Detecção de Misoginia em Redes Sociais

João Rangel[1], Alexander Feitosa[1], Alice Rios[1], Lilian Ferrari[2], Fabio Junior[1], Kele Belloze[1], Eduardo Ogasawara[1], Gustavo Guedes[1]

[1] CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca

[2]UFRJ - Universidade Federal do Rio de Janeiro

{joao.rangel.1, alexander.feitosa,alice.rios}@aluno.cefet-rj.br, lilianferrari@uol.com.br, {fabio.junior, kele.belloze}@cefet-rj.br, {eduardo.ogasawara, gustavo.guedes}@cefet-rj.br

Abstract. *This paper contributes to the creation of linguistic resources for misogyny detection in Brazilian Portuguese. To this end, it combines lexical construction, social media message collection, and human annotation. As a result, the* `LexMis-BR` *lexicon was proposed, guiding message collection and the selection of a subset for annotation. The annotation stage showed that lexical recovery is useful for corpus construction, but insufficient as a final decision criterion, reinforcing the importance of human validation. Using the annotated dataset, supervised classification models were evaluated.*

Resumo. *Este artigo busca contribuir com a criação de recursos linguísticos para a detecção de misoginia em português do Brasil. Para isso, combina construção lexical, coleta de mensagens em redes sociais e anotação humana. Como resultado, foi proposto o léxico* `LexMis-BR`*, que orientou a coleta e a seleção de um subconjunto para anotação. A anotação mostrou que a recuperação lexical é útil para construir o corpus, mas insuficiente como critério de decisão final, reforçando a importância da validação humana. Com esse conjunto, foram avaliados modelos supervisionados de classificação.*

1. Introdução

A misoginia constitui um problema social de alta relevância contemporânea, tanto por sua incidência na vida cotidiana das mulheres quanto por sua expressão em formas diversas de violência. Em sentido amplo, ela pode ser compreendida como ódio, desprezo ou preconceito direcionado às mulheres, manifestando-se por meio de práticas discriminatórias, sexistas e violentas [Srivastava et al., 2017]. No caso brasileiro, a gravidade desse fenômeno torna-se particularmente evidente quando observada à luz de dados oficiais recentes: segundo o Relatório Anual Socioeconômico da Mulher de 2025, o país contabilizou 71.892 casos de estupro em 2024, o equivalente a 196 vítimas por dia, além de 1.450 feminicídios no mesmo ano [Ministério das Mulheres, 2025].

As redes sociais, em particular, têm se consolidado como espaços de formação, circulação e reforço de discursos misóginos, frequentemente associados a comunidades digitais como os grupos *redpill* e *incel*. Em plataformas como o *Twitter*, esse tipo de conteúdo encontra condições favoráveis de difusão, seja sob a forma de insultos diretos, seja por meio da naturalização de visões de mundo que reforçam a submissão feminina

[Lima-Santos and Santos, 2022; Koch et al., 2025]. Embora essas plataformas disponham de políticas de enfrentamento ao discurso de ódio, sua aplicação depende, em grande medida, de mecanismos de denúncia e de revisão manual, restringindo a contenção efetiva e reforçando a necessidade de recursos computacionais para esse enfrentamento.

A relevância do problema também vem sendo reconhecida nos campos institucional, científico e legislativo. No âmbito da Agenda 2030 das Nações Unidas, o Objetivo de Desenvolvimento Sustentável 5 estabelece como meta eliminar todas as formas de discriminação e violência contra mulheres e meninas, inclusive nas esferas pública e privada [Nações Unidas no Brasil, 2026]. Mais recentemente, em março de 2026, o Plenário do Senado aprovou a inclusão da misoginia entre os crimes de preconceito ou discriminação, definindo-a como conduta que exterioriza ódio ou aversão às mulheres e incluindo a expressão "condição de mulher" entre os critérios de interpretação da Lei nº 7.716/1989; a matéria seguiu para a Câmara dos Deputados [Senado Federal, 2026].

A literatura internacional já reúne esforços relevantes voltados à detecção de misoginia, incluindo os *shared tasks* AMI (*Automatic Misogyny Identification*), estudos multilíngues e revisões do estado da arte [Fersini et al., 2018; Pamungkas et al., 2020; Shushkevich and Cardiff, 2019]. No entanto, a adaptação dessa agenda ao português do Brasil ainda enfrenta uma lacuna importante: são escassos estudos que construam e validem, para esse contexto, um léxico expandido, um corpus coletado por recuperação lexical e um conjunto anotado utilizável em experimentação inicial.

Diante desse quadro, este artigo busca contribuir com a criação de recursos linguísticos e anotados para apoiar a detecção de misoginia em português do Brasil. Para isso, o estudo organiza um fluxo no qual o léxico de termos misóginos criado (`LexMis-BR`) orienta a coleta de mensagens em redes sociais. As contribuições centrais do trabalho consistem na produção articulada de um léxico de termos misóginos em português do Brasil, de um conjunto de dados anotado, de um protocolo experimental reprodutível e de uma *baseline* de classificação supervisionada. A etapa de classificação, portanto, não é apresentada como a principal inovação técnica do artigo, mas como uma avaliação inicial da utilidade dos recursos produzidos.

