UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Luiz Henrique Quevedo Lima

Identificação e Caracterização de Conteúdo Tóxico de Usuários em Comunidades Brasileiras no Reddit

Belo Horizonte
2025

Luiz Henrique Quevedo Lima

Identificação e Caracterização de Conteúdo Tóxico de Usuários em Comunidades Brasileiras no Reddit

Versão Final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientadora: Ana Paula Couto da Silva

Belo Horizonte
2025

Lima, Luiz Henrique Quevedo.

L732i Identificação e caracterização de conteúdo tóxico de usuários em comunidades brasileiras no Reddit [recurso eletrônico] / Luiz Henrique Quevedo Lima – 2025.

1 recurso online (84 f. il.) : pdf.

Orientadora: Ana Paula Couto da Silva.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.

Referências: f. 74-84.

1. Computação - Teses. 2. Redes sociais on-line - Teses. I. Silva, Ana Paula Couto da. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. III. Título.

CDU 519.6*75 (043)

Ficha catalográfica elaborada por Célio Resende Diniz, bibliotecário CRB 6/2403 - Universidade Federal de Minas Gerais – ICEx.

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

IDENTIFICAÇÃO E CARACTERIZAÇÃO DE CONTEÚDO TÓXICO DE USUÁRIOS EM COMUNIDADES BRASILEIRAS NO REDDIT

LUIZ HENRIQUE QUEVEDO LIMA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores(a):

Profa. Ana Paula Couto da Silva - Orientadora
Departamento de Ciência da Computação - UFMG

Profa. Mirella Moura Moro
Departamento de Ciência da Computação - UFMG

Profa. Aline Marins Paes Carvalho
Instituto de Computação - UFF

Prof. Evandro Landulfo Teixeira Paradela Cunha
Faculdade de Letras - UFMG

Belo Horizonte, 16 de maio de 2025.

Documento assinado eletronicamente por **Ana Paula Couto da Silva**, **Professora do**

Magistério Superior, em 15/07/2025, às 17:25, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Evandro Landulfo Teixeira Paradela Cunha**, **Professor do Magistério Superior**, em 15/07/2025, às 23:52, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Aline Marins Paes Carvalho**, **Usuária Externa**, em 17/07/2025, às 17:10, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Mirella Moura Moro**, **Professora do Magistério Superior**, em 17/07/2025, às 17:16, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4386637** e o código CRC **B8A2263C**.

Referência: Processo nº 23072.243758/2025-59 SEI nº 4386637

Agradecimentos

Primeiramente, agradeço à minha família por todo o suporte que me deram ao longo da minha jornada acadêmica. À minha esposa, por ter sido paciente e compreensiva durante todo o processo e por ter me auxiliado durante o programa, permitindo que eu pudesse focar na minha pesquisa.

Agradeço à minha orientadora, Ana Paula Couto, por todos os ensinamentos transmitidos durante o programa de mestrado e pela paciência ao revisar meu trabalho inúmeras vezes, além de estar sempre presente e me incentivar a continuar. Estendo o agradecimento à professora Adriana Pagano, que nos acompanhou durante grande parte do processo e foi essencial ao nos conectar com alunos da Faculdade de Letras (UFMG).

Aos professores que marcaram minha jornada acadêmica no DCC. Algumas disciplinas foram essenciais para a minha formação e contribuíram imensamente para esta pesquisa: Aprendizado de Máquina (Adriano Veloso), Mineração de Dados e Interpretabilidade (Wagner Meira) e Métodos Quantitativos (Jussara Almeida), para citar apenas algumas. Com toda certeza, levarei os ensinamentos obtidos no DCC para toda a minha carreira.

Agradeço à UFMG por me permitir desfrutar de conhecimento de ponta, me conectar com outros estudantes incríveis de todas as partes do Brasil e estudar o estado da arte em Computação.

Por fim, agradeço ao financiamento parcial das agências de fomento CNPq, FAPEMIG e CAPES, por terem contribuído com recursos para que esta pesquisa se tornasse realidade.

