Detecção de Discurso de Ódio nas Redes Sociais: Uma Abordagem de Extração de Padrões com Algoritmo Apriori

Gabriel Ikaro Fonseca de Paiva[1], **Hugo Baraky**[1], **João Vítor Vaz**[1], **Victor Hugo Martins**[1], **Vitória Maria Silva Bispo**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

{ikarocomk, hugoparreiras, vitorvaz, victorhm, vitoriamsbispo}@ufmg.br

Abstract. *The rise of hate speech on social media is a growing concern, especially in Portuguese, where research and resources are limited. This paper addresses this challenge by focusing on the automatic identification of hate speech on the social media platform X. It utilizes Natural Language Processing (NLP) techniques, the Apriori algorithm, and classification algorithms. Through text analysis, the study uncovers linguistic patterns that help categorize messages into specific groups, such as "Homophobia" , "Racism" , and "Xenophobia" , achieving an accuracy rate of up to 78%. This research aims to contribute to the development of more effective tools to mitigate the spread of online hate speech, promoting a safer and more tolerant digital environment.*

Resumo. *A proliferação de discurso de ódio nas redes sociais representa um desafio crescente, especialmente em português, onde ainda faltam pesquisas e recursos dedicados ao tema. Este artigo se dedica à identificação automática desse tipo de discurso na plataforma X, utilizando técnicas de Processamento de Linguagem Natural, algoritmo Apriori e algoritmos de classificação. A partir da análise de textos, o estudo identifica padrões linguísticos que permitem categorizar mensagens em classes específicas como "Homofobia" , "Racismo" e "Xenofobia" , com uma acurácia de até 78%. Essa pesquisa contribui para o desenvolvimento de ferramentas mais eficazes no combate à proliferação do discurso de ódio online, buscando um ambiente digital mais seguro e tolerante.*

1. Introdução

Dados estatísticos recentes evidenciam crescimentos expressivos na utilização das redes sociais ao longo dos últimos anos. Em 2023, o Brasil contava com 171 milhões de usuários de redes sociais, e estima-se que esse número aumente para 180 milhões até o fim de 2024 [1]. Esse crescimento é impulsionado por fatores como o aumento do acesso à internet e a popularização de dispositivos móveis [2]. Desenvolvidas com o objetivo de promover a comunicação e o compartilhamento de informações, as redes sociais desempenham um papel fundamental no cotidiano de pessoas em todo o mundo. No entanto, essas plataformas também têm sido cenário para a disseminação de discursos de ódio [3].

O discurso de ódio, definido como insulto, intimidação ou assédio contra indivíduos devido à sua raça, cor, etnia, nacionalidade, sexo ou religião [4], é um problema crescente no Brasil. Segundo o IBGE, 3.924.763 pessoas com 18 anos ou mais foram

ameaçadas, ofendidas, xingadas ou tiveram suas mensagens expostas sem consentimento nas redes sociais nos últimos 12 meses [5].

Esse crescimento pode ser atribuído, em parte, à sensação de anonimato que as redes sociais proporcionam, além da percepção equivocada de que tais discursos estão protegidos pelo direito à liberdade de expressão [6]. Contudo, é fundamental ressaltar que o discurso de ódio representa um sério problema social, frequentemente ligado a casos de *cyberbullying*, ou *bullying* virtual, que podem ter consequências devastadoras para as vítimas, como danos psicológicos, violência e exclusão social, econômica e política [3].

A detecção de discurso de ódio é uma questão emergente nas redes sociais, e técnicas computacionais avançadas são essenciais para detectar e monitorar esses conteúdos nocivos, visando reduzir a proliferação de comentários de ódio. Diversas pesquisas, como [7], [8] e [9] têm investigado estratégias para essa tarefa, incluindo a classificação de diferentes formas de discursos de ódio como racismo, sexismo, homofobia e xenofobia. Compreender os padrões e as formas de manifestação desses discursos é crucial para desenvolver abordagens eficazes que abrangem todas essas categorias de discurso de ódio.

