MultiDomainYT-ToxBR: Um Corpus Multidomínio de Comentários Tóxicos do YouTube em Português Brasileiro

Gabriel Z. de Souza[1], **Isabel H. Manssour**[1]

[1]Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Escola Politécnica
Porto Alegre - RS - Brasil

zurawski.gabriel@edu.pucrs.br, isabel.manssour@pucrs.br

Abstract. *This paper presents MultiDomainYT-ToxBR, a corpus of 5,000 YouTube comments annotated across nine toxicity categories and covering five thematic domains in the Brazilian context. The analysis of annotator agreement highlights inherent challenges, with moderate indices for categories such as racism and xenophobia, whose veiled and context-dependent manifestations hinder consensus. Comparison with ToLD-Br reveals complementary patterns: the new corpus surpasses the previous one in homophobia agreement (α = 0.73 vs. 0.68) and introduces novel categories, such as political and religious intolerance. As validation, BERTweet.BR, after fine-tuning, achieved an F1 of 0.78 in the binary classification.*

Resumo. *Este trabalho apresenta o MultiDomainYT-ToxBR, um corpus de 5.000 comentários do YouTube anotados em nove categorias de toxicidade e distribuídos em cinco domínios temáticos no contexto brasileiro. A análise de concordância entre anotadores evidencia desafios inerentes à tarefa, com índices moderados em categorias como racismo e xenofobia, cujas manifestações veladas e dependentes de contexto dificultam o consenso. A comparação com o ToLD-Br revela padrões complementares: o novo corpus supera o anterior em concordância para homofobia (α = 0,73 vs. 0,68) e introduz categorias inéditas como intolerância política e religiosa. Como validação, o BERTweet.BR, após fine-tuning, atingiu F1 de 0,78 na classificação binária.*

1. Introdução

O crescimento das mídias sociais transformou a comunicação, mas também intensificou o discurso de ódio e o conteúdo tóxico, muitas vezes como mecanismo de engajamento. Evidências indicam que algoritmos de ranqueamento baseados em engajamento amplificam conteúdos emocionalmente carregados e hostis, priorizando a retenção em detrimento da segurança do usuário [Milli et al. 2025]. Esse fenômeno não se restringe ao ambiente digital, pois pesquisas mostram que o ódio disseminado nas redes sociais prediz a ocorrência de crimes violentos contra minorias no mundo real, demonstrando sua influência na propagação de violência offline [Müller and Schwarz 2021].

Casos como a violência contra a minoria muçulmana Rohingya[1] em Mianmar e os motins antimuçulmanos no Sri Lanka[2], ambos associados à amplificação algorítmica de conteúdo hostil no Facebook, ilustram um risco concreto.

[1]https://tinyurl.com/mtrrnxvf
[2]https://tinyurl.com/yetkj4tx

Nesse contexto, a automação da detecção de conteúdo tóxico torna-se uma necessidade. O volume massivo de dados gerados continuamente nas redes sociais inviabiliza a moderação humana, tornando indispensável o desenvolvimento de sistemas de Processamento de Linguagem Natural (PLN) capazes de identificar linguagem nociva em larga escala. Modelos de Linguagem de Grande Escala (LLMs) têm demonstrado potencial promissor nessa tarefa, mas sua eficácia depende da disponibilidade de corpora de treinamento e de avaliação que reflitam a diversidade linguística e cultural de cada contexto.

Este artigo apresenta o MultiDomainYT-ToxBR, um novo corpus de comentários do YouTube em português brasileiro, composto por 5.000 comentários anotados manualmente por múltiplos anotadores em nove categorias de toxicidade, distribuídos em cinco domínios temáticos: política polarizada, futebol, conflitos internacionais, feminismo e direitos das mulheres e conteúdo LGBTQIA+. Este corpus visa suprir a carência de dados do YouTube, complementando os corpora ToLD-Br [Leite et al. 2020] (X, antigo Twitter) e HateBR [Vargas et al. 2022] (Instagram). A especificidade das interações baseadas em vídeos no YouTube justifica a necessidade de um corpus dedicado a compreender suas dinâmicas particulares. As principais contribuições deste trabalho são:

- O MultiDomainYT-ToxBR, um corpus anotado de comentários do YouTube em português, cobrindo múltiplos domínios temáticos;
- Avaliação experimental do corpus, demonstrando sua adequação para classificação binária de toxicidade (F1 de 0,78 com BERTweet.BR, desempenho comparável ao ToLD-Br);
- Análise exploratória da toxicidade no YouTube, evidenciando um perfil mais estrutural e ideológico, com destaque para Intolerância Política;
- Identificação de limitações estruturais da tarefa multilabel, como o impacto do desbalanceamento e da subjetividade na detecção de categorias minoritárias;
- Proposição de uma nova direção de pesquisa, baseada na substituição da toxicidade por tipificação legal, apoiada em corpora anotados por especialistas jurídicos, para reduzir a subjetividade.

