UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

DOUGLAS PEREIRA DE LIMA

IDENTIFICAÇÃO AUTOMÁTICA DE TWEETS HOMOFÓBICOS EM PORTUGUÊS

CAMPINA GRANDE - PB

2022

DOUGLAS PEREIRA DE LIMA

IDENTIFICAÇÃO AUTOMÁTICA DE TWEETS HOMOFÓBICOS EM PORTUGUÊS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Cláudio Elízio Calazans Campelo

CAMPINA GRANDE - PB

2022

DOUGLAS PEREIRA DE LIMA

IDENTIFICAÇÃO AUTOMÁTICA DE TWEETS HOMOFÓBICOS EM PORTUGUÊS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Cláudio Elízio Calazans Campelo

Orientador – UASC/CEEI/UFCG

Fabio Jorge Almeida Morais

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 02 de Setembro de 2022.

CAMPINA GRANDE - PB

RESUMO

Discursos de ódio com conteúdo homofóbico são cada dia mais frequentes nas redes sociais. Muitas dessas plataformas, como o Twitter, disponibilizam algumas ferramentas, como a denúncia, para contornar esses problemas, mas não são efetivas. A sociedade precisa se haver de técnicas, que não dependam dessas plataformas, para lidar com esse tipo de violência e garantir que vidas e corpos distintos sejam respeitados. Uma das ações possíveis, em relação a esse problema, é a detecção automática desse conteúdo. Técnicas de aprendizagem de máquina foram criadas para automatizar essa detecção, mas diversos estudos mostram que essas técnicas podem ser refinadas e melhoradas. Com isso, essa pesquisa propõe utilizar técnicas de aprendizagem de máquina para identificar automaticamente discursos de ódio com conteúdo homofóbico em tweets em português. Os resultados mostram que métricas satisfatórias de classificação automática podem ser atingidas e os modelos produzidos tem potencial, de serem utilizados para auxiliar a população LGBTQIA+, na luta contra a violência em redes sociais.

Identificação automática de tweets homofóbicos em português

Trabalho de Conclusão de Curso

Douglas Pereira de Lima (Aluno)
douglas.lima@ccc.ufcg.edu.br
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

Cláudio Campelo (Orientador)
campelo@dsc.ufcg.edu.br
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

RESUMO

Discursos de ódio com conteúdo homofóbico são cada dia mais frequentes nas redes sociais. Muitas dessas plataformas, como o Twitter, disponibilizam algumas ferramentas, como a denúncia, para contornar esses problemas, mas não são efetivas. A sociedade precisa se haver de técnicas, que não dependam dessas plataformas, para lidar com esse tipo de violência e garantir que vidas e corpos distintos sejam respeitados. Uma das ações possíveis, em relação a esse problema, é a detecção automática desse conteúdo. Técnicas de aprendizagem de máquina foram criadas para automatizar essa detecção, mas diversos estudos mostram que essas técnicas podem ser refinadas e melhoradas. Com isso, essa pesquisa propõe utilizar técnicas de aprendizagem de máquina para identificar automaticamente discursos de ódio com conteúdo homofóbico em tweets em português. Os resultados mostram que métricas satisfatórias de classificação automática podem ser atingidas e os modelos produzidos tem potencial, de serem utilizados para auxiliar a população LGBTQIA+, na luta contra a violência em redes sociais.

KEYWORDS

Processamento de Linguagem Natural, Classificação automática, Discurso de ódio, Homofobia, Lgbtfobia

1 INTRODUÇÃO

O uso das redes sociais aumentou muito nos últimos anos e com isso a propagação de discursos de ódio direcionados às minorias sociais também aumentou [4]. No Twitter, o número de postagens com conteúdos preconceituosos apontam para o fato que muitos usuários não sabem a distinção entre exercer a sua liberdade de expressão e atacar alguém de forma criminosa. A população LGBTQIA+ é uma das principais vítimas desse tipo de prática, em 2019, o Supremo Tribunal Federal brasileiro aprovou a criminalização da homofobia e transfobia; no julgamento da tese o plenário considerou que tais práticas serão enquadradas nos crimes de racismo [3]. Essa decisão marcou uma evolução na garantia de direitos básicos da comunidade LGBTQIA+ no Brasil. Mesmo com essa evolução, ainda podemos perceber que grande parte da população brasileira utiliza as redes sociais para promover ódio e preconceito [5].

Por causa desse cenário de grande número de publicações contendo discursos de ódio, muitas redes sociais voltaram a atenção para construção de ferramentas que possibilitam a diminuição, e no futuro a extinção, desse tipo de prática. Uma das principais ferramentas são as denúncias, onde outros usuários podem identificar conteúdos problemáticos em publicações e apontar isso para a plataforma, para que essa remova a publicação e puna o usuário. Mas a dependência de denúncias feitas por outros usuários não é muito efetiva. Uma possível saída para esse problema é a detecção automática de postagens que contenham discurso de ódio utilizando algoritmos. Algo assim pode ser utilizado pela população em geral, ou por algumas instituições, para construção de uma base de dados que pode fundamentar estatísticas, hoje, escassas em relação a esse tipo de prática online.