Apesar da abundância de discursos de ódio nas redes sociais, as pesquisas focadas em dados brasileiros são limitadas. Consequentemente, as bases de dados disponíveis não são abrangentes, o que representa um grande desafio para pesquisadores da área. Dessa forma, este estudo se justifica pela necessidade de preencher a lacuna existente na literatura e contribuir para potenciais impactos em diferentes áreas, desde o desenvolvimento de tecnologias de detecção mais precisas até a formulação de políticas públicas mais eficazes para a promoção de um ambiente digital mais seguro e respeitoso.

Este estudo propõe realizar uma análise de discursos de ódio em português no X[1] - anteriormente chamado de Twitter, por meio de uma combinação de métodos computacionais. Dados coletados dessa rede social serão utilizados para essa análise, com o objetivo de identificar padrões e características específicas desses discursos. A análise será conduzida utilizando o algoritmo Apriori para descoberta de conjuntos de itens frequentes e os métodos *Naive Bayes* (NB) e Regressão Logística (RL) para classificação.

Assim, este trabalho tem como objetivo principal a investigação de padrões de discurso de ódio em dados do X na língua portuguesa, com o intuito de compreender melhor sua dinâmica e identificar padrões. Os objetivos específicos incluem o pré-processamento e limpeza da base de dados, a identificação de padrões através da mineração de conjuntos de itens frequentes nos dados, e a classificação de novos textos de acordo com as categorias de insultos identificadas. Com isso, pretende-se contribuir para o desenvolvimento de ferramentas mais eficazes para o monitoramento e a moderação de discursos online na língua portuguesa.

2. Trabalhos Relacionados

Identificar discursos de ódio em plataformas online tem sido objeto de diversos estudos, especialmente no X. Cada um desses estudos utiliza diferentes estratégias, técnicas e algoritmos para alcançar seus objetivos. Entre as abordagens mais comuns estão a análise

[1] `https://x.com/`

semântica e sintática do conteúdo, o uso de técnicas de mineração de dados e a aplicação de algoritmos de aprendizado de máquina.

O estudo conduzido por [10] busca aplicar técnicas de aprendizado preditivo para classificar publicações na plataforma X, com o objetivo de identificar se as publicações contêm discursos de ódio. A pesquisa abrangeu textos em português e em inglês, que foram classificados em três categorias: "discurso de ódio" , "ofensivo" e "regular" . Para a análise comparativa, o autor aplicou três métodos de classificação: SVM, *Naive Bayes* e Regressão Logística. Os experimentos foram realizados com diversas combinações de N-gramas, vetorização e *stemming* para cada conjunto de dados. Os resultados indicam que a vetorização e o *stemming* não afetaram significativamente o desempenho dos classificadores. Entre os métodos analisados, o classificador SVM obteve os melhores resultados.

Em relação ao estudo de [11], foi empregada uma abordagem integrada que combinou o Processamento de Linguagem Natural (PLN) e o algoritmo Apriori para decifrar padrões explícitos e implícitos nas transações dos clientes. A análise se estendeu além das *IDs* de transação, incorporando também as avaliações dos clientes. Uma técnica de análise de sentimento foi aplicada a essas avaliações, categorizando as palavras como positivas, negativas ou neutras. Isso proporcionou esclarecimentos sobre produtos implícitos que, apesar de menos frequentes, impactam significativamente na satisfação do cliente. O algoritmo Apriori permitiu a identificação de conjuntos frequentes de itens comprados juntos e a geração de regras de associação. Essas regras foram posteriormente enriquecidas com os resultados da análise de sentimentos para revelar relações implícitas, oferecendo uma visão mais completa do comportamento de compra dos clientes.

Em paralelo, [12] realizou a aplicação do algoritmo Apriori para minerar regras de associação em um conjunto de dados composto por 8.275 publicações postados no X por usuários malaios. Diferentemente de outras abordagens, este estudo não incluiu metadados na análise. As métricas utilizadas para avaliar os resultados foram suporte, confiança e *lift*, que mediram a frequência de ocorrência das palavras-chave e a validade e importância das regras de associação geradas, respectivamente. Os resultados demonstraram a eficácia do algoritmo Apriori na identificação de padrões de *cyberbullying* no X, sendo capaz de identificar 88 regras de associação com altos níveis de suporte e confiança, indicando várias sequências de co-ocorrência de palavras frequentemente associadas ao contexto de *cyberbullying*.