2. Trabalhos Relacionados

O ToLD-Br (*Toxic Language Dataset for Brazilian Portuguese*) [Leite et al. 2020] é o maior conjunto de dados para análise de linguagem tóxica em português, com 21.000 tweets do X anotados em sete categorias: não tóxico, LGBTQfobia+, obsceno, insulto, racismo, misoginia e xenofobia. A anotação foi feita por 42 voluntários com perfis demográficos diversos para reduzir vieses. Modelos BERT monolíngues em português alcançaram até 76% de macro-F1 na classificação binária, superando abordagens multilíngues e reforçando a importância de conjuntos de dados específicos para o idioma.

O HateBR [Vargas et al. 2022] é um corpus de 7.000 comentários do Instagram, coletados de contas públicas de políticos brasileiros, para detecção de linguagem ofensiva e discurso de ódio. As anotações, realizadas por indivíduos com alta escolaridade e perfis diversos, abrangem duas camadas: classificação binária (ofensivo/não ofensivo) e nível de ofensividade (leve, moderado, alto), com alta concordância (Kappa de Cohen = 75%).

Mais recentemente, [Trajano et al. 2024] propuseram o OLID-BR, um corpus de 13.538 comentários coletados do X, do YouTube e de *datasets* relacionados (OffComBR [de Pelle and Moreira 2017], HLPHSD [Fortuna et al. 2019], NCCVG [Trajano 2023], ToLD-Br), anotado em um esquema hierárquico de cinco tarefas,

incluindo detecção de alvo e de spans tóxicos. Essa abordagem é inédita para o português brasileiro nesse nível de granularidade.

[Assis et al. 2024] comparam classificadores baseados em BERT com chatbots generativos para detecção de discurso de ódio em português, mostrando que modelos ajustados com poucos exemplos superam os chatbots, embora estes melhorem significativamente com exemplos contextuais. Complementarmente, [Oliveira et al. 2024b] comparam ChatGPT e MariTalk (Sabiá) na mesma tarefa, com o MariTalk demonstrando melhor compreensão do português coloquial e reforçando a eficácia de modelos monolíngues menores. Já [Oliveira et al. 2024a] investigam o LLaMA 3.1 8B com um método iterativo de evolução de *prompts*, obtendo ganhos de F1 em relação ao BERTimbau e precisão comparável à do GPT-4o mini. Por fim, [Junqueira et al. 2024] avaliam o Sabiá-7B em quatro tarefas de PLN, incluindo detecção de ódio e de ironia, com bom desempenho geral em poucos exemplos, mas com limitações na tarefa de pergunta-resposta.

A Tabela 1 sintetiza os trabalhos relacionados. A análise revela uma lacuna: os corpora multilabel disponíveis publicamente para português concentram-se predominantemente no X [Leite et al. 2020] e no Instagram [Vargas et al. 2022]. O OLID-BR [Trajano et al. 2024] incorpora também comentários do YouTube, mas sem estrutura temática por domínio e sem comparação de perfis de toxicidade entre contextos, uma lacuna que o MultiDomainYT-ToxBR busca preencher.

Tabela 1. Síntese dos trabalhos relacionados.

Trabalho	Ano	Modelos Usados	Corporas Usados
[Assis et al. 2024]	2024	BERTimbau, AlBERTina, ChatGPT, BERTweet.BR, MariTalk	HateBR e ToLD-Br
[Oliveira et al. 2024b]	2024	ChatGPT, MariTalk, BERTimbau	ToLD-Br
[Oliveira et al. 2024a]	2024	LLaMA 3.1 8B, BERTimbau, GPT-4o mini, ChatGPT 3.5 Turbo, Sabiá 3	ToLD-Br
[Junqueira et al. 2024]	2024	Sabiá-7B	ABSAPT 2022, ToLD-BR, IDPT 2021, SQUAD v1.1
[Trajano et al. 2024]	2023	BERTimbau, BERT	OffComBR, HLPHSD, NCCVG, ToLD-Br

3. Construção do Corpus

O corpus foi composto por 5.000 comentários em português brasileiro, distribuídos em cinco domínios temáticos propensos à ocorrência de conteúdo controverso e tóxico: Política Polarizada, Futebol, Conflitos Internacionais, Feminismo e Direitos das Mulheres e Conteúdo LGBTQIA+. A escolha desses domínios foi motivada pela representatividade dos debates de alta polarização no contexto brasileiro, o que garante diversidade temática e maior probabilidade de captar diferentes categorias de toxicidade.

Para cada domínio, foram definidos cinco termos de busca, totalizando 25 termos apresentados na Tabela 2. Parte deles foi selecionada manualmente considerando casos de alta repercussão e potencial de toxicidade na mídia brasileira recente, como “xandão e elon musk” e “caso mari ferrer”, enquanto outros foram sugeridos pelo modelo Gemini para complementar e diversificar a cobertura temática de cada domínio.

A coleta seguiu cinco etapas: (1) recuperação de dois vídeos por termo via API do YouTube, totalizando 50 vídeos, filtrados por relevância temática e volume de co-

Tabela 2. Temas e termos de pesquisa associados.