Em conjunto, esses elementos oferecem uma base empírica inicial para o estudo computacional da misoginia em português do Brasil, com disponibilização dos artefatos possíveis, respeitando as restrições éticas da pesquisa e as regras da plataforma utilizada. Nesse sentido, a pesquisa parte da seguinte questão: em que medida uma abordagem de recuperação lexical combinada à validação humana permite construir recursos úteis para o estudo computacional da misoginia em português do Brasil? A pesquisa foi aprovada por Comitê de Ética em Pesquisa, sob CAEE 75789023.0.0000.5289. Vale destacar que parte dos recursos produzidos neste estudo encontra-se disponibilizada em repositório *on-line*: `https://eic.cefet-rj.br/mmcomp/misoginia2026`.

O restante do artigo está organizado em quatro seções: a Seção 2 apresenta os trabalhos relacionados; a Seção 3, a metodologia; a Seção 4, os resultados e a discussão; e a Seção 5, as conclusões.

2. Trabalhos Relacionados

No cenário internacional, a detecção de misoginia já foi explorada em tarefas compartilhadas, estudos experimentais e trabalhos de síntese. No âmbito das campanhas AMI,

Fersini et al. [2018] descrevem a tarefa de *Automatic Misogyny Identification* com dados de *Twitter* em italiano e inglês, organizada em duas subtarefas: identificação de conteúdo misógino e classificação de comportamento misógino e alvo. Os autores observam que a identificação binária da misoginia foi tratada de forma mais satisfatória pelas equipes participantes, enquanto a classificação do comportamento e do alvo permaneceu mais desafiadora. Em perspectiva de revisão, Shushkevich and Cardiff [2019] sintetizam abordagens para misoginia em redes sociais, sobretudo no *Twitter*, cobrindo modelos clássicos e neurais em inglês, espanhol e italiano. Em complemento, Pamungkas et al. [2020] investigam a tarefa em inglês, italiano e espanhol sob uma perspectiva multilíngue e *cross-domain*, indicando que a misoginia constitui um fenômeno específico no interior da linguagem abusiva e não se confunde diretamente com o sexismo.

No contexto da língua portuguesa, os estudos disponíveis ainda aparecem de forma mais pontual. Braga et al. [2021] apresentam um corpus de discurso sexista em português, enfatizando a construção e a caracterização de uma base anotada para a tarefa. Em seguida, Plath et al. [2022] descrevem a construção da base MINA-BR para detecção de discurso de ódio contra mulheres em português brasileiro e apresentam uma *baseline* de classificação, destacando a escassez de bases em idiomas diferentes do inglês como motivação do estudo. Trata-se de contribuições importantes para o domínio, mas com foco concentrado na construção de corpora anotados para classificação supervisionada. Mais recentemente, Vargas et al. [2025] apresentam recursos contextuais e anotados por especialistas para detecção de discurso de ódio em português brasileiro, reforçando a relevância de recursos sensíveis a contexto e a particularidades culturais. Já Martins et al. [2025] analisam a misoginia no YouTube brasileiro em conteúdos associados à comunidade *Red Pill*, deslocando a discussão para outra plataforma e reforçando a relevância do fenômeno em ecossistemas digitais brasileiros distintos do *Twitter*. Diferentemente dessas propostas, o presente estudo enfatiza a construção de um léxico de termos misóginos em português do Brasil, a recuperação de mensagens orientada por esse léxico, a validação humana e a apresentação de uma *baseline* supervisionada inicial para o recurso produzido.

Em conjunto, essa literatura sugere uma lacuna mais específica: ainda são escassos trabalhos que construam e validem recursos para um fenômeno em que a simples presença de termos ofensivos não basta para caracterizar misoginia. É nesse espaço que se insere o presente trabalho, ao combinar expansão lexical orientada por participantes, coleta de mensagens em rede social, anotação humana e modelagem computacional para produzir um léxico de termos misóginos expandido (`LexMis-BR`), um conjunto de dados anotado e uma base empírica para investigações futuras.

3. Metodologia

Esta seção descreve a abordagem metodológica geral adotada no estudo, contemplando as etapas de construção lexical, coleta de mensagens, anotação humana e avaliação inicial de modelos supervisionados. Os detalhes empíricos da aplicação dessa abordagem, incluindo fontes utilizadas, quantitativos, perfil dos participantes, composição dos conjuntos de dados e resultados obtidos, são apresentados na Seção 4.

O presente estudo tem como objetivo construir e avaliar uma abordagem de recuperação lexical e validação humana para apoiar a detecção de misoginia em mensagens textuais em português, com foco em publicações extraídas de redes sociais. Mais

especificamente, busca-se produzir um léxico expandido, consolidar um conjunto anotado para o domínio e verificar se esse material sustenta uma avaliação inicial de modelos supervisionados. O trabalho foi organizado em quatro etapas (Fig. 1): (i) criação do léxico de termos misóginos; (ii) coleta de mensagens em redes sociais; (iii) anotação dos textos com apoio de avaliadores humanos; e (iv) avaliação inicial de modelos supervisionados de classificação. O desenho parte de duas premissas complementares: a recuperação lexical aumenta a probabilidade de localizar mensagens relevantes para o domínio, mas a caracterização final da misoginia depende de validação humana e contextual.

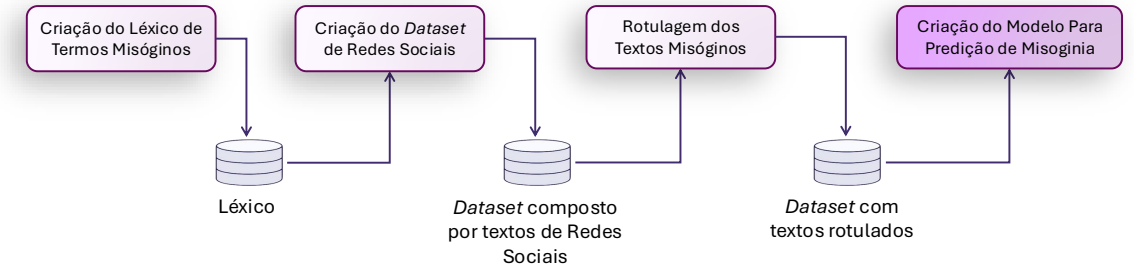

Figura 1. Fluxo adotado no estudo, contemplando a criação do léxico de termos misóginos, a construção do corpus de redes sociais, a anotação dos textos e a avaliação inicial de modelos para detecção de misoginia.