Por fim, o estudo de [13] investiga a aplicação de técnicas de aprendizado de máquina para a classificação de documentos textuais. O autor utiliza o algoritmo Apriori para minerar regras de associação em diferentes categorias de texto. Para validar a relevância dessas regras extraídas, os autores empregaram o método *Naive Bayes* classificando novos documentos com base nas regras geradas. A classificação é realizada de modo que, quanto maior a correspondência entre as palavras do texto e o conjunto de palavras das regras, maior a probabilidade de o texto pertencer àquela categoria. O autor destaca a importância de uma base de dados grande e diversificada, uma vez que esses atributos ajudam a reduzir as possibilidades de falhas na classificação dos documentos.

De modo geral, é possível perceber a variedade de abordagens que possibilitam a detecção de discursos de ódio em redes sociais. Os trabalhos apresentados abrangem métodos de aprendizado preditivo e descritivo e, além disso, apresentam abordagens que

utilizam ambos os métodos para resolver problemas. O trabalho desenvolvido por [10] trata da classificação de publicações através da Regressão Logística, um método de aprendizagem preditiva. Os trabalhos [11] e [12] visam descobrir padrões em textos para compreender comportamentos específicos. Finalmente, o trabalho [13] possui uma metodologia que integra as abordagens mencionadas anteriormente, utilizando os padrões identificados pelo algoritmo Apriori para classificar outros textos com base nesses padrões. A abordagem de [13] é particularmente relevante, pois demonstra uma aplicação prática dos padrões extraídos e destaca a qualidade da metodologia ao possibilitar a classificação de novos documentos.

3. Desenvolvimento

Baseado na pesquisa de [13], este trabalho encontra-se organizado em quatro etapas principais, que serão discutidas ao longo das próximas seções.

3.1. Base de Dados

O conjunto de dados utilizado neste estudo, composto por 21.000 publicações, foi adquirido a partir do trabalho conduzido por [14]. A coleta desses dados teve o intuito de aumentar a probabilidade de capturar publicações com conteúdo tóxico, uma vez que essas representam uma pequena porção do total de postagens nas redes sociais. Inicialmente, o foco estava associado a publicações que continham palavras-chave ou *hashtags* específicas, notadamente associadas a discursos tóxicos no contexto brasileiro da plataforma X. Esta busca incluiu termos pejorativos e ofensivos frequentemente utilizados em contextos de homofobia, machismo e xenofobia. Ademais, a coleta de dados também se voltou para publicações que mencionavam usuários influentes, como o então presidente do Brasil e jogadores de futebol famosos, indivíduos que são frequentemente alvos de discussões inflamadas relacionadas ao tema. Essa estratégia permitiu ampliar o escopo dos dados coletados, possibilitando a captura de uma variedade mais extensa de conteúdos potencialmente tóxicos.

Os dados coletados foram categorizadas manualmente em seis classes distintas: homofobia, obscenos, insultos, racismo, misoginia e xenofobia. Esse processo de classificação foi realizado por voluntários de uma universidade brasileira, selecionados criteriosamente para assegurar a diversidade demográfica e minimizar possíveis vieses [14]. Cada publicação recebeu três anotações independentes, reforçando a robustez e a confiabilidade do conjunto de dados.

A tabela 1 detalha as frequências de palavras ofensivas e preconceituosas encontradas no conjunto de dados de comentários online proposto por [14]. Cada célula exibe a palavra, seguida pelo número de ocorrências entre parênteses.