Categorias	Termos de Pesquisa
Política polarizada	"eleições 2024", "pl das fake news", "xandão e elon musk", "lula e bolsonaro", "reforma tributária"
Futebol	"flamengo vs fluminense pós jogo", "corinthians palmeiras torcedor", "gol anulado", "futebol racismo", "torcida organizada briga"
Conflitos internacionais	"guerra ucrânia", "guerra israel e palestina", "crise migratória na Europa", "conclave para escolha de novo papa", "tarifaço trump"
Feminismo e direitos das mulheres	"marieli franco", "violência contra mulher reportagem", "feminismo debate brasil", "caso mari ferrer", "lei maria da penha"
Conteúdo LGBTQIA+	"parada gay", "casamento homoafetivo", "ideologia de gênero", "pronome neutro debate", "educação sexual"

mentários; (2) coleta de todos os comentários desses vídeos; (3) filtragem de idioma com xlm-roberta-base-language-detection[3]; (4) estimativa de toxicidade com Detoxify[4] para amostragem estratificada; (5) seleção de 100 comentários por vídeo (50% sinalizados como potencialmente tóxicos, 50% aleatórios), resultando em 1.000 comentários por domínio.

A anotação considerou nove categorias de toxicidade, baseadas na taxonomia de [Leite et al. 2020] e expandidas para o contexto brasileiro contemporâneo. A Tabela 3 apresenta as categorias e suas definições. Em relação ao ToLD-Br, com seis categorias, o MultiDomainYT-ToxBR introduz três categorias: Intolerância Política, associada à polarização brasileira [Ortellado et al. 2022]; Intolerância Religiosa, impulsionada pelo crescente papel da pauta religiosa no debate público; e Anticapacitismo, comumente expresso como insultos disfarçados em outros contextos.

Tabela 3. categorias de linguagem ofensiva e discriminação.

Categoria	Descrição
Insulto de baixo calão	Agressão verbal com xingamentos, palavrões ou linguagem vulgar, sem conotação discriminatória específica.
Homofobia	Comentários ofensivos baseados em orientação sexual ou identidade de gênero.
Racismo	Comentários hostis ou estereotipados com base em raça, cor da pele ou etnia.
Misoginia/Machismo	Expressões de ódio, desprezo ou inferiorização da mulher enquanto mulher.
Xenofobia	Hostilidade baseada em origem geográfica, nacionalidade ou região, incluindo ataques a migrantes internos.
Violência/Extremismo	Defesa, incentivo ou apologia ao uso de violência física ou psicológica.
Intolerância Religiosa	Ataques ou zombaria com base em crença religiosa.
Intolerância Política	Ataques ou desumanização com base em orientação política ou filiação partidária.
Anticapacitismo	Comentários que zombam ou desumanizam pessoas com deficiência.

A anotação foi realizada em planilhas do Excel e os anotadores foram instruídos a não discutir suas classificações, de modo que interpretações distintas fossem preservadas. Os 5.000 comentários foram divididos em dois grupos independentes de três anotadores cada: o Grupo 1 anotou os primeiros 2.500 comentários e o Grupo 2 anotou os 2.500

[3]https://huggingface.co/papluca/xlm-roberta-base-language-detection

[4]https://pypi.org/project/detoxify/

restantes. A cada categoria foi atribuída uma pontuação de 0 a 3, correspondente ao número de anotadores que a identificaram. Inspirada no ToLD-Br, essa abordagem define uma probabilidade empírica (0/3 a 3/3) por categoria para cada comentário, capturando a ambiguidade inerente à tarefa.

A qualidade das anotações foi avaliada por meio do Kappa de Fleiss e do Alfa de Krippendorff, duas métricas amplamente utilizadas para mensurar concordância em tarefas de anotação com múltiplos avaliadores. A Tabela 4 apresenta as métricas por categoria para cada grupo. Os resultados revelam uma diferença significativa entre os grupos: o Grupo 1 apresenta concordância moderada ($\kappa = 0,493$), enquanto o Grupo 2 atinge apenas o nível justo ($\kappa = 0,372$), segundo os critérios de Landis e Koch [Landis and Koch 1977]. Essa discrepância pode refletir diferenças de perfil e de sensibilidade cultural entre os anotadores, o que reforça a dificuldade inerente à tarefa.

Tabela 4. Kappa de Fleiss por categoria e grupo de anotadores. Valores de Krippendorff's α diferem em menos de 0,0001 em todas as categorias.

Categorias	Grupo 1	Grupo 2
Insulto de baixo calão	0,481	0,262
Homofobia	0,734	0,442
Racismo	0,422	0,332
Misoginia/Machismo	0,532	0,407
Xenofobia	0,329	0,420
Violência/Extremismo	0,364	0,215
Intolerância Religiosa	0,549	0,555
Intolerância Política	0,584	0,499
Anticapacitismo	0,443	0,218
Média Geral	**0,493**	**0,372**

Homofobia apresenta a maior concordância no Grupo 1 ($\kappa = 0,734$), resultado atribuível à presença de marcadores linguísticos explícitos e amplamente reconhecíveis nessa categoria. Intolerância Política e Intolerância Religiosa também demonstram concordância moderada a substancial em ambos os grupos ($\kappa > 0,49$ e $\kappa > 0,54$, respectivamente). No extremo oposto, Violência/Extremismo e Anticapacitismo apresentam os menores índices no Grupo 2 ($\kappa < 0,22$), possivelmente devido à natureza contextual e frequentemente sarcástica dessas manifestações nas discussões online.