3.1. Criação do Léxico de Termos Misóginos

A primeira etapa da metodologia consiste na construção de um léxico de termos misóginos em português do Brasil a partir da ampliação de um repertório lexical previamente selecionado. Assim, parte-se de um conjunto inicial de termos e expressões potencialmente associados à misoginia. A partir disso, emprega-se um instrumento de coleta voltado à elicitação de novas expressões do domínio. Esse instrumento é estruturado de modo a expor os participantes ao conjunto lexical inicial e, em seguida, solicitar a indicação de outros termos ou expressões misóginas não contemplados nesse conjunto.

A abordagem 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. Com isso, procura-se reduzir a dependência de listas produzidas em outros contextos e ampliar a cobertura de usos aderentes ao português brasileiro contemporâneo. 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, com o objetivo de reduzir a inclusão de formas muito idiossincráticas, ambíguas ou pouco representativas. Assim, os itens só passam a integrar o léxico final quando atingem o ponto de corte de recorrência definido para cada etapa de coleta. Como resultado dessa etapa, obtém-se um léxico expandido de termos e expressões misóginas, posteriormente utilizado para orientar a coleta de mensagens em redes sociais.

3.2. Criação do Corpus de Redes Sociais

A segunda etapa da abordagem consiste na criação de um corpus de mensagens textuais extraídas de redes sociais a partir do léxico expandido obtido na etapa anterior. Para

isso, emprega-se um procedimento de coleta orientado por palavras-chave, no qual os itens do léxico são utilizados como referências para a busca de mensagens publicadas em ambiente digital. O corpus resultante dessa etapa corresponde ao material bruto de recuperação orientada sobre o qual incidem as etapas subsequentes de seleção e anotação.

Após a coleta inicial, o material textual obtido é submetido a uma etapa de organização do corpus, com o objetivo de estruturar o conjunto de dados de forma adequada às etapas subsequentes. Esse processo envolve a consolidação das mensagens recuperadas, a remoção de duplicidades e a padronização do formato de armazenamento dos textos, de modo a garantir maior consistência ao corpus.

Nesta etapa do fluxo, não se pressupõe que toda mensagem recuperada por meio do léxico seja necessariamente misógina, razão pela qual a criação do corpus corresponde a uma etapa de recuperação orientada, e não de classificação final.

3.3. Rotulagem dos Textos Misóginos

A terceira etapa da abordagem consiste na anotação das mensagens coletadas na fase anterior, com o objetivo de produzir um conjunto de dados anotado para o domínio da misoginia. Nessa etapa, parte-se do pressuposto de que a simples presença de termos do léxico não é suficiente para determinar, por si só, o caráter misógino de uma mensagem, uma vez que esses itens podem ocorrer em contextos variados e assumir diferentes funções discursivas.

Para enfrentar essa limitação, adota-se um procedimento de anotação humana, no qual participantes avaliam as mensagens recuperadas e atribuem rótulos de acordo com sua percepção sobre o conteúdo apresentado. Esse procedimento busca incorporar julgamento contextual à classificação dos textos, permitindo distinguir ocorrências efetivamente misóginas de usos não ofensivos, ambíguos ou indeterminados. Para viabilizar essa etapa, o instrumento de coleta é estruturado de modo a apresentar subconjuntos de mensagens aos participantes, equilibrando a distribuição dos textos e reduzindo efeitos de fadiga durante o processo de anotação.

As respostas obtidas são, então, consolidadas para formar o conjunto de dados anotado, associando cada mensagem a uma categoria de análise (*e.g.*, misógina x não misógina). Desse modo, obtém-se um conjunto de dados anotado que reflete a interpretação humana do fenômeno investigado e que, ao mesmo tempo, constitui uma base de referência para a etapa de modelagem computacional. Para a experimentação supervisionada, esse conjunto anotado dá origem a um subconjunto binário composto apenas pelas instâncias classificadas nas categorias decisórias do protocolo.

3.4. Avaliação inicial de modelos para detecção de misoginia

A quarta etapa da abordagem consiste na avaliação inicial de modelos supervisionados de classificação para detecção de misoginia em mensagens textuais. Nessa fase, busca-se verificar em que medida o conjunto de dados anotado produzido nas etapas anteriores é capaz de sustentar uma tarefa computacional de classificação.

Para essa finalidade, emprega-se um processo de preparação textual das mensagens, sua representação em formato numérico e o treinamento de modelos supervisionados de classificação. Esse processo busca converter o conteúdo textual em atributos compatíveis com o processo de aprendizagem algorítmica, permitindo que o modelo

estabeleça relações entre características linguísticas dos textos e os rótulos atribuídos na etapa de anotação.

O processamento inicial envolve a organização e o tratamento dos textos, com o objetivo de reduzir ruídos e padronizar o material de entrada. Em seguida, aplica-se uma técnica de vetorização para representar computacionalmente as mensagens, de modo a preservar informações relevantes para a tarefa de classificação. Essa representação constitui a base sobre a qual os algoritmos são ajustados e avaliados.