Tabela 1. Frequência de palavras ofensivas e preconceituosas

Homofobia	Obscenos	Insultos	Racismo	Misoginia	Xenofobia
viado (59)	porra (332)	puta (221)	nego (6)	putinha (38)	sulista (12)
boiola (15)	caralho (317)	caralho (150)	branco (6)	puta (22)	carioca (7)
viadinho (13)	puta (268)	cara (135)	preto (4)	piranha (19)	fala (4)
sapatão (12)	tomar (136)	porra (122)	nada (4)	mulher (11)	paulista (4)
caralho (11)	fuder (98)	lixo (101)	negão (3)	vagabunda (11)	gente (3)
cara (10)	cara (94)	filho (92)	cara (3)	quer (8)	nordestino (3)
quer (9)	merda (90)	burro (87)	falando (3)	vaca (8)	todo (3)
homem (9)	mano (87)	tomar (86)	vida (3)	fica (6)	ainda (3)
todo (9)	toma (85)	merda (78)	segue (2)	onde (5)	sendo (2)
bicha (9)	fazer (77)	idiota (76)	página (2)	tudo (5)	dança (2)

Fonte: Produzida por [14].

3.2. Processamento de Linguagem Natural

Nesta etapa, o processo segue uma série de etapas fundamentais para preparar os dados textuais para análise e modelagem, conforme destacado por [15]. A primeira etapa envolve a remoção de informações consideradas irrelevantes, como nomes de usuários, sequências de compartilhamento, símbolos, números, pontuações e *links*, visando minimizar ruídos na base de dados, prática também adotada por [7]. Posteriormente, realizou-se a normalização dos textos, convertendo todas as letras para minúsculas para evitar diferentes representações para a mesma palavra, conforme explicado por [16].

A segunda parte do processo inclui a remoção de *stop words*, tais como artigos, conjunções e preposições, que geralmente não carregam significado relevante para a detecção de discurso de ódio, segundo [17]. Por fim, o processo de tokenização divide as frases pré-processadas em unidades individuais de significado, como palavras ou frases, preparando assim os dados para análise computacional [16]. Esta abordagem detalhada garante a eficácia e precisão da análise de discursos de ódio nos dados coletados.

3.3. Aplicação do Algoritmo Apriori

O algoritmo Apriori é eficaz em encontrar padrões e associações frequentes entre palavras e frases [18], permitindo a detecção precisa de diferentes categorias de discurso de ódio. Para complementar essa análise, os métodos NB e RL são aplicados para classificar os textos conforme as categorias identificadas [19]. Esses métodos possibilitam uma análise detalhada e refinada, ajustando-se às nuances dos dados coletados e aprimorando a categorização dos discursos de ódio.

Com o objetivo de aprimorar a eficácia do algoritmo, realizou-se uma etapa adicional de redução de dimensionalidade do vocabulário. Essa etapa visa otimizar o processamento e a análise dos dados, tornando-os mais manejáveis e propícios à aplicação dos métodos subsequentes. De acordo com [13] e [20], o processo tem início com a vetorização e o cálculo das frequências das palavras presentes nos textos, tanto nos dados classificados como ofensivos quanto nos não ofensivos, convertendo textos em uma representação numérica compacta que facilite as manipulações necessárias.

Ainda segundo [20], a escolha da representação depende das unidades significa-

tivas do texto (semântica lexical), mas, normalmente, um texto é representado como um vetor de pesos de termos, onde T é o conjunto de termos que aparecem em pelo menos um documento, e $(tf_{w,d})$ indica a contribuição do termo t_w para a semântica do documento d. As abordagens variam na definição de termos e no cálculo dos pesos.

Para transformar os textos em vetores numéricos, foi utilizada a técnica TF-IDF (*Term Frequency-Inverse Document Frequency*). Essa técnica pondera a frequência de cada palavra em um texto (TF) pela sua importância em todo o conjunto de dados (IDF), gerando representações vetoriais que capturam a relevância de cada termo para o seu contexto. Nesta abordagem, $(tf_{w,d})$ se refere a quantidade de vezes que a palavra w ocorre no texto t. A frequência (df_w) se refere ao número de frases nos quais a palavra w ocorre pelo menos uma vez. Observa-se que, quanto mais vezes um termo aparece em uma frase, maior a probabilidade de que esse termo seja relevante para o tópico principal abordado. No entanto, quanto mais vezes o termo ocorre nas transações gerais, pior ele as diferencia. A frequência inversa de documentos de uma palavra é baixa se ela ocorrer em muitos textos e mais alta se ocorrer em apenas um deles.

