Detecção de Homofobia em Português do Brasil: Construção de Conjunto de Dados e Modelo de Classificação Automática

Vinícius Soares dos Santos[1], **Gustavo Guedes**[1]

[1] CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

vinicius.santos@eic.cefet-rj.br, gustavo.guedes@cefet-rj.br

Abstract. *This article investigates homophobia detection in Brazilian Portuguese through lexicon expansion, labeled dataset construction, and automatic text classification. First, an existing lexical resource is expanded with homophobic expressions provided by human participants. Next, the resulting lexicon guides the retrieval and annotation of social media messages. Finally, supervised models are trained for binary classification. The study produces an expanded lexicon, a labeled dataset in Brazilian Portuguese, and promising classification results (F1 = 0.8688).*

Resumo. *Este artigo investiga a detecção de homofobia em português do Brasil por meio de ampliação lexical, construção de conjunto de dados rotulado e classificação automática de textos. Inicialmente, um recurso lexical prévio é expandido com termos e expressões homofóbicas obtidos com participantes humanos. Em seguida, o léxico resultante orienta a coleta e a rotulagem de mensagens em rede social. Por fim, modelos supervisionados são treinados para classificação binária. O estudo produz um léxico expandido, um conjunto de dados rotulado em português do Brasil e resultados promissores (F1 = 0,8688).*

1. Introdução

O discurso de ódio homofóbico constitui uma forma de violência discursiva que promove discriminação, hostilidade e exclusão de indivíduos com base em sua orientação sexual [Costa and Nardi, 2015]. Em ambientes digitais, esse fenômeno adquire maior alcance e rapidez de propagação, sobretudo em plataformas de redes sociais, nas quais conteúdos ofensivos podem ser publicados, compartilhados e replicados em larga escala em curtos intervalos de tempo. Nesse contexto, a circulação de mensagens homofóbicas ultrapassa a esfera da ofensa individual e passa a reforçar estigmas, legitimar práticas discriminatórias e ampliar a vulnerabilidade de grupos historicamente marginalizados. A visibilidade desse tipo de manifestação em redes sociais também evidencia que o problema não se restringe a episódios isolados, mas se manifesta de forma recorrente.

Nesse contexto, métodos automáticos de análise textual têm se mostrado relevantes para apoiar a identificação de manifestações de ódio em redes sociais. Técnicas de mineração de textos e de aprendizado de máquina podem contribuir para a triagem, o monitoramento e a classificação de grandes volumes de mensagens, reduzindo a dependência exclusiva da revisão humana e tornando mais ágil o processo de moderação. Conforme destacam MacAvaney et al. [2019], esse tipo de abordagem permite direcionar o esforço

humano para análises mais especializadas, ao mesmo tempo em que amplia a capacidade de detecção de conteúdos potencialmente ofensivos em ambientes digitais.

A literatura recente evidencia um avanço internacional na detecção automática de discurso de ódio direcionado à população LGBTQIA+. Locatelli et al. [2023] mostram, em estudo *cross-lingual* no Twitter, que a homotransfobia constitui um fenômeno global, mas linguisticamente e culturalmente situado. Na mesma direção, Chakravarthi et al. [2024] descrevem uma tarefa compartilhada dedicada à detecção de homofobia e transfobia em dez línguas, enquanto Chan et al. [2024] mostram que a detecção multilíngue continua desafiadora, sobretudo quando mediada por tradução automática. Além disso, novos recursos têm sido construídos para línguas de menor cobertura, como o *dataset* apresentado por Kumaresan et al. [2024] para Telugu, Kannada e Gujarati.

Apesar dos avanços no uso de técnicas automáticas para detecção de discurso de ódio, a eficácia desses métodos depende, em grande medida, da disponibilidade de recursos linguísticos adequados ao domínio analisado. No caso da homofobia em português do Brasil, observa-se ainda uma escassez de léxicos especializados que contemplem, de forma mais abrangente, termos e expressões empregados em contextos ofensivos. Essa limitação compromete a cobertura vocabular de abordagens automáticas e dificulta a identificação de mensagens que recorrem a formulações específicas, variações linguísticas e expressões de uso recorrente em contextos discriminatórios.

Um exemplo dessa limitação pode ser observado no *Hatebase* [Hatebase, 2020], vocabulário multilíngue mantido de forma colaborativa e voltado ao mapeamento de termos associados ao discurso de ódio em diferentes idiomas e países. Embora represente uma iniciativa relevante, sua cobertura para o português, no que se refere especificamente à homofobia, ainda é reduzida. A plataforma reúne apenas 28 termos relacionados, o que sugere a necessidade de ampliar os recursos disponíveis para subsidiar pesquisas e aplicações computacionais voltadas à detecção automática de conteúdo homofóbico.

De modo complementar, a limitação de recursos em português também se manifesta na escassez de *datasets* rotulados para tarefas específicas de detecção de homofobia. Conforme apontado por Caseli and Nunes [2024], em 2023 foi lançada uma avaliação conjunta para a detecção de homofobia/transfobia com *datasets* em inglês, espanhol, hindi, tâmil e malaiala, mas não em português, justamente porque, naquele momento, não havia bases disponíveis para essa finalidade. Esse cenário reforça que a carência de recursos no contexto do português do Brasil não se restringe ao plano lexical, mas abrange também a disponibilidade de conjuntos de dados anotados, indispensáveis ao treinamento, à avaliação e à comparação de modelos computacionais.

