, na qual um modelo é treinado com dados rotulados para categorizar novos dados em categorias discretas ou rótulos pré-definidos; a segunda categoria seria a regressão, a qual envolve treinar o modelo com dados rotulados para fazer previsões de valores numéricos contínuos ou escalares. A principal distinção entre essas categorias reside nas saídas a serem previstas, independentemente do tipo de dado de entrada.

Alguns exemplos de aprendizado supervisionado incluem regressão linear, regressão logística, Support Vector Machines (SVM), arvores de decisão e redes neurais.

2.5.2 Aprendizado Não-supervisionado

Diferente do aprendizado supervisionado, nos modelos de aprendizado não supervisionado o modelo é treinado com um conjunto de dado de entrada não rotulados, buscando descobrir padrões e estruturas ocultas nos dados. Exemplos de algoritmos de aprendizado não supervisionado incluem clustering, análise de componentes principais (PCA), associação de regras e redes neurais auto-organizáveis.

2.5.3 Aprendizado Semi-supervisionado

Este tipo de aprendizado é um modelo intermediário entre o supervisionado e não-supervisionado, já que utiliza dados rotulados e não rotulados.O modelo aproveita a informação não rotulada para melhorar seu desempenho. Alguns exemplos de algoritmos semisupervisionados são SEEDED-K-means, COP-k-means, CONSTRAINED-k-means.

2.5.4 Aprendizado por Reforço

Um modelo é treinado sobre aprendizado por reforço quando é forçado a tomar ações e decisões em um ambiente dinâmico e incerto, recebendo um feedback, que pode ser em forma de recompensa ou penalidade, para cada decisão tomada. O objetivo do

modelo será aprender uma estratégica que irá maximizar as recompensas totais obtidas com o tempo. Exemplos de modelos de aprendizado por reforço são Q-leaning, SARSA e métodos de Monte Carlo.

3 TRABALHOS RELACIONADOS

Malmasi and Zampieri (2017) realizaram um estudo para identificar discursos de ódio em redes sociais e distingui-los de discursos gerais contendo vocabulário profano. O objetivo desse estudo foi estabelecer linhas de base lexicais para a tarefa de identificação de discursos de ódio, aplicando métodos de classificação supervisionados.

Para este estudo, os autores utilizaram um dataset criado e disponibilizado por Davdson et al (2017), contendendo aproximadamente 14.000 tweets em idioma inglês. Os tweets foram classificados de acordo a três categorias, sendo elas:

HATE: tweets que contêm conteúdos de discurso de ódio.

OFFENSIVE: tweets que contêm linguagem ofensiva, mas não discursos de ódio.

OK: sem nenhum conteúdo ofensivo.

Cada instância do dataset contém o texto proveniente do tweet postado na rede social e a classificação dentre as três categorias mencionadas. A distribuição dos dados de acordo as classes foi a seguinte:

Class	Texts
HATE	2,399
OFFENSIVE	4,836
OK	7,274
Total	14,509

Figura 3 – Distribuição de dados de acordo as categorias de clases

Como etapa de pré-processamento e preparação dos dados para análise, foram retiradas todas as URLs e emojis dos tweets, e o textos foram transformados em letras minúsculas. Em relação as características, foram usadas duas classes, sendo a primeira N-grams, consistindo em caracteres da ordem de n-grams 2-8 e palavras de ordem n-grams 1-3, todas tokenizadas a minúsculas antes da extração. A segunda classe foi a extração de 1-, 2- e 3-skip word bigrams, escolhidos para aproximar dependências de distancias mais longas entre as palavras. Para a análise dos dados, os autores aplicaram um modelo SVM linear para a classificação multiclasse, usando o pacote LIBLIN-EAR 5, que tem se mostrado um classificador eficaz para a classificação de textos em idiomas nativos, classificação temporal de textos e identificação da variedade linguística. Para avaliação dos modelos foi utilizada uma validação cruzadas de k = 10 (*10-fold crossvalidation*) e o tipo de validação cruzada estratificada para criar as dobras, com o objetivo de garantir a

igualdade de proporção entre as classes dentro de cada partição.

Primeiramente, Malmasi and Zampieri (2017) treinaram um classificador único com uma característica (feature) cada um. Em seguida, treinaram um modelo único combinando todas as features em um mesmo espaço. Os modelos foram comparados com a linha de base da classe majoritária, bem como com o oráculo, obtendo os seguintes resultados:

Feature	**Accuracy (%)**
Majority Class Baseline	50.1
Oracle	91.6
Character bigrams	73.6
Character trigrams	77.2
Character 4-grams	**78.0**
Character 5-grams	77.9
Character 6-grams	77.2
Character 7-grams	76.5
Character 8-grams	75.8
Word unigrams	77.5
Word bigrams	73.8
Word trigrams	67.4
1-skip Word bigrams	74.0
2-skip Word bigrams	73.8
3-skip Word bigrams	73.9
All features combined	77.5

Figura 4 – Acurácia obtida de acordo as diferentes features testadas.

Os resultados obtidos pelos autores mostram que os n-grams funcionaram bem, alcançando seu melhor desempenho com 4-grams. Os unigramas de palavras também tiveram uma boa performance, sendo que o desempenho foi mais baixo com bigramas, trigramas e skip-grams. Os skip-grams poderiam estar capturando dependências de longa distância que fornecem informações complementares aos outros tipos de features. Em tarefas que dependem de informações estilísticas, foi demonstrado que os skip-grams capturaram informações muito semelhantes às dependências sintáticas. A combinação de todos os recursos não atingiu o desempenho de um modelo de caracteres 4-grams, causando um aumento de dimensionalidade, com um total de 5.5 milhões de features. Para os autores, não ficou claro se este modelo seria capaz de capturar corretamente as diferentes informações fornecidas pelos três tipos de classificação (hate, offensive, ok), uma vez que foi incluído mais modelos de n-grams de caracteres que modelos baseados em palavras. Os maiores graus de confusão estão entre discurso de ódio e material ofensivo, sendo o discurso de ódio mais frequentemente confundido com conteúdo ofensivo. Uma quantidade substancial de conteúdo ofensivo também é erroneamente classificada como não ofensiva. A classe não ofensiva alcançou o melhor resultado, com a grande maioria das amostras classificadas corretamente. Os autores concluíram que distinguir palavras ofensivas com discursos de ódio é uma tarefa desafiadora.

4 METODOLOGIA

Neste capítulo, detalhes sobre a metodologia adotada para a realização do projeto de pesquisa serão discutidos e apresentados. O objetivo deste projeto é desenvolver um modelo capaz de classificar o conteúdo de um tweet como discursos de ódio ou não discursos de ódio, usando algoritmos clássicos de de aprendizado de máquina supervisionado.

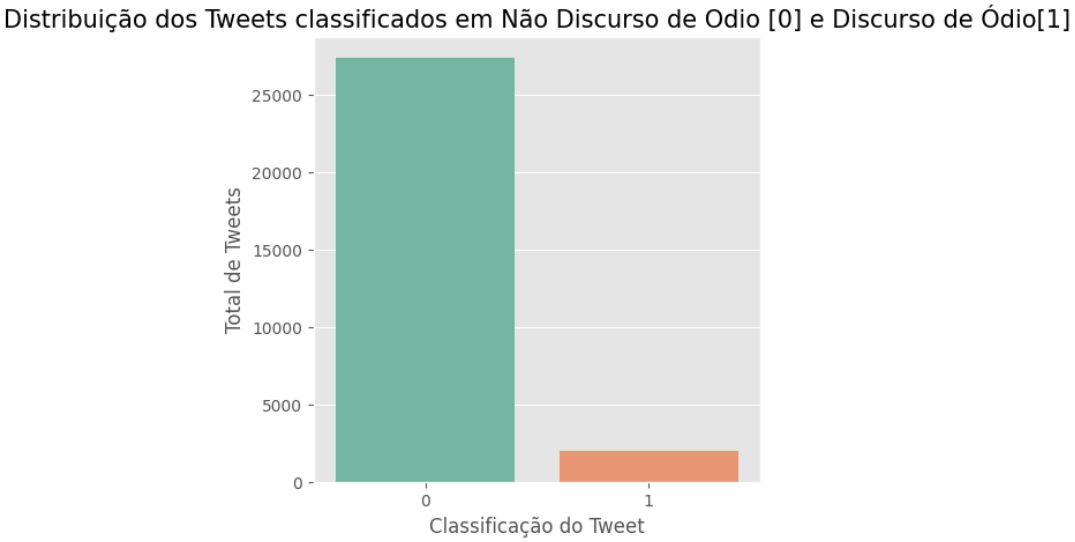
4.1 Coleta do Dataset

Para a coleta do dataset, realizou-se uma busca no Kaggle, uma plataforma popular para compartilhamento de conjuntos de dados. Utilizamos termos de busca relevantes, como *"Twitter hate speech"* e *"discrimination tweets"*. Selecionamos o dataset intitulado *"Twitter Hate Speech"* que contém aproximadamente 32.000 tweets no idioma inglês, extraídos da plataforma Twitter, rotulados como discursos de ódio e não ofensivos. Essa fonte de dados sintética foi escolhida por sua adequação aos objetivos da pesquisa e pela disponibilidade dos rótulos necessários para o treinamento do modelo.

4.2 Análise Exploratória de Dados

Primeiramente foi realizado uma análise exploratória detalhada do dataset com o objetivo de explorar a estrutura dos dados, verificar a distribuição dos rótulos, analisar características dos tweets, como informações de usuário, *hashtags* e palavras-chave relevantes. Essa análise exploratória permite uma melhor compreensão dos dados e a identificação de padrões e tendências.