Os dados ofensivos específicos de cada classe foram preparados para a aplicação do algoritmo Apriori, após passarem por um processo de limpeza textual. Este procedimento foi realizado em duas etapas fundamentais: primeiramente, os termos mais frequentes e não relevantes foram identificados no texto. Posteriormente, somente as palavras que são distintivamente ofensivas foram preservadas nos textos. Este processo tem como objetivo realizar a filtragem dos textos de modo a reter apenas as palavras que possuem potencial ofensivo dentro do vocabulário.

Com o intuito de identificar padrões em de uma classe específica, a estratégia usada foi dividir a base de dados entre as palavras frequentes entre esta classe e as palavras frequentes nos demais textos, de forma que, através do *TF-IDF*, remove-se as palavras mais comuns inclusas em todos os textos que não pertençam àquela categoria específica. Em outras palavras, subtrai-se os termos mais comuns em seu complemento na base de dados.

As palavras removidas são consideradas menos úteis para distinguir entre textos ofensivos e seu complemento, pois aparecem frequentemente em ambos os conjuntos e se mostram pouco significativas para a detecção de padrões e posterior. Tal processo ajuda a destacar palavras significativamente discriminativas para o nível de ofensividade de uma frase, potencializando os resultados gerados pelo Apriori e análises subsequentes.

Para fins de exemplificação, a Tabela 2 exibe os 5 termos mais frequentes no subconjunto de dados homofóbicos e os 5 conjuntos de dados mais frequentes em seu complemento, com respectiva frequência normalizada entre parênteses. A intuição é que, apesar de seu complemento apresentar palavras ofensivas, elas não são identificadores próprios de homofobia.

O algoritmo Apriori visa encontrar e armazenar os conjuntos de palavras frequentes nos dados devidamente preparados. Assim como em [12], foram definidos limites mínimos de suporte, confiança e *lift*. Após testes empíricos, estes hiperparâmetros foram definidos, respectivamente, como 0.1, 0.2 e 1.

Abaixo, a Tabela 3 exemplifica os conjuntos de dados de tamanho 2 gerados para os dados do subconjunto Homofobia.

Tabela 2. Termos frequentes na classe homofobia e complemento

Homofobia	Complemento
viado (1)	porra (1)
boiola (0.36)	pqp (0.92)
sapatão (0.32)	pra (0.9)
bicha (0.28)	caralho (0.89)
gay (0.26)	vai (0.64)

Fonte: Produzida pelos autores.

Tabela 3. Conjuntos de itens gerados para a classe homofobia

Conjunto de itens	Suporte (%)
mundo, todo	1.16
ver, pode	1.16
viado, pode	1.16
viado, pq	1.17
viado, quer	1.17

Fonte: Produzida pelos autores.

3.4. Classificação

A mineração dos conjuntos de itens frequentes de cada categoria resultou em uma unificação dos resultados, formando uma única lista com 680 subconjuntos únicos. Esta lista foi utilizada para formar um *one-hot-encode* na lista de publicações originais, indicando quais conjuntos cada uma delas possui ou não.

Os modelos foram projetados com o intuito de identificar qual categoria, entre seis: “Homofobia” , “Obscenos” , “Insultos” , “Racismo” , “Misoginia” ou “Xenofobia” , cada publicação pertence através do uso de técnicas de classificação multiclasse. Dois algoritmos foram empregados nesta tarefa, RL e NB, em ambos é calculada a probabilidade de cada publicação sendo analisado pertencer a cada classe e a previsão final é considerada a classe com maior probabilidade. Estes algoritmos foram escolhidos devido a sua pouca complexidade e baixa demanda computacional, como indicado em [21], o que nos permite julgar o desempenho da aplicação dos subconjuntos em si.