Diante dessa lacuna, o presente trabalho se articula ao Objetivo de Desenvolvimento Sustentável 16 da ONU, voltado à promoção de sociedades pacíficas e inclusivas e à redução de formas de violência. Nesse contexto, investiga-se a detecção de homofobia em português do Brasil por meio de uma abordagem integrada, que envolve a construção de um léxico especializado, um *dataset* rotulado e um modelo de classificação automática. O estudo amplia o repertório de termos do domínio, coleta e anota mensagens potencialmente homofóbicas de redes sociais e avalia modelos supervisionados na classificação binária entre conteúdos homofóbicos e não homofóbicos. A pesquisa foi aprovada pelo Comitê de Ética em Pesquisa (CEP) sob o CAAE 59293922.0.0000.5289.

Além desta introdução, a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve a metodologia usada na construção do léxico, base de dados e modelo de classificação; a Seção 4 discute os resultados; por fim, a Seção 5 apresenta as conclusões.

2. Trabalhos Relacionados

Segundo de Pelle and Moreira [2017], o aprendizado supervisionado para a classificação de discurso de ódio depende de conjuntos de dados rotulados. Devido à ausência de tais conjuntos em português, os autores criaram dois *datasets*[1] a partir de comentários do portal G1[2]. O primeiro *dataset*, OFFCOMBR-2, contém 1.250 comentários anotados por três juízes, com rótulo final definido por concordância de pelo menos dois avaliadores; o segundo, OFFCOMBR-3, contém 1.033 comentários nos quais houve concordância unânime quanto à classe ofensiva ou não ofensiva. Embora o OFFCOMBR-2 reúna 419 comentários ofensivos, apenas 14 foram associados à categoria homofobia por pelo menos dois juízes; no critério unânime, esse número cai para 9 comentários.

Utilizando como referência o trabalho de de Pelle and Moreira [2017], Fortuna et al. [2019] expandiram a construção de um *dataset* em português para discurso de ódio. Os autores selecionaram 29 perfis do Twitter e, a partir deles, realizaram buscas com 19 termos e 10 *hashtags* para compor um conjunto de dados contendo discurso de ódio. Foram filtrados *tweets* em português contendo ao menos 3 palavras. Além disso, eliminaram repetições e *retweets* para evitar duplicações, removeram *HTML*, *tags* e outros conteúdos não textuais. Após a coleta, as mensagens extraídas foram submetidas ao julgamento de duas pessoas, sendo utilizada a métrica Cohen's Kappa para avaliar a concordância entre as classes homofobia, racismo, religião, imigrantes e outros. No total, o conjunto contém 5.668 *tweets*, sendo apenas 365 associados ao discurso de ódio homofóbico.

Em uma perspectiva mais exploratória, Santos et al. [2022] investigam mensagens homofóbicas no Twitter por meio de técnicas de mineração textual, incluindo *Bag of Words*, análise de bigramas, agrupamento e visualização de dados. O estudo parte de termos do *Hatebase* e analisa 2.597 mensagens únicas, oriundas de respostas a apenas uma postagem específica, totalizando 827 usuários distintos. Seus resultados mostram que a homofobia textual pode se organizar em diferentes agrupamentos léxico-discursivos, indicando que o fenômeno não se reduz à presença isolada de palavras ofensivas, mas envolve combinações contextuais e padrões de uso.

Antunes et al. [2023] investigam a predição de homofobia no Twitter com técnicas de processamento de linguagem natural e aprendizado de máquina, utilizando uma base com mais de 3.000 *tweets*. Embora o estudo reporte resultados promissores, não explicita os termos que orientaram a coleta das mensagens. O presente trabalho avança nesse ponto ao adotar uma metodologia sistematizada para a construção de um léxico especializado, utilizado de forma explícita na recuperação dos textos e na criação do *dataset* rotulado.

A escassez de trabalhos e *datasets* relacionados à detecção de homofobia em português do Brasil continua sendo uma limitação importante que este estudo busca superar. Apesar das contribuições de de Pelle and Moreira [2017], Fortuna et al. [2019], Santos et al. [2022] e Antunes et al. [2023], ainda persistem desafios relevantes, como o número

[1] `https://github.com/rogersdepelle/OffComBR`
[2] `g1.globo.com`

reduzido de textos explicitamente homofóbicos em algumas bases. Portanto, permanece necessária a ampliação de conjuntos de dados em português do Brasil que articulem, de forma integrada, léxico especializado, mensagens rotuladas e modelos de classificação.

3. Metodologia

A metodologia deste estudo organiza-se em cinco etapas articuladas com o objetivo de apoiar a detecção de homofobia em português do Brasil por meio da construção de recursos linguísticos e computacionais. Na primeira etapa, a metodologia utiliza como ponto de partida um léxico previamente existente e busca ampliá-lo com novas expressões associadas à homofobia em português do Brasil. Na segunda, emprega o léxico expandido para orientar a coleta de mensagens em rede social potencialmente relacionadas ao domínio investigado. Na terceira, realiza o pré-processamento e a seleção dos textos coletados, de modo a torná-los adequados à anotação humana. Na quarta, promove a rotulagem das mensagens quanto à presença de conteúdo homofóbico, permitindo a construção de um conjunto de dados anotado. Por fim, a quinta etapa utiliza o conjunto de dados rotulado para treinar e avaliar modelos de classificação automática. A Figura 1 apresenta uma visão geral do fluxo metodológico adotado no estudo.