Durante essa análise, observou-se que o conjunto de dados apresentava um desequilíbrio significativo entre as classes. Especificamente, uma classe estava representada de forma desproporcional em relação à outra, o que poderia afetar a eficácia dos modelos de aprendizado de máquina a serem construídos. A distribuição inicial dos dados entre as classes discurso de ódio e conteúdo não ofensivo pode ser visualizada na imagem abaixo.



[h]

Figura 5 – Distribuição das classes de acordo a classificação dos conteúdos dos tweets em Discurso de Ódio ou Não Discurso de Ódio

4.3 Pré-processamento de Texto

Antes de prosseguir com a modelagem de aprendizado de máquina, foram realizadas etapas de pré-processamento de texto para preparar os tweets do conjunto de dados para análise. Essas etapas visam limpar e transformar o texto bruto dos tweets em uma forma mais adequada para a modelagem.

As principais etapas de pré-processamento de texto incluídas são:

4.3.1 Remoção de Caracteres Especiais e Pontuação

Nesta etapa, removemos caracteres especiais e pontuações, como símbolos, emojis e caracteres especiais de codificação dos tweets. Utilizamos expressões regulares para identificar e remover esses caracteres, pois eles geralmente não contêm informações relevantes para a presente análise.

4.3.2 Remoção de Stopwords

Stopwords são palavras comuns que não possuem um significado distinto e não contribuem significativamente para a análise textual, como "a", "e", "o", "de", em português. Essas palavras geralmente ocorrem com frequência nos tweets, mas podem ser removidas com segurança, uma vez que não trazem informações relevantes para a classificação de discursos de ódio. Considerando que o conjunto de dados utilizado para o presente estudo possui tweets no idioma inglês, utilizamos uma lista predefinida de stopwords do idioma referido da biblioteca NLTK (Natural Language Toolkit) para remover essas palavras dos tokens. Algumas palavras do idioma mencionado que foram retiradas incluem “and”, “before”, “most”, “once”, entre outras.

4.3.3 Normalização de Texto

Essa etapa do pré-processamento textual visa aprimorar a qualidade dos dados ao reduzir variações e inconsistências linguísticas. Um dos aspectos fundamentais nesse processo é a lematização, uma técnica linguística amplamente aplicada que tem como objetivo reduzir as palavras do corpus aos seus lemas ou formas base, eliminando flexões e conjugações. No contexto específico da detecção de discurso de ódio, a lematização apoia o papel da normalização do vocabulário e da agregação de termos relacionados. Sua relevancia provém de que o discurso de ódio pode adotar diferentes formas e empregar palavras ofensivas com variações morfológicas. Ao aplicar a lematização, é possível agrupar palavras similares, contribuindo para uma identificação mais precisa e abrangente do discurso de ódio no Twitter.

A partir das etapas de pré-processamento dos dados textuais, foi possível extrair as palavras relevantes. No contexto deste estudo, o método de nuvem de palavras foi aplicado como uma técnica complementar para visualizar e resumir a frequência de palavras-chave presentes nos textos analisados. A visualização por meio da nuvem de palavras facilita a identificação de termos-chaves e contribui para uma análise exploratória inicial dos dados textuais de maneira intuitiva e acessível.

Foram geradas nuvens em que o tamanho das palavras reflete sua frequência de ocorrência nos textos. Essa abordagem permitiu identificar os termos mais relevantes e recorrentes relacionados ao problema em estudo.Em uma primeira instancia de análise do conteúdo dos tweets utilizando o método de nuvem de palavras, identificou-se que a palavra de maior frequencia em ambas clases do conjunto de dados correspondia a palavra *"user"*, fazendo referencia a outro usuário da plataforma para responder ou direcionar um determinado tweet. Durante o desenvolvimento deste projeto científico, optou-se por excluir a palavra *"user"* dos tweets coletados. Essa decisão baseou-se na consideração de que a palavra não apresenta relevância sintática significativa na análise do conteúdo do texto. Ao remover essa palavra, foi possível direcionar o foco para os elementos linguísticos mais relevantes que contribuem para a detecção e classificação de discursos de ódio, como se pode observar nas imagens seguintes.

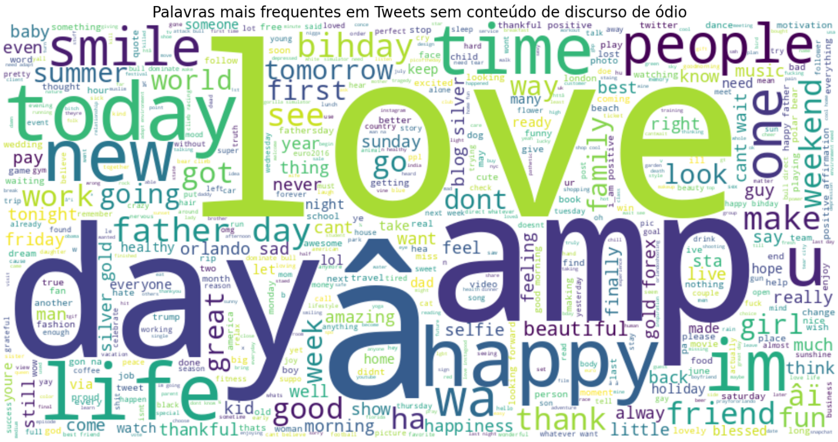


Figura 6 – Palavras mais frequentes identificadas em tweets classificados como sem conteúdo de discurso de ódio

Figura 7 – Palavras mais frequentes identificadas em tweets classificados com conteúdo de discurso de ódio

4.3.4 Tokenização

Técnica linguística que consiste na segmentação do texto em unidades léxicas menores, denominadas tokens, tais como palavras, frases ou caracteres. Especificamente para a detecção de discurso de ódio, a tokenização permite a identificação de elementos linguísticos relevantes, como palavras-chave, sequências de caracteres, hashtags e até mesmo nomes de usuários que podem ser considerados potenciais indicadores de conteúdo ofensivo. A função aplicada nesse contexto foi a *Tokenize* da biblioteca NLTK *(Natural Language Toolkit)*, um algoritmo de tokenização de texto que divide as sentenças em palavras individuais, levando em consideração pontuações e espaços em branco. A partir dessa representação tokenizada, tornou-se viável extrair informações relevantes, como a contagem de termos e a presença de palavras-chave específicas que podem corresponder a uma linguagem ofensiva ou discriminatória.

4.3.5 Vetorização

Para realizar o processo de vetorização no conjunto de dados, utilizou-se a técnica *TfidfVectorizer*, uma combinação das técnicas de contagem de termos *(CountVectorizer)* e ponderação TF-IDF *(Term Frequency-Inverse Document Frequency)*. O *TfidfVectorizer* atribui um valor numérico a cada termo presente nos tweets, levando em consideração tanto a frequência do termo em um tweet específico quanto a sua relevância global no conjunto de dados. Essa abordagem visa destacar termos que são frequentes em um tweet específico, mas que são raros em outros, considerando a importância desses termos para a classificação dos tweets como discurso de ódio ou não. Após a aplicação do *TfidfVectorizer*, os tweets foram convertidos em vetores numéricos, onde cada componente do vetor representa a importância de um termo específico no tweet. Essa representação vetorial permite a utilização de algoritmos de aprendizado de máquina para a classificação dos tweets.

4.4 Divisão do Dataset

O dataset foi dividido em dois conjuntos distintos: um conjunto de treinamento e um conjunto de teste. A divisão foi feita de forma estratificada, garantindo que a distribuição dos rótulos de discursos de ódio e não ofensivo fosse preservada em ambos os conjuntos. A proporção escolhida foi de 0,8 para o conjunto de treinamento e 0,2 para o conjunto de teste. Essa divisão foi feita para avaliar o desempenho do modelo em dados não vistos durante o treinamento.

Como citado anteriormente, no presente estudo foi observado um desequilíbrio significativo entre as classes no conjunto de dados, o que poderia impactar a performance dos modelos de aprendizado de máquina. A presença desse desequilíbrio pode afetar negativamente a performance dos modelos de aprendizado de máquina, já que a classe minoritária está sub-representada em relação à classe majoritária.

Dado esse cenário e a dificuldade de obter mais exemplos de dados rotulados para o treinamento dos modelos,optou-se por adotar a abordagem de *Data Augmentation* (aumento de dados) para gerar dados sintéticos. Esses novos dados gerados artificialmente possuem rótulos conhecidos e podem ser utilizados no processo de treinamento dos modelos (Feng *et al.* (2021)). Optou-se por aplicar a técnica SMOTE *(Synthetic Minority Over-sampling Technique)* para lidar com a desproporção entre as classes no conjunto de dados selecionado. O SMOTE é uma técnica de *oversampling* que visa equilibrar a distribuição das classes, sintetizando novas instâncias da classe minoritária. Ao gerar dados sintéticos, o SMOTE se baseia na interpolação dos atributos dos vizinhos mais próximos da classe minoritária. Dessa forma, são criadas instâncias adicionais, que mantêm as características e padrões da classe minoritária original, tornando-as mais representativas no conjunto de dados. Importante ressaltar que essa técnica foi aplicada apenas para o conjunto de treino da divisão do conjunto de dados. O conjunto de teste permaneceu com suas amostras

originais.

A aplicação do SMOTE resultou em um conjunto de dados de treino balanceado, no qual as classes minoritárias foram aumentadas para atingir um número mais próximo da classe majoritária.