A modelagem é conduzida por classificação supervisionada, na qual diferentes algoritmos podem ser empregados e comparados em termos de desempenho. Para isso, utiliza-se um protocolo de avaliação que permite estimar a capacidade de generalização dos modelos, bem como identificar a configuração mais adequada para a tarefa proposta.

4. Resultados e Discussão

Esta seção apresenta os principais resultados obtidos nas diferentes etapas do estudo, contemplando a construção do léxico de termos misóginos, a coleta e a anotação de mensagens do *Twitter*, além da avaliação inicial de modelos supervisionados de classificação. Em conjunto, esses resultados evidenciam a viabilidade do fluxo adotado para a geração de um recurso lexical expandido, um conjunto de dados anotado e uma evidência experimental inicial para o estudo computacional da misoginia em português do Brasil.

4.1. Construção do léxico de termos misóginos

Como ponto de partida para a ampliação lexical proposta neste estudo, utilizou-se um repertório inicial de termos e expressões potencialmente associados à misoginia em português, composto a partir de diferentes fontes. Esse repertório reuniu 11 termos extraídos do *Hatebase*[1] e cinco termos de dois estudos anteriores sobre linguagem, gênero e xingamentos no contexto brasileiro. A Tabela 1 apresenta os termos utilizados como ponto de partida. Os termos com asterisco foram extraídos dos trabalhos de Zanello et al. [2011] e Zanello and Romero [2012] (*i.e., puta*, *prostituta*, *piriguete*, *piranha* e *galinha*). Os demais termos foram extraídos do *Hatebase*.

Tabela 1. Termos inicialmente selecionados.

Puta*	Prostituta*	Piriguete*	Piranha*	Galinha*	Baleia	Baranga	Cabra
Camafeu	Fufa	Galdéria	Mulata	Pega	Perua	Rameira	Sapatão

A etapa de ampliação do léxico de termos misóginos foi realizada por meio de formulário eletrônico, aplicado entre março e junho de 2024. Ao término da coleta, obteve-se a participação de 143 respondentes. A amostra da criação do léxico foi composta majoritariamente por residentes do estado do Rio de Janeiro (83%) e por mulheres cisgênero (68,5%); quanto à faixa etária, a maior parte dos participantes tinha entre 18 e 30 anos (53%). Esse perfil é particularmente relevante, uma vez que o estudo incide sobre manifestações discursivas que afetam diretamente mulheres em ambientes digitais.

Na etapa de validação do repertório inicial, os participantes avaliaram os 16 termos apresentados inicialmente, indicando se os consideravam ou não misóginos. Para selecionar os itens mais representativos dessa etapa, adotou-se um critério inspirado na

[1]Os termos no *Hatebase* estão em *Advanced Search*, opções *'female'* no gênero e *'Portuguese'* na língua.

lógica de Pareto (80/20), que aponta que 80% dos resultados são gerados por 20% das causas. Assim, foram mantidos os itens classificados como misóginos por pelo menos 20% dos participantes, o que corresponde a, no mínimo, 29 ocorrências. Com base nesse critério, 12 dos 16 termos iniciais foram incorporados ao léxico final: *piranha* (129), *puta* (126), *galinha* (118), *baranga* (117), *piriguete* (115), *baleia* (93), *sapatão* (91), *perua* (76), *prostituta* (68), *mulata* (67), *rameira* (63) e *cabra* (37). Por outro lado, *camafeu*, *fufa*, *galdéria* e *pega* não atingiram esse ponto de corte (*i.e.*, 29), o que sugere baixa circulação no português brasileiro contemporâneo ou uso mais restrito a variedades regionais e outros contextos.

Na etapa de sugestões abertas, os participantes indicaram novos termos e expressões associados à misoginia. Essas contribuições textuais foram submetidas a pré-processamento, envolvendo segmentação, normalização e contagem de ocorrências. Como critério de inclusão, consideraram-se apenas os itens mencionados ao menos três vezes, valor superior à média geral de recorrência observada entre os termos sugeridos (2,12 aparições por item). Com base nesse procedimento, 20 novos termos foram incorporados ao repertório final.

A Tabela 2 apresenta o léxico final (`LexMis-BR`), composto por 32 termos e expressões organizados em ordem decrescente de ocorrência. Os termos marcados com asterisco pertencem ao repertório inicial, enquanto os demais foram sugeridos espontaneamente pelos participantes. Em comparação com a lista inicial de 16 itens, o recurso obtido amplia substancialmente a cobertura lexical do domínio e incorpora formas mais correntes no português do Brasil, especialmente no contexto de interações em redes sociais. Os números entre parênteses indicam a quantidade de participantes que reconheceram cada termo como misógino. Naturalmente, os itens marcados com asterisco tendem a apresentar frequências mais elevadas, uma vez que já integravam o formulário inicial de avaliação. Em contraste, os termos sem asterisco tendem a exibir frequências menores por terem sido mencionados em campo aberto, no qual os participantes podiam sugerir livremente novas expressões. Ainda assim, observa-se que o termo *vadia*, com 30 menções, aproxima-se da frequência de *cabra*, identificado como misógino por 37 participantes.

