Proposta recebida em Outubro 2024 e aceite para publicação em Dezembro 2024.

Rotulação e Caracterização de Conteúdo Tóxico de Comunidades do Reddit no Brasil

Toxic Content Detection in Online Social Networks: A New Dataset from Brazilian Reddit Communities

Luiz Henrique Quevedo Lima
Universidade Federal de Minas Gerais

Ana Clara Souza Pagano
Universidade Federal de Minas Gerais

Adriana Silvina Pagano
Universidade Federal de Minas Gerais

Ana Paula Couto da Silva
Universidade Federal de Minas Gerais

Resumo

A ausência de dados de qualidade em idiomas com baixa disponibilidade de recursos, como o Português brasileiro, é um desafio significativo para a moderação automatizada de conteúdo online. Nos últimos anos, a proliferação de interações sociais online e o crescimento de conteúdo gerado por usuários trouxeram à tona a questão crescente da linguagem tóxica. Embora modelos automáticos de aprendizado de máquina tenham sido eficazes na moderação do vasto volume de dados nas redes sociais, ferramentas eficientes para esses idiomas ainda são escassas. Neste trabalho, tratamos essa lacuna criando um conjunto de dados de alta qualidade, coletado de algumas das comunidades brasileiras mais populares da plataforma Reddit. Para isso, rotulamos manualmente um conjunto de 2.500 comentários extraídos das comunidades com maior engajamento e número de inscritos. Realizamos uma análise exploratória para encontrar achados valiosos sobre a linguagem de conteúdo *tóxico* e *não-tóxico*. Nossos resultados mostram um nível moderado de concordância entre os anotadores, validando a relevância desse conjunto de dados para diversas tarefas de aprendizado de máquina. Esta pesquisa busca contribuir para a criação de um ambiente online mais seguro para os usuários que participam de discussões virtuais, além de abrir caminho para o desenvolvimento de ferramentas de moderação automática mais eficazes baseadas em aprendizado de máquina.

Palavras chave

toxicidade; português; conjunto de dados

Abstract

The proliferation of online social interactions in recent years, with the consequent growth in user-generated content, has brought the escalating issue of toxic language. While automatic machine learning models have been effective in moderating the vast amount of data on online social networks, low-resource languages, such as Brazilian Portuguese, still lack efficient automated moderation tools. We address this gap by creating a novel dataset collected from some of the most popular Brazilian Reddit communities. To that end, we manually labeled a sample dataset of 2,500 comments extracted from the most engaging communities. We conducted an in-depth exploratory analysis to gain valuable insights into the language of *toxic* and *non-toxic* content. Our results show a high level of agreement among annotators, attesting to the suitability of this dataset for various downstream machine learning tasks. This research offers a significant contribution to the creation of a safer online environment for users engaging in discussions in Portuguese and paves the way for more effective automatic moderation tools using machine learning.

Keywords

toxicity; portuguese; dataset

1. Introdução

O aumento no número de plataformas de redes sociais estimula cada vez mais as interações de usuários nestas mídias. De acordo com Statista (2022), o número total de usuários das diferentes redes sociais já alcançou os 4 bilhões de pessoas. Esse número sinaliza o nível de importância e onipresença dessas plataformas na sociedade e seu impacto, nem sempre benéfico, na vida das pessoas. De acordo com Vogels (2021), um estudo realizado em 2020 com adultos norte-americanos descobriu que cerca de 41% dos entrevistados haviam experimentado alguma forma de assédio online. Além disso, comentários abusivos em discussões propagam a toxicidade, levando à radicalização das discussões (Salehabadi et al., 2022).

DOI: 10.21814/lm.16.2.459

As consequências dessas interações transcendem o mundo virtual, afetando seriamente a vida dos usuários no mundo real. De acordo com Vogels (2021), 18% dos usuários que participaram de uma enquete sofreram algum tipo de abuso, considerado grave, fora do ambiente online, incluindo ameaças físicas e perseguição.

A moderação manual do conteúdo gerado por usuários tem sido considerada a principal abordagem para atenuar o impacto negativo das interações tóxicas. No entanto, a escala e a velocidade com que o conteúdo é gerado tornam a moderação manual impraticável, o que leva à necessidade de soluções automatizadas (Gillespie, 2020). Os modelos de aprendizado de máquina surgiram como uma alternativa promissora para automatizar a moderação de conteúdo criado online. Esses modelos podem identificar conteúdo potencialmente prejudicial, permitindo que as plataformas tomem medidas proativas, como banir usuários e remover conteúdo prejudicial. Embora os modelos de aprendizado de máquina tenham se mostrado eficazes em vários idiomas (Google, 2022b), seu desempenho em línguas que possuem menos recursos, como o português brasileiro, ainda é insuficiente.

Face a esses desafios, neste trabalho apresentamos um novo conjunto de dados destinado à detecção de toxicidade em português brasileiro. Os textos anotados foram extraídos de uma das redes sociais online mais numerosas —Reddit—, que possui cerca de 1,5 bilhões de usuários registrados e 430 milhões de usuários ativos (Wise, 2023). Reddit é uma comunidade que permite que os usuários interajam por meio de postagens anônimas e comentários. Os usuários se agrupam em comunidades (subreddits) que escolhem por serem mais mais alinhadas com seus tópicos de interesse. Visando propor novos modelos de detecção de toxicidade e aprimorar os já existentes para características específicas da língua portuguesa, realizamos a coleta e a anotação de dados. Nosso conjunto de dados é adaptado para dados de redes sociais online, especificamente com dados do Reddit, uma plataforma que possui características bem distintas das comumente consideradas, como X e Instagram. Além disso, apesar da existência de mecanismos de moderação, muitas publicações ainda contêm conteúdo tóxico, o que reforça a relevância de nossa contribuição no preenchimento da lacuna de dados e modelos disponíveis treinados para o português nesse domínio.

Este artigo está organizado da seguinte forma. Na Seção 2, apresentamos uma síntese da literatura disponível sobre detecção de toxicidade em português. Em seguida, na Seção 3, apresentamos nossa metodologia de coleta e anotação de dados, bem como as técnicas utilizadas para caracterizar a linguagem das postagens coletadas. Na Seção 4, apresentamos os resultados do nosso estudo. Primeiramente, fazemos uma descrição geral do conjunto de dados, para depois relatar nosso experimento no qual comparamos a anotação humana dos dados com a anotação automática por meio da API Perspective e de dois grandes modelos de linguagem (LLMs). Em seguida, caracterizamos a linguagem usada nos comentários *tóxicos* e *não-tóxicos*. A Seção encerra-se com a discussão dos principais achados e seu potencial impacto, sobretudo no que diz respeito ao uso deste conjunto de dados para ajustar os modelos de classificação de toxicidade existentes. Apresentamos também as limitações do nosso estudo. Por último, a Seção 5 apresenta as conclusões do nosso estudo, no escopo de iniciativas que visam aprimorar a moderação automática de conteúdo em um ambiente online em constante crescimento. Ao abordar as deficiências dos recursos existentes, nosso objetivo é contribuir com os esforços para tornar as redes sociais online mais seguras e inclusivas para todos. Para permitir a reprodutibilidade e promover estudos de acompanhamento, publicamos o conjunto de dados anotados para acesso público.[1]

2. Artigos relacionados

Os estudos sobre a detecção automática de comentários tóxicos em idiomas como o português brasileiro são escassos, bem como os conjuntos de dados anotados manualmente e de livre acesso para uso público e estudos de acompanhamento.

de Pelle & Moreira (2017) disponibilizam um conjunto de dados com 1.250 comentários, extraídos de sessões de comentários do site `https://g1.globo.com` e anotados com as categorias ofensivo e não ofensivo, sendo 32,5% do total rotulado como ofensivo. A classe de comentários ofensivos foi subdividida em *racismo, sexismo, homofobia, xenofobia, intolerância religiosa e xingamentos*. Os xingamentos, incluindo linguagem vulgar, foram a categoria mais frequente de comentários ofensivos, presentes em quase 70% dos comentários considerados ofensivos.

Fortuna et al. (2019) descrevem um conjunto de dados com 5.668 tweets, rotulados com um esquema de anotação hierárquica por anotadores com diferentes níveis de expertise. Os anotadores

[1] O conjunto de dados está disponível em `https://github.com/luizhenriqueds/reddit-br-toxicity-dataset/`.

Trabalho	Conjunto de dados	# exemplos	# classes	Anotadores	Especialistas	Tarefa
de Pelle & Moreira (2017)	Comentários de notícias	1,250	2	*Crowdsourcing*	✗	Detecção de linguagem ofensiva
Fortuna et al. (2019)	Twitter	5,668	2+	Especialistas e não especialistas	✓	Classificação binária; Comentários são classificados em 81 diferentes subclasses por especialistas
Leite et al. (2020)	Twitter	21,000	6	42 estudantes voluntários	✗	Multiclasse: insulto, misogenia, racismo, etc.
Vargas et al. (2022)	Instagram	7,000	2+	Não informado	✓	Detecção de linguagem ofensiva e nível de ofensa (se aplicável)
Trajano et al. (2023)	YouTube e Twitter	7,943	5	5 anotadores contratados	✗	Detecção de discurso de ódio; detecção de toxicidade

Tabela 1: Resumo das características dos trabalhos relacionados à construção de conjuntos de dados em Português do Brasil. Trabalhos que estão marcados com *2+* classes indicam estudos de anotação hierárquica, onde uma tarefa inicial de classificação binária (*ofensivo* vs *não-ofensivo*, *tóxico* vs *não-tóxico*, por exemplo) é decomposta em esquemas de anotação mais detalhados e de grão fino.

não especialistas anotaram os tweets com rótulos binários (*ódio vs. não-ódio*).Em seguida, os anotadores mais experientes utilizaram um esquema hierárquico para classificar com maior granularidade os tweets, fazendo uso de 81 categorias relativas ao discurso do ódio.