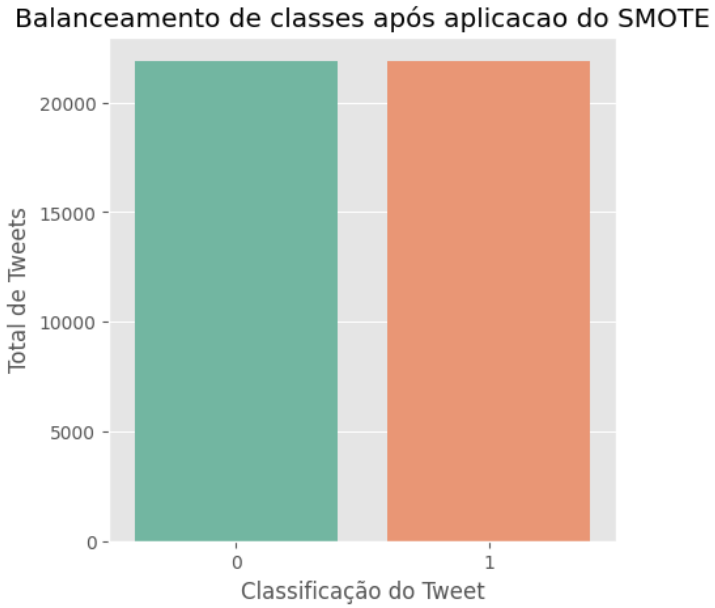


Figura 8 – Nova distribuição das classes após aplicação da técnica SMOTE (Synthetic Minority Over-sampling Technique

4.5 Modelagem de Aprendizado de Máquina

Foram aplicados três modelos de aprendizado de máquina supervisionado: *Support Vector Machine* (SVM), *Logistic Regression* e *Naive Bayes*, com o objetivo de realizar a classificação no conjunto de dados selecionado. Esses modelos foram selecionados devido às suas capacidades de lidar com problemas de classificação binária e multiclasse.

Cada modelo passou por um processo de validação cruzada *(*cross-validation) durante a etapa de treinamento. A validação cruzada é recomendada porque ela permite uma avaliação mais robusta do desempenho do modelo, ajudando a reduzir o impacto da variância do conjunto de treinamento e minimizando o risco de superajuste (overfitting). Ao dividir o conjunto de dados em diferentes conjuntos de treinamento e teste, a validação cruzada permite que o modelo seja avaliado em múltiplas iterações, garantindo uma análise mais abrangente de sua capacidade de generalização para diferentes combinações de dados de treinamento e teste. Essa abordagem oferece uma estimativa mais precisa do desempenho real do modelo em novos dados e ajuda a selecionar o melhor modelo com os hiperparâmetros mais adequados para a tarefa de classificação.

Inicialmente, o modelo SVM foi empregado, aproveitando sua capacidade de encon-

trar um hiperplano de separação ótimo que maximize a margem entre as classes. Além das características lineares presentes nos tweets de discurso de ódio, como palavras-chave ofensivas ou frases explicitamente agressivas, é importante destacar as características não lineares que podem estar presentes nesse tipo de texto. Os discursos de ódio frequentemente envolvem a utilização de sarcasmo, ironia, metáforas e outros recursos linguísticos não literais, que podem dificultar a classificação linear dos tweets e apresentam um desafio nas tarefas de classificação de textos. Uma das formas de ajudar a identificar tais características não lineares pode ser através do uso de kernels não lineares, como o kernel RBF *(Radial Basis Function)*, do algoritmo SVM. O kernel RBF permite mapear as palavras e frases para um espaço de dimensionalidade infinita, onde as relações complexas e não lineares entre as características podem ser capturadas. Dessa forma, é possível considerar contextos sutis, nuances e ambiguidades presentes nos tweets de discurso de ódio, aprimorando a capacidade de detecção desses conteúdos ofensivos. A inclusão dessas características não lineares na modelagem do SVM é indicada para lidar com a complexidade semântica dos discursos de ódio e melhorar a acurácia da detecção desses conteúdos. Por esta razão, no presente estudo o kernels RBF foi utilizado para treinamento do modelo.

No contexto do SVM com kernel RBF, a gamma é um parâmetro crítico que controla a influência dos pontos de treinamento vizinhos na decisão da superfície de separação. Para encontrar a gamma mais adequada, utilizou-se a *cross validation* (validação cruzada), técnica na qual o conjunto de dados é dividido em k partições *(folds)* de tamanhos iguals. O modelo SVM com diferentes valores de gamma é treinado k vezes, onde em cada iteração, *k-1 folds* são utilizados para treinamento e o *fold* restante é utilizado para teste. As métricas de desempenho calculadas para cada iteração determina o valor de gamma ideal para se aplicar ao modelo.

Um modelo de *Logistic Regression* foi igualmente aplicado no conjunto de dados, visando a classificação binária dos textos nas categorias de discurso de ódio ou não ofensivo. Neste contexto, a variável dependente é a presença ou ausência de discurso de ódio, enquanto as variáveis independentes são características relevantes extraídas dos textos analisados e podem incluir características linguísticas, palavras-chave relevantes e outros atributos que possam influenciar na presença de conteúdo ofensivo no tweet.

O modelo de regressão logística adotado neste estudo utilizou como variáveis independentes características dos discursos de ódio que foram analisadas por meio de tokenização e vetorização dos tweets. A etapa de tokenização consistiu em dividir os tweets em unidades de texto significativas, como palavras ou n-gramas (n=3), permitindo assim a identificação de padrões linguísticos relevantes. Em seguida, a vetorização foi realizada para transformar essas unidades de texto em vetores numéricos, que representaram as características linguísticas dos discursos.

As variáveis independentes foram construídas a partir desses vetores, considerando diferentes abordagens, como a contagem de palavras e a frequência dos termos. Essa vetorização permitiu capturar as nuances linguísticas presentes nos tweets e transformá-las em características numéricas que puderam ser utilizadas como preditores no modelo de regressão logística.

O modelo *Naive Bayes* foi também adotado como parte da metodologia deste projeto de pesquisa, em específico o modelo Multinomial, com o objetivo de identificar discursos de ódio. A escolha do modelo Multinomial Naive Bayes se deu pela sua adequação para lidar com variáveis discretas, como a contagem de palavras em textos curtos como os tweets. Para construir o modelo, as variáveis independentes igualmente passaram pelo processo de tokenização e vetorização dos textos dos discursos. Nesse modelo, é assumida uma independência condicional entre as variáveis independentes, o que simplifica o cálculo da probabilidade de um discurso pertencer à classe de discurso de ódio, dado um conjunto de características. As características vetorizadas dos discursos foram usadas para estimar a probabilidade condicional de pertencer à classe de discurso de ódio ou não.

A aplicação do modelo *Multinomial Naive Bayes* para a identificação de discursos de ódio em tweets envolveu a configuração de parâmetros específicos, como o parametro de suavização de Laplace. A suavização de Laplace foi aplicada para evitar probabilidades nulas e melhorar a generalização do modelo. Um valor comumente utilizado para o parâmetro de suavização é 1. Essa adição suaviza as contagens dos termos, considerando-os como se tivessem ocorrido uma vez a mais do que realmente ocorreram. A aplicação de *grid-search* para encontrar o melhor valor de suavização através da comparação dos resultados obtidos pelos diferentes modelos busca identificar a abordagem que obteve o melhor desempenho na identificação e classificação de tweets com conteúdo de ódio.

4.6 Validação e Ajuste do Modelo

Após o treinamento do modelo, avaliamos seu desempenho utilizando o conjunto de teste. Utilizamos as seguintes métricas de avaliação:

Precisão *(Precision)*: é a proporção de exemplos classificados corretamente como discursos de ódio em relação ao total de exemplos classificados como discursos de ódio (positivos). A fórmula da precisão é dada por:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Recall (Revocação ou Sensibilidade): é a proporção de exemplos classificados corretamente como discursos de ódio em relação ao total de exemplos que são realmente discursos de ódio. A fórmula do recall é dada por:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

F1-Score: é uma métrica que combina precisão e recall em um único valor, fornecendo uma medida de desempenho geral. O F1-score é a média harmônica entre a precisão e o recall, e a fórmula é dada por:

$$F1 - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

Matriz de Confusão *(Confusion Matrix)*: é uma tabela que mostra a contagem de verdadeiros positivos *(True Positives)*, verdadeiros negativos *(True Negatives)*, falsos positivos *(False Positives)* e falsos negativos *(False Negatives)*. A matriz de confusão fornece uma visão mais detalhada do desempenho do modelo em relação a cada classe.

Os resultados obtidos por cada modelo aplicado serão discutidos no capítulo seguinte do presente estudo.

5 AVALIAÇÃO EXPERIMENTAL

Neste projeto científico, investigou-se a detecção de discursos de ódio em tweets coletados no idioma inglês, empregando uma abordagem de processamento de linguagem natural e algoritmos clássicos de aprendizado de máquina. Os experimentos foram divididos em dois cenários principais: um no qual o modelo foi treinado com dados originais e dados sintéticos da classe alvo de discursos de ódio, gerados a partir da aplicacao da técnica SMOTE (Modelo-A) e outro no qual o modelo foi treinado apenas com os dados originais provenientes do dataset utilizado (Modelo-B). Ambos cenários ocuparam o mesmo conjunto de dados e passaram pelas mesmas técnicas de pré-processamento, a exceção da técnica de data augmentation (SMOTE). A seguir, se apresenta uma análise detalhada dos resultados obtidos para cada modelo nos dois cenários especificados.