Tabela 2. Léxico final (`LexMis-BR`) em ordem decrescente de ocorrência.

piranha* (129)	puta* (126)	galinha* (118)	baranga* (117)	piriguete* (115)
baleia* (93)	sapatão* (91)	perua* (76)	prostituta* (68)	mulata* (67)
rameira* (63)	cabra* (37)	vadia (30)	vagabunda (22)	vaca (12)
histérica (9)	cachorra (9)	louca (8)	mulherzinha (8)	burra (4)
rapariga (4)	desequilibrada (4)	biscate (4)	mal comida (4)	mal amada (4)
maluca (4)	quenga (3)	mulher fácil (3)	cadela (3)	velha (3)
descontrolada (3)	doida (3)			

Cabe destacar que os itens do `LexMis-BR` não devem ser interpretados como misóginos em qualquer ocorrência isolada. Muitos termos do léxico são polissêmicos, dependem do contexto discursivo e podem variar regionalmente em circulação, intensidade ofensiva e alvo social. Por essa razão, o léxico foi utilizado neste estudo como instrumento de recuperação de mensagens potencialmente relevantes, e não como mecanismo automático de rotulagem.

Além do aumento no número de entradas, o léxico final evidencia regularidades semânticas importantes. Observa-se concentração de insultos relacionados à sexualidade

e ao comportamento feminino, como *puta*, *piriguete*, *piranha*, *vadia* e *vagabunda*, de termos associados à desqualificação psicológica, como *histérica*, *louca*, *maluca*, *desequilibrada* e *descontrolada*, e de formas de animalização simbólica. Também se destaca um padrão recorrente de animalização visível em usos associados à metáfora "Ser humano é animal" [Lakoff and Johnson, 1980]. Mais especificamente, a metáfora "Mulher é animal" tem se tornado recorrente no português brasileiro [Marques, 2025], ocorrendo em formas como *galinha*, *baleia*, *vaca*, *cachorra* e *piranha*. Em linhas gerais, esses usos metafóricos constituem estratégias discursivas de desumanização dirigidas às mulheres.

Em conjunto, os resultados dessa etapa indicam que a participação humana foi decisiva para ampliar a cobertura lexical do recurso, incorporando termos e expressões mais aderentes ao uso real do português do Brasil. Isso é particularmente relevante em um domínio como o da misoginia, no qual insultos, gírias, metáforas depreciativas e variações linguísticas desempenham papel central na manifestação do discurso ofensivo. Assim, o léxico resultante não apenas amplia a lista inicial, mas também reforça a necessidade de recursos sensíveis a particularidades sociolinguísticas do contexto brasileiro.

4.2. Coleta e caracterização dos textos

Com base no léxico `LexMis-BR`, realizou-se a coleta de mensagens na rede social *Twitter*, com o objetivo de constituir um conjunto de dados textual potencialmente relacionado ao domínio da misoginia. Para essa etapa, empregou-se um processo automatizado de extração, configurado a partir dos 32 termos e expressões apresentados na Tabela 2. A extração foi limitada a 100 *tweets* para cada item lexical do `LexMis-BR`, considerando apenas publicações em língua portuguesa, com no mínimo cinco curtidas, excluindo-se links e respostas. Além disso, restringiu-se a coleta ao período compreendido entre janeiro e junho de 2024. O conjunto de dados desenvolvido nesta etapa foi denominado `MisoginiaBR-Coleta`, formado por 3200 (*i.e.*, 100 para cada um dos 32 termos) mensagens recuperadas a partir de consultas orientadas pelo léxico. Esse procedimento buscou equilibrar a representatividade dos diferentes termos do `LexMis-BR` com a viabilidade das etapas posteriores de anotação manual.

Após a extração, as mensagens foram organizadas em formato tabular, associando cada texto ao termo lexical que motivou sua recuperação e a metadados mínimos necessários ao controle da coleta. Em seguida, foram removidas duplicidades e registros incompatíveis com os critérios definidos. O procedimento parte do pressuposto de que a presença de termos do léxico aumenta a probabilidade de recuperação de mensagens potencialmente misóginas, mas não garante, por si só, que todas as ocorrências correspondam efetivamente a discursos de ódio contra mulheres. Assim, o conjunto `MisoginiaBR-Coleta` deve ser compreendido como um corpus inicial de mensagens potencialmente relevantes para o domínio, cuja interpretação final depende da etapa seguinte de anotação humana.

4.3. Anotação dos textos e composição do conjunto de dados anotado

Considerando que etapas de anotação com participação humana demandam tempo, coordenação e esforço interpretativo consideráveis, optou-se, neste estudo, por concentrar a anotação em um subconjunto lexical mais representativo do domínio. Em vez de submeter à anotação manual todas as ocorrências recuperadas a partir dos 32 termos do

`LexMis-BR`, priorizaram-se os itens de maior relevância em duas frentes complementares: (i) os cinco termos mais recorrentes do repertório inicial já documentado na literatura, a saber, *piranha*, *puta*, *galinha*, *baranga* e *piriguete*; e (ii) os cinco termos mais recorrentes sugeridos espontaneamente pelos participantes na etapa de expansão lexical, a saber, *vadia*, *vagabunda*, *vaca*, *histérica* e *cachorra*.

Esse recorte busca equilibrar viabilidade metodológica e relevância empírica. Por um lado, contempla itens já consolidados em estudos anteriores sobre misoginia e xingamentos de gênero; por outro, incorpora formas lexicais emergentes da percepção dos próprios respondentes, mais diretamente associadas ao uso corrente em português do Brasil. Com isso, a anotação concentra-se em publicações com maior probabilidade de configurar discurso misógino, preservando a viabilidade da etapa de anotação.