Leite et al. (2020) apresentam ToLD-Br: um conjunto de dados para a classificação de comentários tóxicos no Twitter em português brasileiro. Um total de 21 mil tweets foi anotado manualmente com sete categorias: *não tóxico, LGBTQ+fobia, obsceno, insulto, racismo, misoginia e xenofobia*. Cada tweet foi rotulado por três anotadores voluntários de uma universidade brasileira. Por meio de uma análise ampla e abrangente, eles comprovam a necessidade de se desenvolver conjuntos de dados específicos por língua para estudos de classificação automática de comentários tóxicos.

O desempenho da API Perspective no português brasileiro é avaliado por Kobellarz & Silva (2022). Os comentários de dois sites brasileiros de mídia de notícias foram traduzidos para o inglês e sua toxicidade foi rotulada pela Perspective tanto em sua versão original em português e sua tradução para o inglês. Os resultados da rotulação automática foram comparados com a anotação humana. A comparação evidenciou melhor desempenho da API em textos em seu idioma original.

O corpus HateBR é apresentado por Vargas et al. (2022). O corpus abrange 7.000 comentários de contas de Instagram de políticos brasileiros, anotados manualmente por especialistas, com uma alta concordância entre os anotadores. Os documentos possuem três tipos de anotação: uma classificação binária (comentários ofensivos versus não ofensivos), nível de ofensividade (altamente, moderadamente e levemente ofensivo) e nove categorias de discurso de ódio (*xenofobia, racismo, homofobia, sexismo, intolerância religiosa, partidarismo, apologia à ditadura, antissemitismo e gordofobia*).

Trajano et al. (2023) apresentam OLID-BR, um conjunto de dados de alta qualidade para detecção de linguagem ofensiva. O conjunto de dados contém 6.354 (extensível a 13.538) comentários rotulados usando um esquema de anotação de três camadas compatível com conjuntos de dados em outros idiomas, o que permite o treinamento de modelos multilíngues.

O trabalho de Oliveira et al. (2023) comparou o desempenho do ChatGPT usando a abordagem *zero-shot* na tarefa de detecção de discurso de ódio em tweets em Português. Os conjuntos de dados utilizados foram o ToLD.BR e o HLPHSP. Os modelos comparados incluem: BERTimbau, ChatGPT-3.5 com variação de *prompts*, DistilBERT e um modelo linear. Os resultados mostraram que o ChatGPT pode superar modelos treinados especificamente para a tarefa, alcançando *score* F1 de 73% e 74% nos conjuntos de dados, respectivamente. Além disso, os resultados indicaram que o desempenho é muito sensível à engenharia de prompts.

A Tabela 1 sumariza as principais características dos trabalhos relacionados apresentados nesta seção. Nosso trabalho contribui para os estudos sobre caracterização de conteúdo tóxico ao investigar comentários em português brasileiro publicados em redes sociais online. Este é o primeiro estudo que enfoca a criação e caracterização de um corpus do Reddit em português brasileiro, anotado manualmente quanto à toxicidade.

3. Metodologia

Nesta seção, primeiro descrevemos nossa metodologia de coleta do corpus. Em seguida, descrevemos o processo para rotular manualmente uma amostra de comentários como *tóxico* e *não-tóxico*. Por fim, apresentamos os métodos usados para analisar a linguagem dos comentários rotulados.

Subreddit	Inscritos	Postagens	Commentários
r/brasil	1,516,433	110,829	2,382,928
r/desabafos	490,049	115,876	1,487,076
r/futebol	369,925	35,826	1,272,009
r/saopaulo	358,681	7,308	88,894
r/eu_nvr	308,064	12,631	221,348
r/botecodoreddit	270,451	7,059	62,999
r/conversas	247,545	21,967	355,761
r/investimentos	232,485	9,756	156,695
r/tiodopave	219,926	2,371	12,106
r/brasilivre	210,582	67,301	1,308,441
	Total	390,924	7,348,257

Tabela 2: Subreddits selecionados, número de assinantes, postagens e comentários do ano 2022.

3.1. Coleta de dados

Reddit é uma rede social online multilíngue fundada em 2005 e organizada em subcomunidades por áreas de interesse (subreddits). Nosso conjunto de dados consiste em atividades de usuários (postagens e comentários) entre janeiro e dezembro de 2022 nos 10 principais subreddits brasileiros com o maior número de assinantes[2] e com uma vida útil de pelo menos cinco anos, o que atesta sua importância nessa rede social online.

A Tabela 2 apresenta os subreddits selecionados e algumas estatísticas descritivas. Coletamos um total de 7.348.257 comentários e 390.924 postagens por meio do Pushshift, uma API de terceiros que agrega comentários e publicações do Reddit (Baumgartner et al., 2020). Em nossa análise, utilizaremos a denominação *comentários* para referir-nos tanto aos comentários quanto às postagens feitas pelos usuários.

Nosso conjunto de dados inclui comentários apenas em português, sendo excluídos comentários de comunidades que permitem discussões em vários idiomas. Aproximadamente 600 mil comentários, nos quais o texto foi substituído por *deletado* ou *removido*, foram excluídos da análise, bem como comentários contendo apenas emojis ou símbolos, URLs, caracteres não alfanuméricos e reações de texto apenas de risada.[3] Por fim, também excluímos comentários gerados por contas de automoderadores e bots que detectamos em nossos dados. Esses filtros reduziram nosso corpus para aproximadamente 6,6 milhões de comentários. A Tabela 3 apresenta algumas estatísticas para os subreddits analisados após a aplicação dos filtros.

Subreddit	Postagens	Comentários
r/brasil	115,876	2,136,866
r/desabafos	115,876	1,211,643
r/futebol	35,826	1,214,412
r/saopaulo	7,308	81,969
r/eu_nvr	12,631	188,620
r/botecodoreddit	7,059	57,298
r/conversas	21,967	326,061
r/investimentos	9,756	141,823
r/tiodopave	2,371	11,584
r/brasilivre	67,301	1,219,265
Total	390,924	6,589,541

Tabela 3: Subreddits selecionados e total de postagens e comentários (2022) após a filtragem.

3.2. Processo de anotação

Primeiramente, extraímos uma amostra de 2.500 comentários do nosso corpus após aplicados os filtros, usando um processo de amostragem estratificada que preservou a distribuição original do número total de comentários por mês em cada subreddit. Essa amostra de comentários foi dividida em 5 Lotes de 500 textos cada. Em seguida, recrutamos 12 alunos de graduação e pós-graduação dos cursos de Ciência da Computação e Estudos da Linguagem de uma universidade brasileira para participarem como anotadores. Os alunos foram divididos em 4 grupos e receberam instruções sobre como rotular cada comentário do Reddit com uma de quatro categorias: *Tóxico*, *Não-tóxico*, *Não sei* ou *Informações insuficientes para rotular o conteúdo*.[4] Para fins de anotação, consideramos linguagem tóxica todo comentário *rude, desrespeitoso ou irracional com*

[2] De acordo com o ranking de Almerekhi et al. (2019).

[3] Em português, textos de risadas são representados pela sequência de caracteres "*kkkkk*"

[4] *Informações insuficientes para rotular o conteúdo* foi uma categoria incluída para futura investigação em um estudo sobre propagação de toxicidade no Reddit.

grande probabilidade de fazer com que alguém saia de uma discussão, conforme definido pela API Perspective. Cada grupo recebeu um Lote e cada comentário foi rotulado por três anotadores independentes. Um dos grupos recebeu um Lote adicional de comentários, dada a alta qualidade da anotação realizada por eles, conforme será discutido na Seção 4.1.

Para a rotulação final do conjunto de dados, a cada comentário do Reddit é atribuída uma das categorias quando há um consenso majoritário entre os anotadores. Aplicamos três métricas para medir a concordância entre os avaliadores: a estatística Kappa de Fleiss, o Alfa de Krippendorf e a Concordância Observada.

3.3. Caracterização da linguagem

Para investigar padrões na linguagem dos comentários de conteúdo tóxico em língua portuguesa, adotamos os seguintes procedimentos na análise do conjunto de dados anotados manualmente.

3.3.1. Identificação automática de comentários tóxicos

Para medir a correlação entre a identificação automática e manual de conteúdo tóxico na amostra de comentários do Reddit, selecionamos três diferentes modelos de aprendizado de máquina. O primeiro deles é o modelo implementado pela API Perspective (Google, 2022b). A API Perspective é um conjunto de classificadores de toxicidade prontos para uso do Google Jigsaw, que foi amplamente usado em pesquisas anteriores (Almerekhi et al., 2020; Salehabadi et al., 2022; Zannettou et al., 2020; ElSherief et al., 2018). A API recebe um comentário como entrada e retorna uma pontuação de 0 a 1 para vários classificadores (por exemplo, palavrões, ameaças, ataques à identidade, toxicidade geral). Com relação ao idioma português, Google (2022a) relatam uma área sob a curva ROC (AUC) de 0,89 para a tarefa de classificação do modelo. Para enriquecer a análise de concordância entre classificação automática e classificação manual, selecionamos dois modelos representantes dos grandes modelos de linguagem (em inglês, *Large Language Models* ou LLMs): o chat GPT-3.5 (Brown et al., 2020) e o Sabia-2 (Almeida et al., 2024). Estes modelos têm sido cada vez mais utilizados para realizar diferentes tarefas de processamento de linguagem (Kukreja et al., 2024).