No conjunto de dados, 80% das informações *one-hot-encode* foram utilizadas para treinar os algoritmos, enquanto os 20% restantes foram empregados para testes. Para avaliar os modelos, foram exploradas as seguintes métricas: precisão, que mede a proporção de predições positivas corretas; *recall*, que quantifica o número de casos positivos previstos corretamente pelo modelo; e F1-*Score*, que é a média harmônica entre precisão e *recall*. Além disso, a acurácia foi utilizada como métrica geral do algoritmo. Essas métricas foram aplicadas para análise individual das classes, conforme descrito em [22].

Os testes iniciais foram realizados utilizando todas as categorias presentes na base de dados. Em relação aos algoritmos, foram realizados ajustes minuciosos para encontrar valores para seus hiperparâmetros que produzissem os melhores resultados. As melhores versões alcançadas para cada modelo tiveram desempenhos semelhantes, mostrando

resultados melhores ao se tratar das categorias “Obscenos” e “Insultos” e tendo um desempenho inferior com categorias como Racismo, Misoginia e Xenofobia. Além disso, foram observados desempenhos ligeiramente superiores com o algoritmo RL com 75% de acurácia contra 72% do algoritmo NB.

As tabelas 4 e 5 apresentam resultados detalhados.

Tabela 4. Desempenho do algoritmo de RL em porcentagem.

Classe	Precisão	*Recall*	F1-*Score*
Homofobia	79,1	65,9	71,0
Obscenos	78,1	90,8	84,1
Insultos	65,9	53,1	58,9
Racismo	85,7	15,7	26,6
Misoginia	64,8	29,2	40,3
Xenofobia	63,6	26,9	37,8
Média Pond.	74,4	75,3	73,5

Fonte: Produzida pelos autores.

Tabela 5. Desempenho do algoritmo NB em porcentagem.

Classe	Precisão	*Recall*	F1-*Score*
Homofobia	75,4	67,8	71,4
Obscenos	77,7	87,5	82,3
Insultos	58,9	48,6	53,3
Racismo	60,0	23,7	33,9
Misoginia	45,4	30,5	36,5
Xenofobia	66,6	38,4	48,7
Média Pond.	70,9	72,4	71,1

Fonte: Produzida pelos autores.

Os resultados dos modelos de classificação e dos subconjuntos minerados de cada categoria foram avaliados, indicando que as categorias “Obscenos” e “Insultos” poderiam estar impactando negativamente o desempenho geral dos modelos. Essa situação pode ter ocorrido devido à expressiva quantidade de exemplos presentes nessas categorias e aos conjuntos de itens específicos que são gerados a partir delas. Notou-se que muitos dos itens extraídos dessas duas categorias são palavrões ou termos usados como expletivos, que também aparecem frequentemente em outras categorias. Essa sobreposição leva a uma classificação imprecisa, onde várias postagens que deveriam ser classificadas sob um tipo específico de discurso de ódio são erroneamente classificadas apenas como “Obscenos” ou “Insultos” .

Com base nessas observações, uma nova série de testes foram realizadas, excluindo as categorias “Obscenos” e “Insultos” . Nessa nova análise, foram observadas pequenos avanços no desempenho dos modelos na tarefa de classificar as publicações em cada tipo de discurso de ódio. O algoritmo RL manteve seu desempenho superior, com 78% de acurácia, contra 76% do NB. As métricas específicas de cada classe podem ser analisadas nas Tabelas 6 e 7.

Tabela 6. Desempenho do algoritmo de RL em porcentagem com 4 classes.

Classe	Precisão	*Recall*	F1-*Score*
Homofobia	87,8	82,8	85,2
Racismo	58,3	36,8	45,1
Misoginia	73,8	90,8	81,4
Xenofobia	75	48	58,5
Média Pond.	77,4	77,6	76,5

Fonte: Produzida pelos autores.

Tabela 7. Desempenho do algoritmo NB em porcentagem com 4 classes.

Classe	Precisão	*Recall*	F1-*Score*
Homofobia	83,3	78,5	80,8
Racismo	64,2	47,3	54,5
Misoginia	73,2	85	78,7
Xenofobia	70	56	62,2
Média Pond.	75,5	75,6	75,1

Fonte: Produzida pelos autores.