Em comparação com o ToLD-Br, cujo α médio é de 0,55, o MultiDomainYT-ToxBR apresenta concordância levemente inferior ($\alpha = 0,493$ no G1), uma diferença de aproximadamente 0,06 pontos, atribuível, em parte, à maior granularidade taxonômica do novo corpus. Homofobia no MultiDomainYT-ToxBR supera a categoria análoga no ToLD-Br ($\alpha = 0,734$ vs. $\alpha = 0,68$), enquanto Racismo e Xenofobia apresentam índices inferiores ($\alpha = 0,422$ e $\alpha = 0,330$ vs. $\alpha = 0,48$ e $\alpha = 0,57$), sugerindo que essas formas de toxicidade se manifestam de modo mais velado no contexto do YouTube do que no X — um achado com implicações diretas para a modelagem.

4. Experimento de Validação

Para verificar a usabilidade do MultiDomainYT-ToxBR como recurso para treinamento supervisionado, foi conduzido um experimento de *fine-tuning* com o modelo

BERTweet.BR [Carneiro et al. 2025], pré-treinado com tweets em português brasileiro, com aproximadamente 110 milhões de parâmetros. A escolha se justifica por seu desempenho superior, reportado em tarefas de detecção de toxicidade em português, quando comparado a alternativas como BERTimbau e ALBERTina [Assis et al. 2024]. Os modelos resultantes desse experimento podem ser encontrados publicamente no Hugging Face[5]. Os *scripts* usados para o *fine-tuning*, bem como os utilizados para a coleta, o tratamento e a análise do corpus, estão disponíveis no GitHub[6].

Os dados foram divididos em 80% treino, 10% validação e 10% teste, por *Hold-Out* estratificado, mantendo a proporção das categorias em cada partição. O tokenizador usado foi o do próprio BERTweet.BR, com limite de 128 tokens. O *fine-tuning* foi conduzido por até 15 épocas com *batches* de 16 exemplos, utilizando o otimizador AdamW com taxa de aprendizado de $2 * 10^{-}5$ e um agendador linear com *warmup* de 500 passos. Foram treinados dois modelos sobre o MultiDomainYT-ToxBR: um para a tarefa de classificação binária (tóxico/não tóxico), com função de ativação *softmax* na inferência, e outro para a tarefa de classificação multilabel (nove categorias), com função sigmoide e limiar de 0,5 para cada categoria. A escolha desses hiperparâmetros baseou-se nos resultados do trabalho de [Assis et al. 2024], que apresentou bons resultados na classificação binária ao ajustar os modelos com o ToLD-Br e o HateBR.

Os resultados da classificação binária são apresentados na Tabela 5. O modelo atingiu F1 de 0,78 no conjunto de teste, com acurácia, precisão e revocação igualmente balanceadas, convergindo já na primeira época. Para fins de referência, o mesmo modelo treinado sobre o HateBR (corpus balanceado artificialmente) atingiu F1 de 0,92, enquanto o treinamento sobre o ToLD-Br resultou em F1 de 0,78. O desempenho inferior do MultiDomainYT-ToxBR em relação ao HateBR é consistente com dois fatores: (i) o desbalanceamento de classes, que reduz a exposição do modelo a exemplos tóxicos durante o treinamento - em contraste com o balanceamento perfeito (50%/50%) do HateBR -; e (ii) a menor concordância entre anotadores observada na Seção 3 ($\kappa = 0,493$ e 0,372, frente ao Kappa de 0,75 do HateBR), que introduz ruído no rótulo-alvo e limita o desempenho máximo alcançável por qualquer classificador supervisionado. Adicionalmente, a maior heterogeneidade temática do corpus (cinco domínios distintos) pode dificultar a generalização, em comparação com o contexto único do HateBR.

Tabela 5. Desempenho do BERTweet.BR na classificação binária sobre o MultiDomainYT-ToxBR.

Corpus	Melhor Época	Acurácia	Precisão	Revocação	F1	Suporte
MultiDomainYT-ToxBR (binário)	1	0,81	0,77	0,79	0,78	500

A tarefa multilabel revelou-se mais desafiadora, conforme apresentado na Tabela 6. O modelo obteve um F1 médio de apenas 0,11, com desempenho concentrado nas duas categorias majoritárias: Insulto de baixo calão (F1 = 0,50) e Intolerância Política (F1 = 0,52). Para as demais sete categorias, o modelo obteve F1 nulo (0,00), falhando em detectá-las no conjunto de teste. Esse resultado era esperado devido ao forte desbalanceamento: categorias como Racismo e Anticapacitismo têm apenas 2 e 3 exemplos no

[5]https://huggingface.co/zurawski

[6]https://github.com/DAVINTLAB/MultiDomainYT-ToxBR

conjunto de teste, tornando inviável o aprendizado de padrões discriminativos por modelos supervisionados.

Tabela 6. Desempenho do BERTweet.BR na classificação multilabel sobre o MultiDomainYT-ToxBR.