3.3.2. Análise de classe de palavras

Para caracterizar a linguagem dos comentários *tóxicos* e *não-tóxicos*, exploramos a frequência das palavras utilizadas e sua classe (Part-of-Speech). Para etiquetar a classe de palavra (Petrov et al., 2011), usamos um modelo pré-treinado (spaCy, 2022) e um banco de árvores em português (treebank) anotado com sintaxe de dependência de acordo com as diretrizes das Dependências Universais (Rademaker et al., 2017). O modelo selecionado possui uma precisão de mais de 97%.

Calculamos a frequência das etiquetas de classe de palavra nos comentários *tóxicos* e *não-tóxicos* a fim de descobrir se essa poderia ser uma característica distintiva dos dois tipos de comentários.

3.3.3. Razão type-token e extensão dos comentários

Para computar a razão type-token, dividimos o número total de palavras não repetidas ("types") pelo número total de palavras "tokens"). Também comparamos o tamanho dos comentários *tóxicos* e *não-tóxicos*. Diferentemente de outras redes sociais online, no Reddit não há restrições para o tamanho do comentário; portanto, essa medida permite calcular a probabilidade de os usuários publicarem um texto curto ou longo na plataforma.

3.3.4. Análise de tópicos

Para extrair os tópicos dos comentários em que os anotadores mais concordam ou discordam, executamos o modelo BERTopic (Grootendorst, 2022), que se baseia em uma representação vetorial para agrupar documentos semelhantes.

3.3.5. Análise de n-gramas

N-gramas são sequências de n tokens consecutivos de um determinado conjunto de dados, usados para representação da ocorrência de termos. Formalmente, definimos um n-grama como uma sequência de palavras $w_1, w_2, \ldots, w_n$, onde cada w_i representa um token no corpus. Neste artigo, apresentamos resultados para unigramas ($n = 1$) e para bigramas ($n = 2$).

3.3.6. Grafos de coocorrência

Uma análise complementar de um corpus pode ser feita por meio da observação de redes de coocorrência. Para obter essa rede, contamos o

número de coocorrências de cada par de palavras encontradas nos comentários analisados. Em seguida, definimos um grafo onde os vértices representam as palavras e as arestas indicam se existe coocorrência nos comentários coletados. Para a construção do grafo, consideramos os pares de palavras que ocorreram pelo menos duas vezes. Os nós do grafo são termos mencionados pelo menos 50 vezes nos comentários analisados.

3.3.7. Reconhecimento de entidades nomeadas

Investigamos entidades nomeadas nos comentários do Reddit com base em um modelo pré-treinado do Spacy para reconhecimento de entidades nomeadas (NER). O modelo usado foi treinado para o português brasileiro usando o conjunto de dados WikiNER (Nothman et al., 2013) e classifica as entidades em três categorias predefinidas: PESSOA, LOCALIZAÇÃO e ORGANIZAÇÃO. As entidades não definidas são classificadas como DIVERSAS.

4. Resultados

Nesta seção, apresentamos os principais resultados obtidos com a avaliação e a caracterização do conjunto de dados anotado manualmente.

4.1. Concordância entre anotadores

Primeiramente, calculamos a concordância global entre os anotadores nos comentários do Reddit rotulados manualmente, cujos resultados são mostrados na Tabela 4.

Métrica	Global	Rótulos binários (*Não-tóxico* ou *Tóxico*)
Kappa de Fleiss	0,31	0,46
Alfa de Krippendorff	0,35	0,46
Concordância Observada	0,64	0,80

Tabela 4: Concordância entre anotadores.

Como esperado, a métrica de *concordância observada* obteve os valores mais altos, pois essa medida não leva em conta a possibilidade de a concordância ocorrer por acaso. Houve total concordância e total discordância em 1.594 e 107 comentários, respectivamente. Um exemplo de concordância total sobre um comentário considerado *tóxico* é: *"Como assim? Eu nem sou o OP. Só tô dizendo que ele é retardado de seguir a medicina de gado"*. Por outro lado, um exemplo de discordância total é um comentário polêmico como: *"[..] é o lugar do Brasil que mais tem neonazi mesmo ué"*, o que aponta para o alto nível de subjetividade da tarefa de classificação.

Com relação às métricas *Kappa de Fleiss* e *Alfa de Krippendorff*, os valores indicam concordância de razoável a moderada no pior dos casos. Por fim, a toxicidade geral classificada pelos anotadores foi de 11,28%, com 88,7% de comentários *não-tóxicos*, o que é consistente com a natureza desbalanceada do problema.

Em seguida, calculamos a concordância entre os anotadores de cada grupo, denominados A, B, C e D, para os Lotes de comentários, numerados como 1, 2, 3, 4 e 5. Os Lotes 3 e 5 foram anotados pelo grupo C, enquanto os Lotes 1, 2 e 4 foram anotados pelos grupos A, B e D, respectivamente. O Lote 5 foi rotulado em uma segunda rodada de anotação pelo grupo C, selecionado para isso por ser o grupo que obteve o maior valor de *Kappa de Fleiss* e *Alfa de Krippendorff* para a concordância entre anotadores na primeira rodada. A Tabela 5 mostra os resultados. Com exceção do Grupo D, que obteve de leve a nenhuma concordância, os grupos A, B e C obtiveram uma concordância de razoável a moderada.

Em seguida, examinamos a rotulagem feita por cada anotador, cujos resultados são mostrados na Tabela 6. O grupo A rotulou a menor porcentagem de comentários como *tóxico*. Já o Grupo B apresenta a maior variabilidade na rotulagem de conteúdo tóxico, sendo o anotador 2 o que rotulou mais de 21% dos comentários como *tóxicos*. Assim como o Grupo B, o Grupo D atingiu um nível não negligenciável de incerteza na tarefa de classificação, sendo que o anotador 2 tendeu a ser mais tolerante com o conteúdo *tóxico* em potencial. Para fins de ilustração, o comentário *"Vamos fingir que não é (você) que posta que quer morrer por ser depressivo. Pick me boy"* foi classificado como *tóxico* pelos anotadores 1 e 3 e como *não-tóxico* pelo anotador 2. Os anotadores do Grupo C, que receberam os Lotes 3 e 5, são os que apresentam o menor grau de incerteza.

Em geral, nossos resultados corroboram o alto nível de subjetividade inerente à tarefa de classificar conteúdo como *tóxico* ou *não-tóxico*. Isso está de acordo com os resultados da literatura sobre como a percepção do grau de severidade do conteúdo nocivo é afetada por valores individuais e culturais (Jiang et al., 2021).

Métrica	Lote 1	Lote 2	Lote 3	Lote 4	Lote 5
Kappa de Fleiss	0,46	0,33	0,51	0,17	0,54
Alfa de Krippendorff	0,46	0,33	0,51	0,17	0,54
Concordância observada	0,87	0,81	0,76	0,78	0,77

Tabela 5: Métricas de concordância entre anotadores por Lote de anotação.

	Lote 1 (Grupo A)			Lote 2 (Grupo B)			Lote 3 (Grupo C)		
	Anot. 1	Anot. 2	Anot. 3	Anot. 1	Anot. 2	Anot. 3	Anot. 1	Anot. 2	Anot. 3
Não-tóxico	84,60%	88,96%	90,60%	83,17%	69,48%	74,95%	75,90%	68,01%	78,51%
Tóxico	9,40%	9,84%	7,40%	7,82%	21,29%	4,81%	19,28%	21,73%	17,87%
Não sei dizer	0,60%	1,00%	0,00%	3,81%	3,82%	2,81%	4,02%	7,65%	3,01%
Informação insuficiente	5,40%	0,20%	2,00%	5,21%	5,42%	17,43%	0,80%	2,62%	0,60%

	Lote 4 (Grupo D)			Lote 5 (Grupo C)		
	Anot. 1	Anot. 2	Anot. 3	Anot. 1	Anot. 2	Anot. 3
Não-tóxico	72,80%	93,59%	69,14%	84,51%	68,60%	75,20%
Tóxico	11,60%	5,21%	9,02%	14,49%	25,00%	19,72%
Não sei dizer	4,20%	0,80%	6,41%	1,01%	4,20%	4,67%
Informação insuficiente	11,40%	0,40%	15,43%	0,00%	2,20%	0,41%

Tabela 6: Distribuição dos rótulos de anotação para cada grupo de anotadores.

4.2. Comparação entre a rotulagem manual e as rotulagens automáticas

4.2.1. API Perspective

Comparamos nossa anotação manual de dados com a realizada pela API Perspective. Consideramos tóxicos os comentários aos quais a API Perspective atribuiu uma pontuação de **toxicidade grave** acima de 0,7. Essa decisão prioriza um bom equilíbrio entre precisão e revocação, pois nossa intenção é compreender melhor os principais motivos de concordância e discordância na classificação de conteúdo *tóxico* e *não-tóxico*. Um valor limite de 0,9 resulta na seleção de apenas 3% dos comentários tóxicos para comparação. Em contrapartida, um valor de 0,7 retorna aproximadamente 10% dos comentários como tóxicos, uma porcentagem semelhante àquela rotulada por nossos anotadores.

Limiar	Precisão	Revocação	F1	#Tóxico
0,5	0,65	0,69	0,67	92
0,6	0,69	0,62	0,65	78
0,7	0,8	0,41	0,55	45
0,8	0,81	0,4	0,54	43
0,9	1,00	0,15	0,26	13

Tabela 7: Desempenho da API Perspective no conjunto de dados de teste com distintos limiares de escore de toxicidade.