4. Considerações Finais

Nas próximas seções, serão expostas as considerações finais deste estudo, além de sugestões para direcionamentos de pesquisas futuras.

4.1. Conclusão

Este estudo[1] explorou a detecção de discurso de ódio em português, utilizando técnicas PLN, mineração de subconjuntos de itens frequentes com o algoritmo Apriori, e técnicas de filtragem de vocabulário baseadas em TF-IDF. A análise se concentrou em comentários da plataforma X, demonstrando que a identificação de padrões linguísticos específicos para cada categoria de discurso de ódio pode ser aprimorada ao remover termos genéricos e frequentes em diversas classes.

Os resultados obtidos com classificadores de baixa complexidade, como Regressão Logística e *Naive Bayes*, alcançaram uma acurácia de até 78% na classificação de comentários em categorias como “Homofobia” , “Racismo” , “Misoginia” e “Xenofobia” . No entanto, a pesquisa também revelou um ponto importante: a presença de categorias muito abrangentes pode diluir a precisão da detecção em classes mais específicas, destacando a importância da curadoria e especificidade na construção de modelos de detecção de discurso de ódio.

Houveram limitações, incluindo um volume relativamente limitado e o desequilíbrio entre as categorias de discurso de ódio na base de dados utilizada. Além disso, a utilização de um modelo de classificação “*single-label*” simplifica a análise, ignorando a natureza multifacetada que o discurso de ódio pode assumir em um único comentário.

[1] O algoritmo desenvolvido pelos autores está disponível em: `https://github.com/VictorHugoMartins/deteccao_odio`

Em suma, este estudo representa um passo crucial na luta contra a proliferação de discursos de ódio, pavimentando o caminho para que as redes sociais sejam ambiente de interações sociais seguras.

4.2. Trabalhos Futuros

Esta pesquisa oferece uma contribuição significativa para preencher as lacunas na literatura existente sobre o assunto em discussão. Com o objetivo de expandir o alcance deste estudo, as seguintes oportunidades para futuras pesquisas são identificadas:

- **Construção de um corpus robusto:** A criação de um corpus extensivo e balanceado, dedicado à língua portuguesa e abrangendo diferentes plataformas online, é crucial para o desenvolvimento de modelos mais robustos e generalizáveis. Este corpus deve contemplar a diversidade do discurso de ódio online, incluindo variações regionais, gírias e diferentes comunidades online, bem como uma quantidade superior de literatura, de modo que expanda o estado da arte e possa ser utilizado em combinação com bases já existentes.
- **Abordagem *multilabel*:** O discurso de ódio pode assumir diversas formas em um único comentário. Portanto, a implementação de modelos de classificação *multi-label* permitirá capturar a complexidade do discurso de ódio, reconhecendo que um único comentário pode expressar diferentes tipos de preconceito simultaneamente.
- **Exploração de algoritmos avançados:** A investigação de algoritmos de mineração de padrões mais sofisticados, como SSD++ e Prim, em conjunto com modelos de classificação mais poderosos, como Redes Neurais Profundas, poderá revelar padrões complexos e impulsionar a acurácia da detecção.
- **Análise de diferentes plataformas:** A aplicação da metodologia em outras plataformas online, como comentários em portais de notícias, fóruns de discussão e plataformas de jogos online, permitirá analisar a dinâmica do discurso de ódio em diferentes contextos e comunidades online.

As direções futuras desta pesquisa abrem um leque de possibilidades para o desenvolvimento de ferramentas de detecção de discurso de ódio mais eficazes. Aprofundar o conhecimento nessa área permitirá explorar novas técnicas e abordagens, com potencial para impactar significativamente a forma como o problema é combatido.

Referências

[1] Statista. Number of social media users in brazil from 2020 to 2029, 2024. URL `https://www.statista.com/statistics/278408/number-of-social-network-users-in-brazil/`. Acesso em: 29 de Julho de 2024.

[2] Cetic. Acesso a tecnologias da informção e comunicação (tic), 2023. URL `https://cetic.br/pt/tics/domicilios/2023/domicilios`. Acesso em: 29 de Julho de 2024.