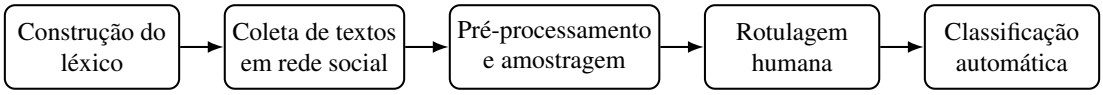

Figura 1. Fluxo metodológico adotado no estudo.

3.1. Construção do léxico homofóbico

A primeira etapa da metodologia consiste na construção de um léxico de termos homofóbicos em português do Brasil a partir da ampliação de uma base lexical previamente existente. Essa etapa parte do pressuposto de que a cobertura lexical é um componente relevante na recuperação de mensagens potencialmente ofensivas e, consequentemente, na construção de recursos voltados à detecção automática de homofobia.

Como ponto de partida, utiliza-se um conjunto inicial de termos e expressões já documentados em recurso lexical anterior. A partir desse repertório de referência, emprega-se um instrumento de coleta voltado à elicitação de novas expressões do domínio. Esse instrumento é estruturado de modo a expor os participantes a um conjunto lexical inicial e, em seguida, solicitar a indicação de outros termos ou expressões homofóbicas não contemplados na base de partida.

A estratégia adotada busca combinar um repertório lexical previamente disponível com o conhecimento linguístico dos participantes, favorecendo a incorporação de formas de uso mais correntes no contexto do português do Brasil. As contribuições textuais obtidas nessa etapa são então submetidas a um processo de consolidação lexical, que envolve segmentação, normalização e contagem de ocorrências.

Após esse tratamento inicial, aplica-se um critério de filtragem com base na recorrência dos itens informados, com o objetivo de reduzir a inclusão de formas muito idiossincráticas, ambíguas ou pouco representativas. Como resultado dessa etapa, obtém-se um léxico expandido de termos e expressões homofóbicas, posteriormente utilizado para orientar a coleta de mensagens em rede social.

3.2. Coleta dos textos em rede social

A etapa seguinte da metodologia consiste na coleta de mensagens em rede social com o objetivo de reunir textos potencialmente relacionados à homofobia em português do Brasil. Essa etapa utiliza o léxico expandido como mecanismo de orientação da busca, de modo a concentrar a recuperação em ocorrências associadas ao domínio investigado.

A coleta é realizada por meio de consultas automatizadas a dados públicos da plataforma, implementadas em ambiente computacional apropriado para recuperação e organização textual. As consultas são formuladas a partir dos termos e expressões do léxico construído na etapa anterior, permitindo que o repertório lexical funcione como ponto de entrada para a composição de um corpus inicial de mensagens.

Como critérios gerais, consideram-se apenas mensagens em português. Além disso, descartam-se ocorrências que introduzam redundância ou que não contribuam diretamente para a análise do conteúdo textual como republicações automáticas, duplicações e casos em que os termos de interesse aparecem apenas em menções ou identificadores, sem relação efetiva com a mensagem em si.

Essa etapa não assume que a presença de um termo do léxico seja suficiente para caracterizar homofobia. Seu papel consiste em restringir o espaço de busca a mensagens potencialmente relevantes para o fenômeno estudado, produzindo um corpus inicial que será posteriormente submetido a pré-processamento, seleção e rotulagem humana.

3.3. Pré-processamento e amostragem

Após a coleta, as mensagens passam por uma etapa de pré-processamento com o objetivo de reduzir ruídos e adequar os textos às etapas posteriores de anotação e classificação. Nessa etapa, são removidos ou normalizados elementos não centrais à análise do conteúdo linguístico, como marcas estruturais da plataforma, além de registros duplicados, vazios ou sem conteúdo textual aproveitável.

Em seguida, realiza-se uma etapa de amostragem sobre o material pré-processado, com a finalidade de compor um subconjunto viável para rotulagem humana. Como a coleta orientada pelo léxico pode recuperar quantidades distintas de mensagens para diferentes termos, a amostragem busca equilibrar representatividade lexical e viabilidade operacional, preservando diversidade de contextos de uso sem tornar a etapa de anotação excessivamente onerosa.

O produto dessa etapa consiste em um conjunto de mensagens textuais selecionadas e preparadas para julgamento humano, servindo como base para a construção do *dataset* rotulado utilizado nas etapas subsequentes do estudo.

3.4. Rotulagem dos textos e geração do conjunto de dados final

Esta etapa da metodologia consiste na rotulagem humana das mensagens previamente selecionadas, com o objetivo de construir um conjunto de dados anotado quanto à presença de conteúdo homofóbico. O objetivo é transformar o corpus textual inicial em um recurso supervisionado, adequado tanto à análise do fenômeno quanto ao treinamento e à avaliação de modelos de classificação automática.

A rotulagem é conduzida em ambiente *online*, por meio de uma interface experimental que apresenta as mensagens aos avaliadores de forma individual. Para cada texto,

solicita-se um julgamento categórico sobre a presença de conteúdo homofóbico, a partir de um esquema de três rótulos: "sim", "não" e "não tem certeza". A inclusão da terceira categoria busca contemplar casos ambíguos ou limítrofes, evitando que situações de incerteza sejam artificialmente forçadas para uma das classes principais.