A detecção de conteúdos em texto é uma técnica dentre as várias inclusas na área de Processamento de Linguagem Natural. Hoje, essa área é bem explorada por pesquisadores, devido às suas aplicações, como chats de atendimento automático em lojas virtuais e robótica. A predição de conteúdos em dados textuais vem evoluindo, conforme os algoritmos de Processamento de Linguagem Natural evoluem. Muitos estudos nos mostram o direcionamento dessas técnicas de detecção e neles podemos perceber que avanços ainda maiores na capacidade de predição são possíveis [1, 2, 8, 9].

Tendo em vista a pouca quantidade de estudos sobre o desenvolvimento de técnicas de classificação para textos em português e a necessidade de construção de ferramentas que auxiliem no combate à homofobia, esse artigo se propõe criar, configurar e avaliar, experimentalmente, modelos que usam algoritmos de aprendizagem de máquina supervisionada para classificar automaticamente um texto como homofóbico ou não homofóbico. O Twitter foi escolhido como objeto desse trabalho por ser umas das redes sociais mais utilizadas no Brasil e, diferente das outras redes sociais mais usadas, o principal formato de interação entre seus usuários é o textual.

Para o treinamento dos modelos, foram coletados tweets publicados no dia da eleição presidencial brasileira de 2018. A escolha desse período se deu pois, acredita-se que pelo perfil dos principais candidatos dessa eleição, a incidência desse tipo de conteúdo foi maior que em outros períodos. De acordo com os indicadores da Central Nacional de Denúncias de Crimes Cibernéticos da ONG SaferNet [7], no mês de outubro de 2018 a ONG recebeu mais que o dobro da quantidade de denúncias que qualquer outro mês de 2018 e essa diferença entre os meses não foi verificada nos anos seguintes, reforçando a possibilidade de que nesse período houveram mais publicações com discurso de ódio.

Além do período, outro fator foi incluído para estabelecer quais tweets seriam coletados, a presença de termos com conotação homofóbica. De acordo com Coutinho "[...] dependendo do contexto,

uma palavra pode ter um sentido ofensivo ou não, assim, a presença de uma palavra não determina o caráter homofóbico de uma mensagem." [2, p. 2]. Considerando a afirmação de Coutinho, os tweets coletados foram rotulados utilizando um classificador de discurso de ódio, aqueles classificados como contendo discurso de ódio foram separados e rotulados manualmente de acordo com a presença de teor homofóbico. Devido a falta de uma base de dados já rotulada disponível para o uso nesse contexto, a rotulagem manual se fez necessária para a construção dos dados de treinamento dos modelos de aprendizagem de máquina produzidos nessa pesquisa.

Ao todo 70 modelos foram criados e muitos apresentaram boas métricas de classificação. Em comparação às pesquisas encontradas nessa área, os resultados de classificação dos modelos foram satisfatórios, apontando para um caminho possível de utilização dessas técnicas na diminuição dos casos de homofobia em redes sociais.

O artigo é disposto da seguinte forma: na Seção 2 temos a fundamentação teórica, que explicitará os conceitos necessários para o entendimento dessa pesquisa; na Seção 3 a metodologia, que mostrará como a pesquisa foi feita; a Seção 4 mostra os resultados da pesquisa e na Seção 5 temos uma discussão conclusiva acerca dos resultados da pesquisa e indicações para possíveis trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Alguns conceitos e procedimentos foram utilizados para a construção dessa pesquisa e devem ser explicitados para a melhor compreensão dela, dentre eles estão: discurso de ódio e preconceito, aspectos do processamento de linguagem natural, algoritmos de aprendizagem de máquina supervisionada e métricas de avaliação de modelos de classificação.

2.1 Discurso de ódio

O discurso de ódio é uma forma de expressão caraterizada por insultar, invalidar ou deteriorar um grupo ou pessoa marcados por aspectos minorizados socialmente. Na última década, discussões acerca do tema se mostraram mais presentes na sociedade, com isso, surgiu a necessidade de estabelecer um conceito do que seria discurso de ódio. Schäfer et al. publicaram um artigo na Revista De Informação Legislativa do Senado Federal Brasileiro que tinha como objetivo construir um conceito normativo para o discurso de ódio, para eles

> [...] discurso de ódio consiste na manifestação de ideias intolerantes, preconceituosas e discriminatórias contra indivíduos ou grupos vulneráveis, com a intenção de ofender-lhes a dignidade e incitar o ódio em razão dos seguintes critérios: idade, sexo, orientação sexual, identidade e expressão de gênero, idioma, religião, identidade cultural, opinião política ou de outra natureza, origem social, posição socioeconômica, nível educacional, condição de migrante, refugiado, repatriado, apátrida ou deslocado interno, deficiência, característica genética, estado de saúde física ou mental, inclusive infectocontagioso, e condição psíquica incapacitante, ou qualquer outra condição. [11, p. 149–150]

O ponto evidente em uma expressão que tem discurso de ódio, é o direcionamento a algo diferente de forma negativa, quem profere esse tipo de discurso, considera alguém ou um grupo distinto de si e inferior. Mas, além disso, o discurso de ódio deve ser distinguindo de uma fala qualquer discriminatória, pois ele é caracterizado por ser direcionado a alguém que faz parte de alguma minoria social, a lgbtfobia, por exemplo, é considerada um tipo de discurso de ódio. Discursos preconceituosos ou discriminatórios dirigidos a pessoas que não fazem parte de alguma minoria, não são considerados discursos de ódio.