[3] Organização das Nações Unidas. *Informe de Política para a Nossa Agenda Comum: Integridade da Informação nas Plataformas Digitais*. ONU, 2023. URL `https://brasil.un.org/sites/default/files/2023-10/ONU_Integridade_Informacao_Plataformas_Digitais_Informe-Secretario-Geral_2023.pdf`.

[4] Winfried Brugger. Proibição ou proteção do discurso do Ódio? algumas observações sobre o direito alemão e o americano. *Direito Público*, 4(15), 02 2010. URL https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/1418.

[5] Instituto Brasileiro de Geografia e Estatística. Pesquisa nacional de saúde - pns 2019, 2019. URL https://sidra.ibge.gov.br/pesquisa/pns/pns-2019.

[6] Tatiana Stroppa and Walter Claudius Rothenburg. Liberdade de expressÃo e discurso do Ódio: O conflito discursivo nas redes sociais. *Revista Eletrônica do Curso de Direito da UFSM*, 10(2):450–468, 12 2015. doi: 10.5902/1981369419463. URL https://periodicos.ufsm.br/revistadireito/article/view/19463.

[7] João Vítor dos Santos Vaz. Detecção de discursos racistas: uma abordagem baseada em processamento de linguagem natural e aprendizado de máquina. 2024.

[8] Rogers de Pelle and Viviane Moreira. Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, Porto Alegre, RS, Brasil, 2017. SBC. doi: 10.5753/brasnam.2017.3260. URL https://sol.sbc.org.br/index.php/brasnam/article/view/3260.

[9] Rogers Prates De Pelle and Viviane P Moreira. Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC, 2017.

[10] Robson Murilo Ferreira do Nascimento. Classificação automática de discursos de ódio em textos do twitter. B.S. thesis, Brasil, 2019.

[11] T. Velmurugan and B.Hemalatha. Mining implicit and explicit rules for customer data using natural language processing and apriori algorithm. In A. B. Smith-Jones, editor, *Advances in Computer Science*, pages 487–499. International Journal of Advanced Science and Technology, Vol. 29, No. 9s, 2020.

[12] Zuraini Zainol, Sharyar Wani, Puteri N.E. Nohuddin, Wan M.U. Noormanshah, and Syahaneim Marzukhi. Association analysis of cyberbullying on social media using apriori algorithm. *International Journal of Engineering Technology*, 7 (4.29):72–75, 11 2018. doi: 10.14419/ijet.v7i4.29.21847. URL https://www.sciencepubco.com/index.php/ijet/article/view/21847.

[13] Chowdhury Mofizur Rahman, Ferdous Ahmed Sohel, Parvez Naushad, and S M Kamruzzaman. Text classification using the concept of association rule of data mining. *Proc. International Conference on Information Technology*, pages 234–241, 03 2003.

[14] João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-main.91.

[15] Ariel da S. Dias. *Processamento de Linguagem Natural*. Editora Saraiva, e-book edition, 2021. ISBN 9786589881995. Acesso em: 21 jul. 2023.

[16] Rafael Anchiêta, Francisco AR Neto, Jeziel C Marinho, and Raimundo Moura. Pln:

Das técnicas tradicionais aos modelos de deep learning. *Sociedade Brasileira de Computação*, 2021.

[17] Hugo Honda Ferreira. Processamento de linguagem natural e classificação de textos em sistemas modulares. 2019.

[18] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In A. B. Smith-Jones, editor, *Advances in Computer Science*, pages 487–499. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, 1994.

[19] Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*, 2018.

[20] Fabrizio Sebastiani. Machine learning in automated text categorization. *Consiglio Nazionale delle Ricerche*, pages 11–14, 10 2001.

[21] Hao Chen, Susan McKeever, and Susan Jane Delany. *Lecture Notes in Computer Science*, chapter A Comparison of Classical Versus Deep Learning Techniques for Abusive Content Detection on Social Media Sites. Springer, 2018.

[22] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Lecture Notes in Computer Science*, 2005.