Adicionalmente, é relevante mencionar que os experimentos foram executados com uma abordagem de validação cruzada com k=5 e a vetorização dos dados foi realizada utilizando a técnica TF-IDF com n-gramas de tamanho 3. A seguir, proporcionamos uma análise aprofundada dos resultados obtidos para cada um dos modelos nos dois cenários previamente especificados.

Support Vector Machine (SVM)

Avaliando o desempenho do modelo SVM (Support Vector Machine) com kernel RBF (Radial Basis Function) na tarefa proposta, utilizou-se o método de Grid Search, que possibilitou explorar uma variedade de valores do hiperparâmetro gamma, o que abrange tanto valores pequenos quanto grandes, sendo esses valores entre 0.001 e 10. O valor ótimo encontrado foi 1 para ambos modelos (SVM-A E SVM-B), valor este que foi usado nos modelos finais de treinamento.

A análise da matriz de confusão do modelo SVM-A, treinado a partir dos dados aumentados com exemplos sintéticos da classe de discurso de ódio (processo de data augmentation), mostra que foram corretamente identificados 5934 casos de discursos não ofensivos e 151 casos de discursos de ódio. No entanto, o modelo cometeu alguns erros, classificando erroneamente 3 casos de conteúdo não ofensivo como discursos de ódio e deixando passar 305 casos de discursos de ódio, classificando-os como conteúdo não ofensivo. Esses resultados destacam que o modelo foi capaz de identificar discursos de ódio, mas apresentou mais dificuldades com amostras da classe alvo que com a classe de discursos não ofensivos, apontando uma necessidade de aprimoramentos para reduzir os falsos positivos e falsos negativos, tornando-o mais robusto e preciso na detecção do conteúdo ofensivo.

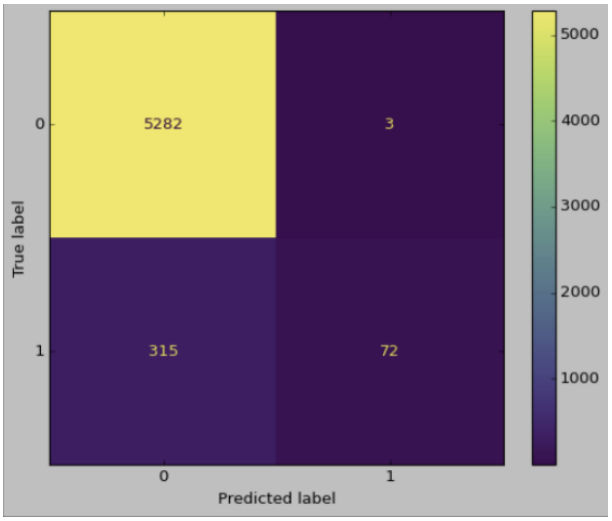


Figura 9 – Matriz de confusao do modelo SVM-A

Os resultados das métricas de avaliação desse modelo mostram uma precisão média para a classe 0 (não discurso de ódio) de 0.95, o que indica uma alta capacidade de acerto na classificação das instâncias dessa classe. Para a classe 1 (discurso de ódio), a precisão média foi de 0.98.

Entretanto, ao analisar o recall para a classe 1, identificamos um valor de 0.19, o que indica que o modelo teve dificuldades em recuperar corretamente todas as instâncias dessa classe. É importante resaltar que, para a tarefa proposta de identificacao de discursos de ódio, o recall é considerado um parametro significativo, uma vez que o objetivo principal do modelo é não deixar de identificar e classificar corretamente tweets que contenham conteúdos ofensivos. Com base no anteriormente mencionado, o score de recall do modelo indica que ele não foi capaz de identificar adequadamente todas as ocorrências de discurso de ódio presentes no conjunto de dados, não apresentando um desempenho satisfatório na tarefa proposta. A métrica F1-score, que considera o equilíbrio entre precisão e recall, resultou em 0.31 para a classe de discursos de ódio. Essa pontuação sugere um baixo desempenho na harmonização dessas métricas para a classe de discurso de ódio, apontando também a necessidade de aprimorar a capacidade do modelo em identificar corretamente todos os casos positivos.

	precision	recall	f1-score	support
0	0.94	1.00	0.97	5285
1	0.96	0.19	0.31	387
accuracy			0.94	5672
macro avg	0.95	0.59	0.64	5672
weighted avg	0.94	0.94	0.93	5672

Figura 10 – Resultados obtidos pelo modelo SVM-A nas métricas de avaliação.

Embora a acurácia geral foi de 0.94, essa métrica não é mais adequada para avaliar a performance do modelo, devido a que mesmo com um alto score, observou-se que o modelo SVM-A classificou erroneamente um número considerável de amostras da classe alvo (discurso de ódio), evidenciado pelo seu baixo score de recall (19%). Em suma, os resultados evidenciam um bom desempenho do modelo SVM-A com kernel RBF em termos de precisão, mas revelam a necessidade de melhorar o recall e o F1-score para a classe de interesse (discurso de ódio).

Já os resultados para o modelo SVM-B, treinado com os dados originais (sem passar pelo processo de data augmentation), mostraram uma acurácia de 94%. Na matriz de confusão desse modelo é possível observar seu alto desempenho em classificar tweets nao ofensivos, tendo categorizado todas as amostras dessa classe de forma correta. Porém, para a classe alvo de tweets contendo discursos de ódio, o modelo apresentou muita dificuldade na classificação, categorizando correramente apenas 39 das 387 amostras totais dessa classe.

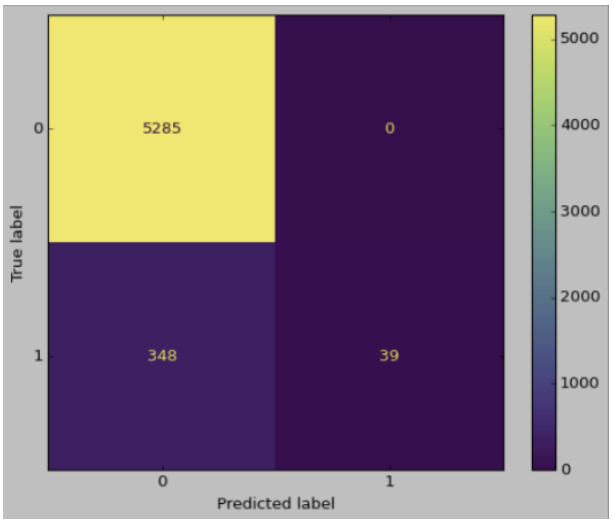


Figura 11 – Matriz de confusão do modelo SVM-B

As conclusões obtidas a partir da matriz de confusão para o modelo SVM-B são corroboradas pelos scores obtidos pelo modelo nas métricas de avaliação propostas. Seus resultados mostraram uma precisão de 94% para a classe de tweets não ofensivos e um alto recall de 100%, indicando que quase todos os tweets dessa classe foram identificados corretamente.

Por outro lado, o modelo obteve um recall de apenas 10% para a classe alvo de discursos de ódio, indicando dificuldades em classificar uma parte significativa dos tweets ofensivos, resultando em falsos negativos.A precisão para essa classe foi de 100%, o que pode indicar um possível overfitting ou a escassez de amostras para essa classe específica.

O F1-Score, que combina precisão e recall, foi baixo para a classe alvo, com apenas 18%. Esses resultados sugerem que o modelo apresenta um desempenho satisfatório na

	precision	recall	f1-score	support
0	0.94	1.00	0.97	5285
1	1.00	0.10	0.18	387
accuracy			0.94	5672
macro avg	0.97	0.55	0.58	5672
weighted avg	0.94	0.94	0.91	5672

Figura 12 – Resultados obtidos pelo modelo SVM-B nas métricas de avaliação.

detecção de tweets não ofensivos, mas enfrenta desafios significativos na identificação de discursos ofensivos. A disparidade entre as métricas de precisão e recall pode ser atribuída ao desequilíbrio das classes no conjunto de dados, uma vez que há uma quantidade consideravelmente menor de tweets ofensivos.

Os resultados obtidos pelos modelos não apresentaram uma diferença significativa, embora o modelo SVM-A tenha mostrado um desempenho ligeiramente superior na identificação e classificação de tweets com conteúdo de ódio em comparação com o modelo SVM-B. Como mencionado anteriormente, essa diferença pode ser explicada pelo fato de que o modelo SVM-A foi submetido a um processo de data augmentation e balanceamento das classes de discursos de ódio e conteúdo não ofensivo. Esse processo aumentou a quantidade de amostras da classe de discursos de ódio, criando exemplos sintéticos a partir dos tweets originais.

O aumento da quantidade de amostras da classe de discursos de ódio no modelo SVM-A permitiu que ele se tornasse mais sensível à identificação desses discursos, melhorando seu desempenho nessa categoria específica. Em contraste, o modelo SVM-B pode ter sido prejudicado pelo desequilíbrio entre as classes, o que impactou sua capacidade de classificar corretamente os tweets com conteúdo de ódio.

Esses resultados destacam a importância do balanceamento das classes e da utilização de técnicas de data augmentation para melhorar o desempenho de modelos de classificação em cenários de desequilíbrio de dados. Ao lidar com discursos de ódio em redes sociais, essas abordagens podem ser cruciais para tornar o modelo mais robusto e eficaz na identificação desses discursos. A análise comparativa entre os dois modelos reforça a relevância de técnicas que levem em conta o desequilíbrio de classes ao desenvolver soluções para problemas de classificação em cenários do mundo real.