2.2 Web Scraping

O Web Scraping ou "raspagem web"é uma técnica de coleta e estruturação de informações, geralmente não estruturadas, contidas em páginas da web para algum uso específico. Esse uso pode ser uma análise de preços de produtos em lojas diferentes, a uma busca extensa sobre um tópico na web.

O script que realiza o scraping, vai buscar informações em elementos HTML ou regras de CSS da página. E utilizando regex, ou outra forma de identificar cadeias de caracteres, ele coleta informações e as estruturam de uma forma que outros sistemas ou scripts possam consumir.

2.3 Processamento de linguagem natural

Processamento de linguagem natural, é uma ramificação da área de inteligência artificial que lida com dados textuais na forma de uso do dia a dia, ou seja, não estruturada para um computador interpretar. Essa área auxilia a preencher a lacuna entre a comunicação humana e o modo que os computadores processam as informações. Muitas tarefas usam técnicas de PLN para alcançar seus objetivos: análise de sentimentos em textos, geração automática de textos, tradução de textos para outros idiomas, construção de chatbots, identificação automática de temas, entre outras tarefas.

2.3.1 Pré Processamento de dados textuais. O pré processamento dos dados é uma etapa essencial na área de PLN, tendo em vista que os dados utilizados são geralmente não estruturados. Esse pré processamento consiste em limpar, organizar e estruturar os textos para que esses atendam às especificações das técnicas utilizadas.

2.3.2 Tf e Tf-idf. Outros dados usados por algoritmos de inteligência artificial, quando estão sendo executados em um contexto textual, ŝão os valores Tf e Tf-IDF. O valor Tf é uma estatística que informa a frequência de um termo em um texto, ou seja, quantas vezes aquela palavra, bigrama ou trigrama, apareceu no texto. Já o TF-IDF é uma estatística, que foi criada com o intuito de penalizar termos que são frequentes em muitos textos, e, por isso, não contribuem a entender o sentido do texto ou a diferencia-los.

2.4 Aprendizagem de máquina supervisionada

De acordo com Monard e Baranauskas:

> Aprendizado de Máquina é uma área de IA cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática.[6]

Por sua vez, a aprendizagem de máquina supervisionada, é um conjunto de técnicas que realiza essa aprendizagem utilizando dados para o treinamento que estão classificados. Vale ressaltar, que

modelos de aprendizagem de máquina supervisionada podem ser empregados em problemas de classificação, onde o objetivo é atribuir uma classe para o dado de entrada, mas também em problemas de regressão, onde o objetivo é atribuir um valor numérico para os dados de entrada.

Dentre vários algoritmos de aprendizagem de máquina supervisionada alguns são mais relevantes para essa pesquisa:

2.4.1 ***Regressão logística***. É um algoritmo que calcula a probabilidade de uma variável dependente ser de uma classe, dado informações de outras variáveis independentes. Nele se utiliza dados que já estão com a suas respectivas probabilidades para o treinamento de modelos. Esses modelos, por sua vez, irão utilizar os valores das variáveis independentes para calcular a probabilidade de um resultado em novos dados, atribuindo ao dado a classe que ele atingiu a maior probabilidade de pertencer.

2.4.2 ***Naive bayes***. Também é um algoritmo que calcula a probabilidade de algo, mas tem como princípio fundamental o Teorema de Bayes e considera sempre que as variáveis que determinam a probabilidade, a ser calculada, sempre são independentes entre si, é dessa característica do algoritmo que vem o termo naive.

2.4.3 ***SVM***. É um algoritmo de classificação que considera os dados como pontos no espaço e constrói uma reta, ou hiperplano, que separa esses dados em dois conjuntos distintos.

2.4.4 ***Árvores de Decisão***. Um algoritmo de classificação que utiliza uma estrutura de árvore para estabelecer qual classe um dado pertence. A classificação funciona seguindo o fluxo da árvore, indo do nó raiz, até um dos nós folha, onde cada nó tem uma condição que dirá, de acordo com alguma informação do dado, para qual nós abaixo na hierarquia o fluxo deve seguir. O fluxo termina ao chegar em algum nó folha, que contém a classe a qual o dado será atribuído.

3 METODOLOGIA

Para a realização desse trabalho foi realizado um conjunto de procedimentos: a coleta de dados, o processamento dos dados coletados, o processo de rotulagem necessário para utilizar as técnicas de aprendizagem de máquina supervisionada, e a construção e avaliação dos modelos.

3.1 Coleta de dados

Para realização deste estudo, coletamos dados da rede social Twitter[1], que foi escolhida por causa do formato de interação de seus usuários, que se dá predominantemente por texto. A coleta dos dados foi realizada utilizando o SNScrape[2], que é um scraper feito em python, especializado em redes sociais. Com essa ferramenta, podemos coletar tweets de qualquer período e outros metadados atrelados a eles: usuário que fez o tweet e informações desse usuário, idioma do tweet, número de likes e retweets, informações de alguma mídia caso o tweet contenha, entre outros. Para esse trabalho foram coletados: conteúdo textual do tweet, o idioma dele, uma variável que indica se o tweet é um retweet ou não e a descrição da biografia (bio) do usuário que o criou.