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. Primeiramente, tratamos essa lacuna criando um conjunto de dados de alta qualidade, coletado de algumas das comunidades brasileiras mais populares da plataforma Reddit. A partir desse conjunto de dados, propomos o uso de modelos de Aprendizado de Máquina, abertos e fechados, treinados para a tarefa de classificação de toxicidade. Por fim, exploramos o uso de grandes modelos de linguagem (LLMs) para assistir no processo de geração de dados sintéticos de qualidade. Nossos principais achados mostram que comentários tóxicos em comunidades brasileiras apresentam padrões linguísticos distintos, e que modelos de linguagem pré-treinados (como LLMs e Transformers) são fundamentais para a classificação automática eficiente e escalável. Esses resultados destacam a importância de incorporar conhecimento externo dos modelos pré-treinados para serem aplicados em tarefas específicas. Com essa pesquisa, buscamos contribuir com a importante tarefa de moderação automática nas redes sociais, promovendo um ambiente online mais seguro e inclusivo para todos.

Palavras-chave: Processamento de Linguagem Natural, Toxicidade, Conjunto de Dados, Redes Sociais Online, Reddit

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. Using manually labeled data, we propose the use of both open and closed machine learning models trained for the task of toxicity classification. We also explore the use of Large Language Models (LLMs) to assist in generating high-quality synthetic data. Our main findings show that toxic comments in Brazilian communities exhibit distinct linguistic patterns, and that pre-trained language models (such as LLMs and Transformers) are essential for scalable and effective automated toxicity detection. These results highlight the importance of leveraging external knowledge learned by pre-trained models for application in specific tasks. With this research, we aim to contribute to the critical task of automated moderation on online social media, promoting a safer and more inclusive online environment for all.

Keywords: Natural Language Processing, Toxicity, Datasets, Online Social Networks, Reddit

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 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 sofreram alguma forma de assédio online. Além disso, comentários abusivos em discussões propagam a toxicidade, levando à radicalização das discussões Salehabadi et al. (2022). As consequências dessas interações transcendem o mundo virtual, afetando seriamente a vida dos usuários no mundo real. Ainda segundo Vogels (2021), 18% dos usuários que participaram de uma enquete sofreram algum tipo de abuso, considerado grave, iniciado no 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 considerado nocivo. Embora os modelos de aprendizado de máquina tenham se mostrado eficazes em vários idiomas Perspective (2022), seu desempenho em línguas que possuem menos recursos, como o português brasileiro, ainda é insuficiente.

Face a esses desafios, esta dissertação tem como principal objetivo a detecção de toxicidade em conteúdo de redes sociais online escrito em português brasileiro. Neste trabalho, similar a Perspective (2022), definimos comentário tóxico como um *comentário rude, desrespeitoso ou insensível que provavelmente fará alguém abandonar uma discussão.* Os dados analisados foram extraídos de uma das maiores redes sociais online - 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 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 (Twitter) e Instagram Boulianne et al. (2024). 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.

1.1 Motivações e desafios

Muitos trabalhos na literatura que abordam o problema de discurso tóxico no Reddit focam em idiomas que possuem muitos recursos disponíveis para detecção automática de toxicidade, como a disponibilidade de vastos conjuntos de dados anotados e modelos linguísticos, além de analisarem comunidades de temas específicos Chong and Kwak (2022); Almerekhi et al. (2022b, 2019, 2022a); Görzig and Ólafsson (2013). Várias abordagens são exploradas para detecção de toxicidade. Em Chong and Kwak (2022); Almerekhi et al. (2022b, 2019), os autores usam a Perspective API para identificar comentários tóxicos no Reddit, enquanto em Guimarães et al., a API é utilizada para classificar conteúdo gerado no Facebook. O trabalho em Almerekhi et al. (2022a) conduz um experimento com mais de 2 bilhões de comentários e publicações a partir de um modelo treinado com dados anotados. Alguns trabalhos propõem a utilização de técnicas de aprendizado de máquina e aprendizado profundo para detecção automática de discurso tóxico Davidson et al. (2017); Chakrabarty (2020); d'Sa et al. (2020), enquanto ElSherief et al. (2018) apresenta uma análise do discurso de ódio generalizado e direcionado a terceiros.

Outras abordagens também são estudadas na literatura para detecção de toxicidade. O trabalho em Machová et al. (2022) utiliza um léxico linguístico para detectar toxicidade em redes sociais para o idioma eslovaco, enquanto em Larochelle and Khoury (2020), múltiplos conjuntos de dados anotados de comentários abusivos são utilizados para avaliar a generalização do modelo treinado para identificar casos de *cyberbullying*. Para o contexto brasileiro, modelos pré-treinados e comerciais podem ser usados em conjunto com características textuais para detecção automática de toxicidade.