A partir desses dez termos, selecionou-se um subconjunto de 300 *tweets*, com 30 mensagens para cada termo. Esse subconjunto foi selecionado aleatoriamente no interior do corpus previamente coletado, de modo a preservar a diversidade contextual e, simultaneamente, tornar a etapa de anotação humanamente viável. O conjunto resultante foi denominado `MisoginiaBR-10-300` (10 termos do léxico e 300 mensagens).

Para a etapa de avaliação, o `MisoginiaBR-10-300` foi segmentado em dez listas contendo 30 *tweets* cada, distribuídas aleatoriamente para reduzir a fadiga dos participantes e mitigar possíveis vieses de ordem. Essa organização teve como objetivo tornar viável a análise integral de cada lista por um mesmo respondente. Após a consolidação desse subconjunto, elaborou-se um novo instrumento de coleta por meio do *Google Forms*. Inicialmente, cada participante informava dados sociodemográficos, como gênero, idade e localização, além de confirmar o Termo de Consentimento Livre e Esclarecido. Em seguida, o respondente era direcionado à avaliação de uma lista contendo 30 textos selecionados aleatoriamente. Conforme as listas atingiam cinco respostas, eram desabilitadas, de maneira que outras listas fossem priorizadas.

Durante a anotação, os participantes rotularam cada mensagem em uma entre três categorias: "Considero misógina", "Não considero misógina" e "Não sei definir". A presença dessa terceira categoria buscou contemplar casos ambíguos, limítrofes ou insuficientemente claros, evitando forçar decisões categóricas em situações incertas. Para a consolidação do conjunto anotado, o rótulo final de cada mensagem foi definido pela categoria majoritária entre os avaliadores. Quando não houve maioria simples entre as respostas, não foi realizado desempate forçado; nesses casos, a mensagem foi mantida separadamente como empate e excluída da etapa binária de classificação supervisionada.

Ao final do período de coleta, foram obtidas 54 respostas válidas, assegurando uma participação mínima de cinco avaliadores por texto. A amostra da rotulação foi composta por 67% de mulheres cisgênero e 33% de homens cisgênero; quanto à faixa etária, 61% dos participantes tinham entre 18 e 30 anos. Em relação à localização, 93% declararam residir no estado do Rio de Janeiro. Esse procedimento permitiu compor um conjunto de dados anotado com base em julgamento humano, constituindo uma referência empírica para a etapa de avaliação de modelos supervisionados de classificação. O conjunto de dados rotulado foi denominado `MisoginiaBR-300-rot`.

Embora a participação não tenha sido restrita exclusivamente a mulheres ou a especialistas no tema, a predominância de mulheres cisgênero no grupo de avaliadores é

relevante para o domínio investigado, uma vez que a misoginia afeta diretamente mulheres em ambientes digitais. Além disso, a coleta de múltiplas avaliações independentes por mensagem buscou reduzir a dependência de julgamentos individuais e incorporar diferentes leituras contextuais sobre o conteúdo analisado.

A distribuição final dos rótulos atribuídos às 300 mensagens do conjunto `MisoginiaBR-300-rot` indica que a maior parte se concentra nas classes "Não considero misógina" (142 textos; 47,3%) e "Considero misógina" (122 textos; 40,7%), que juntas correspondem a 88,0% do conjunto anotado. Em contrapartida, 10 textos (3,3%) foram classificados como "Não sei definir", e 26 (8,7%) permaneceram em situação de empate entre duas das três categorias de anotação. Embora a estratégia de seleção lexical tenha favorecido a recuperação de mensagens relevantes para o domínio, a simples presença dos termos não implica, necessariamente, a ocorrência de misoginia. Em outras palavras, a recuperação lexical mostrou-se útil para orientar a coleta, mas insuficiente como critério de decisão final. A frequência ligeiramente superior da classe "Não considero misógina", somada à existência de casos de empate e de mensagens classificadas como "Não sei definir", reforça a importância da anotação humana para distinguir usos efetivamente misóginos, usos não ofensivos e ocorrências ambíguas. Em conjunto, esses achados respondem diretamente à pergunta de pesquisa ao indicar que a recuperação lexical constitui uma etapa útil de construção do corpus, mas que a utilidade dos recursos produzidos depende da validação humana para consolidar um conjunto anotado confiável.

4.4. Classificação supervisionada dos textos

A etapa de classificação supervisionada foi conduzida como uma *baseline* experimental, com o objetivo de verificar se o conjunto de dados anotado produzido neste estudo sustenta uma tarefa inicial de detecção automática de misoginia. Para isso, foram consideradas apenas as instâncias pertencentes às classes "Considero misógina" e "Não considero misógina". Desse modo, casos marcados como empate ou "Não sei definir" não integraram a tarefa binária final. A partir dessa seleção, obteve-se um conjunto final de 264 textos, dos quais 53 instâncias (20%) foram reservadas para teste final em *holdout* e 211 instâncias (80%) foram utilizadas para treinamento e validação interna. Essa configuração permitiu conduzir a tarefa como um problema de classificação binária supervisionada, com avaliação separada entre ajuste do modelo e teste final de generalização.

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 foi orientada pela métrica F1-macro, por se tratar de uma medida mais apropriada para avaliar, de forma equilibrada, o desempenho do classificador nas duas classes de interesse e reduzir uma leitura excessivamente dependente da classe mais frequente. Além disso, foram também computadas as métricas de precisão e revocação, a fim de permitir uma análise mais detalhada.

Foram avaliados quatro algoritmos amplamente empregados em tarefas de classificação de textos: *Logistic Regression*, *LinearSVC*, *SGDClassifier* e *Complement Naive Bayes*. Os três primeiros representam alternativas lineares consolidadas para dados textuais esparsos e de alta dimensionalidade, enquanto o último permanece relevante em representações baseadas em frequência. A Tabela 3 resume os classificadores e os respectivos espaços de busca explorados, com nomenclatura simplificada dos parâmetros.