Porcentagem de toxicidade Primeiro, analisamos a porcentagem de comentários anotados como *tóxicos* por nossos voluntários e a porcentagem rotulada pela API Perspective. O Grupo A (Lote 1) anotou menos comentários tóxicos do que a API Perspective, enquanto um anotador do Grupo B (Lote 2) classificou uma porcentagem muito maior de comentários como *tóxicos*. O Grupo C (Lotes 3 e 5) evidencia uma anotação consistente de maior número de comentários tóxicos do que API Perspective. O Grupo D (Lote 4), apesar de mostrar grande discordância entre anotadores, também evidenciou um número menor de comentários tóxicos do que a API. A Tabela 7 mostra o desempenho da API Perspective em uma amostra de teste rotulada pelo Grupo C. O objetivo da análise não é comparar diretamente a concordância entre anotadores humanos e a API Perspective, mas, sim, avaliar a qualidade das previsões da API em diferentes limiares em um conjunto de teste com curadoria. Os resultados indicam uma clara compensação (trade-off) no desempenho entre precisão e revocação. Na prática, ao escolher um limiar de precisão alto, estamos compensando uma grande parte do desempenho de revocação. Portanto, o modelo treinado do Perspective tem uma grande margem de melhoria para textos em português do Brasil, considerando os limiares selecionados. Combinando as métricas de revocação e precisão, obtemos um escore F1 máximo de 0,67.

Correlação entre as anotações tóxicas Calculamos como os rótulos manuais e os rótulos da API Perspective se correlacionam entre si. A correlação geral de Benesty et al. (2009) na amostra de teste é de 0,51, comparando o rótulo do voto majoritário para cada comentário. Também calculamos a correlação entre grupos de anotado-

res e as previsões da API Perspective. Os anotadores dos Lotes 1, 2 e 3 apresentaram uma correlação moderada consistente com a API, enquanto os anotadores do Lote 4 apresentaram uma correlação fraca. Por fim, os anotadores do Lote 5 apresentaram uma correlação consistente e forte com a API.

4.2.2. Grandes modelos de linguagem

Para comparar os resultados de grandes modelos de linguagens (LLMs) com as anotações manuais, classificamos o mesmo conjunto de dados previamente anotado usando os modelos GPT-3.5 e Sabia-2. O primeiro modelo é treinado com dados de diversos idiomas (inclusive o Português do Brasil), enquanto o segundo é uma versão ajustada (*fine-tuned*) para dados em Português. Para ambos os modelos, definimos o *prompt* que indica, de forma bem estruturada e clara, as principais instruções para a realização da tarefa de classificação de toxicidade. A Figura 1 mostra o *prompt* utilizado para a tarefa de classificação. Para reproduzir as mesmas condições do experimento com anotadores, essas instruções foram baseadas nas mesmas instruções fornecidas aos anotadores para fazer a anotação manual do conjunto de dados. Por fim, para ambos os modelos, consideramos o tamanho máximo de tokens de entrada de cada modelo igual a 500 e temperatura[5] igual a zero.

> Você é um assistente e deve atuar como moderador de conteúdo de redes sociais, responsável por classificar comentários do Reddit em português brasileiro (PT-BR) como *tóxicos* ou *não-tóxicos*.
>
> Siga as instruções abaixo para realizar a tarefa:
>
> **1.** Um comentário é considerado *tóxico* se for um texto ofensivo ou rude em relação a outra pessoa em publicações de redes sociais;
>
> **2.** Você receberá o texto de um comentário, e sua tarefa é classificar a toxicidade do texto fornecido. Use apenas essas informações para fazer a previsão;
>
> **3.** Para cada comentário, classifique-o em uma das seguintes categorias: *Não-tóxico* ou *Tóxico*.

Figura 1: Instruções fornecidas ao assistente para classificação de comentários no Reddit usando a abordagem *zero-shot*.

[5]Temperatura é um parâmetro que permite modificar a saída de um modelo de linguagem, tornando-as mais previsível ou criativa.

Porcentagem de toxicidade Ao analisar a proporção de comentários classificados como *tóxico* pelos modelos, observamos que o modelo Sabiá-2 se aproxima da distribuição do percentual de toxicidade por Lotes observada para a Perspective API, variando entre 9,7% para o Lote 4 e 12,09% para o Lote 5, com média 11,45%. Em contrapartida, as classificações de toxicidade do ChatGPT foram maiores em média (28,40%), variando entre 29,2% para o Lote 2 e 31,4% para o Lote 3. Esse resultado evidencia uma maior sensibilidade do ChatGPT ao teor de toxicidade dos comentários analisados no período, enquanto o Sabiá-2 é menos influenciado por nuances na semântica das discussões, uma vez que o percentual de classificação se aproxima da distribuição das anotações manuais. Como exemplo, o seguinte comentário foi classificado como *tóxico* pelo ChatGPT e *não-tóxico* pelos anotadores *"Isso é bobagem, país nenhum tá livre de corrupção, só algum paiseco inexpressivo talvez, tipo a nova Zelândia. O que o povo quer de verdade é uma vida de qualidade e com segurança, e isso nem Bolsonaro nem Lula chegam perto de promover"*, enquanto o mesmo acontece com o Sabiá-2 classificando o seguinte comentário *"Fala que roubado é mais gostoso"*.

Para analisar a sensibilidade dos modelos a um comentário que foi classificado como *tóxico* de forma unânime, submetemos o seguinte comentário *"Posso fazer a piada do bebe morto?"* aos LLMs e tanto o ChatGPT quanto o Sabia-2 classificaram o teor deste comentário como *tóxico*, em contraste com a Perspective API que o classificou como *não-tóxico*.

Os resultados da comparação entre a anotação manual e a anotação automática revelam diferenças relevantes no comportamento dos modelos Sabiá-2 e ChatGPT em relação à classificação de toxicidade. A Tabela 8 mostra que o modelo Sabiá-2 apresentou um equilíbrio entre falsos positivos (128) e falsos negativos (126), demonstrando uma abordagem mais balanceada. Em contrapartida, o ChatGPT teve um número consideravelmente maior de falsos positivos (454), mas um número significativamente menor de falsos negativos (43), indicando que, embora mais propenso a classificar erroneamente comentários *não-tóxicos* como *tóxicos*, o ChatGPT foi mais eficaz na identificação de comentários realmente *tóxicos*. Esses resultados corroboram com a análise anterior de que o Sabia-2 é menos conservador, enquanto o ChatGPT é mais sensível à presença de toxicidade.

		Lt 1	Lt 2	Lt 3	Lt 4	Lt 5	Total
Sabia-2	**FP**	33	30	18	36	11	**128**
	FN	16	11	46	11	42	**126**
ChatGPT	**FP**	106	88	82	113	65	**454**
	FN	6	4	15	5	13	**43**
	Total	**161**	**133**	**161**	**165**	**131**	

Tabela 8: Falso positivos (FP) e falso negativos (FN) por grupo.

Essas características, embora diferentes, podem ser aplicadas a depender do cenário em questão. Por exemplo, se o interesse for minimizar a presença de falsos negativos, pode ser interessante empregar um modelo que seja mais sensível à presença de toxicidade e, por sua vez, consiga capturar esses casos com mais precisão. Entretanto, pensando na quantidade massiva de comentários em redes sociais online, uma abordagem híbrida que alcance o maior equilíbrio entre os dois tipos de erro pode ser mais eficiente.

Correlação entre as anotações tóxicas Ao comparar a classificação dos anotadores com os modelos, observa-se uma correlação moderada entre as classes geradas pelos modelos e anotadores. A correlação entre os anotadores e o Sabiá-2 é de 0,43, enquanto o ChatGPT teve uma correlação um pouco menor (0,41). Ao analisar a correlação por Lotes do conjunto de dados, tivemos a menor correlação para o Lote 4 (Sabia-2: 0,22, ChatGPT: 0,59) e a maior correlação para o Lote 5 (Sabia-2: 0,21, ChatGPT: 0,58), reforçando a alta concordância do Grupo C com modelos de classificação automática. Para explorarmos a concordância entre as classes, calculamos o Kappa de Cohen entre as anotações e as classificações dos modelos. A Tabela 9 mostra o Kappa score para cada grupo. Observa-se, novamente, menor concordância entre o Grupo D e maior concordância com os Grupo C.

	Lt 1	Lt 2	Lt 3	Lt 4	Lt 5
Sabiá-2	0,3964	0,3583	0,4396	0,2072	0,5779
ChatGPT-3.5	0,2628	0,2418	0,4511	0,1277	0,5616

Tabela 9: *Kappa score* entre anotações e modelos de linguagem (PT-BR) instruídos para executar a tarefa de classificação de toxicidade.

Por fim, comparamos o resultado dos modelos de classificação automática. O modelo pré-treinado da Perspective API nos permite controlar a flexibilidade das classificações, uma vez que um *score* de toxicidade é calculado. No limiar de 0,7, o modelo alcançou precisão de 0,8 e *recall* de 0,41. Nesse limiar, o modelo é capaz de classificar comentários como *tóxico* com mais precisão, ao custo de deixar de recuperar algumas instâncias. Para os modelos de linguagem, não temos acesso ao *score* com a confiança do modelo. Entretanto, através de engenharia de *prompt*, instruímos os LLMs a favorecer a precisão (ou seja, favorecer casos em que o modelo tem alta confiança de que um comentário é realmente *tóxico*). O Sabia-2 apresentou equilíbrio entre as métricas de precisão e revocação (precisão: 0,71; revocação: 0,72), enquanto o ChatGPT favoreceu muito mais a métrica de precisão (precisão: 0,8; revocação: 0,64). Todos os modelos de classificação automática foram influenciados pela presença de gíriais locais e palavrões, como discutido anteriormente.