Embora alguns trabalhos abordem a criação de conjunto de dados de comentários

tóxicos e detecção de toxicidade para o Português Leite et al. (2020); Fortuna et al. (2019); de Pelle and Moreira (2017), até o momento, não foram encontrados trabalhos na literatura que abordem a detecção de toxicidade no contexto das comunidades brasileiras no Reddit. Além disso, existe o desafio de utilizar modelos pré-treinados que possuem limitações para a língua portuguesa. Por esse motivo, o trabalho proposto tem o objetivo de detectar comentários tóxicos para a língua portuguesa, em um primeiro momento utilizando a API Perspective para caracterizar os comentários das maiores comunidades brasileiras em número de usuários ativos. Posteriormente, conduzimos um estudo comparativo envolvendo diversos modelos de aprendizado de máquina, tanto de código aberto quanto comerciais, que abrangeu desde abordagens lineares até as mais avançadas LLMs. Além disso, investigamos técnicas específicas para mitigar o desbalanceamento das classes presentes no conjunto de dados original. Por fim, validamos a eficácia dessa classificação e propomos a implementação de um modelo adaptado para o contexto das discussões online em Português a partir de um conjunto de dados manualmente anotado com dados do Português Brasileiro.

1.2 Objetivos

O principal objetivo dessa dissertação é analisar e caracterizar o conteúdo tóxico nas comunidades brasileiras no Reddit, juntamente com a proposta de modelos de classificação automática de toxicidade para conteúdo gerado nesta rede. Para alcançar este objetivo, este trabalho é estruturado em duas questões de pesquisa (QPs):

QP1: *É possível distinguir conteúdo tóxico e não-tóxico a partir de diferentes características linguísticas encontradas em seus textos?* Para investigar padrões na linguagem dos comentários de conteúdo tóxico anotado manualmente por 12 voluntários, selecionamos as seguintes análises linguísticas: análise de distribuição do tamanho dos comentários tóxicos e não tóxicos e a razão *type-token*, análise de classe de palavras (*POS tagging*) Petrov et al. (2012); Rademaker et al. (2017), análise de tópicos Grootendorst (2022) e entidades nomeadas Nothman et al. (2013) e análise de termos frequentes usando n-gramas e grafo de coocorrência. Essas técnicas foram utilizadas para fazer a caracterização do estilo linguístico de comentários considerados tóxicos comparados com comentários não tóxicos.

QP2: *Quais são os principais aspectos que devem ser considerados para a proposta de modelos acurados para a classificação automática de conteúdo tóxico no Reddit, considerando a língua portuguesa? Quais classes de modelos de classificação são as mais adequadas para esta tarefa de aprendizado?* Para treinar modelos de classificação para

a tarefa de detecção de toxicidade, comparamos diversas abordagens: modelos lineares, variantes de modelos transformers de código aberto — como BERT e RoBERTa Kenton and Toutanova (2019); Souza et al. (2020); Zhuang et al. (2021) — e grandes modelos de linguagem (LLMs). Para tratar o desbalanceamento de classes do conjunto de dados original, utilizamos técnicas clássicas de reamostragem dos dados, bem como a utilização LLMs de código aberto para rotular sinteticamente novos exemplos de comentários com o objetivo de aumentar a proporção da classe minoritária. Por fim, investigamos o impacto da representação dos dados no desempenho do modelo, comparando a eficiência de abordagens que utilizam representações esparsas (como TF-IDF) com aquelas baseadas em representações vetoriais, obtidas por meio de word embeddings.