Categoria	Precisão	Revocação	F1	Suporte
Insulto de baixo calão	0,63	0,41	0,50	93
Homofobia	0,00	0,00	0,00	15
Racismo	0,00	0,00	0,00	2
Misoginia/Machismo	0,00	0,00	0,00	10
Xenofobia	0,00	0,00	0,00	6
Violência/Extremismo	0,00	0,00	0,00	5
Intolerância Religiosa	0,00	0,00	0,00	7
Intolerância Política	0,69	0,42	0,52	26
Anticapacitismo	0,00	0,00	0,00	3
Média	**0,15**	**0,09**	**0,11**	**167**

5. Análise Exploratória e Discussão

A análise dos corpora MultiDomainYT-ToxBR, ToLD-Br e HateBR revela padrões complementares de toxicidade em português, organizados em quatro eixos descritos a seguir.

5.1. Perfis de Toxicidade por Plataforma

A distribuição das categorias nos três corpora (Figura 1) reflete as características das plataformas de origem e as metodologias de construção. O HateBR apresenta distribuição perfeitamente balanceada (50% tóxico/50% não tóxico), resultado de uma estratégia deliberada de construção que, embora ideal para certas tarefas de treinamento supervisionado, não reflete a prevalência natural do conteúdo tóxico em ambientes reais.

Os demais corpora são desbalanceados e apresentam perfis de toxicidade distintos. No ToLD-Br, a toxicidade é dominada por linguagem explícita: Obsceno (6.652 ocorrências) e Insulto (4.385 ocorrências) concentram a maioria das ocorrências, enquanto Racismo (138) e Xenofobia (151) são marginais. Esse padrão reflete a natureza impulsiva e concisa da comunicação no X, em que a restrição de caracteres favorece ofensas diretas em detrimento de argumentos discriminatórios mais elaborados.

O MultiDomainYT-ToxBR apresenta perfil distinto. Embora Insulto de baixo calão seja a categoria mais frequente (18,5%), o novo corpus se destaca pela maior relevância de toxicidade ideológica e identitária: Intolerância Política surge como segunda categoria mais frequente (5,1%), seguida de Homofobia (3,1%). Esse padrão sugere que, no YouTube, a toxicidade emerge em contextos de debate temático, favorecendo formas mais estruturais e argumentativas de discurso hostil. A ausência de Intolerância Política no ToLD-Br (2019) mostra que corpora antigos podem não capturar dinâmicas sociais recentes, o que torna modelos treinados neles desatualizados.

5.2. Categorias de Baixa Prevalência e Desafios de Detecção

No ToLD-Br, Racismo e Xenofobia representam menos de 1,4% do corpus combinado. No MultiDomainYT-ToxBR, Racismo aparece em apenas 17 comentários (0,34%), Anticapacitismo em 32 (0,64%) e Violência/Extremismo em 57 (1,14%). Essa baixa prevalência não decorre exclusivamente de viés de amostragem. A baixa concordância entre