O procedimento de anotação é estruturado de modo a distribuir os textos entre os participantes e viabilizar o julgamento humano em condições operacionais adequadas. Cada mensagem é submetida à avaliação de mais de um julgador, permitindo posterior comparação entre respostas e consolidação dos rótulos. Essa estratégia é importante porque a identificação de homofobia em textos curtos e descontextualizados pode envolver interpretações divergentes, o que torna desejável a existência de múltiplos julgamentos para uma mesma instância.

Após a coleta das respostas, aplica-se um critério de consolidação para atribuição do rótulo final de cada mensagem. De forma geral, esse processo considera a convergência entre os julgamentos realizados, de modo a produzir um conjunto de dados anotado com base em consenso ou maioria entre avaliadores. Como resultado dessa etapa, obtém-se um *dataset* rotulado de mensagens em português do Brasil, posteriormente utilizado tanto para análise descritiva quanto para a etapa de classificação automática.

3.5. Classificação automática dos textos

A etapa final da metodologia consiste na classificação automática das mensagens rotuladas, com o objetivo de avaliar em que medida o conjunto de dados construído no estudo pode sustentar a detecção computacional de homofobia em português do Brasil. Para essa etapa, consideram-se apenas as instâncias associadas às classes "sim" e "não", configurando uma tarefa supervisionada de classificação binária.

Inicialmente, os textos passam por uma etapa adicional de limpeza e normalização, com o propósito de reduzir ruídos de formatação e tornar as instâncias adequadas à modelagem computacional. Em seguida, as mensagens são representadas por meio de TF-IDF (*Term Frequency–Inverse Document Frequency*), contemplando descritores baseados em palavras e em caracteres, com diferentes abrangências de *n-gramas*.

Para avaliação, o conjunto de dados é inicialmente dividido em duas partições: treinamento e teste. A partição de treinamento é utilizada para seleção de modelos e ajuste de hiperparâmetros por meio de validação cruzada estratificada repetida, reduzindo a sensibilidade dos resultados a uma única divisão dos dados. A escolha da melhor configuração é orientada pela métrica F1-macro, por ser adequada à análise equilibrada do desempenho entre as classes.

Após a seleção da melhor configuração, o modelo correspondente é reestimado com toda a partição de treinamento e avaliado no conjunto de teste reservado, que não é utilizado durante a etapa de ajuste. O desempenho final é então examinado por meio de métricas de classificação e matriz de confusão, permitindo verificar a viabilidade da classificação automática no conjunto de dados produzido.

4. Resultados e Discussão

Esta seção apresenta os principais resultados obtidos nas diferentes etapas do estudo, contemplando a construção do léxico homofóbico, a coleta e rotulagem de mensagens em

rede social e construção/avaliação de modelos de classificação automática. Em conjunto, esses resultados evidenciam a viabilidade da metodologia proposta para a geração de um recurso lexical expandido, de um conjunto de dados anotado e de um modelo computacional para detecção de homofobia em português do Brasil.

4.1. Construção do léxico homofóbico

Como ponto de partida para a ampliação lexical proposta neste estudo, utilizou-se um recurso previamente disponível, o *Hatebase* [Hatebase, 2020], que reunia 28 termos e expressões associados à homofobia em português do Brasil. A Tabela 1 apresenta esse repertório inicial, adotado como base de referência para a etapa de expansão lexical.

Tabela 1. Termos homofóbicos do *Hatebase* e utilizados como base inicial.

Termo	Termo	Termo	Termo	Termo	Termo	Termo
Baitola	Bambi	Bambis	Biba	Bicha	Bichinha	Bichona
Boiola	Boiolas	Desviada	Desviadas	Desviado	Desviados	Escorregar no quiabo
Fufa	Fufas	Maricas	Mulherzinha	Mulherzinhas	Paneleira	Paneleiras
Paneleiro	Paneleiros	Rabeta	Rabetas	Sapatão	Sapatões	Viado

A etapa de ampliação do léxico homofóbico foi realizada por meio de formulário eletrônico, aplicado entre 01 e 10 de julho de 2022. Ao término da coleta, obteve-se a participação de 487 respondentes, cujas contribuições textuais serviram de base para a expansão do repertório lexical adotado no estudo.

A partir das respostas abertas, foram inicialmente identificados 136 termos ou expressões potencialmente associados à homofobia. Em seguida, os itens passaram por segmentação, normalização, contagem de ocorrências e filtragem por recorrência, com o objetivo de reduzir formas muito idiossincráticas ou pouco representativas. Após esse processo, 54 termos ou expressões foram selecionados para compor o léxico final (Tabela 2), incluindo os itens já presentes na base lexical de partida. Em comparação com os 28 termos inicialmente disponíveis no *Hatebase*, o léxico resultante representa um aumento relevante na cobertura lexical do recurso utilizado como referência.

Tabela 2. Termos homofóbicos resultantes desta pesquisa.

Termo	Termo	Termo	Termo	Termo	Termo
afeminada	caminhoneira	entendida	guena	mulher macho	sabonete
afetada	coca cola é fanta	escorregar no quiabo	indeciso	mulherzinha	sapa
agasalha croquete	cola velcro	fancha	kitty	paneleira	sapatão
aidetica	dá a rosca	flozô	mamador	pão com ovo	tchola
baitola	dá cu	franga	manja rola	passiva	transviado
bambi	dá ré no kibe	fresco	marica	princesa	traveco
biba	desmunheca	fruta	maricona	qualira	viado
bicha	desviada	fufa	mocinha	queima rosca	xibungo
boiola	enrustida	gay	morde a fronha	rabeta	xinxa

Além do aumento no número de entradas, o léxico final incorpora expressões coloquiais e variantes linguísticas que ajudam a representar, de forma mais ampla, manifestações homofóbicas no contexto do português do Brasil. Esse achado é relevante porque a cobertura lexical constitui um componente importante na recuperação de mensagens potencialmente ofensivas e, por consequência, na construção de conjuntos de dados voltados à detecção automática de homofobia.