1.3 Contribuições

As principais contribuições desta dissertação são detalhadas a seguir:

- **Disponibilização de dados anotados manualmente com conteúdo tóxico:** A coleta dos dados seguiu uma metodologia para selecionar as maiores comunidades brasileiras no Reddit, além de pré-processamento dos dados para reduzir ruídos e amostragem do conjunto para posterior rotulação manual de toxicidade. O conjunto de dados original possui 6.589.541 comentários e 390.924 postagens das 10 maiores comunidades brasileiras no Reddit. Para fomentar a pesquisa no idioma português, este trabalho disponibiliza o conjunto de dados com 2.500 amostras de comentários anotados manualmente por 12 anotadores voluntários, estudantes de graduação e pós-graduação dos cursos de Letras e Ciência da Computação da UFMG.[1]

- **Caracterização linguística das discussões nas comunidades analisadas:** Nesse estudo, buscamos entender como as diferentes características linguísticas variam a depender da toxicidade observada dos comentários. As análises linguísticas realizadas para caracterização do conteúdo tóxico foram executadas a partir do conjunto de dados anotado manualmente pelos anotadores voluntários. Além disso, realizamos um estudo de tópicos para identificar os principais assuntos discutidos nas comunidades e os termos mais frequentes associados a cada categoria de toxicidade. Para extrair os tópicos, utilizamos o modelo pré-treinado BERTopic.

- **Definição e comparação entre diferentes modelos de aprendizado para classificação de toxicidade:** Novos modelos de aprendizado de máquina foram

[1] https://github.com/luizhenriqueds/reddit-br-toxicity-dataset/

treinados a partir do conjunto de dados anotado. Para lidar com o problema de desbalanceamento dos dados desta tarefa, propomos um *framework* para fazer a anotação de uma nova amostra de dados usando modelos de linguagem abertos (*open-source*). Nossos resultados indicam que o desbalanceamento dos dados do conjunto de dados original é o principal fator que afeta o desempenho dos modelos. Além disso, a representação textual dos dados fornecidos ao modelo possui papel importante e ajuda os modelos a classificar corretamente os comentários. Em nossos experimentos, vimos que ao combinar técnicas para rotular sinteticamente novos comentários em conjunto com a representação vetorial (word embeddings), mesmo a Regressão Logística é capaz de alcançar desempenho similar aos LLMs comerciais estados da arte. Por fim, propomos uma metodologia que integra diversos modelos e técnicas, permitindo o treinamento de modelos escaláveis e eficientes, possibilitando o processamento de grandes volumes de dados com um custo de predição reduzido.

As principais contribuições desta dissertação estão publicadas em Lima et al. (2024a,b). Como contribuição decorrente dos dados coletados e da metodologia para caracterização linguística proposta nesta dissertação, o trabalho publicado em Piorino et al. (2024) aborda a tarefa de análise de sentimentos no mesmo contexto (Reddit) desta dissertação. Entre as limitações deste estudo, destacam-se a subjetividade inerente à anotação de toxicidade e os possíveis vieses associados ao período de coleta dos dados (2022).

1.4 Organização

Esta dissertação está organizada da seguinte forma. No Capítulo 2, apresentamos uma síntese da literatura disponível sobre detecção de toxicidade em português e em outras línguas. Em seguida, o Capítulo 3 descreve detalhadamente nossa metodologia de coleta e anotação dos dados, as técnicas empregadas para a caracterização linguística dos comentários e os métodos utilizados no treinamento dos modelos de detecção automática. O Capítulo 4 explora a caracterização linguística do conteúdo tóxico, demonstrando características que auxiliam na identificação de conteúdo nocivo na plataforma. Na sequência, o Capítulo 5 apresenta o treinamento de diferentes categorias de modelos de aprendizado de máquina. Além disso, exploramos técnicas para lidar com o desbalanceamento de classes e representação textual, evidenciando como essas abordagens influenciam no desempenho final do modelo. Por fim, o Capítulo 6 conclui o estudo, apresentando os principais achados e sugerindo possíveis direções para pesquisas futuras.

Capítulo 2

Trabalhos Relacionados

Este capítulo apresenta os trabalhos relacionados a esta dissertação de mestrado. Na Seção 2.1, são abordados estudos sobre a detecção de toxicidade em idiomas e línguas sub-representadas, como eslovaco e nepali. A Seção 2.2 discute pesquisas focadas na detecção de toxicidade na língua portuguesa. A Seção 2.3 discute trabalhos na literatura que exploram o uso de LLMs para tarefas de classificação e anotação sintética de dados. Por fim, a Seção 2.4 discute o enquadramento do trabalho proposto no estado da arte.