Logistic Regression

Após a análise dos resultados, constatou-se que o modelo de Regressão Logística alcançou uma acurácia de 95% para o modelo treinado com a aplicação do processo de data augmentation (modelo LR-A). Isso implica que, para este resultado, foram consideradas quantidades iguais de dados classificados como discurso de ódio e não discurso de ódio. A maioria das amostras de teste foi classificada corretamente pelo modelo.

Observando a matriz de confusão gerada para o modelo LR-A, observou-se que para a classe de discursos não ofensivos, o modelo apresentou um alto número de verdadeiros negativos (5197), indicando que a grande maioria das amostras foi corretamente classificada como discursos sem conteúdo de ódio. No entanto, ao analisar os falsos positivos, nota-se que algumas amostras (99) foram incorretamente classificadas como discursos de ódio quando, na verdade, não continham conteúdo ofensivo. Quanto à classe de discursos de ódio, o modelo classificou corretamente 232 de 388 amostras para a classe 1, o que significa que algumas amostras foram classificadas erroneamente como não ofensivas. Esta classificação equivocada é o principal ponto de atenção para o experimento, uma vez que o objetivo do estudo é justamente identificar e classificar corretamente esse tipo de conteúdo para combater discursos de ódio de forma eficaz.

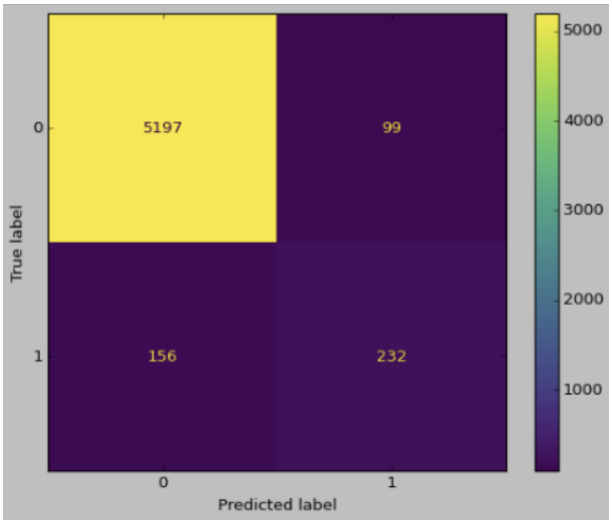


Figura 13 – Matriz de confusão do modelo LR-A

Em relação às métricas de avaliação, o modelo LR-A demonstrou uma precisão de 0.68 na classificação de discursos de ódio. No entanto, notamos que o modelo teve uma quantidade significativa de falsos positivos. Isso quer dizer que algumas amostras foram incorretamente classificadas como discursos de ódio quando, na verdade, não pertenciam a essa categoria. Essa alta taxa de falsos positivos pode sugerir que o modelo apresenta dificuldades em distinguir entre tweets agressivos e discursos de ódio genuínos.

	precision	recall	f1-score	support
0	0.97	0.98	0.98	5285
1	0.68	0.60	0.64	387
accuracy			0.95	5672
macro avg	0.83	0.79	0.81	5672
weighted avg	0.95	0.95	0.95	5672

Figura 14 – Resultados obtidos pelo modelo LR-A nas métricas de avaliação.

No que diz respeito ao recall, o modelo LR-A obteve um valor razoável de 0.6 para a classe de discursos de ódio, indicando que o modelo foi capaz de identificar corretamente 60% das amostras que realmente pertenciam à classe de discursos de ódio, em relação ao total de amostras de discursos de ódio presentes no conjunto de teste. O modelo apresentou um F1-Score de 0,64 para a classe de discursos de ódio, o que sugere um equilíbrio moderado entre sua capacidade de identificar corretamente os discursos de ódio e sua propensão a gerar tanto falsos positivos quanto falsos negativos. Essa métrica indica que o modelo alcança um desempenho razoável entre a precisão de suas previsões e sua capacidade de abranger efetivamente as instâncias reais de discurso de ódio, mas também deixa espaço para melhorias visando a redução de falsos positivos e negativos.

Já o modelo LG-B, treinado com os dados originais provenientes do dataset, sem passar pelo processo de data augmentation, apresentou uma acurácia de 93,3%. Em relação a matriz de confusão desse modelo, se pode observar o seguinte comportamento:

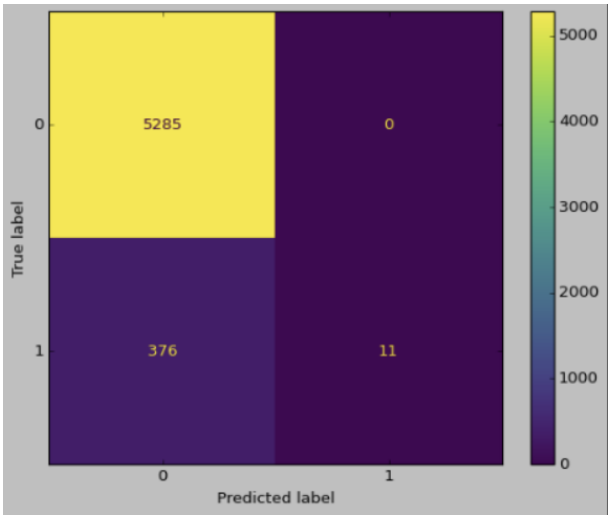


Figura 15 – Matriz de confusão do modelo LR-B

Como observado na imagem acima, o modelo LR-B obteve um excelente resultado na identificação de tweets não ofensivos, classificando corretamente todas as amostras dessa classe. No entanto, o modelo apresentou um desempenho muito baixo na identificação de

discursos de ódio (classe 1). Das 387 amostras dessa classe, apenas 11 foram corretamente identificadas como discursos de ódio (verdadeiros positivos). Esse baixo desempenho é comprovado pelo score de recall da classe alvo, de apenas 0.03. Isso significa que o modelo foi capaz de identificar corretamente apenas 3% das amostras dessa classe.

A baixa pontuação de 0.06 de F1-Score para o modelo RL-B também pode ser atribuída ao desequilíbrio das classes no conjunto de dados. O F1-Score é uma métrica sensível ao desbalanceamento, e quando as classes têm distribuições desiguais, ela pode ser afetada negativamente. No caso em questão, a classe de discursos de ódio é menos representativa em comparação com a classe de discursos não ofensivos. Devido a esse desequilíbrio, o modelo pode favorecer a classificação da classe majoritária (discursos não ofensivos), resultando em uma alta precisão para essa classe (97%). Como o F1-Score considera tanto a precisão quanto o recall, ele reflete o equilíbrio entre essas duas métricas. Nesse contexto, um baixo F1-Score indica que o modelo não está conseguindo generalizar adequadamente para ambas as classes, especialmente devido ao desequilíbrio presente no conjunto de dados.

	precision	recall	f1-score	support
0	0.93	1.00	0.97	5285
1	1.00	0.03	0.06	387
accuracy			0.93	5672
macro avg	0.97	0.51	0.51	5672
weighted avg	0.94	0.93	0.90	5672

Figura 16 – Resultados obtidos pelo modelo LR-B nas métricas de avaliação.

A melhoria nos resultados do modelo treinado com a técnica de data augmentation (modelo LR-A) em comparação com o modelo LG-B também pode ser explicada em base a que o desequilíbrio de classes é um desafio comum em conjuntos de dados para detecção de discursos de ódio, no qual a classe de discursos de ódio é minoritária em comparação a classe de discursos não ofensivos. Esse desequilíbrio pode gerar um viés no treinamento do modelo, levando-o a favorecer a classe majoritária. A técnica de data augmentation, ao abordar essa disparidade, gera exemplos sintéticos da classe minoritária, equilibrando, assim, a distribuição das classes no conjunto de dados de treinamento.

Ao aumentar a representação da classe minoritária, o modelo treinado com SMOTE (modelo RL-A) é exposto a uma variedade mais ampla de exemplos de discursos de ódio, o que permite uma melhor aprendizagem das características distintivas dessa classe. Como resultado, o modelo demonstra uma maior capacidade discriminativa entre discursos de ódio e não discursos de ódio, levando a melhorias em métricas de avaliação, como precisão, recall, F1-score e acurácia.

Naive Bayes

Durante a busca em grade com validação cruzada para encontrar os melhores hiperparâmetros para o modelo Naive Bayes, observou-se que o valor ótimo para o parâmetro *alpha* é 0.01, com base na métrica de acurácia do modelo. O *alpha* é um parâmetro de suavização Laplace utilizado no algoritmo Naive Bayes Multinomial para evitar a probabilidade zero de palavras não vistas no conjunto de treinamento, melhorando a capacidade de generalização e eficiência do modelo na identificação de discursos de ódio.

Os resultados para esse modelo treinado com exemplos sintéticos, obtidos a partir do processo de data augmentation, mostraram uma acurácia de 92%, indicando a habilidade do modelo em classificar corretamente tweets de ambas as classes. Porém, analisando o recall para a classe alvo de discursos de ódio, observou-se un recall de 0.8, indicando que o modelo foi capaz de identificar e recuperar 80% das amostras que realmente pertencem a essa classe, em relação ao total de amostras de discursos de ódio presentes no conjunto de teste.