Tabela 3. Classificadores e grades de parâmetros dos experimentos.

Algoritmo	Grade de parâmetros avaliada
Logistic Regression	analisador $\in$ {`word`, `char_wb`}; faixa de n-gramas $\in$ $\{(1,1),(1,2),(1,3),(3,5),(3,6),(4,6)\}$; frequência documental mínima $\in$ $\{1,2,3\}$; escala sublinear de frequência $\in$ {`True`, `False`}; $C \in \{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$; ponderação de classes $\in$ {`None`, `balanced`}
Complement Naive Bayes	analisador = `word`; faixa de n-gramas = $(1,2)$; frequência documental mínima $\in \{1,2\}$; escala sublinear de frequência $\in$ {`True`, `False`}; $\alpha \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$
SGD Classifier	analisador $\in$ {`word`, `char_wb`}; faixa de n-gramas $\in$ $\{(1,2),(3,5)\}$; frequência documental mínima = 2; escala sublinear de frequência $\in$ {`True`, `False`}; função de perda $\in$ {`hinge`, `log_loss`}; $\alpha \in \{10^{-5}, 10^{-4}, 10^{-3}\}$; ponderação de classes $\in$ {`None`, `balanced`}
Linear SVC	analisador = `word`; faixa de n-gramas = $(1,2)$; frequência documental mínima $\in \{1,2\}$; escala sublinear de frequência $\in$ {`True`, `False`}; $C \in \{0.01, 0.05, 0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$; ponderação de classes $\in$ {`None`, `balanced`}

A Tabela 4 apresenta, lado a lado, os melhores resultados na validação cruzada e no teste *holdout* para cada algoritmo. A *Logistic Regression* obteve o melhor desempenho médio na validação cruzada, seguida pelo *LinearSVC* e *SGDClassifier*. O *Complement Naive Bayes*, embora com resultados inferiores, também apresentou resultado consistente.

Tabela 4. Desempenho dos algoritmos em validação cruzada e *holdout*.

Algoritmo	Treino - validação cruzada				Teste - *holdout*		
	Precisão	Revocação	F1-macro	Desv. pad.	Precisão	Revocação	F1-macro
Logistic Regression	0,801	0,798	0,798	0,078	0,735	0,730	0,731
LinearSVC	0,793	0,788	0,788	0,090	0,618	0,612	0,611
SGDClassifier	0,789	0,781	0,783	0,073	0,696	0,692	0,693
Complement Naive Bayes	0,781	0,773	0,772	0,076	0,618	0,612	0,611

A superioridade da *Logistic Regression* em validação cruzada é particularmente relevante porque sua melhor configuração foi baseada em TF-IDF com n-gramas de caracteres do tipo *char_wb*, indicando que padrões sublexicais tiveram papel importante na discriminação entre mensagens misóginas e não misóginas. O melhor resultado foi obtido com $C = 1.0$, ponderação balanceada de classes, frequência documental mínima igual a 3, escala sublinear de frequência ativada e representação TF-IDF com n-gramas de caracteres de 4 a 6 posições.

Após a seleção da melhor configuração de cada algoritmo, procedeu-se à avaliação no conjunto de teste em *holdout*. A Tabela 4 mostra que a *Logistic Regression* permaneceu como o modelo de melhor desempenho, com F1-macro de 0,731, seguida por *SGDClassifier* (0,693). *LinearSVC* e *Complement Naive Bayes* apresentaram desempenho inferior e idêntico no teste, ambos com F1-macro de 0,611. Em relação à validação cruzada, o *holdout* oferece aqui uma estimativa mais conservadora da capacidade de generalização, especialmente diante do tamanho do conjunto avaliado.

O resultado detalhado da *Logistic Regression*, modelo com melhor desempenho no conjunto de teste, é apresentado na Tabela 5. No *holdout*, o modelo alcançou acurácia de 0,736 e F1-macro de 0,731. Para a classe "Considero misógina", obteve-se precisão de 0,727, revocação de 0,667 e F1-score de 0,696. Para a classe "Não considero misógina", os valores correspondentes foram 0,742, 0,793 e 0,767. Esses resultados mostram que o desempenho na classe "Considero misógina" foi inferior ao observado para a classe "Não considero misógina", sobretudo em termos de revocação. Os componentes da matriz de confusão por classe também são apresentados nessa tabela. Esses resultados indicam que o classificador apresentou desempenho superior na identificação das instâncias não misóginas, particularmente em termos de revocação.

Tabela 5. Desempenho da *Logistic Regression* no teste *holdout*.

Classe	Precisão	Revocação	F1-score	VP	FP	FN
Considero misógina	0,727	0,667	0,696	16	6	8
Não considero misógina	0,742	0,793	0,767	23	8	6
Média macro	0,735	0,730	0,731	-	-	-

Nota-se que, entre as 29 instâncias reais da classe "Não considero misógina", 23 foram corretamente classificadas e 6 foram erroneamente atribuídas à classe oposta. Já entre as 24 instâncias reais da classe "Considero misógina", 16 foram corretamente identificadas e 8 foram classificadas como não misóginas.