[1]Twitter - https://twitter.com/
[2]SNScrape - https://github.com/JustAnotherArchivist/snscrape

Foi realizada uma busca com SNScrape dos tweets criados em 7 de outubro de 2018 (dia da eleição presidencial brasileira), que continham pelo menos uma das palavras do conjunto proposto por Pereira [9]. Esse conjunto, é constituído por 14 termos referentes à população LGBTQIA+ e que possivelmente são usados com caráter discriminatório. Desses 14 termos, dois deles não foram usados, 'sapatão' e 'traveco', pois esse trabalho tem como objetivo diminuir o escopo e focar em um dos grupos da população LGBTQIA+, os gays. A Tabela 1 mostra as palavras utilizadas e a quantidade de tweets coletados para cada uma delas.

Tabela 1: Quantidade de tweets coletados para cada palavra.

Palavra	Quantidade
baitola	115
bicha	1966
bichona	60
bixa	1000
boiola	161
gayzada	45
gayzismo	11
homossexualismo	176
viadagem	459
viadao	174
viadinho	743
viado	9718

Nenhum dos tweets coletados era um retweet, logo, nenhum foi removido por esta razão. Porém, o conteúdo textual de todos os tweets passou por uma remoção, de menções e de links, que foi realizada através de um script implementado com Python 3. Dos 14.628 tweets, 9.718 continham a palavra 'viado', ou seja, entre os 12 termos do conjunto, um único termo estava relacionado a aproximadamente 67% dos dados. Por isso, foi criado outro conjunto, executando uma nova coleta, também de tweets criados no dia da eleição presidencial brasileira, mas retirando o termo 'viado' da busca. Essa nova coleta gerou um conjunto de 4.910 tweets. Para melhorar a referência aos dois conjuntos o conjunto com 14.628 será chamado de base1 e o conjunto com 4.910 tweets será chamado de base2.

3.2 Rotulagem

Para a construção de modelos de aprendizagem de máquina supervisionada, dados já classificados devem ser utilizados para o treinamento, mas nenhuma base que continha tweets classificados como homofóbicos, ou não homofóbicos, foi encontrada, por isso, uma rotulagem manual deveria ser feita para a construção da base de treinamento. Antes de iniciar a rotulagem manual, foi feito uma filtragem dos tweets com o objetivo de agilizar esse processo. Um modelo encontrado no Github[3], que apresentava boas métricas de desempenho de classificação de discurso de ódio em dados textuais, foi utilizado para classificar o conteúdo textual dos tweets da base2. Após a classificação, 1.713 dados foram rotulados como contendo

[3]Link para o repositório - https://github.com/DarlanNoetzold/HateSpeech-portuguese

discurso de ódio e separados para compor uma terceira base de dados que será chamada de base3.

A base3 foi usada na rotulagem manual que se deu pela classificação de cada tweet, como homofóbico ou não homofóbico de acordo com o senso comum e com o seguinte conceito de homofobia proposto por Rios [10, 73]:

> "A homofobia, como expressão discriminatória intensa e cotidiana, ocorre sempre que distinções, exclusões, restrições ou preferências anulam ou prejudicam o reconhecimento, o gozo ou o exercício em pé de igualdade de direitos humanos e liberdades fundamentais nos campos econômico, social, cultural ou em qualquer campo da vida pública."

Mesmo contendo termos considerados socialmente como pejorativos à população gay, dependendo do contexto, um tweet pode não conter conteúdo homofóbico. Na Tabela 2, podemos observar dois exemplos de tweets que foram coletados por conter o termo 'bicha'. O primeiro tweet não foi rotulado como homofóbico, nele o uso do termo é feito de forma não agressiva, algo recorrente em redes sociais como o Twitter. Já o segundo, recebeu o rótulo de homofóbico por utilizar o termo 'bicha' de forma ofensiva e pela presença do termo 'bambi', uma palavra conhecida como insulto à alguém gay.

Rótulo	Tweet
Não homofóbico	'eu tbm bicha, amo que é só atravessar a rua pra votar kkkk'
Homofóbico	'Uiiiiiii a bicha tá brava. CHUPA BAMBI.'

Tabela 2: Exemplos da rotulagem dos tweets que contêm o termo 'bicha'

Cada dado utilizado nesse trabalho contém, pelo menos, um termo considerado como ofensa à população gay. Essa característica dos dados, reverbera em um aspecto importante dos modelos que serão criados, a classificação dos posts será realizada em dados textuais que já contêm um termo conhecido como insultuoso. Logo, esse trabalho gera modelos que classificam posts como homofóbicos ou não, que tenham essa particularidade de já conter um termo ofensivo, mas, dependendo do contexto, podem não ser ofensivos.

Dos 1.713 tweets da base3, 372 foram rotulados como homofóbicos e 1341 como não homofóbicos, na Tabela 3 podemos ver a distribuição dos rótulos entre as palavras. Nela observamos que o termo 'gayzismo' não está presente, isso aconteceu na filtragem realizada pelo modelo de discurso de ódio, que não atribuiu a esses tweets essa classe. Também percebemos, que mesmo não colocando o termo 'viado' na nova coleta, ele apareceu em 19 tweets. Isso se deu, pois alguns tweets podem conter mais de um dos termos.

Além dessa rotulagem manual e a rotulagem sobre discurso de ódio, mais duas foram realizadas nos dados da base3, dessa forma, outras features, além da frequência das palavras, poderiam ser usadas para o treinamento dos modelos.

Uma das rotulagens foi sobre a presença de linguagem tóxica nos tweets, para realizá-la foi utilizado um modelo de classificação

Tabela 3: Distribuição dos rótulos entre os termos.