4.2. Coleta e caracterização dos textos

A coleta dos textos em rede social foi realizada por meio da *Twitter API v2*, quando a plataforma ainda operava sob a denominação *Twitter*. As consultas foram automatizadas em *Python*, com apoio das bibliotecas *tweepy*, para autenticação e recuperação dos dados, e *pandas*, para organização e armazenamento das mensagens extraídas. O desenvolvimento e a execução das rotinas de coleta ocorreram em ambiente *Google Colab*.

As consultas foram formuladas a partir dos 54 termos e expressões do léxico final construído na etapa anterior. Como critérios de coleta, consideraram-se apenas mensagens em português, publicadas no Brasil, no período de 01/01/2019 a 30/06/2022. A extração foi realizada de forma segmentada ao longo desse intervalo, de modo a contornar limitações operacionais associadas ao volume de requisições. Além disso, foram excluídos *retweets*, mensagens duplicadas e ocorrências em que os termos de interesse apareciam apenas em menções ou identificadores de perfis, sem relação efetiva com o conteúdo textual da postagem.

Ao final dessa etapa, foi constituído o conjunto de dados `Homofensa-Coleta`, composto por 1.620 textos, correspondentes a 30 mensagens recuperadas para cada um dos 54 termos do léxico final. Esse procedimento buscou equilibrar a representatividade lexical das consultas e a viabilidade operacional das etapas posteriores de anotação.

Após o tratamento inicial do corpus, procedeu-se à caracterização descritiva do tamanho das mensagens em número de palavras. Os textos apresentam média de 18,81 palavras e mediana de 15, o que indica predominância de mensagens curtas, compatíveis com a dinâmica da plataforma utilizada. Ao mesmo tempo, o desvio padrão de 12,57 e a amplitude entre os valores mínimo e máximo sugerem heterogeneidade relevante no tamanho das postagens. Tamanho mínimo e máximo foram, respectivamente, dois e 57.

Esse resultado indica que a coleta orientada pelo léxico produziu um corpus textual variado, com mensagens curtas e diretas, mas também com ocorrências mais extensas e contextualmente mais elaboradas. Tal heterogeneidade é relevante porque amplia a diversidade de contextos linguísticos em que os termos do léxico aparecem. Ao mesmo tempo, a constituição desse corpus reforça uma característica importante do procedimento adotado: a presença de termos associados à homofobia aumenta a probabilidade de recuperação de mensagens relevantes para o domínio, mas não garante, por si só, que as ocorrências sejam efetivamente conteúdo homofóbico. Por essa razão, o conjunto `Homofensa-Coleta` deve ser entendido como um corpus inicial de mensagens potencialmente relevantes, cuja interpretação final depende da etapa de rotulagem humana.

4.3. Rotulagem dos textos e composição do conjunto de dados final

A etapa de rotulagem foi conduzida por meio da plataforma *PCIBEX*[3]. Os 1.620 textos do conjunto `Homofensa-Coleta` foram distribuídos em 14 questionários, sendo 13 com 120 mensagens e 1 com 60 mensagens, de forma a tornar a tarefa mais viável para os participantes e reduzir o risco de abandono durante a anotação.

Três julgadores responderam integralmente ao conjunto de mensagens, atribuindo a cada texto um dos seguintes rótulos: “sim”, “não” e “não tem certeza”, conforme a

[3] `https://www.pcibex.net/`

presença percebida de conteúdo homofóbico. A presença de três opções de resposta permitiu distinguir não apenas entre ocorrências positivas e negativas, mas também entre casos em que o conteúdo se mostrou ambíguo ou insuficientemente claro para uma decisão. Os três julgadores apresentavam perfis demográficos distintos: uma mulher heterossexual com mais de 60 anos, um homem heterossexual entre 21 e 25 anos e um homem homossexual entre 36 e 40 anos. Não foi adotada uma rubrica formal extensa de anotação, sendo apresentados aos julgadores o Termo de Consentimento Livre e Esclarecido (TCLE), os rótulos disponíveis e a tarefa de julgar a presença de conteúdo homofóbico.

Ao final da coleta dos julgamentos, o rótulo final de cada mensagem foi definido por maioria simples entre os três avaliadores. Nos casos em que não houve convergência entre os julgamentos (*i.e.,* cada julgador atribuiu um dos três rótulos) , a mensagem foi atribuída à categoria "não tem certeza". A distribuição final dos rótulos é apresentada na Tabela 3. O conjunto resultante dessa etapa foi denominado `Homofensa-PTbr`.

Rótulo final	**Quantidade**	**Percentual**
Sim	723	45%
Não	683	42%
Não tem certeza	214	13%
Total	1.620	100%

Tabela 3. Distribuição dos rótulos finais no conjunto `Homofensa-PTbr`.