Esse resultado sugere que a classe "Considero misógina" apresenta maior grau de dificuldade para a classificação supervisionada. Isso decorre do padrão observado na matriz de confusão, na qual os falsos negativos se concentram mais nessa classe do que na classe oposta, e é compatível com a presença de enunciados menos prototípicos, mais ambíguos ou mais dependentes de inferência contextual. Essa leitura converge com a literatura discutida na Seção 2, que já apontava a dependência de contexto como obstáculo para a caracterização automática da misoginia [Shushkevich and Cardiff, 2019; MacAvaney et al., 2019]. Nesse sentido, a assimetria observada reforça não apenas a dificuldade da tarefa, mas também a pertinência do desenho metodológico adotado neste trabalho, no qual a recuperação lexical é seguida de validação humana antes da etapa de classificação.

Em conjunto, os resultados indicam que o `LexMis-BR` ampliou a cobertura lexical do domínio, a coleta orientada permitiu reunir um corpus inicial relevante, a anotação humana evidenciou os limites de uma decisão baseada apenas em termos lexicais e a classificação supervisionada mostrou que o conjunto consolidado é adequado para avaliação supervisionada inicial. De modo geral, abordagens lineares tradicionais, em especial a *Logistic Regression* e o *SGDClassifier*, apresentaram os resultados mais favoráveis neste conjunto de dados.

5. Conclusão

Este trabalho apresentou a construção de recursos linguísticos e anotados voltados à detecção de misoginia em textos curtos em português do Brasil. As principais contribuições do estudo são a construção do léxico `LexMis-BR`, a consolidação de um conjunto de dados anotado para misoginia em português brasileiro e a proposição de um fluxo metodológico reprodutível em um cenário ainda marcado pela escassez de recursos linguísticos voltados a esse fenômeno. Os resultados também mostram que a recuperação lexical é eficaz para orientar a coleta, mas não substitui a validação humana, uma vez que muitos termos dependem de contexto para serem interpretados como misóginos.

A avaliação supervisionada foi conduzida como uma *baseline* inicial, com o objetivo de verificar a utilidade computacional do conjunto anotado produzido. Os resultados obtidos indicam que o recurso sustenta experimentos de classificação, embora o tamanho reduzido do conjunto anotado ainda limite conclusões mais amplas sobre desempenho e generalização. Como limitações, reconhece-se que a amostra de participantes apresenta concentração geográfica no estado do Rio de Janeiro. Além disso, embora a participação de mulheres tenha sido majoritária nas etapas de coleta, trabalhos futuros podem adotar critérios mais específicos para a seleção de avaliadores, incluindo experiência acadêmica ou profissional.

Referências

Braga, R., Moniz, H., Figueira, A., and Batista, F. (2021). Creation and characterization of a sexist discourse corpus in portuguese. *iSys - Brazilian Journal of Information Systems*, 14(4):34–57.

Fersini, E., Nozza, D., and Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (ami). In *Proceedings of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*. CEUR-WS.

Koch, L., Ghawi, R., Pfeffer, J., and Steinert, J. I. (2025). Online misogyny against female candidates in the 2022 brazilian elections: A threat to women's political representation? *Information, Communication & Society*.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press, Chicago.

Lima-Santos, A. V. d. S. and Santos, M. A. d. (2022). Incels e misoginia on-line em tempos de cultura digital. *Estudos e Pesquisas em Psicologia*, 22(3):1081–1102.

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.

Marques, E. C. A. (2025). Mesclagem conceptual e construção de novos significados: a metáfora "ser humano é animal" no português brasileiro. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Programa de Pós-Graduação em Linguística, Rio de Janeiro.

Martins, V., Serafim, S. E. V., Pereira, L. C. R., Alves, A. F., Ferreira, C. H. G., and Almeida, J. (2025). A misoginia no youtube brasileiro: Um estudo de caso sobre o conteúdo produzido pela comunidade red pill. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*.

Ministério das Mulheres (2025). *Raseam 2025 - Relatório Anual Socioeconômico da Mulher*. Ministério das Mulheres, Brasília.

Nações Unidas no Brasil (2026). Objetivo de desenvolvimento sustentável 5: Igualdade de gênero. `https://brasil.un.org/pt-br/sdgs/5`.

Pamungkas, E. W., Basile, V., and Patti, V. (2020). Misogyny detection in twitter: A multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Plath, H. O., Paiva, M. E. O., Pinto, D. L., and Costa, P. D. P. (2022). Detecção de discurso de Ódio contra mulheres em textos em português brasileiro: Construção da base mina-br e modelo de classificação. *Revista Eletrônica de Iniciação Científica em Computação*, 20(3).

Senado Federal (2026). Inclusão da misoginia como crime de preconceito é aprovada e vai à câmara. `https://www12.senado.leg.br/noticias/materias/2026/03/24/inclusao-da-misoginia-como-crime-de-preconceito-e-aprovada-e-vai-a-camara`.

Shushkevich, E. and Cardiff, J. (2019). Automatic misogyny detection in social media: A survey. *Computación y Sistemas*, 23(4).

Srivastava, K., Chaudhury, S., Bhat, P. S., and Sahu, S. (2017). Misogyny, feminism, and sexual harassment. *Industrial Psychiatry Journal*, 26(2):111–113.

Vargas, F., Carvalho, I., Pardo, T. A. S., and Benevenuto, F. (2025). Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, 31:435–456.

Zanello, V., Bukowitz, B., and Coelho, E. (2011). Xingamentos entre adolescentes em brasília: linguagem, gênero e poder. *Revista Interacções*, 7(17).

Zanello, V. and Romero, A. C. (2012). "Vagabundo" ou "vagabunda"? Xingamentos e relações de gênero. `http://www.labrys.net.br/labrys22/libre/valeskapt.htm`.