4.3. Caracterização da linguagem do conteúdo tóxico e não tóxico

Comparamos os padrões na linguagem do conteúdo *tóxico* e *não-tóxico* para entender melhor como os falantes de português utilizam a linguagem para gerar conteúdo tóxico.

4.3.1. Análise da extensão dos comentários e razão type-token

Com relação à extensão dos comentários, o número médio de tokens e o intervalo de confiança de 95% para comentários *não-tóxicos* é 26,34 [24,68, 28,19]. Para os comentários *tóxicos* a média é de 35,54 [29,41, 42,87]. Portanto, os comentários *tóxicos* são, em média, mais longos do que os *não-tóxicos* (p-valor $< 0,05$). A distribuição do comprimento nos comentários *tóxicos* tem um intervalo maior, o que pode indicar diferenças dentro dos próprios subreddits.

A média da razão type-token e o intervalo de confiança para os comentários *não-tóxicos* é 0,78 [0,78, 0,79], enquanto para os comentários *tóxicos* é 0,83 [0,82, 0,84]. Os resultados apontam para significância estatística, tendo os comentários *tóxicos* mais diversidade lexical. Isso pode variar entre os subreddits, pois algumas das comunidades são mais propensas a ter postagens mais verbosas.

4.3.2. Análise de etiquetas de POS

A diversidade de etiquetas de classe de palavra (POS) para comentários *não-tóxicos* tem uma média de 0,51 [0,50, 0,52], enquanto que para textos *tóxicos* rotulados a média é de 0,46 [0,43, 0,48]. Embora os comentários *tóxicos* sejam mais longos, eles geralmente são menos diversificados em termos de etiquetas de POS.

Para investigar melhor as etiquetas de POS, comparamos a distribuição de algumas etiquetas específicas. Primeiro, comparamos os adjetivos (ADJ) com uma média de 1,68 [1,55, 1,81] para comentários *não-tóxicos* e 2,14 [1,71, 2,66] para comentários *tóxicos*. Como os intervalos de confiança se sobrepõem entre as classes, realizamos um teste estatístico Mann-Whitney para comparar as diferenças nas distribuições. O uso da etiqueta ADJ é estatisticamente diferente entre as classes, com um valor de $p < 0{,}01$.

Da mesma forma, realizamos o mesmo teste para a etiqueta NOUN, da classe de palavra substantivo. O uso médio em comentários *não-tóxicos* é de 5,43 [5,07, 5,83], enquanto nos comentários *tóxicos* a média é de 7,44 [6,15, 8,94]. Essa diferença é novamente validada pelo teste Mann-Whitney, com um valor de $p < 0{,}01$.

Uma análise da distribuição de etiquetas de POS nos comentários é essencial para entender as características do texto gerado pelos usuários do Reddit nas maiores comunidades brasileiras. Para isso, usamos o marcador de POS pré-treinado do Spacy para o português brasileiro. Cada token em uma frase foi classificado com uma das etiquetas de POS existentes. À lista de etiquetas de POS, foram adicionadas outras classes específicas para o problema de classificação de etiquetas, como SYM, SPACE e X para denotar "símbolos", "espaço em branco" e "outros", respectivamente, com a observação de que, como esse é um modelo de aprendizado de máquina treinado em corpora pertencentes a outros domínios, a classificação pode resultar em falsos positivos.

As duas etiquetas de POS mais comuns para comentários *tóxicos* e *não-tóxicos* são NOUN (substantivo) e VERB (verbo). Os comentários *não-tóxicos* usam mais etiquetas PROPN (nome próprio), enquanto uma alta porcentagem de tokens de comentários *tóxicos* foi etiquetada como PUNCT (sinal de pontuação). Além disso, os comentários *tóxicos* usam mais interjeições, etiquetadas como INTJ. Também comparamos as distribuições de etiquetas de POS de ambas as classes por meio de um teste de qui-quadrado. Os resultados indicam que a diferença observada entre a distribuição das etiquetas de POS é significativa (p-valor $< 0{,}05$).

Para uma análise mais detalhada das diferenças no uso de palavras nos comentários *tóxicos* e *não-tóxicos*, extraímos as palavras mais frequentes por classe de comentário para as etiquetas de POS mais frequentes, ou seja, as etiquetas ADJ, NOUN e PROPN. Os resultados são mostrados na Tabela 10. Uma descoberta relevante é o termo *mulher* em comentários *tóxicos*. De fato, realizamos um teste de qui-quadrado para comparar a associação desse termo com comentários *tóxicos* e *não-tóxicos* Os resultados indicam uma associação positiva para alguns dos subreddits brasileiros (como o r/desabafos) com valor de $p < 0{,}05$. Esse resultado pode sugerir a presença de comportamento misógino associado a alguns tópicos e comunidades nas redes sociais. Exemplos de comentários direcionados a mulheres nas discussões das comunidades podem ser encontrados na (Tabela 14).

4.3.3. Análise de tópicos

Investigamos os comentários sobre os quais os anotadores discordaram totalmente, especialmente em relação aos principais tópicos extraídos com o modelo BERTopic. Eles dizem respeito a discussões relacionadas a grupos específicos (mulheres, homens) e abrangem vários temas, incluindo finanças, guerra, governo e relacionamentos (Tabela 11). As palavras no tópico 0 (*feedback, remoção*) revelam que alguns comentários foram moderados anteriormente pela DMCA (Digital Millennium Copyright Act, (US Congress, 1998)).

É interessante notar que os principais tópicos dos comentários sobre os quais os anotadores concordaram totalmente também discutem os mesmos temas (Tabela 12). Entretanto, os descritores de tópicos incluem mais termos ofensivos (como palavrões) e ideologicamente carregados.

Com relação aos tópicos extraídos dos comentários em que os três anotadores concordaram no rótulo *tóxico* e que a API Perspective classificou como *não-tóxico* (Tabela 13), os principais têm a ver com política, liberdade, discriminação e grupos-alvo. Os resultados indicam que a API é menos sensível ao contexto para essa tarefa específica do português brasileiro. Por exemplo, o comentário a seguir foi rotulado como *tóxico* por todos os três anotadores, mas foi classificado como *não-tóxico* pelo modelo de aprendizado de máquina *"Posso fazer a piada do bebe morto?"*. Ao comparar com LLMs, vemos um comportamento similar. Por exemplo, o seguinte comentário foi classificado como *tóxico* pelos anotadores e *não-tóxico* pelo ChatGPT: *"Não, acho que deveriam fazer rinha de fake news, f*da-se, no fim, 'todo mundo' recebe e sabe discernir e prefere a informação verdadeira."*, enquanto o comentário *"Nojento! E bizarro o quão frequente isso é. Só demonstra que a nossa sociedade é machista e doente."* foi classificado como *tóxico* pelo processo de anotação e *não-tóxico* pelo Sabia-2.

	ADJ	**NOUN**	**PROPN**
Não-tóxico	bom, melhor, mesmo, grande, mesma, pior, fácil, diferente	cara, gente, pessoas, coisa, tempo, anos, vida, mundo, dinheiro	Brasil, Lula, Bolsonaro, OP, Deus, Flamengo, Landau, Ciro, PT, STF
Tóxico	melhor, mesmo, pobre, ruim, primeiro, forte, diferente, social, capaz, política, rico	pessoas, cara, mundo, mulher, c*, casa, homem, m**da, pai	Lula, Bolsonaro, Brasil, OP, Ciro, Ucrânia, Flamengo, FDP, Liberdade, Rússia, Paris

Tabela 10: Palavras mais frequentes por etiqueta de POS e classe de comentário.

Tópico	**Descritores**
0	video, mulher, opinião, homem, dinheiro, beleza, burro , padrão, feedback, removal
1	guerra, liberdade, post, motivo, país, massacres, massa, atrocidades, históricas, democracia, governo, xenofóbico
2	m**da, sexo, maluco, apoiadores, preocupado, machão, malditos, insegurança, op (original poster)

Tabela 11: Tópicos e palavras-chave relevantes em comentários em que os três anotadores discordaram (desacordo total).

A Tabela 14 mostra exemplos de comentários direcionados à *mulher* diretamente em comentários *tóxicos*. Alguns dos comentários causaram total desacordo entre os anotadores. Por exemplo, o comentário *"Pelo direito de bater na própria mulher! Uow"* foi rotulado como 'Não sei', 'Tóxico' e 'Não tóxico'. Uma hipótese é que o texto seja interpretado como um comentário sarcástico ou irônico. Um experimento adicional com mais contexto (como fornecer a sequência da conversa ao anotador) pode mitigar o desacordo nesses casos. Comentários que exigem detalhes contextuais são difíceis de rotular, mesmo para anotadores humanos, e ainda mais para modelos de aprendizado de máquina treinados em corpora que não se assemelham a interações em redes sociais. Na verdade, para esse comentário específico, a API do Perspective previu como *não-tóxico* com as configurações de parâmetros padrão. Em relação aos LLMs, o ChatGPT classificou corretamente os dois primeiros exemplos da tabela como *tóxico* e o Sabiá-2 classificou os três comentários como *não-tóxico*.

4.3.4. Análise de n-gramas

A Figura 2 apresenta os termos (unigramas) mais frequentes considerando todos os comentários do conjunto de dados. Termos como *pessoas, melhor, gente, ano, mundo, Brasil* aparecem em grande destaque, refletindo o foco nas discussões sobre pessoas, eventos globais e relacionados ao país. Além desses termos, encontramos termos como *vida, problema, caso, amigo*, que sugerem discussões sobre questões pessoais, desafios e relacionamentos, assuntos que são frequentemente discutidos em algumas das principais comunidades, como r/desabafos e r/conversas. Essas comunidades incentivam a troca de ideias acerca de assuntos do dia-a-dia e discussões sobre relacionamentos. Por fim, a presença de termos como *Lula, Bolsonaro* mostra que a política está entre os assuntos mais discutidos, refletindo a polarização e o engajamento nas questões de política interna, principalmente considerando que os dados foram coletados em 2022, ano de eleição presidencial no Brasil.