Palavra	Não Homofóbicos	Homofóbicos
baitola	29	20
bicha	902	64
bichona	9	9
bixa	232	24
boiola	18	27
gayzada	4	3
homossexualismo	9	7
viadagem	63	54
viadao	11	32
viadinho	64	132
viado	7	12

automática desse tipo de conteúdo, encontrado no Github[4]. A última rotulagem, atribuída à base3, foi relacionada à descrição da bio do usuário criador do tweet. Nela foi atribuída, a cada tweet, a quantidade de termos presentes na descrição da bio do usuário, que integravam outro conjunto de termos também estruturado por Pereira [9]. Esse conjunto foi construído, verificando quais unigramas eram os mais presentes nas descrições dos usuários dos tweets classificados como homofóbicos na pesquisa de Pereira.

Foram gerados os vetores de valores TF, dos unigramas do corpo textual dos tweets da base3 e juntos aos valores das rotulagens de discuso de ódio, de linguagem tóxica e dos termos da descrição do usuário, esse foi o conjunto de features de entrada usadas no treinamento dos modelos descritos na seção 3.3.

3.3 Modelos

De posse dos conjuntos de dados o próximo passo foi a criação dos modelos, que foi realizada utilizando a SciKit-Learn[5], que é uma biblioteca em Python especializada em inteligência artificial. Foram utilizadas 14 configurações de algoritmos para a construção dos modelos, descritas a seguir:

- Três variações do algoritmo de regressão logística, mudando o algoritmo de regularização, que são ElasticNet, L1 e L2 - essas variações serão chamadas de regLogE, regLogL1 e regLogL2 respectivamente.
- Três variações do algoritmo Multinomial Naive Bayes, utilizando os seguintes valores de suavização 1, 0.5 e 1^−10 (o valor 1^−10 é o mínimo indicado para evitar erros) - essas variações serão chamadas de multiNB1, multiNB2 e multiNB3 respectivamente.
- Três variações do algoritmo Bernouli Naive Bayes, seguindo as mesmas mudanças do algoritmo Multinomial - essas variações serão chamadas de bernouNB1, bernouNB2 e bernouNB3 respectivamente.
- Duas variações do algoritmo SVM Linear, variando o algoritmo de regularização, L1 e L2 - essas variações serão chamadas de linearSVML1 e linearSVML2 respectivamente.
- Duas variações do algoritmo Random Forest, variando o parâmetro que define qual critério determina a qualidade de

[4]Link para o repositório - https://github.com/JAugusto97/ToLD-Br
[5]SciKit-Learn - https://scikit-learn.org/stable/

um split na árvore, utilizamos o critério de impureza Gini e a entropia - essas variações serão chamadas de randomFG e randomFE respectivamente.
- E uma configuração de XGBoost, utilizando os parâmetros padrões da biblioteca - essa configuração será chamada de xgboost.

Os dados da base3 foram separados em dados de treino e teste, sendo 25% dos dados para teste. Os dados do conjunto de treinamento foram usados para treinar todas as variações dos algoritmos citados acima, tendo um total de 14 modelos construídos.

Com o objetivo de ter um conjunto maior para treinamento, foi utilizado o modelo com a melhor F1 Score entre os 14 citados anteriormente, para classificar os dados da base1 como homofóbicos ou não, e utilizá-la como conjunto de treinamento de mais modelos. Como a rotulagem da base1 foi feita por um modelo, existe uma porcentagem de erro atrelada a ela. Para mitigar esse erro foram criados 3 conjuntos a partir da base1 classificada, considerando o limiar de probabilidade de classificação (essa probabilidade é um valor retornado pelo modelo, após classificar um dado, demonstrando qual a chance desse dado ser de uma determinada classe). Os limiares utilizados foram 70%, 80% e 90%, dessa forma, os dados dos conjuntos estruturados a partir de cada limiar tem probabilidade alta de serem de alguma das classes. Chamaremos esses conjuntos de base70p, base80p e base90p. A Tabela 4 mostra quantos tweets ficaram em cada conjunto:

Tabela 4: Quantidade de tweets coletados para cada limiar.

Limiar	Quantidade
70%	8049
80%	5936
90%	3993

Os 25% dos dados da base3, separados anteriormente para teste, foram utilizados para testar todos os modelos. Dessa forma, podemos garantir uma avaliação mais coerente, já que a base3 é a única cujo rótulo objeto da classificação não é resultado de uma classificação feita por modelos. Além dos 75% da base3, os conjuntos, base70p, base80p, base90p e base1 também foram usados para treinar os modelos. Portando, com os cinco conjuntos usados em cada uma das 14 variações dos algoritmos, foram criados 70 modelos, que serão comparados de acordo com os valores de acurácia, recall, precisão e F1 score para escolhermos os melhores.

4 RESULTADOS

Os resultados de classificação dos 70 modelos construídos estão descritos a seguir, separados por cada conjunto de dado utilizado para o treinamento. Para estabelecermos os melhores modelos, foi considerado os modelos com maior F1, por essa ser uma boa métrica de avaliação e abranger outras duas, o Recall e a Precisão. Mas, mesmo assim, a acurácia será considerada nas discussões por ser uma métrica mais fácil de se interpretar, pois aponta a porcentagem de acertos que um modelo teve, nos dados de teste.