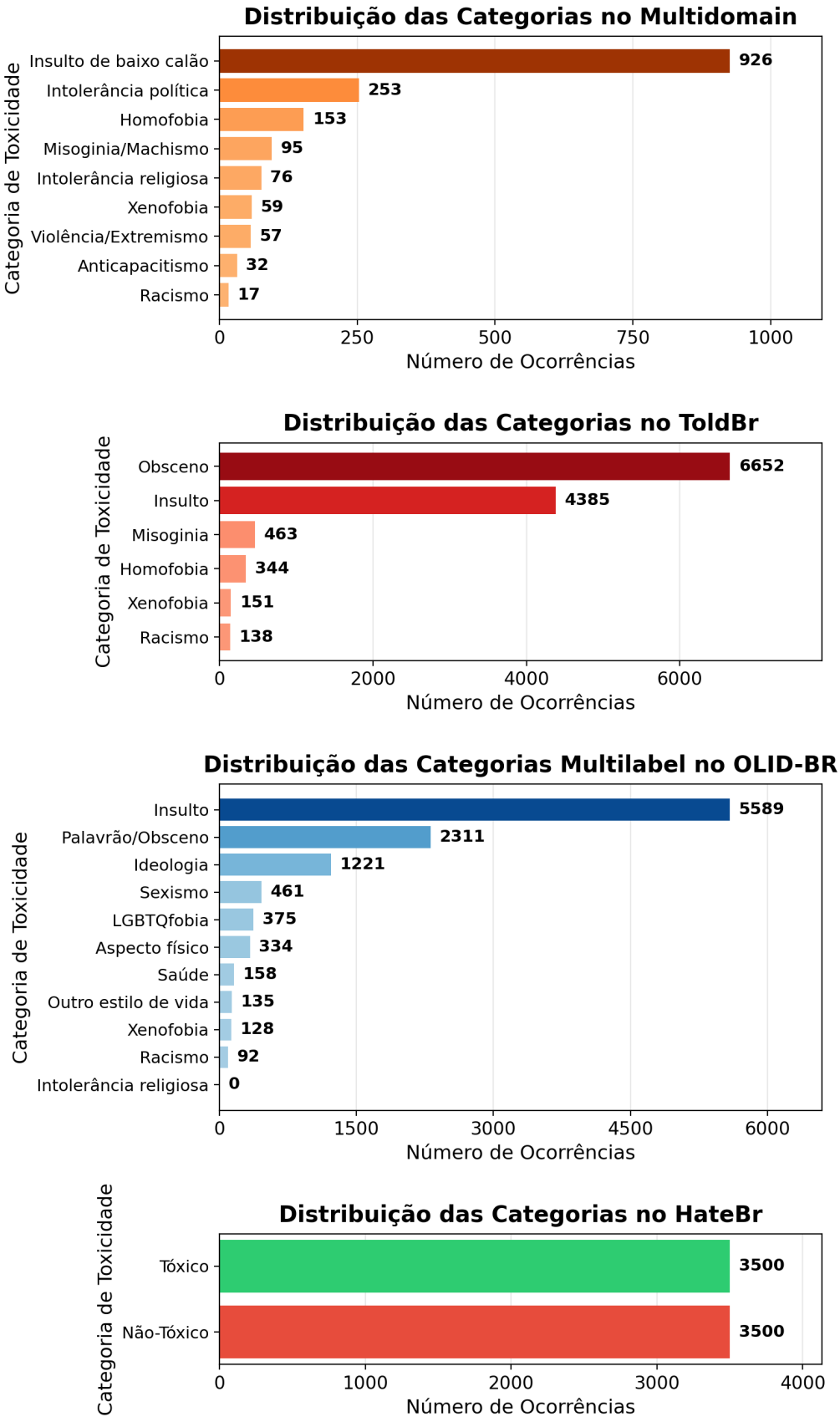

Figura 1. Distribuição das categorias de cada corpus.

anotadores nestas categorias sugere que essas formas de toxicidade se manifestam de modo suficientemente velado ou dependente de contexto para resistir ao reconhecimento consistente por parte de anotadores humanos. Essa combinação de baixa frequência e baixa concordância constitui um obstáculo estrutural à detecção automatizada. Conjuntos de dados com categorias extremamente desbalanceadas levam modelos supervisionados a favorecer as categorias majoritárias e ignorar manifestações tóxicas menos frequentes, mas social e legalmente graves.

5.3. Toxicidade Explícita versus Implícita e Marcadores Linguísticos

A análise de frequência de palavras por categoria[7], descritas nas Tabelas 7 e 8, revela a distinção entre toxicidade explícita e implícita, com implicações para a modelagem. No ToLD-Br, categorias como Homofobia e Misoginia apresentam vocabulário explícito: termos como "viado", "sapatão", "puta" e "piranha" deixam pouca margem para ambiguidade interpretativa. O MultiDomainYT-ToxBR revela um cenário mais complexo. Na categoria Xenofobia, as palavras mais frequentes, "imigrantes", "muçulmanos", "europeus" e "país", são termos neutros que adquirem conotação tóxica apenas em contextos específicos. De modo semelhante, Anticapacitismo apresenta termos explícitos como "retardado", mas também palavras genéricas como "povo", "mundo" e "coisa", indicando que a identificação dessa categoria requer análise contextual que vai além da correspondência por palavras-chave.

Tabela 7. Palavras mais frequentes por categoria de toxicidade - ToldBr.

Homofobia	Obsceno	Insulto	Racismo	Misoginia	Xenofobia
viado (78)	porra (997)	vai (563)	branco (14)	puta (57)	sulista (26)
sapatão (27)	caralho (825)	puta (360)	nego (14)	putinha (56)	carioca (13)
boiola (23)	vai (735)	cara (294)	@user (12)	vai (54)	todo (10)
gay (20)	pqp (710)	porra (272)	vai (10)	piranha (53)	nada (9)
vai (20)	puta (662)	caralho (263)	cara (10)	mulher (30)	gente (8)
caralho (18)	pau (308)	tomar (174)	porra (8)	vagabunda (30)	porra (8)
bicha (17)	tomar (281)	fdp (173)	pau (7)	tudo (23)	vai (8)
cara (15)	vou (273)	gente (161)	ter (7)	caralho (21)	brasil (7)
sapatao (15)	cara (256)	pau (160)	negão (7)	quer (21)	fala (6)
bixa (14)	mano (220)	lixo (154)	agora (6)	fala (18)	ainda (6)

Cada categoria apresenta marcadores linguísticos que atuam como indicadores semânticos. Em Homofobia, termos como "bicha" e "todes", este último frequentemente usado em deboche à linguagem neutra, constituem um vocabulário discriminatório. Em Misoginia, o ToLD-Br apresenta termos sexualizado: "puta", "piranha", "vagabunda"; enquanto no MultiDomainYT-ToxBR emergem termos mais contextuais, como "maria", "penha" e "lei", em referência à Lei Maria da Penha, sugerindo que no YouTube a misoginia se manifesta mais frequentemente em debates sobre direitos das mulheres do que em ofensas diretas. Em Intolerância Política, marcadores como "esquerda", "lula", "direita" e "esquerdista" revelam a forte polarização do contexto brasileiro, com discussões sobre identidade nacional servindo de campo fértil para hostilidade política.