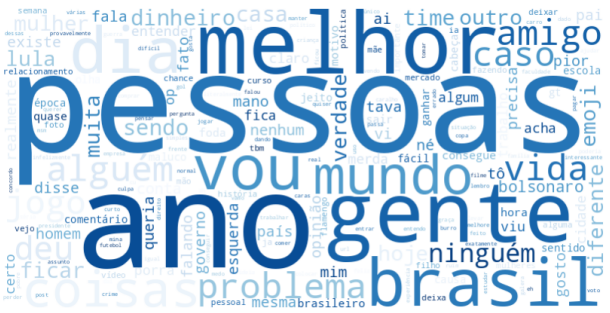

Figura 2: Unigramas mais frequentes encontrados nos comentários das principais comunidades brasileiras no Reddit.

Já a Figura 3 mostra os unigramas mais frequentes divididos por classe de comentários. Nos comentários *tóxicos*, palavras como *Lula, Bolsonaro, m*a, p*a, burro* se destacam, muitas vezes associadas a discussões sobre política e insultos. É possível notar que a polarização política é predominante nas discussões —uma vez que os dois principais candidatos à pre-

Tópico	Descritores
0	burro, homem, p**ra, c*, mulher, mercado, gente, anos, país, b**ta, criança, ódio, sentido
1	guerra, bolsonaro, ucrânia, realidade, putin, intervenção, pobre, nuclear, bandido, vergonha, russia
2	ideologia, liberdade, política, mundo, cancelamento, expressão, op (original poster), preconceito, oprimidos, vagabundo, família

Tabela 12: Tópicos e palavras-chave relevantes em comentários em que todos os três anotadores classificaram como *tóxico*.

Tópico	Descritores
0	ideologia, política, liberdade, mundo, pessoas, expressão, mulheres, preconceito, bolsonaro, esquerdistas, apolíticos, piada, realidade, oprimidos, opiniões

Tabela 13: Palavras-chaves extraídas de comentários em que todos os três anotadores classificaram como *tóxicos* e que a API do Perspective previu como *não-tóxicos* (falsos negativos).

sidência em 2022 são mencionados— e insultos direcionados a outros usuários da plataforma. Além disso, é possível observar a presença de termos que indicam discussões mais ideológicas ou culturais, como *governo, ideologia, política*. Por fim, o termo *guerra* também é mencionado com frequência nos comentários classificados como *tóxicos*, o que possivelmente foi causado pelo estopim da guerra Rússia-Ucrânia no período de coleta dos dados, revelando o uso difundido dessa rede social para discutir assuntos inerentes da globalização mundial.

Para uma análise simplificada de contexto, a Tabela 15 apresenta os bigramas mais recorrentes nas comunidades analisadas neste artigo. Conforme mencionado anteriormente, durante o período de coleta dos dados houve a eleição presidencial no Brasil, o que explica a alta frequência de termos como *votar lula*, *liberdade expressão* em comentários classificados como tóxicos, revelando o cenário de polarização interna (*"Eu deixo claro minha opinião: sou contra liberdade de expressão(tem gente que não sabe o que diz) e sou contra a democracia (tem gente que não sabe votar)... kkkkk mas respeito ambas as coisas por que sei que a b*s*a deste mundo não é perfeito"*). Além disso, observa-se a presença da expressão *tio sam*, usada como referência aos Estados Unidos. Por fim, há uma forte ocorrência de palavrões e insultos, reforçando a associação deste tipo de linguagem com toxicidade percebida por humanos, como, por exemplo, o comentário *"fit cultural é de cair o c* da bunda. Desculpa para qualquer coisa p*p"*.

No caso dos comentários *não-tóxicos*, a análise dos bigramas revela uma variedade de temas presentes nas principais comunidades brasileiras do Reddit. Nos subreddits r/desabafos e r/brasil, os termos mais frequentes estão relacionados a reflexões pessoais e discussões construtivas sobre temas sociais e políticos. Bigramas como *muita gente*, *ler escrever* e *ficar sozinha* refletem interações mais introspectivas e emocionais, típicas de conversas sobre experiências individuais e desabafos, por exemplo: *"Não, nem um pouco. Você não chorar não significa que não amava a pessoa ou não a considerava. Tem muita gente que não consegue chorar com essas coisas; e isso é completamente normal"*. Já em r/brasil, aparecem tópicos mais amplos como *rock rio*, *redes sociais*, e figuras políticas como *ciro gomes*.

No subreddit r/brasilivre, observa-se uma concentração de termos como *processo eleitoral* e *opinião pública*, sugerindo debates politizados e possíveis temas controversos. Além disso, é possível observar uma grande prevalência de comentários removidos nessa comunidade, indicados pelos termos *view, removal* e *removal, request*. Em r/futebol, temos bigramas como *time jogando* e *flamengo culpa*, esperados dada a natureza do subreddit. No entanto, ainda há a presença de termos relacionados às eleições e políticas externas, como, por exemplo *ministério público* e *guerra*.

Por fim, no subreddit r/investimentos, os bigramas revelam uma ênfase em tópicos educacionais e profissionais, com termos como *lógica matemática*, *atividades extracurriculares*, e *curso voltado*, indicando discussões focadas no desenvolvimento pessoal e financeiro. Na comunidade

Comentário	Texto
1	*"A minoria quer realmente ser independente - mas como o universo do /r/brasil é majoritariamente progressista, não irão concordar - as demais estão entre o "mulher tem que ser mulher" e aquelas que usam o discurso de independência, mas acham que quem tem que pagar as coisas é o homem."*
2	*"O mundo é assim, do mesmo jeito que você não quer uma mulher feia, uma mulher não vai querer alguém feio ou sem status, não cai nesse papo de que aparência não importa que em rede social só tem alienado, veja você mesmo pesquisas relacionadas ao assunto ou se tiver coragem crie um perfil com a foto de alguém bonito e veja como as pessoas te tratam diferente."*
3	*"[..] Mas o homem casa com quem ele quiser. A mulher casa com quem ela consegue."*

Tabela 14: Exemplos de comentários mencionando a palavra "mulher" em comentários *tóxicos*.

Figura 3: Termos frequentes em comentários classificados como *tóxico* (esquerda) e comentários classificados como *não-tóxicos* (direita).

r/conversas, também temos a prevalência da moderação de conteúdo, onde comentários são frequentemente removidos pela moderação por provavelmente infringirem direitos autorais ou regras da comunidade.

Em resumo, é possível observar que termos políticos dominam os assuntos discutidos e estão presentes em praticamente todas as comunidades. Além disso, em comunidades consideradas mais liberais, temos maior incidência de moderação ou conteúdo removido por direitos autorais. Consideramos comunidades liberais, comunidades que discutem majoritariamente assuntos considerados polêmicos, o que pode influenciar na tendência dos usuários a tecer comentários tóxicos e/ou ataques direcionados. Nesse tipo de discussão, pode ter uma influência maior do uso de termos pejorativos ou palavrões e isso influenciar na decisão de modelos automáticos, por exemplo. Além disso, algumas comunidades - como a r/brasilivre - incentivam abertamente o debate livre entre os participantes. Segundo a sua definição no Reddit: *O subreddit brasileiro mais livre do reddit. Nosso objetivo é proporcionar a todos um espaço aberto e livre para debates.*

4.3.5. Grafo de coocorrência

O grafo de coocorrência extraído dos comentários possui o total de 9373 nós e 41844 arestas, com grau médio das arestas igual a 6,11 para termos associados com comentários *tóxicos* e 10,33 para termos associados com comentários *não-tóxico*. A Figura 4 apresenta a métrica de centralidade de nó baseada no grau do nó do grafo. Esta métrica é definida por $C_D(v) = \deg(v)$ e mede o número de conexões que um nó (termo) tem com outros nós no grafo. Em uma matriz de coocorrência, os termos influentes são termos que coocorrem com muitos outros termos no conjunto. O grau de cada nó a partir da matriz de adjacências é definido pela soma $\deg(v) = \sum_{u \in V} A_{vu}$.

Para os comentários *tóxicos*, *mundo, pessoas, gente* aparecem com alta centralidade, refletindo seu papel central nas discussões polarizadas, frequentemente ligadas a temas políticos, como demonstrado pela presença de *esquerda, Bolsonaro, Lula*. Esses termos indicam que debates políticos intensos geram alta toxicidade, especialmente em torno de figuras públicas e ideologias.