Observa-se que 87% das mensagens puderam ser alocadas nas duas classes principais de interesse do estudo, com 723 textos rotulados como homofóbicos e 683 como não homofóbicos. Esse resultado indica que o procedimento de coleta orientado pelo léxico foi eficaz para recuperar mensagens relevantes ao domínio, mas também mostra que a simples presença de termos potencialmente ofensivos não é suficiente, por si só, para caracterizar homofobia. Em outras palavras, a etapa de julgamento humano revelou-se necessária para distinguir entre uso efetivamente discriminatório, uso ambíguo e usos não ofensivos dos termos recuperados.

Como medida adicional de concordância entre os três avaliadores, calculou-se o coeficiente Kappa de Fleiss, cujo valor obtido foi de 0,2654. Esse resultado indica baixa concordância entre os julgadores, sugerindo que a identificação de homofobia em mensagens curtas e contextualmente limitadas é uma tarefa complexa e sujeita a ambiguidades interpretativas. Tal achado reforça a importância da anotação humana na construção de bases de dados para esse domínio, bem como da manutenção de uma categoria de incerteza para lidar com casos fronteiriços.

Em conjunto, os resultados dessa etapa evidenciam a consolidação de um conjunto de dados rotulado em português do Brasil, construído especificamente para apoiar investigações sobre detecção automática de homofobia. O `Homofensa-PTbr` constitui, assim, um recurso empírico relevante tanto para análises descritivas quanto para o treinamento e a avaliação de modelos supervisionados.

4.4. Classificação automática dos textos

A etapa de classificação automática foi realizada com base no conjunto de dados `Homofensa-PTbr`, considerando apenas as instâncias rotuladas nas classes "sim" e "não" (Kappa de Fleiss = 0,3281). Com a exclusão dos casos marcados como "não tem

certeza", o conjunto final utilizado na tarefa de classificação totalizou 1.406 textos, distribuídos em 723 instâncias da classe "sim" e 683 da classe "não". Essa distribuição relativamente equilibrada permitiu conduzir a modelagem supervisionada sem necessidade de técnicas adicionais de balanceamento externo.

Para a configuração experimental, o conjunto de dados foi dividido em 80% para treinamento e validação interna e 20% para teste final, resultando em 282 instâncias reservadas para avaliação em *holdout*. Sobre a partição de treinamento, aplicou-se validação cruzada estratificada repetida com 5 partições e 3 repetições, totalizando 15 execuções por configuração avaliada. A seleção das melhores configurações utilizou o F1-macro.

Foram avaliados cinco algoritmos de classificação supervisionada, escolhidos por representarem diferentes famílias de modelos tradicionalmente empregadas em classificação de textos. A *Logistic Regression* foi utilizada por constituir uma referência sólida para dados textuais esparsos e de alta dimensionalidade. O *LinearSVC* foi incluído por sua reconhecida eficácia em problemas lineares com grande número de atributos. O *SGDClassifier* foi considerado por oferecer uma alternativa linear eficiente e escalável, com diferentes funções de perda. O *Complement Naive Bayes* foi adotado por sua adequação a representações baseadas em frequência e por seu uso recorrente em tarefas de mineração de textos. Por fim, o *Random Forest* foi incluído como modelo de *ensemble*, permitindo avaliar uma abordagem não linear; nesse caso, a representação textual foi precedida por redução de dimensionalidade via *Truncated SVD*, a fim de tornar a modelagem mais adequada ao espaço vetorial derivado do TF-IDF. A Tabela 4 resume os classificadores considerados e os respectivos espaços de busca explorados.

Algoritmo	**Grade de parâmetros avaliada**
Logistic Regression	$tfidf\_\_analyzer \in \{\texttt{word}, \texttt{char\_wb}\}$; $tfidf\_\_ngram\_range \in \{(1,1),(1,2),(1,3),(3,5),(3,6),(4,6)\}$; $tfidf\_\_min\_df \in \{1,2,3\}$; $tfidf\_\_sublinear\_tf \in \{\texttt{True}, \texttt{False}\}$; $clf\_\_C \in \{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$; $clf\_\_solver = \texttt{liblinear}$; $clf\_\_class\_weight \in \{\texttt{None}, \texttt{balanced}\}$
Complement Naive Bayes	$tfidf\_\_analyzer = \texttt{word}$; $tfidf\_\_ngram\_range = (1,2)$; $tfidf\_\_min\_df \in \{1,2\}$; $tfidf\_\_sublinear\_tf \in \{\texttt{True}, \texttt{False}\}$; $clf\_\_alpha \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$
SGDClassifier	$tfidf\_\_analyzer \in \{\texttt{word}, \texttt{char\_wb}\}$; $tfidf\_\_ngram\_range \in \{(1,2),(3,5)\}$; $tfidf\_\_min\_df = 2$; $tfidf\_\_sublinear\_tf \in \{\texttt{True}, \texttt{False}\}$; $clf\_\_loss \in \{\texttt{hinge}, \texttt{log\_loss}\}$; $clf\_\_alpha \in \{10^{-5}, 10^{-4}, 10^{-3}\}$; $clf\_\_class\_weight \in \{\texttt{None}, \texttt{balanced}\}$
LinearSVC	$tfidf\_\_analyzer \in \{\texttt{word}, \texttt{char\_wb}\}$; $tfidf\_\_ngram\_range \in \{(1,2),(3,5),(3,6)\}$; $tfidf\_\_min\_df \in \{1,2,3\}$; $tfidf\_\_sublinear\_tf \in \{\texttt{True}, \texttt{False}\}$; $clf\_\_C \in \{0.01, 0.05, 0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$; $clf\_\_class\_weight \in \{\texttt{None}, \texttt{balanced}\}$
Random Forest	$svd\_\_n\_components \in \{100, 200, 400\}$; $clf\_\_n\_estimators \in \{200, 500\}$; $clf\_\_max\_depth \in \{\texttt{None}, 20, 40\}$; $clf\_\_min\_samples\_split \in \{2, 5\}$; $clf\_\_min\_samples\_leaf \in \{1, 2, 4\}$; $clf\_\_max\_features \in \{\texttt{sqrt}, \texttt{log2}\}$; $clf\_\_class\_weight \in \{\texttt{None}, \texttt{balanced}\}$