Após conduzir um novo treinamento com o modelo, fixando o parâmetro *alpha* em 1.0, observou-se uma acurácia de 90% e um recall para a classe alvo de 0.88, atingindo um valor ligeiramente superior ao do modelo original. Dado que o foco principal deste estudo reside na identificação e classificação precisa de discursos de ódio, a decisão foi tomada em favor deste modelo devido ao seu recall mais elevado. Nesse contexto, este modelo foi designado como NB-A. A matriz de confusão resultante desse modelo pode ser observada a seguir:

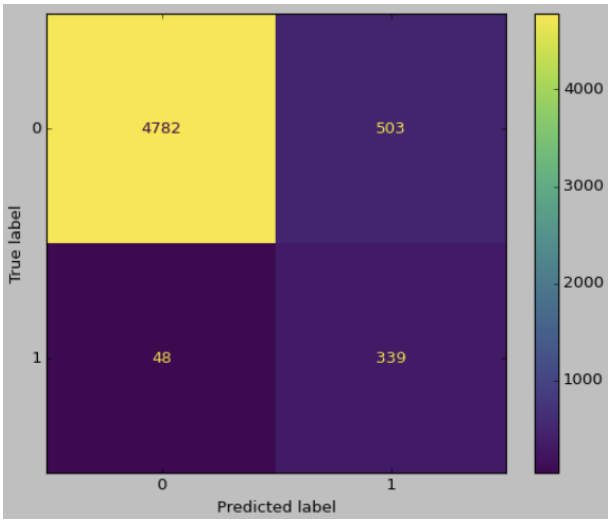


Figura 17 – Matriz de confusão do modelo NB-A

O modelo NB-A demonstrou a capacidade de identificar com precisão 339 amostras de discursos de ódio, das 387 presentes no conjunto de dados. Além disso, obteve uma correta identificação de 4782 amostras de discursos não ofensivos. No entanto, é importante notar que este modelo também apresentou um número consideravelmente maior de falsos

positivos, classificando incorretamente 503 amostras como discursos de ódio, quando, na realidade, não se enquadravam nessa categoria.

No que diz respeito às outras métricas de avaliação do modelo, especificamente para a classe alvo de discursos de ódio, a precisão atingiu o valor de 0.40, contrastando com a elevada precisão de 0.99 alcançada para a classe de discursos não ofensivos. No que tange aos valores do F1-score, que considera a harmonização de precisão e recall, os resultados mostraram um valor de 0.55 para a classe alvo, em comparação com o valor substancialmente mais elevado de 0.95 para a classe de discursos não ofensivos. Essa análise ressalta a disparidade na capacidade do modelo em equilibrar precisão e recall entre as duas classes, similar aos resultados obtidos pelos outros modelos usados no presente estudo.

	precision	recall	f1-score	support
0	0.99	0.90	0.95	5285
1	0.40	0.88	0.55	387
accuracy			0.90	5672
macro avg	0.70	0.89	0.75	5672
weighted avg	0.95	0.90	0.92	5672

Figura 18 – Resultados obtidos pelo modelo NB-A nas métricas de avaliação.

No caso do modelo treinado com os dados originais, o modelo NB-B, também foi conduzido um processo de treinamento com validação cruzada a fim de identificar o valor ótimo para o parâmetro *alpha*. O valor ideal determinado foi 0.1. Subsequentemente, ao realizar um novo treinamento com *alpha* igual a 1.0 para avaliar eventuais diferenças no desempenho, não foram constatadas quaisquer discrepâncias. Ambos os valores de *alpha* resultaram nos mesmos escores, apontando para uma estabilidade de desempenho entre as duas configurações.

Os resultados do modelo NB-B evidenciaram um aumento marginal na acurácia, atingindo um valor de 93%. Assim como o modelo treinado com os dados sintéticos, o modelo exibiu uma alta precisão de 93% ao classificar tweets como não ofensivos (classe 0), refletindo sua capacidade de minimizar a ocorrência de falsos positivos, ou seja, classificações incorretas de tweets inofensivos como discursos de ódio. Em particular, o recall para essa classe alcançou o valor máximo de 1.00, o que indica que não houve ocorrência de falsos positivos nessa categoria.

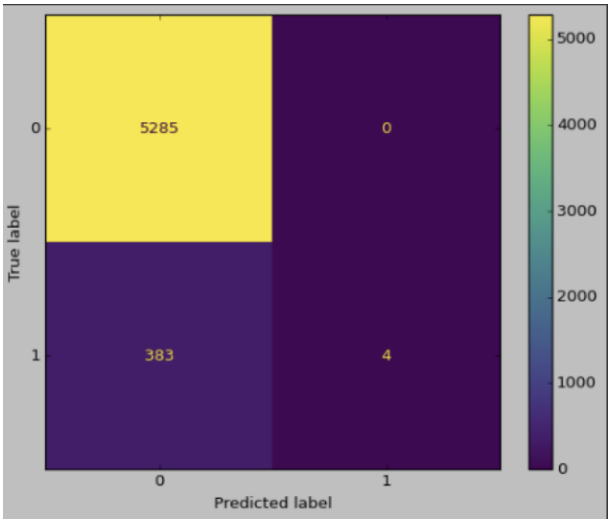


Figura 19 – Matriz de confusão do modelo NB-B

Entretanto, a métrica de recall para a classe de discursos de ódio (classe 1) revelou um valor significativamente mais baixo que o modelo NB-A, totalizando apenas 0.01, indicando que o modelo apresentou grandes dificuldades em identificar e recuperar adequadamente os tweets realmente ofensivos, resultando em uma alta taxa de falsos negativos. Igualmente aos outros modelos, estes resultados podem estar relacionados com o número limitado de amostras de tweets ofensivos quando comparado ao total de amostras de tweets não ofensivos do dataset original.

	precision	recall	f1-score	support
0	0.93	1.00	0.97	5285
1	1.00	0.01	0.02	387
accuracy			0.93	5672
macro avg	0.97	0.51	0.49	5672
weighted avg	0.94	0.93	0.90	5672

Figura 20 – Resultados obtidos pelo modelo NB-B nas métricas de avaliação.

Em resumo, na tabela abaixo é apresentada uma comparação dos escores das métricas de avaliação obtidos pelos três modelos distintos utilizados na tarefa de identificação e classificação de discursos de ódio. Esses modelos foram treinados em dois cenários diferentes: utilizando dados sintéticos gerados por meio da aplicação de data augmentation (SMOTE) e utilizando os dados originais.

	Accuracy	Precision	Recall	F1-Score
SVM-A	0.94	0.96	0.19	0.31
SVM-B	0.94	1.00	0.10	0.18
LR-A	0.95	0.68	0.60	0.64
LR-B	0.93	1.00	0.03	0.06
NB-A	0.90	0.40	0.88	0.55
NB-B	0.93	1.00	0.01	0.02

Tabela 1 – Comparação dos resultados das métricas de avaliação obtidos pelos três modelos nos dois cenários de teste.

A análise comparativa dos resultados obtidos pelos três modelos em avaliação, destaca-se o desempenho do modelo NB-A, treinado com o algoritmo Naive Bayes, na tarefa de classificação em questão, particularmente em relação à métrica de recall. Apesar de todos os modelos terem alcançado acurácias consideráveis, a ênfase no valor de recall é justificada pela sua capacidade de refletir o quão eficaz o modelo é na identificação precisa e classificação de conteúdos ofensivos presentes nos tweets.

O modelo NB-A demonstrou um desempenho superior ao alcançar um recall de 88%, destacando sua habilidade em identificar a grande maioria das ocorrências de discurso de ódio, seguido pelo modelo LR-A, treinado com Regressão Logística, que obteve um recall de 60%. Porém,, ao avaliar as demais métricas, em especial o F1-score, que combina precisão e recall, o modelo LR-A se sobressai, já que suas métricas indicam uma capacidade mais equilibrada em identificar tanto as instâncias positivas quanto as negativas de discurso de ódio. Assim, apesar do alto score de recall do modelo NB-A, o modelo LR-A oferece um desempenho global superior na detecção de discursos de ódio devido à sua capacidade de equilibrar precisão e recall de forma mais eficaz.

No entanto, mesmo com um altos scores de recall e f1-score, uma análise minuciosa da matriz de confusão de ambos modelos revela a necessidade de aprimorar sua capacidade de classificação. A dificuldade em identificar esses padrões e realizar categorizações mais precisas é evidenciada pela ocorrência de falsos negativos, isto é, a classificação incorreta de tweets ofensivos como não ofensivos, além da presença significativa de falsos positivos.

Como mencionado anteriormente, os três modelos não apresentaram grandes disparidades em relação a suas métricas de acurácia, mas mostraram diferenças significativas se tratando de recall e F1-score. Para visualizar melhor os dados obtidos, foi utilizado tambem a Curva ROC (Receiver Operating Characteristic). A Curva ROC é uma representação bidimensional que ilustra como um modelo é capaz de discriminar entre duas classes, representando graficamente a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos (especificidade) em pontos de corte ou limiares de classificação. Quanto mais próxima a curva estiver do canto superior esquerdo do gráfico, melhor será o desempenho do modelo, indicando maior capacidade de distinguir corretamente entre

as duas classes. A área sob a Curva ROC (AUC) é uma métrica resumida usada para quantificar a eficácia global do modelo, onde um valor de AUC próximo a 1 indica um desempenho excelente, enquanto um valor próximo a 0,5 sugere um desempenho similar ao de um classificador aleatório.