4.1 Base3

Os modelos criados com os dados da base3 apresentaram métricas boas. Dos 14, somente 5 tiveram F1 menor que 60%, tendo só 1 muito baixo que foi o modelo bernouNB1, cuja F1 foi de 25.60%. Os modelos de Regressão logística se destacaram com os melhores valores de F1 e também de acurácia. O modelo regLogL2 foi o melhor para esse conjunto de dados de treinamento e de acordo com a sua acurácia ele acertou a classificação de 87.88% dos dados. Na Tabela 5 podemos visualizar os valores das métricas de todos os modelos.

Tabela 5: Métricas dos modelos criados com os dados da base3

Modelo	Acurácia	Precisão	Recall	F1 Score
regLogL2-base3	87.88%	68.48%	73.26%	70.79%
regLogL1-base3	87.41%	72.83%	69.79%	71.28%
regLogE-base3	86.95%	66.30%	70.93%	68.54%
multiNB1-base3	84.62%	50.00%	69.70%	58.23%
multiNB2-base3	84.85%	70.65%	63.11%	66.67%
multiNB3-base3	79.49%	57.61%	51.96%	54.64%
bernouNB1-base3	78.32%	17.39%	48.48%	25.60%
bernouNB2-base3	79.72%	39.13%	53.73%	45.28%
bernouNB3-base3	80.89%	56.52%	55.32%	55.91%
linearSVML1-base3	86.01%	71.74%	66.00%	68.75%
linearSVML2-base3	84.62%	69.57%	62.75%	65.98%
randomFG-base3	85.55%	63.04%	67.44%	65.17%
randomFE-base3	85.78%	65.22%	67.42%	66.30%
xgboost-base3	85.55%	67.39%	65.96%	66.67%

4.2 Base1

Com a base1, sendo ela a maior entre as 5, os modelos que a utilizaram para o treinamento não tiveram muita diferença entre si, mas foram os melhores em relação aos outros conjuntos. Todos tiveram F1 entre 67% e 72%, e acurácia entre 85% e 88%. O modelo com o maior valor de F1 foi o linearSVML1, sendo de 71.96%, mas a sua acurácia, de 87.65%, foi a mesma dos modelos: regLogL2, randomFG e randomFE. Na Tabela 6 podemos visualizar o quão próximo os valores desse conjunto de dados ficaram.

4.3 Base90p

O conjunto base90p gerou só 1 modelo com F1 maior que 70% e 5 com F1 menor que 60%. Mesmo não demonstrando os maiores valores de F1, todos os modelos tiveram acurácia próxima ou maior que 80%, reafirmando a necessidade de observar mais métricas de avaliação além da acurácia. Os demais valores podem ser observados na Tabela 7.

4.4 Base80p

Na Tabela 8, podemos visualizar as métricas dos modelos criados com a base80p e nela podemos perceber que os valores são bem próximos do caso da base90p, onde só um modelo teve F1 maior que 70% e todos tiveram acurácia maior que 80%.

Tabela 6: Métricas dos modelos criados com os dados da base1

Modelo	Acurácia	Precisão	Recall	F1 Score
regLogL2-base1	87.65%	72.83%	70.53%	71.66%
regLogL1-base1	87.41%	72.83%	69.79%	71.28%
regLogE-base1	86.95%	69.57%	69.57%	69.57%
multiNB1-base1	85.31%	71.74%	64.08%	67.69%
multiNB2-base1	85.31%	70.65%	64.36%	67.36%
multiNB3-base1	85.78%	70.65%	65.66%	68.06%
bernouNB1-base1	86.25%	72.83%	66.34%	69.43%
bernouNB2-base1	86.48%	72.83%	67.00%	69.79%
bernouNB3-base1	86.71%	73.91%	67.33%	70.47%
linearSVML1-base1	87.65%	73.91%	70.10%	71.96%
linearSVML2-base1	87.41%	72.83%	69.79%	71.28%
randomFG-base1	87.65%	72.83%	70.53%	71.66%
randomFE-base1	87.65%	72.83%	70.53%	71.66%
xgboost-base1	86.01%	66.30%	67.78%	67.03%

Tabela 7: Métricas dos modelos criados com os dados da base90p

Modelo	Acurácia	Precisão	Recall	F1 Score
regLogL2-base90p	86.01%	50.00%	76.67%	60.53%
regLogL1-base90p	86.95%	64.13%	71.95%	67.82%
regLogE-base90p	86.95%	57.61%	75.71%	65.43%
multiNB1-base90p	85.08%	58.70%	67.50%	62.79%
multiNB2-base90p	86.01%	61.96%	69.51%	65.52%
multiNB3-base90p	85.78%	47.83%	77.19%	59.06%
bernouNB1-base90p	79.25%	17.39%	55.17%	26.45%
bernouNB2-base90p	81.35%	35.87%	61.11%	45.21%
bernouNB3-base90p	84.62%	39.13%	78.26%	52.17%
linearSVML1-base90p	87.18%	64.13%	72.84%	68.21%
linearSVML2-base90p	87.88%	69.57%	72.73%	71.11%
randomFG-base90p	86.48%	53.26%	76.56%	62.82%
randomFE-base90p	86.95%	53.26%	79.03%	63.64%
xgboost-base90p	85.31%	51.09%	72.31%	59.87%

4.5 Base70p

As métricas dos modelos feitos com os dados da base70p podem ser verificados na Tabela 9 e nela é percebido que não houve valor de F1 menor que 65%, com o destaque para o modelo bernouNB1 que teve o maior F1, 71.20%.