Essa distinção entre toxicidade explícita e implícita tem consequências diretas para o desenvolvimento de sistemas de moderação: abordagens baseadas em listas de

[7]Os termos apresentados nesta seção refletem fielmente o conteúdo dos dados e podem incluir linguagem ofensiva.

Tabela 8. Palavras mais frequentes por categoria de toxicidade - MultiDomainYT-ToxBR.

Insulto	Homofobia	Racismo	Misoginia	Xenofobia	Violência	Into. Relig.	Into. Pol.	Anticap.
vai (103)	mulher (34)	vai (4)	mulher (43)	imigrantes (9)	cara (10)	deus (23)	esquerda (57)	retardado (8)
cara (96)	homem (33)	raça (3)	mulheres (16)	muçulmanos (8)	dessa (7)	religião (21)	lula (25)	povo (4)
merda (74)	vai (19)	macaco (3)	maria (15)	povo (7)	pro (7)	evangélicos (16)	direita (24)	faz (4)
lixo (70)	todes (16)	situação (3)	lei (11)	europeus (7)	vai (7)	sabe (15)	brasil (22)	mundo (4)
mulher (67)	coisa (15)	galo (3)	vida (10)	tudo (7)	inferno (6)	querem (9)	tudo (19)	mulher (4)
brasil (52)	pessoas (15)	racismo (3)	cara (10)	homem (9)	mulher (6)	nada (8)	sempre (15)	retardados (4)
ter (49)	nao (15)	prisão (3)	porque (9)	quer (6)	mano (5)	onde (8)	povo (15)	sabe (4)
faz (46)	deus (15)	egito (2)	penha (9)	serem (6)	mundo (5)	igreja (8)	quer (15)	todes (3)
tudo (45)	gay (14)	agora (2)	feminista (7)	muitos (6)	ter (5)	pessoas (8)	mundo (15)	coisa (3)
todos (44)	pronome (13)	fazer (2)	vida (10)	vai (6)	queimar (4)	vai (7)	esquerdista (15)	ver (3)

palavras-chave ofensivas capturam adequadamente categorias como Insulto e Homofobia, mas falham em categorias cuja toxicidade é construída discursivamente, precisamente aquelas com maior relevância jurídica e social, como Racismo e Xenofobia.

5.4. Implicações Metodológicas

Os resultados obtidos evidenciam limitações que vão além do MultiDomainYT-ToxBR na detecção de toxicidade em português como um todo. Nenhum corpus isolado captura a complexidade da toxicidade em suas múltiplas dimensões: o ToLD-Br, o HateBR e o OLID-BR oferecem cobertura sólida de toxicidade explícita, enquanto o MultiDomainYT-ToxBR contribui com toxicidade estrutural e politizada, mas nenhum dos corpora multilabel apresenta exemplos suficientes de Racismo e Xenofobia. Os resultados também sugerem que a tarefa de classificação baseada em "toxicidade" apresenta uma limitação fundamental: trata-se de um conceito subjetivo, influenciado pelo *background* cultural e pelo nível de tolerância dos anotadores. Assim, os modelos tendem a reproduzir percepções subjetivas de ofensa, em vez de identificar padrões objetivos de conduta ilícita, penalizando contradiscursos legítimos tanto quanto o discurso de ódio que visam combater. Trabalhos futuros devem considerar uma reorientação de paradigma: em vez de classificadores treinados sobre anotações genéricas de toxicidade, reconhecidamente suscetíveis a ruído e viés cultural [Zufall et al. 2022], adotar LLMs instruídos a tipificar textos conforme o Código Penal Brasileiro, distinguindo figuras como injúria racial, apologia ao crime e difamação, validados por corpora com anotações especializadas. [Luo et al. 2023] mostram que esse alinhamento entre PLN e o direito penal não é apenas factível, mas também necessário para que a moderação automatizada de conteúdo tenha respaldo jurídico sólido.

6. Conclusão

Este artigo apresentou o MultiDomainYT-ToxBR, um corpus de 5.000 comentários do YouTube em português, anotados em nove categorias de toxicidade e distribuídos em cinco domínios temáticos. O corpus preenche uma lacuna na literatura ao capturar um perfil de toxicidade mais estrutural, político e ideológico, refletindo a natureza temática dos comentários do YouTube e as dinâmicas sociais brasileiras recentes. A análise comparativa revelou que nenhum corpus isolado captura plenamente a complexidade da toxicidade em português, reforçando a necessidade de avaliações com múltiplos recursos para o desenvolvimento de sistemas robustos. A validação com o BERTweet.BR confirma a adequação do corpus para classificação binária (F1 = 0,78), mas expõe o desbalanceamento das categorias minoritárias como um obstáculo à classificação multilabel, um problema recorrente na área.