[h]

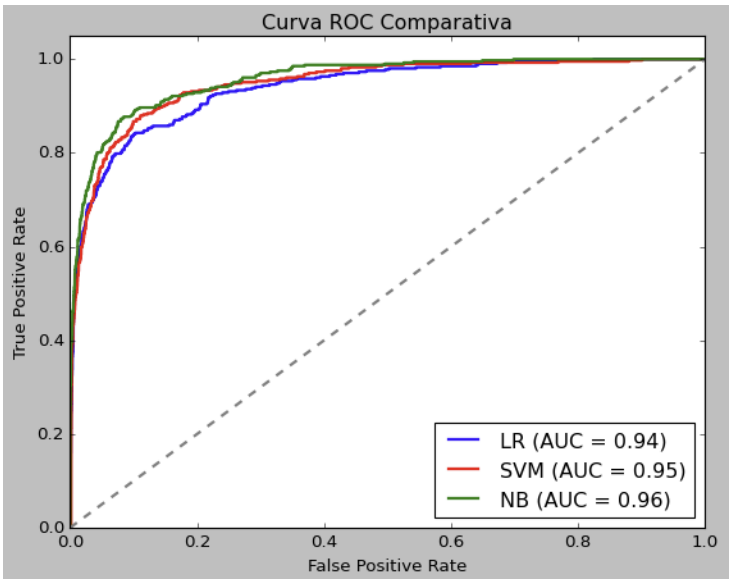
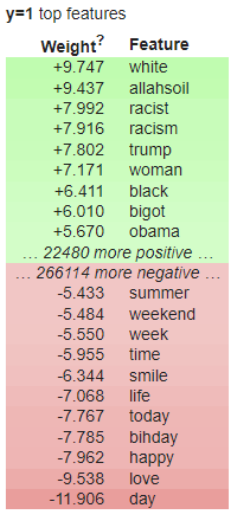


Figura 21 – Comparativo dos valores de AUC obtidos pelos modelos treinados com amostras sintéticas

Na figura acima, uma comparação das métricas de Área sob a Curva ROC (AUC) obtidas pelos três modelos empregados neste estudo é apresentada. Os valores de AUC de 0,94 para LR-A, 0,96 para NB-A e 0,95 para SVM-A, revelam que todos os modelos possuem capacidade de discriminar com certa precisão entre discursos de ódio e conteúdos não ofensivo. Os valores de AUC próximos de 1 indicam que os modelos conseguem alcançar boas taxas de verdadeiros positivos (discursos de ódio identificados corretamente) enquanto mantêm taxas relativamente baixas de falsos positivos (conteúdo não ofensivo erroneamente classificado como discurso de ódio). É importante ressaltar que o modelo com desempenho ligeiramente superior em términos de AUC, modelo NB-A, reforça seus bons resultados obtidos em termos de recall e sua capacidade de identificar corretamente amostras contendo discursos de ódio.

Para aprimorar a compreensão dos mecanismos subjacentes aos modelos de detecção de discursos de ódio apresentados, adotou-se uma abordagem elucidativa, focalizando exclusivamente nos modelos treinados com dados sintéticos. Esta escolha se fundamenta nos resultados superiores obtidos por esses modelos em comparação aos modelos treinados exclusivamente com os dados originais. As análises e explicações subsequentes se concentrarão em compreender as decisões e características subjacentes desses modelos, delineando um quadro mais esclarecedor das capacidades e limitações dos sistemas de detecção de discursos de ódio.

A aplicação do algoritmo ELI5 (Explain Like I'm 5) no âmbito deste estudo científico proporcionou uma ferramenta de interpretabilidade para o entendimento dos resultados obtidos pelo modelo LR-A, na tarefa proposta. O ELI5 opera como uma técnica de análise pós-processamento que visa explicitar as relações de peso entre recursos (features) e as decisões de classificação do modelo. Os resultados revelam a relevância discriminante de certas palavras ou termos específicos na determinação da presença de discursos de ódio, como se pode observar na imagem a seguir:



y=1 top features

Weight[?]	Feature
+9.747	white
+9.437	allahsoil
+7.992	racist
+7.916	racism
+7.802	trump
+7.171	woman
+6.411	black
+6.010	bigot
+5.670	obama
… 22480 more positive …	
… 266114 more negative …	
-5.433	summer
-5.484	weekend
-5.550	week
-5.955	time
-6.344	smile
-7.068	life
-7.767	today
-7.785	bihday
-7.962	happy
-9.538	love
-11.906	day

[h]

Figura 22 – Análise de importância das 20 principais features identificadas pelo modelo LR-A para a classicação de discursos de ódio com ELI5

Os recursos identificados com os pesos mais elevados positivos, como as palavras "white,allahsoil,racist,"e outros, destacaram-se como indicadores preponderantes da categoria de discurso de ódio. Em contrapartida, recursos com pesos negativos, como "love, happy,"e "life,"denotaram a importância de termos que aludem a conteúdos não ofensivos. Esse nível de interpretabilidade aprofundada promovido pelo ELI5 proporciona uma visão mais detalhada do funcionamento do modelo, contribuindo com uma apreciação mais precisa das nuances lingüísticas e semânticas inerentes à classificação automatizada de conteúdo ofensivo.

Porém, o algoritmo ELI5 não pode ser aplicado diretamente aos outros algoritmos utilizados, como SVM e Naive Bayes, devido à sua natureza intrínseca de interpretação, sendo mais adequado para modelos baseados em coeficientes, como a regressão logística. Esses outros algoritmos, como o SVM, dependem da separação de dados em espaços multidimensionais ou na probabilidade condicional, tornando mais complexa a atribuição direta de pesos interpretáveis a recursos específicos. Portanto, a interpretação proporcionada pelo ELI5, que é baseada na análise de coeficientes, não se estende de maneira direta a esses algoritmos. Assim, recorreu-se à técnica LIME (Local Interpretable Model-Agnostic Explanations), que oferece uma abordagem mais adaptável e geral para interpretar e

comparar os resultados dos três modelos apresentados neste estudo.

O LIME, ou "Local Interpretable Model-Agnostic Explanations,"é uma abordagem em interpretabilidade de modelos de aprendizado de máquina. No contexto de detecção de discursos de ódio, o LIME permite a compreensão detalhada do funcionamento dos modelos complexos, incluindo aqueles baseados em algoritmos como SVM e Naive Bayes. O LIME opera através da criação de modelos locais interpretáveis que explicam as decisões de classificação do modelo principal em nível individual para instâncias de dados específicas. Assim como o ELI5, o LIME será aplicado apenas nos modelos treinados com amostras sintéticas (modelos A).

Considerando o tweet extraído do conjunto de teste com o conteúdo original "#zaitoon wishes you father's"day #chennai", observa-se que, à primeira vista, essa sentença pode ser prontamente interpretada como uma expressão de conteúdo não ofensivo por um observador humano (em contexto, "Zaitoon"parece se referir ao nome de um restaurante situado na cidade de Chennai, no sul da Índia). No entanto, esta tarefa de classificação apresenta maior complexidade para máquinas.Assim, a aplicação da técnica LIME permite uma análise mais detalhada das decisões do modelo em relação à classificação deste tweet como discurso de ódio ou não. Nas figuras abaixo se pode observar os resultados obtidos pelo LIME na interpretação da classificação, pelos três modelos distintos, do tweet analisado:

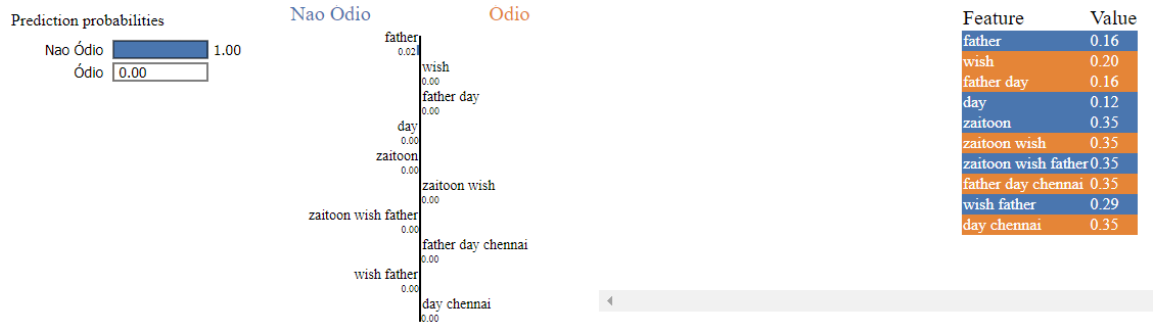


Figura 23 – Interpretação da classificação do Modelo SVM-A com o Auxílio do LIME

Figura 24 – Interpretação da classificação do Modelo LR-A com o Auxílio do LIME

Figura 25 – Interpretação da classificação do Modelo NB-A com o Auxílio do LIME

Conforme observado, os três modelos - SVM-A, LR-A e NB-A - concordaram na classificação do tweet como conteúdo não ofensivo, atribuindo probabilidades de 100%, 93% e 86%, respectivamente, a essa classificação. Para uma análise mais detalhada, os termos foram coloridos de acordo com sua associação provável à classificação, levando em consideração a vetorização dos termos com n-gram igual a 3, o que permite a combinação de até três termos das palavras que compõem a frase em análise. Os termos em cor laranja indicam uma associação mais forte com a classificação de "ÓDIO", enquanto aqueles destacados em cor azul sugerem uma associação com a classificação de "não ódio". Portanto, palavras como "day"e "father,"por exemplo, estão mais fortemente associadas ao conteúdo classificado como "não ódio", enquanto palavras como "chennai"e "zaitoon"sugerem uma associação com o conteúdo classificado como "ÓDIO". Essa análise granular ressalta a importância desses termos específicos na tomada de decisão do modelo. Uma possível explicação para a tendência de associar as palavras "chennai"e "zaitoon"com discursos de ódio, ainda que ao parecer apenas fazem referência a nomes de cidade e restaurante, respectivamente, pode ser atribuída ao fato de que essas palavras são menos comuns e, portanto, o modelo não foi exposto a um número significativo de exemplos que permitiriam compreender seu significado e os contextos em que essas palavras foram aplicadas. Em contraste, termos como "day"e "father"são mais frequentes e provavelmente ocorrem com maior frequência no conjunto de dados utilizado neste estudo. Isso sugere que o modelo tem uma familiaridade maior com o uso dessas palavras comuns, o que pode explicar sua forte associação com conteúdo não ofensivo. Apesar dessa associação aparentemente equivocada, destaca-se que o modelo realizou a classificação correta do tweet em questão, considerando-o livre de conteúdo ofensivo e evidenciando sua capacidade de realizar um discernimento preciso.. Essa análise reforça também a importância da disponibilidade de dados representativos para o treinamento de modelos para a detecção de discursos de ódio, especialmente quando se lida com termos menos frequentes ou contextos específicos.