4.6 Avaliação Geral

Analisando o conjunto total de modelos, podemos perceber que as métricas atingidas foram satisfatórias. Na revisão bibliográfica feita sobre o assunto, as métricas dos melhores modelos se aproximam das métricas dos modelos construídos nessa pesquisa. Considerando que um F1 alto é de 70% ou maior, 19 dos 70 modelos tiveram F1 alto. Os dez melhores modelos, baseado no seus valores de F1, podem ser visualizados no gráfico da Figura 1. O melhor modelo foi o linearSVML1-base1, sendo o modelo criado, utilizando o algoritmo de SVM Linear, tendo como dados de treinamento a base1.

Tabela 8: Métricas dos modelos criados com os dados da base80p

Modelo	Acurácia	Precisão	Recall	F1 Score
regLogL2-base80p	86.71%	63.04%	71.60%	67.05%
regLogL1-base80p	86.25%	69.57%	67.37%	68.45%
regLogE-base80p	86.95%	68.48%	70.00%	69.23%
multiNB1-base80p	86.01%	66.30%	67.78%	67.03%
multiNB2-base80p	86.25%	68.48%	67.74%	68.11%
multiNB3-base80p	84.85%	55.43%	68.00%	61.08%
bernouNB1-base80p	83.68%	55.43%	63.75%	59.30%
bernouNB2-base80p	85.78%	65.22%	67.42%	66.30%
bernouNB3-base80p	85.55%	55.43%	70.83%	62.20%
linearSVML1-base80p	86.25%	70.65%	67.01%	68.78%
linearSVML2-base80p	86.95%	71.74%	68.75%	70.21%
randomFG-base80p	86.48%	57.61%	73.61%	64.63%
randomFE-base80p	87.41%	61.96%	75.00%	67.86%
xgboost-base80p	85.78%	56.52%	71.23%	63.03%

Tabela 9: Métricas dos modelos criados com os dados da base70p

Modelo	Acurácia	Precisão	Recall	F1 Score
regLogL2-base70p	87.41%	69.57%	71.11%	70.33%
regLogL1-base70p	86.95%	71.74%	68.75%	70.21%
regLogE-base70p	86.95%	69.57%	69.57%	69.57%
multiNB1-base70p	86.95%	75.00%	67.65%	71.13%
multiNB2-base70p	86.48%	73.91%	66.67%	70.10%
multiNB3-base70p	85.31%	65.22%	65.93%	65.57%
bernouNB1-base70p	87.18%	73.91%	68.69%	71.20%
bernouNB2-base70p	86.48%	71.74%	67.35%	69.47%
bernouNB3-base70p	85.08%	66.30%	64.89%	65.59%
linearSVML1-base70p	87.41%	70.65%	70.65%	70.65%
linearSVML2-base70p	86.95%	71.74%	68.75%	70.21%
randomFG-base70p	87.41%	69.57%	71.11%	70.33%
randomFE-base70p	87.41%	66.30%	72.62%	69.32%
xgboost-base70p	86.48%	67.39%	68.89%	68.13%

A base1 se destacou, sendo a que gerou os melhores modelos. Isso aponta para o benefício alcançado, ao utilizar a base3, rotulada manualmente, para gerar modelos que rotularam automaticamente a base1. Dessa forma, foi possível reduzir o tempo dispensado na rotulagem manual, rotulando manualmente uma base menor, mas ainda assim, conseguir utilizar uma base maior para o treinamento dos modelos.

Observando os piores modelos, podemos perceber no gráfico da Figura 2, que todos foram criados utilizando uma variação de algum algoritmo Naive Bayes. Entretanto, o uso desse algoritmo unicamente, não indica o motivo do baixo desempenho. Além disso, outra informação que pode ser percebida é que os conjuntos de treinamentos que foram utilizados nesses modelos foram os de menor tamanho. Os dois maiores conjuntos, base1 e base70p, não estão entre os conjuntos de treinamentos dos piores modelos. Isso pode indicar que a junção, do uso desses algoritmos, com um conjunto

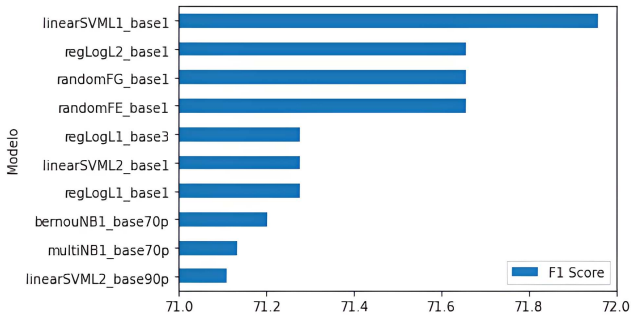

Figura 1: Melhores modelos de acordo com seus valores de F1 Score

pequeno de dados, gere modelos com uma capacidade baixa de classificação nesse contexto.