7. Limitações

O corpus reflete o debate público brasileiro até 2025, e novos padrões de toxicidade podem não estar representados. Os cinco domínios temáticos, embora diversificados, não cobrem a totalidade dos contextos em que a toxicidade ocorre no YouTube brasileiro. A concordância moderada entre anotadores, especialmente em categorias como Xenofobia e Violência/Extremismo, reflete a dificuldade inerente à tarefa e introduz ruído nos rótulos. Adicionalmente, a divisão dos 5.000 comentários em dois grupos independentes de três anotadores, necessária para reduzir a carga de trabalho voluntário, implica que nenhum anotador avaliou o corpus completo, o que reduz a robustez da anotação global e impede o cálculo de métricas de concordância unificadas entre todos os avaliadores.

Além disso, a etapa de pré-filtragem com a biblioteca Detoxify, utilizada para obter uma amostragem significativa de conteúdo tóxico, introduz um viés sistemático relevante: por ser baseada em um modelo BERT treinado predominantemente em dados em inglês, o Detoxify pode apresentar menor sensibilidade a formas de toxicidade culturalmente específicas do português brasileiro, favorecendo a seleção de conteúdo com marcadores linguísticos mais explícitos e universais em detrimento de manifestações tóxicas mais sutis. Por fim, o experimento se limita ao *fine-tuning* do BERTweet.BR sem técnicas de balanceamento de classes. Arquiteturas alternativas e estratégias, como *oversampling*, estão fora do escopo deste artigo.

8. Considerações Éticas e Uso de IA Generativa

Todos os comentários foram coletados por meio da API oficial do YouTube, em conformidade com seus Termos de Serviço. Os dados são públicos e nenhuma informação privada foi armazenada. No que diz respeito aos anotadores, os seis participantes foram previamente informados sobre a natureza explícita e potencialmente perturbadora do conteúdo a ser classificado, e a participação foi voluntária, podendo ser interrompida a qualquer momento sem qualquer consequência. Quanto ao uso do corpus, este é disponibilizado exclusivamente para fins acadêmicos. Reconhecemos ainda que modelos treinados com dados de toxicidade podem reproduzir vieses presentes nesses dados e recomendamos cautela na implantação em ambientes de produção. Por fim, este artigo contém exemplos de linguagem explícita e ofensiva coletados de redes sociais, apresentados exclusivamente com finalidade analítica e que não representam a opinião dos autores.

Durante a revisão deste artigo, utilizamos o ChatGPT e o Grammarly para aprimorar a clareza e a correção gramatical do texto. Os autores são integralmente responsáveis pelo conteúdo e pela precisão técnica do artigo.

Agradecimentos

Manssour agradece ao CNPq pelo apoio financeiro (processo nº 303208/2023-6).

Referências

Assis, G., Amorim, A., Carvalho, J., de Oliveira, D., Vianna, D., and Paes, A. (2024). Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of*

the 16th International Conference on Computational Processing of Portuguese - Vol. 1, pages 301–311, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Carneiro, F., Vianna, D., Carvalho, J., Plastino, A., and Paes, A. (2025). BERTweet.BR: a pre-trained language model for tweets in Portuguese. *Neural Computing and Applications*, 37(6):4363–4385.

de Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., and Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In Roberts, S. T., Tetreault, J., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Junqueira, J. d. R., Lopes, , M., C. L. D. S., da Silva, F. L. V., Carvalho, E. A., de Freitas, L. A., and Corrêa, U. B. (2024). Sabiá in action: An investigation of its abilities in aspect-based sentiment analysis, hate speech detection, irony detection, and question-answering. In *IJCNN*, pages 1–8.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Wong, K.-F., Knight, K., and Wu, H., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Luo, C., Bhambhoria, R., Dahan, S., and Zhu, X. (2023). Legally enforceable hate speech detection for public forums. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10948–10963, Singapore. Association for Computational Linguistics.

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., and Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3):pgaf062.

Müller, K. and Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.

Oliveira, A., Silva, P. H., Santos, V., Moreira, G., Freitas, V. L., and Luz, E. J. (2024a). Toxic text classification in portuguese: Is llama 3.1 8b all you need? In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 57–66, Porto Alegre, RS, Brasil. SBC.

Oliveira, A. d. S., Cecote, T. d. C., Alvarenga, J. P. R., de Souza Freitas, V. L., and da Silva Luz, E. J. (2024b). Toxic speech detection in Portuguese: A comparative study of large language models. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th Internati-*

onal Conference on Computational Processing of Portuguese - Vol. 1, pages 108–116, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Ortellado, P., Ribeiro, M. M., and Zeine, L. (2022). Existe polarização política no brasil? análise das evidências em duas séries de pesquisas de opinião. *Opinião Pública*, 28(1):62–91.

Trajano, D., Bordini, R. H., and Vieira, R. (2024). Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, 58(4):1263–1289.

Trajano, D. d. O. (2023). Detecção de linguagem tóxica aplicada a textos em português. Dissertação (mestrado em ciência da computação), Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brasil. Orientador: Rafael Heitor Bordini.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Zufall, F., Hamacher, M., Kloppenborg, K., and Zesch, T. (2022). A legal approach to hate speech – operationalizing the EU's legal framework against the expression of hatred as an NLP task. In Aletras, N., Chalkidis, I., Barrett, L., Goanță, C., and Preoțiuc-Pietro, D., editors, *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 53–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.