Para os comentários classificados com *não-tóxicos*, como *coisas, dia, pessoas, hoje*, possuem alta centralidade de grau e representam discussões mais cotidianas e neutras, voltadas a in-

	Tóxico	**Não-tóxico**
r/desabafos	{preferir, ficar}, {compra, carro}, {olhar, ser}, {tomar, c*}, {compra, carro}, {liberdade, expressão}, {existem, pessoas}, {ato, libidinoso} , {pessoa, tratar}, {querer, alguém}, {maltratar, animal}	{muita, gente}, {ler, escrever}, {ficar, sozinha}, {material, genético}, {possa, ajudar}, {mano, foda}, {carne, magra}, {diferente, alguém}, {auto, estima}, {hoje, dia}, {queria, alguém}, {tô, falando}
r/brasil	{liberdade, expressão}, {pé, chão}, {sonho, americano}, {virar, prof}, {sonhar, trabalho}, {hipócrita, dissimular}, {opinião, diferente}, {mulheres, pretos}	{muita, gente}, {hoje, dia}, {dia, seguinte}, {rock, rio}, {álcool, isopropílico}, {durar, anos}, {redes, sociais}, {segue, vida}, {gente, falando}, {ciro, gomes}, {vale, pena}, {nota, fiscal}
r/brasilivre	{existir, pessoa}, {ir, votar}, {sociedade, viver}, {tio, sam}, {mínimo, respeito}, {chupa, c*}, {pessoa, preferir}	{muita, gente}, {view, removal}, {removal, request}, {request, video}, {processo, eleitoral}, {juros, dividas}, {velho, testamento}, {esquerda, autoritária}, {pior, esquerda}, {opinião, pública}, {vamos, ignorar}, {carga, viral}
r/futebol	{tomar, c*}, {entender, futebol}, {comparação, sentido}, {ficar, calado}, {mano, neymar}, {andar, campo}	{time, jogando}, {deus, existe}, {funções, diferentes}, {flamengo, culpa}, (graças, deus}, {bola, parada}, {coréia, sul}, {fisica, nuclear}, {jogo, grêmio}, {viu, guerra}, {viu, surgimento}, {viu, queda}
r/investimentos	{operar, vender}, {ironia, lula}, {lula, vencedor}, {precificado, fato}, {verdade, operar}, {rir, toa}	{lógica, matemática}, {engenharias, vida}, {curso, voltado}, {atividades, extracurriculares}, {boas, indicações}, {manter, dinheiro}, {opções, risco}, {celebração, imigração}, {imigração, japonesa}, {clt, real}, {real, empresa}, {empresa, brasileira}
r/conversas	{mau, aluno}, {ir, fingir}, {postar, morte}, {morrer, depressivo}, {pick, boy}, {pick, m*rda}	{envie, mensagem}, {post, comentário}, {url, remoção}, {remoção, envie}, {mensagem, apelar}, {apelar, post}, {comentário, removido}, {removido, diretório}, {diretório, envie}, {pede, sair}

Tabela 15: Bigramas mais frequentes entre as principais comunidades brasileiras no Reddit. Os conjuntos de termos estão segmentados entre os termos associados com comentários classificados como *tóxicos* e *não-tóxicos*.

terações descritivas e informais. Termos como *melhor, vida, Brasil* também se destacam, sugerindo que a *não-toxicidade* está associada a conversas mais propositivas e positivas. Assim, grafos de coocorrência podem auxiliar na moderação automatizada de conteúdo tóxico, pois revelam, com uma modelagem simples, palavras que frenquentemente são utilizadas em conjunto nestes tipos de discussão.

4.3.6. Reconhecimento de entidades nomeadas (NER)

A Tabela 16 apresenta os resultados da análise de entidades nomeadas realizada em nosso conjunto de dados. A entidade nomeada mais comum em ambas as classes é PESSOA, representando mais de 31% de todos os tokens classificados em comentários tóxicos. A segunda entidade mencionada com mais frequência, LOCALIZAÇÃO, é igualmente predominante em ambas as classes. Embora tanto os comentários *tóxicos* quanto os *não-tóxicos* mencionem essas entidades, seu uso é diferente. Realizamos um teste de qui-quadrado para comparar a distribuição de etiquetas de POS para comentários em que pelo menos uma entidade nomeada é mencionada. O resultado indica uma diferença significativa na distribuição das etiquetas de POS (p-valor $< 0{,}01$). Os comentários *tóxicos* por exemplo, usam mais tokens etiquetados como VERB e NOUN. O comentário a seguir exemplifica entidades nomeadas sendo mencionadas na discussão dos usuários: *"Mais sério que esse tweet só a guerra na Ucrânia"*.

Conteúdo	**PER**	**ORG**	**LOC**	**MISC**
Não-tóxico	28,49%	20,26%	26,35%	24,88%
Tóxico	31,33%	16,23%	27,92%	24,5%

Tabela 16: Percentagem de menções a entidades dos tipos PESSOA (PER), ORGANIZAÇÃO (ORG), LOCALIZAÇÃO (LOC) e MIS (MISCELÂNEA).

É sabido que as redes sociais online são usadas como um meio de discutir eventos da vida real. Investigamos, também, se nossos dados revelam esse comportamento. Para tanto, mostramos a

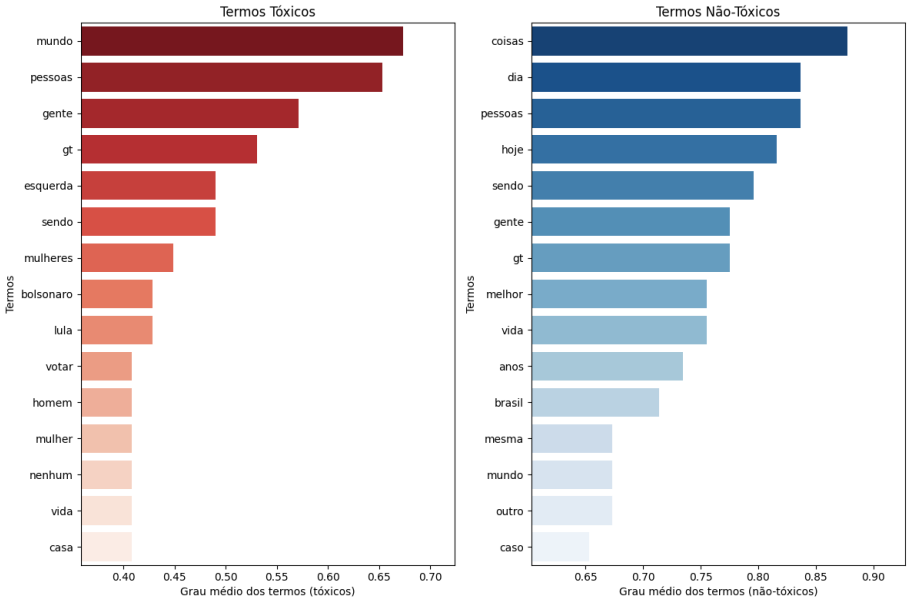

Figura 4: Top-15 termos *tóxicos* e *não-tóxicos* mais influentes extraídos a partir do grafo de coocorrência. No eixo-x, temos o grau médio normalizado dos termos associados com os comentários de cada classe.

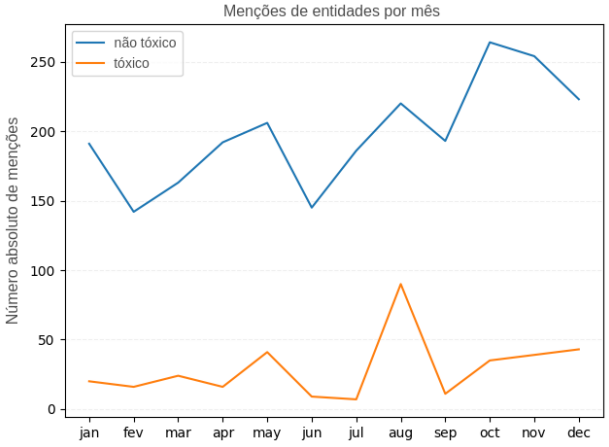

Figura 5: Série temporal mensal de menções de entidades nomeadas.

série temporal mensal dos números de citações de entidades na Figura 5.[6] Há picos significativos no volume de menções em agosto e outubro, que coincidem com o mês de abertura e os dois turnos das eleições brasileiras de 2022. Alguns comentários rotulados como *tóxicos* mencionaram os candidatos presidenciais: *"Vocês são demasiadamente burros! Esse idiota do Bolsonaro pode até "dar um golpe", eu quero ver sustentar esse ato infame, pois, vejamos na década de 60, por exemplo, o Brasil teve essa porcaria de intervenção graças ao apoio do Tio Sam. [..]"*, *"O Lula não vai conseguir ver, pois ele está morto"*.

[6]MISCELÂNEA foi excluída.

4.4. Principais achados

A seguir, resumimos os principais achados do nosso estudo.

4.4.1. Qualidade da anotação

Avaliamos a qualidade do conjunto de dados calculando a concordância entre anotadores, com resultados que corroboram trabalhos semelhantes (Google, 2022b). No entanto, dividimos os anotadores em grupos e nossos resultados mostram que alguns grupos são mais sensíveis a comentários de toxicidade e também apresentam diferentes níveis de qualidade. A forte concordância entre os anotadores do grupo C aponta suas anotações como uma amostra de ouro para avaliar técnicas distintas para o ajuste fino de modelos de aprendizado de máquina de detecção de toxicidade em textos em português brasileiro. Por fim, destacamos a importância da condução de um processo de anotação de dados bem estruturado para a criação de conjuntos de dados de qualidade para o idioma.

4.4.2. Concordância com modelos de aprendizado

Nossa comparação da anotação manual com as classificações da API Perspective mostra que alguns anotadores anotam menor quantidade de comentários tóxicos, enquanto outros são mais sensíveis ao conteúdo *tóxico* gerado. Em geral, a porcentagem média de comentários *tóxicos*

é próxima daquela classificada pela API (entre 10% a 11%). No entanto, a API Perspective é mais sensível a palavrões e não tem o contexto dos tópicos que estão sendo discutidos. Além disso, a API não consegue detectar tipos muito específicos e nuances de ataques direcionados em português (por exemplo, quando grupos específicos são alvo de ofensas na forma de sarcasmo ou ironia). O modelo de linguagem Sabia-2 teve maior correlação com as anotações manuais no geral e também alcançou maior compromisso entre falsos positivos e falsos negativos. O Sabia-2 classificou um percentual de 10,83% de comentários tóxicos —similar ao percentual de comentários tóxicos classificados pelos anotadores—, enquanto o ChatGPT classificou 28,46% dos comentários como *tóxico*. Esse resultado indica um possível ganho de fazer o processo de ajuste fino (*fine-tuning*) em modelos de linguagem usando dados do idioma Português.