Tabela 4. Classificadores e grades de parâmetros avaliados nos experimentos.

A Tabela 5 apresenta o melhor resultado em validação cruzada para cada algoritmo, considerando a melhor configuração obtida na busca em grade. Observa-se que a *Logistic Regression* alcançou o melhor desempenho médio em treinamento/CV, com F1-macro de 0,8649, seguida pelo *SGDClassifier* (0,8599) e pelo *LinearSVC* (0,8483). Em conjunto, os resultados indicam melhor desempenho das abordagens lineares em relação ao *Random Forest*, que obteve F1-macro de 0,8107. A melhor configuração da *Logistic Regression* combinou representação TF-IDF baseada em *character n-grams* do tipo *char_wb*, com faixa de 3 a 6 caracteres, frequência documental mínima igual a 3, frequência sublinear ativada, $C = 2.0$, *solver liblinear* e ponderação de classes *balanced*.

Algoritmo	Precisão (CV)	Revocação (CV)	F1-macro (CV)	Desv. pad.
Logistic Regression	0,8662	0,8655	0,8649	0,0283
SGDClassifier	0,8608	0,8603	0,8599	0,0260
LinearSVC	0,8499	0,8490	0,8483	0,0289
Complement Naive Bayes	0,8461	0,8452	0,8445	0,0236
Random Forest	—	—	0,8107	0,0294

Tabela 5. Melhor resultado em validação cruzada para cada algoritmo.

O resultado em validação cruzada é relevante por indicar que, no treinamento, uma representação baseada em padrões de caracteres foi mais eficaz do que configurações centradas exclusivamente em palavras. Esse comportamento sugere que, para a detecção de homofobia em textos curtos de redes sociais, variações ortográficas, abreviações e formas informais podem ser capturadas de modo mais eficiente por *character n-grams*.

Após a seleção da melhor configuração de cada algoritmo, procedeu-se à avaliação no conjunto de teste em *holdout*. A Tabela 6 apresenta os resultados comparativos. Diferentemente do observado em validação cruzada, o maior valor de F1-macro no conjunto de teste foi obtido pelo *LinearSVC*, com 0,8688. No entanto, os modelos lineares apresentaram desempenhos muito próximos, com F1-macro entre 0,8652 e 0,8688. Assim, embora o *LinearSVC* tenha obtido o maior valor no teste, a diferença em relação aos demais modelos lineares é pequena e deve ser interpretada com cautela.

Algoritmo	Precisão	Revocação	F1-macro
LinearSVC	0,8693	0,8694	0,8688
SGDClassifier	0,8655	0,8657	0,8652
Complement Naive Bayes	0,8655	0,8657	0,8652
Logistic Regression	0,8652	0,8655	0,8652
Random Forest	0,8232	0,8207	0,8190

Tabela 6. Comparação dos algoritmos no conjunto de teste (*holdout*), considerando a melhor configuração de cada modelo.

O resultado detalhado do *LinearSVC* é apresentado na Tabela 7. A melhor configuração desse modelo utilizou $C = 0.5$, sem ponderação de classes, com representação TF-IDF baseada em palavras, *n-gramas* de 1 a 2, frequência documental mínima igual a 1 e frequência sublinear ativada. No conjunto de teste, o modelo alcançou acurácia de 0,8688 e F1-macro de 0,8688. Para a classe "não", obteve-se precisão de 0,8472, revocação de 0,8905 e F1-score de 0,8683. Para a classe "sim", os valores correspondentes foram 0,8913, 0,8483 e 0,8693.

Classe	Precisão	Revocação	F1-score
Não	0,8472	0,8905	0,8683
Sim	0,8913	0,8483	0,8693
Macro avg	0,8693	0,8694	0,8688

Tabela 7. Desempenho do *LinearSVC*, melhor modelo no conjunto de teste.

A matriz de confusão correspondente pode ser observada na Tabela 8. A partir dos valores de precisão e revocação obtidos, verifica-se que o modelo apresentou desempenho equilibrado entre as duas classes, com leve vantagem de revocação para a classe "não" e de

precisão para a classe "sim". Esse comportamento sugere que o classificador foi capaz de generalizar de forma consistente em ambas as direções da decisão binária, sem concentrar excessivamente os acertos em apenas uma das classes.

Real / Previsto	**Não**	**Sim**
Não	122	15
Sim	22	123

Tabela 8. Matriz de confusão do *LinearSVC*.