É crucial ressaltar que a classificação de tweets ou comentários de ódio em redes sociais é uma tarefa desafiadora, pois muitos desses discursos não são escritos de forma direta e clara e podem apresentar palavras ou expressões pouco comuns, como visto

no exemplo acima. Ao contrário, eles frequentemente envolvem nuances e ambiguidades que exigem a consideração cuidadosa do contexto para uma classificação precisa. A linguagem utilizada em redes sociais pode ser altamente variável e subjetiva, tornando difícil identificar automaticamente discursos de ódio. Além disso, os discursos ofensivos muitas vezes se escondem por trás de linguagem codificada, sarcasmo ou ironia, o que pode ser dificíl para um modelo identificar, levando a interpretações equivocadas se não for devidamente analisado. A falta de contexto adequado pode levar a classificações incorretas, seja classificando um discurso inofensivo como ofensivo ou, o que é ainda mais problemático, deixando passar discursos de ódio genuínos. Essas nuances tornam essencial o desenvolvimento de modelos de classificação que considerem o contexto mais amplo de cada mensagem para uma tomada de decisão mais precisa.

Essa limitação é um aspecto crucial a ser abordado em futuras melhorias do modelo, uma vez que a detecção precisa de discursos de ódio é essencial para a implementação eficaz de ações preventivas e de combate a esse tipo de conteúdo nas redes sociais.

6 CONCLUSÕES

Com base na tarefa de identificação de discursos de ódio em tweets e nos scores de avaliação obtidos, os modelos alcançaram valores elevados de precisão, o que é considerado satisfatório para tarefas de classificação. No entanto, uma análise detalhada dos resultados revelou uma discrepância significativa no desempenho dos modelos na classificação de tweets entre as duas classes - discursos de ódio e não ofensivos.

Os modelos demonstraram um desempenho superior na classificação de tweets não ofensivos (classe 0), indicando sua habilidade em identificar corretamente a maioria desses tweets. No entanto, eles enfrentaram dificuldades na identificação e categorização de tweets com conteúdo ofensivo (classe 1), resultando em altas taxas de falsos negativos, ou seja, tweets ofensivos classificados erroneamente como não ofensivos.

Além disso, os resultados evidenciam que os modelos treinados mediante a aplicação da técnica de data augmentation obtiveram desempenhos superiores em comparação aos modelos instruídos exclusivamente com dados originais. Cumpre, entretanto, salientar que os dados sintéticos dificilmente podem servir como substitutos de dados reais, visto que apresentam limitações, tais como pouca variabilidade ou uma representação insuficiente do conjunto total de dados (Shorten, Khoshgoftaar and Furht (2021)).

Dado o objetivo central deste estudo, que é identificar e combater discursos de ódio nas redes sociais, é crucial realizar ajustes e melhorias nos modelos para aperfeiçoar sua capacidade de identificar corretamente tweets ofensivos. Para isso, podem ser exploradas estratégias como técnicas de pré-processamento mais adequadas para dados textuais, outras abordagens de balanceamento de classes para lidar com o desbalanceamento presente nos dados e a investigação de diferentes arquiteturas de modelos, incluindo arquiteturas de modelos baseadas em embeddings pré-treinados em corpus maiores para tarefas específicas de classificação de texto. Essas ações visam melhorar o desempenho dos modelos e aumentar sua sensibilidade na detecção de discursos de ódio, contribuindo para uma abordagem mais efetiva no combate a essa problemática nas plataformas de mídias sociais.

Uma desvantagem associada ao uso de algoritmos de classificação em tarefas de identificação de discursos de ódio é a exigência de dados rotulados para o treinamento. Embora haja uma ampla quantidade de dados textuais disponíveis atualmente para pesquisa, a rotulagem manual desses dados demanda um esforço considerável, o que pode ser difícil de obter. Também é importante destacar que, mesmo quando há uma quantidade relativamente grande de dados rotulados disponíveis, não há garantia de que esses dados sejam verdadeiramente representativos do problema real que está sendo estudado. No contexto deste estudo, a quantidade de dados rotulados pode não ter sido adequada para

realizar um treinamento eficaz do modelo, o que pode ter impactado negativamente na precisão dos resultados e na capacidade de representação precisa do problema em análise (Zhu *et al.* (2009)). Em base a esta observação, seria interessante explorar abordagens de aprendizado automático semi ou não supervisionado, eliminando assim a necessidade de depender exclusivamente de uma grande quantidade de dados rotulados.

É importante mencionar que limitações técnicas relacionadas ao custo computacional para processamento dos textos e dos modelos de classificação também foram identificadas como fatores limitantes neste estudo. O processamento de texto demanda recursos computacionais significativos, especialmente quando lidamos com grandes volumes de dados textuais. O aumento no tamanho do conjunto de dados também pode resultar em tempos de treinamento mais longos e inviáveis em termos de custo computacional e tempo de execução. Para mitigar essas limitações, em estudos futuros pode-se considerar abordagens como o uso de técnicas de redução de dimensionalidade, amostragem estratificada e otimização de hiperparâmetros para obter um equilíbrio entre o custo computacional e a capacidade de classificação precisa. O investimento em infraestrutura de computação, como o uso de ambientes de computação em nuvem ou clusters de alto desempenho, também pode ser uma alternativa para enfrentar os desafios associados ao processamento de grandes volumes de dados textuais e treinamento de modelos de classificação mais complexos. A consideração cuidadosa dessas limitações é essencial ao lidar com projetos que envolvam análise de textos em escala, visando alcançar um equilíbrio entre desempenho e custo computacional.

É pertinente destacar que a identificação precisa de discursos de ódio é um desafio significativo devido à sua natureza subjetiva e evolutiva. Dessa forma, o desenvolvimento de um modelo robusto requer uma análise profunda do contexto, a seleção criteriosa de recursos e a constante adaptação do algoritmo para enfrentar os desafios intrínsecos ao tipo de dado em questão. A colaboração entre especialistas em linguística e profissionais de ciências sociais, juntamente a especialistas em aprendizado de máquina pode ser importante para abordar de forma adequada os desafios inerentes à classificação de discursos de ódio em ambientes online. Tais aprimoramentos são de suma importância para assegurar que os modelos contribuam eficazmente para a mitigação de discursos de ódio nas redes sociais e para promover um ambiente digital mais seguro e respeitoso.

REFERÊNCIAS

AGARWAL, R. **Twitter Hate Speech**. 2018. Available at: <https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech>.

FELDMAN, R. Techniques and applications for sentiment analysis. **Communications of the ACM**, ACM New York, NY, USA, v. 56, n. 4, p. 82–89, 2013.

FENG, S. Y. *et al.* A survey of data augmentation approaches for nlp. **arXiv preprint arXiv:2105.03075**, 2021.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data: an introduction to cluster analysis**. [*S.l.: s.n.*]: John Wiley & Sons, 2009.

KOWSARI, J. M.; HEIDARYSAFA, M. Barnes, and brown,". **Text classification algorithms: A survey," Information**, v. 10, n. 4, p. 150, 2019.

MAGU, R.; JOSHI, K.; LUO, J. Detecting the hate code on social media. *In*: **Proceedings of the international AAAI conference on web and social media**. [*S.l.: s.n.*], 2017. v. 11, n. 1, p. 608–611.

MALMASI, S.; ZAMPIERI, M. Detecting hate speech in social media. **arXiv preprint arXiv:1712.06427**, 2017.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007.

ORGANIZATION for Security and Cooperation in Europe (OSCE) ODIHR - Hate Crime Reporting. 2020. Available at: <https://hatecrime.osce.org/>.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. **Revista de Sistemas de Informacao da FSMA n**, v. 7, p. 7–21, 2011.

SHORTEN, C.; KHOSHGOFTAAR, T. M.; FURHT, B. Text data augmentation for deep learning. **Journal of big Data**, Springer, v. 8, p. 1–34, 2021.

SILVA, A.; ROMAN, N. Hate speech detection in portuguese with naïve bayes, svm, mlp and logistic regression. *In*: **Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional**. Porto Alegre, RS, Brasil: SBC, 2020. p. 1–12. ISSN 2763-9061. Available at: <https://sol.sbc.org.br/index.php/eniac/article/view/12112>.

SILVA, A. d. S. R. d.; ROMAN, N. T. Estudo de modelos distribucionais para detecção de discurso de ódio em português. 2021.

SINOARA, R. A.; MARCACINI, R. M.; REZENDE, S. O. Mineraçao de textos e semântica: desafios, abordagens e aplicaçoes. **Revista de Sistemas de Informaçao da FSMA**, v. 27, n. ja-ju 2021, p. 41–53, 2021.

ZHU, X. *et al.* Synthesis lectures on artificial intelligence and machine learning. *In*: **Introduction to Semi-Supervised Learning**. [*S.l.: s.n.*]: Morgan & Claypool, 2009. v. 3, n. 1, p. 1–130.