4.4.3. Caracterização da linguagem

Os comentários *tóxicos* são, em média, mais longos. Embora tenham uma proporção de etiquetas de POS semelhante à dos *não-tóxicos*, os substantivos e adjetivos mais frequentes apresentam diferenças. Uma clara tendência de aumento nas menções de entidades nomeadas nos subreddits ao longo dos meses, especialmente próximo do período eleitoral brasileiro, mostra o impacto de eventos externos nas interações dos usuários. Isso deve ser considerado ao usar esse conjunto de dados para classificação de textos e criação de modelos, pois o modelo resultante pode ser muito sensível à janela de tempo de dados disponível. Na modelagem de tópicos, identificamos que tópicos em que tivemos discordância total abrangem assuntos relacionados à política, ataques a grupos minorizados e palavrões. Para tópicos em que os comentários foram classificados como *tóxicos* por todos os anotadores, temos a ocorrência de termos relacionados a ideologia, guerra e palavrões. Na análise de bi-gramas, vimos que os termos mais relevantes variam a depender da comunidade (*subreddit*). Entretanto, alguns termos relacionados à política são comuns a todas as comunidades, reforçando a influência dos eventos externos a discussões online. Por fim, a análise de termos influentes usando grafos de coocorrências revelou um padrão similar à análise de bi-gramas, com termos relacionados à política, ideologia e sexo sendo mais influentes no grafo com comentários *tóxicos*, enquanto termos mais genéricos como *Brasil* e *mundo* são mais influentes para comentários classificados como *não-tóxicos*.

Nossos resultados atestam o potencial do nosso conjunto de dados para o ajuste fino de um modelo de aprendizado de máquina em uma tarefa posterior. A alta concordância observada entre os anotadores atesta a consistência dos rótulos. Com os nossos dados, pretendemos fornecer exemplos mais diversificados de textos *tóxicos* de interações de redes sociais online para incentivar o desenvolvimento de modelos de aprendizado de máquina mais robustos, capazes de atenuar comportamentos ofensivos online.

4.4.4. Limitações

Com relação às limitações de nosso estudo, reconhecemos o desafio inerente e a subjetividade da tarefa de rotular conteúdo *tóxico* em um ambiente contextualmente limitado de redes sociais online. Para atenuar esse problema, planejamos repetir o experimento de rotulagem, fornecendo especificamente informações adicionais de contexto para comentários com contexto local ou limitado. Além disso, cabe observar que nosso procedimento de amostragem pode apresentar um viés em relação a tópicos externos específicos que tiveram importância significativa tanto local quanto globalmente durante o período de coleta de dados. Por fim, outro aspecto a se considerar é a diversidade do perfil dos anotadores que participaram do processo de anotação do conjunto de dados. Em experimentos futuros, buscaremos mapear os perfis para garantir maior representatividade dos participantes.

5. Conclusão

Embora os modelos de aprendizado de máquina tenham sido implementados com sucesso como ferramentas de moderação automática para alguns idiomas, ainda não temos suporte para idiomas com poucos recursos, como o português brasileiro. Nosso artigo apresenta um novo conjunto de dados com anotações manuais de comentários tóxicos nas interações de usuários do Reddit dos dez maiores subreddits do Brasil. Nossos resultados indicam uma concordância substancial entre os anotadores e um forte alinhamento com modelos externos pré-treinados para o português, o que apoia a utilização desses dados para tarefas posteriores de aprendizado de máquina.

Em trabalhos futuros, pretendemos integrar esse novo conjunto de dados a modelos de aprendizado de máquina pré-treinados para fornecer ao modelo dados de interações reais em redes sociais. Além disso, pretendemos aproveitar esse conjunto de dados para tarefas mais complexas,

como a detecção de gatilhos de toxicidade em conversas online, a fim de sermos proativos em intervenções de moderação. Por fim, pretendemos realizar uma análise de polaridade para contrastar a caracterização e os resultados obtidos na tarefa de toxicidade.

Agradecimentos

Esta pesquisa foi parcialmente apoiada pelas agências de pesquisa brasileiras CNPq, FAPEMIG e CAPES.

Referências

Almeida, Thales Sales, Hugo Abonizio, Rodrigo Nogueira & Ramon Pires. 2024. Sabiá-2: A new generation of Portuguese large language models. ArXiv:2403.09887 [cs.CL]. 10.48550/arXiv.2403.09887

Almerekhi, Hind, Haewoon Kwak, Bernard J Jansen & Joni Salminen. 2019. Detecting toxicity triggers in online discussions. Em *30th ACM Conference on Hypertext and Social Media*, 291–292. 10.1145/3342220.3344933

Almerekhi, Hind, Haewoon Kwak, Joni Salminen & Bernard J Jansen. 2020. Are these comments triggering? predicting triggers of toxicity in online discussions. Em *Proceedings of The Web Conference (WWW)*, 3033–3040. 10.1145/3366423.3380074

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire & Jeremy Blackburn. 2020. The pushshift Reddit dataset. Em *International AAAI Conference on Web and Social Media*, vol. 14, 830–839. 10.1609/icwsm.v14i1.7347

Benesty, Jacob, Jingdong Chen, Yiteng Huang & Israel Cohen. 2009. Pearson correlation coefficient. Em *Noise reduction in speech processing*, 1–4. Springer Berlin Heidelberg. 10.1007/978-3-642-00296-0_5

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. Em *Neural Information Processing Systems*, 1877–1901.

ElSherief, Mai, Vivek Kulkarni, Dana Nguyen, William Yang Wang & Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. Em *International AAAI Conference on Web and Social Media*, vol. 12 1, 10.1609/icwsm.v12i1.15041

Fortuna, Paula, João Rocha da Silva, Leo Wanner & Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. Em *3rd Workshop on Abusive Language Online*, 94–104. 10.18653/v1/W19-3510

Gillespie, Tarleton. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7(2). 10.1177/2053951720943234

Google. 2022a. Perspective API model cards. Acessado em: 10/12/2023.

Google. 2022b. Perspective API training data. Acessado em: 08/04/2023.

Grootendorst, Maarten. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv [cs.CL]. 10.48550/arXiv.2203.05794

Jiang, Jialun Aaron, Morgan Klaus Scheuerman, Casey Fiesler & Jed R Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PloS One* 16(8). e0256762. 10.1371/journal.pone.0256762

Kobellarz, Jordan & Thiago Silva. 2022. Should we translate? evaluating toxicity in online comments when translating from Portuguese to English. Em *28th Simpósio Brasileiro de Sistemas Multimídia e Web*, 95–104.

Kukreja, Sanjay, Tarun Kumar, Amit Purohit, Abhijit Dasgupta & Debashis Guha. 2024. A literature survey on open source large language models. Em *7th International Conference on Computers in Management and Business*, 133–143. 10.1145/3647782.3647803

Leite, João Augusto, Diego Silva, Kalina Bontcheva & Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. Em *1st Conference of the Asia-Pacific and the 10th International Joint Conference on Natural Language Processing*, 914–924. 10.18653/v1/2020.aacl-main.91

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy & James R Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194. 151–175. 10.1016/j.artint.2012.03.006

Oliveira, Amanda S, Thiago C Cecote, Pedro HL Silva, Jadson C Gertrudes, Vander LS Freitas & Eduardo JS Luz. 2023. How good is ChatGPT for detecting hate speech in Portuguese? Em *14th Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 94–103. 10.5753/stil.2023.233943

de Pelle, Rogers Prates & Viviane P Moreira. 2017. Offensive comments in the Brazilian web: a dataset and baseline results. Em *6th Brazilian Workshop on Social Network Analysis and Mining*, 510–519. 10.5753/brasnam.2017.3260

Petrov, Slav, Dipanjan Das & Ryan McDonald. 2011. A universal part-of-speech tagset. ArXiv [cs.CL]. 10.48550/arXiv.1104.2086

Rademaker, Alexandre, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick & Valeria de Paiva. 2017. Universal dependencies for Portuguese. Em *4th International Conference on Dependency Linguistics (DepLing)*, 197–206.

Salehabadi, Nazanin, Anne Groggel, Mohit Singhal, Sayak Saha Roy & Shirin Nilizadeh. 2022. User engagement and the toxicity of Tweets. ArXiv [cs.SI/cs.CY]. 10.48550/arXiv.2211.03856

spaCy. 2022. Portuguese models. Acessado em: 11/04/2023.

Statista. 2022. Number of social media users worldwide from 2017 to 2027. Acessado em: 08/04/2023.

Trajano, Douglas, Rafael Bordini & Renata Vieira. 2023. OLID-BR: offensive language identification dataset for Brazilian Portuguese. *Lang Resources & Evaluation* 58(4). 1263–1289. 10.1007/s10579-023-09657-0

US Congress. 1998. Digital millennium copyright act. *Public Law* 105(304). 112

Vargas, Francielle, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo & Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. Em *13th Language Resources and Evaluation Conference (LREC)*, 7174–7183.

Vogels, Emily A. 2021. The state of online harassment. Relatório técnico. Pew Research Center.

Wise, Jason. 2023. Reddit users: how many people use Reddit in 2023? Acessado em: 08/04/2023.

Zannettou, Savvas, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh & Gianluca Stringhini. 2020. Measuring and characterizing hate speech on news websites. Em *12th ACM Conference on Web Science*, 125–134. 10.1145/3394231.3397902