Em conjunto, esses resultados indicam que o conjunto `Homofensa-PTbr` não apenas viabiliza a análise e a anotação de mensagens potencialmente homofóbicas, mas também sustenta o treinamento de modelos com desempenho promissor para a tarefa de classificação binária. De modo geral, os resultados mostram que abordagens lineares tradicionais, como *LinearSVC*, *Logistic Regression* e *SGDClassifier*, apresentaram desempenho superior ao do *Random Forest*, sugerindo que representações esparsas de texto combinadas a classificadores lineares constituem uma estratégia particularmente adequada para esse domínio. Entre os modelos avaliados, o *LinearSVC* apresentou o melhor desempenho no conjunto de teste, com F1-macro de 0,8688.

5. Conclusão

Este estudo investigou a detecção de homofobia em português do Brasil por meio de uma abordagem integrada, contemplando a ampliação de um léxico especializado, a construção de um conjunto de dados rotulado e a avaliação de modelos supervisionados de classificação automática. Como principais contribuições, o trabalho expandiu um repertório lexical previamente disponível, incorporando expressões coloquiais e variantes linguísticas associadas à homofobia, e produziu o conjunto `Homofensa-PTbr`, composto por mensagens coletadas em rede social e anotadas por julgadores humanos.

Os resultados indicam que o conjunto de dados construído sustenta a classificação binária entre conteúdos homofóbicos e não homofóbicos, com desempenho promissor dos modelos avaliados. As abordagens lineares apresentaram resultados próximos no conjunto de teste, com F1-macro entre 0,8652 e 0,8688, sendo o maior valor obtido pelo *LinearSVC*. Dessa forma, mais do que indicar vantagem expressiva de um classificador específico, os resultados sugerem que representações TF-IDF combinadas a modelos lineares constituem uma estratégia viável para a tarefa analisada.

Entre as limitações do estudo, destacam-se a dependência de uma coleta inicialmente guiada por léxico, o que pode restringir a recuperação de formas implícitas ou não lexicais de homofobia, e a baixa concordância entre os anotadores, possivelmente associada à subjetividade da tarefa, à ausência de contexto nas mensagens e à inexistência de uma rubrica formal extensa de anotação. Essas limitações também indicam a necessidade de avaliar o modelo em conjuntos externos ou coletados por estratégias não lexicais, a fim de verificar sua capacidade de generalização para manifestações de homofobia menos dependentes dos termos usados na coleta. Como trabalhos futuros, pretende-se ampliar o conjunto de dados com novas fontes textuais, elaborar diretrizes de anotação mais detalhadas, analisar qualitativamente os casos de discordância e os erros dos modelos, além de investigar abordagens baseadas em arquiteturas neurais e modelos de linguagem contextualizados para o domínio da detecção de homofobia em português do Brasil.

Referências

Antunes, M. B. A., Issa, M. d. F., and Hoed, R. M. (2023). TÉcnicas de machine learning aplicada a mineraÇÃo de dados e anÁlise de sentimentos para prediÇÃo de homofobia no twitter. *REVISTA FOCO*, 16(1):e853.

Caseli, H. d. M. and Nunes, M. d. G. V. (2024). Processamento de linguagem natural: conceitos, técnicas e aplicações em português.

Chakravarthi, B. R., Kumaresan, P. K., Priyadharshini, R., Buitelaar, P., Hegde, A., Shashirekha, H. L., Rajiakodi, S., García-Cumbreras, M. Á., Jiménez-Zafra, S. M., García-Díaz, J. A., Valencia-García, R., Ponnusamy, K. K., Shetty, P., and García-Baena, D. (2024). Overview of third shared task on homophobia and transphobia detection in social media comments. In Chakravarthi, B. R., B, B., Buitelaar, P., Durairaj, T., Kovács, G., and García Cumbreras, M. Á., editors, *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian's, Malta. Association for Computational Linguistics.

Chan, F. L., Nguyen, D., and Joshi, A. (2024). "is hate lost in translation?": Evaluation of multilingual LGBTQIA+ hate speech detection. In Baldwin, T., Rodríguez Méndez, S. J., and Kuo, N., editors, *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 146–152, Canberra, Australia. Association for Computational Linguistics.

Costa, A. B. and Nardi, H. C. (2015). Homofobia e preconceito contra diversidade sexual: debate conceitual. *Temas em Psicologia*, 23:715 – 726.

de Pelle, R. and Moreira, V. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, pages 510–519, Porto Alegre, RS, Brasil. SBC.

Fortuna, P., Rocha, J., Soler Company, J., Wanner, L., and Nunes, S. (2019). A hierarchically-labeled portuguese hate speech dataset. pages 94–104.

Hatebase, I. (2020). How it works? `https://hatebase.org/how_it_works`. Acessado em 13 de setembro 2020.

Kumaresan, P. K., Ponnusamy, R., Sharma, D., Buitelaar, P., and Chakravarthi, B. R. (2024). Dataset for identification of homophobia and transphobia for Telugu, Kannada, and Gujarati. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4404–4411, Torino, Italia. ELRA and ICCL.

Locatelli, D., Damo, G., and Nozza, D. (2023). A cross-lingual study of homotransphobia on Twitter. In Dev, S., Prabhakaran, V., Adelani, D. I., Hovy, D., and Benotti, L., editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 16–24, Dubrovnik, Croatia. Association for Computational Linguistics.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., and Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):1–16. `https://doi.org/10.1371/journal.pone.0221152`. Acessado em: 2020-09-21.

Santos, V., Henriques, F., and Guedes, G. (2022). O discurso de Ódio homofóbico no twitter a partir da análise de dados. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 109–120, Porto Alegre, RS, Brasil. SBC.