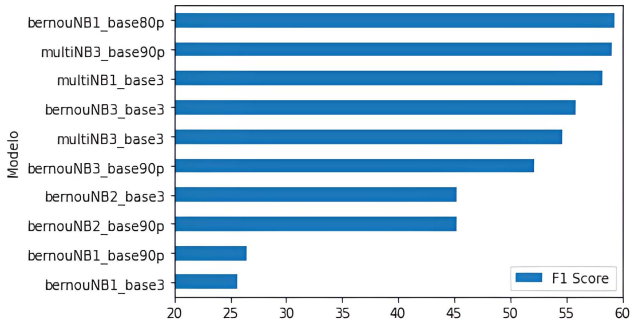

Figura 2: Piores modelos de acordo com seus valores de F1 Score

Mesmo com alguns modelos demonstrando baixa capacidade de classificação, as métricas atingidas nesse trabalho são um indicativo que o método utilizado é promissor, na construção de modelos de classificação de conteúdo homofóbico em tweets. E os modelos gerados podem ser utilizados em tarefas de classificação, já que se mostraram eficazes nisso.

5 CONCLUSÕES E TRABALHOS FUTUROS

No mundo atualmente ser uma pessoa LGBTQIA+ é significado de medo e sofrimento. Essa população é vítima de todos os tipos de violências diárias, sejam físicas ou virtuais. Com a massificação do uso das redes sociais, a violência virtual direcionada a essas pessoas está se tornando banalizada e quem às comete permanece sem punição. As grandes corporações, donas das redes sociais, lucram cada dia mais com o aumento das interações nesses espaços, sejam essa interações criminosas ou não. A população deve se haver de ferramentas que possam, pelo menos, fornecer estatísticas sobre esse problema, algo que também não temos, pois casos de homofobia são sub notificados ou totalmente apagados das estatísticas de violência no Brasil.

Tendo como objetivo a construção de ferramentas que auxiliem na obtenção dessas estatísticas, os resultados dessa pesquisa podem apoiar essa construção de diversas formas. Seja na utilização dos modelos, com boas métricas de classificação, para a identificação automática, e assim mais rápida, de conteúdos homofóbicos em redes sociais. Ou o uso dos conjuntos de dados, criados nessa pesquisa, para melhoria e avanço nas técnicas de predição de conteúdo homofóbico.

Possíveis trabalhos futuros relacionados a esse tema podem focar na construção de uma maior base de dados, utilizando das mesmas técnicas de rotulagem manual. Utilizando técnicas de aprendizagem de máquina não supervisionada, grupos de dados podem ser criados, baseados na semelhança entre os textos. A partir desse agrupamento, a rotulagem manual pode ser realizada em algum dos grupos que demonstre características mais propensas a serem homofóbicas. Acredita-se que a utilização dessas técnicas pode ser de grande contribuição para a construção de uma significativa base de dados destinada a esse fim.

REFERÊNCIAS

[1] Iann Carvalho BARBOSA et al. 2021. Reconhecimento de mensagens com teor transfóbico no twitter. (2021).

[2] Vinicius Matheus de Medeiros Silva Coutinho and Yuri Malheiros. 2020. Detecçao de mensagens homofóbicas em português no twitter usando análise de sentimentos. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*. SBC, 1–12.

[3] Portal de processos do Supremo Tribunal Federal. 2019. AÇÃO DIRETA DE INCONSTITUCIONALIDADE POR OMISSÃO 26. Retrieved 14 de Março de 2022 from https://portal.stf.jus.br/processos/detalhe.asp?incidente=4515053

[4] G1. 2021. Denúncias contra homofobia na internet crescem 106% nos primeiros seis meses de 2021. Retrieved 14 de Março de 2022 from https://g1.globo.com/economia/tecnologia/noticia/2021/06/17/denuncias-contra-homofobia-na-internet-crescem-106percent-nos-primeiros-seis-meses-de-2021.ghtml

[5] Migalhas. 2021. Casos de LGBTfobia na internet crescem no mês do Orgulho LGBTQIA+. Retrieved 14 de Março de 2022 from https://www.migalhas.com.br/quentes/347457/casos-de-lgbtfobia-na-internet-crescem-no-mes-do-orgulho-lgbtqia

[6] Maria Carolina Monard and José Augusto Baranauskas. 2003. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações* 1, 1 (2003), 32.

[7] Safer Net. [n. d.]. Indicadores da Central Nacional de Denúncias de Crimes Cibernéticos. Retrieved 14 de Março de 2022 from https://indicadores.safernet.org.br/index.html

[8] Peter Dias Paiva, Vanecy Matias da Silva, and Raimundo Santos Moura. 2019. Detecção automática de discurso de ódio em comentários online. In *Anais da VII Escola Regional de Computação Aplicada à Saúde*. SBC, 157–162.

[9] Vinicius Gomes Pereira. 2018. *Using supervised machine learning and sentiment analysis techniques to predict homophobia in portuguese tweets.* Master's thesis. Escola de Matemática Aplicada. Fundação Getulio Vargas, Rio de Janeiro.

[10] Roger RAUPP and R JUNQUEIRA. 2009. Homofobia na Perspectiva dos Direitos Humanos e no Contexto dos Estudos sobre Preconceito e Discriminação. *Diversidade Sexual na Educação* (2009).

[11] Gilberto Schäfer, Paulo Gilberto Cogo Leivas, and Rodrigo Hamilton dos Santos. 2015. Discurso de ódio: da abordagem conceitual ao discurso parlamentar. *Revista de informação legislativa* 52, 207 (2015), 143–158.