2.1 Detecção de Toxicidade em Outras Línguas

Em Machová et al. (2022), os autores propõem a detecção de níveis de toxicidade de comentários de redes sociais para a língua eslovaca. Por existirem poucos recursos para este idioma, foi necessário construir um conjunto de dados para condução dos experimentos. Os dados foram coletados da rede social Facebook através de uma extração manual usando uma ferramenta de coleta de dados e também a partir de técnicas de raspagem de dados (*web scraping*). A anotação do conjunto de dados foi feita de forma automática usando um léxico de termos tóxicos para a linguagem eslovaca. Este léxico [1] foi construído a partir de uma tradução de termos para o Inglês juntamente com a inclusão de termos específicos, similar ao disponível em Tuckwood (2017). Cada termo no léxico foi categorizado em "pouco tóxico", "moderadamente tóxico" e "muito tóxico" e seus valores foram codificados de acordo com o grau de toxicidade (1 = leve, 2 = moderado, 3 = alto). Ao todo, 809 termos formam o léxico, sendo 224 termos nível 1, 243 nível 2 e 342 nível 3. Ao todo, 3092 exemplos foram anotados para construir o conjunto de dados.

Para anotar o conjunto de dados automaticamente, duas abordagens foram utilizadas. Primeiro, para cada comentário, foram identificados os termos tóxicos da mesma, realizando a soma dos termos encontrados. A segunda abordagem identificou o termo mais tóxico de cada frase e atribuiu esse valor como toxicidade do texto. A partir dessa

[1] https://kristina.machova.website.tuke.sk/useful/Lexicon_of_toxic_words.json

estratégia, comentários com o valor 0 foram classificados como “neutros”, com o valor 1 classificados como “levemente tóxicos”, comentários com o valor 2 “moderadamente tóxicos” e, por fim, comentários com valor 3 ou mais foram classificados como “altamente tóxicos”. As principais contribuições deste artigo foram propor uma abordagem baseada em léxico para a língua eslovaca e o processo para anotação automática de um novo conjunto de dados de comentários.

Em Wich et al. (2022), os autores propõem um framework para identificar canais (grupos) que apresentam linguagem abusiva no Telegram para a língua alemã. Entretanto, os dados usados para treinar os modelos de aprendizado de máquina foram coletados de outras plataformas, como o Twitter. O estudo investiga se é possível utilizar dados de outras fontes para classificar mensagens específicas do Telegram. Os autores definem o termo linguagem abusiva como “qualquer forma de insulto, abuso, ódio, degradação, ataque de identidade ou ameaça de violência direcionada a um indivíduo ou grupo”. Os dados foram coletados de diversos *datasets* de linguagem abusiva em alemão (especialmente Twitter) e combinados para fazer a classificação de canais com linguagem abusiva no Telegram. O modelo com melhor performance foi uma combinação de modelos treinados em datasets individuais e aplicando a técnica de voto majoritário, tendo resultado levemente superior ao da API Perspective.

Em Singh et al. (2020), os autores apresentam um estudo para detecção de sentimento abusivo baseado em aspecto para a língua nepali. O conjunto de dados consiste em 3.068 exemplos extraídos de comentários de vídeos populares no Youtube. O dataset foi anotado com a categoria (aspecto), como “Violência” e “Sarcasmo” e o alvo “Pessoa” ou “Organização”, por exemplo. O objetivo deste trabalho é classificar o tipo de categoria de comentário abusivo e a qual entidade nomeada o conteúdo é direcionado. Além disso, o estudo detalha a criação de um novo conjunto de dados para classificação de detecção de sentimento abusivo baseado em aspecto para a língua nepali. Como contribuição deste estudo, além da criação do conjunto de dados anotado, os autores sugerem a implementação de modelos para identificar os aspectos envolvidos em um comentário com sentimento abusivo.

O trabalho em Lobo et al. (2022) explora a utilização de grafos de conhecimento para analisar possíveis vieses de anotação em conjuntos de dados de conteúdo tóxico. O conjunto de dados em questão utilizado é o Jigsaw Toxicity 448k, que inclui anotações com atributos demográficos específicos identificando se um texto se refere a um indivíduo ou grupo que possui essas características. Para suportar a análise com conhecimento de domínio, foi utilizada a ontologia GSSO, que possui definições semânticas de cerca de 14 mil termos associados a características protegidas como raça, deficiência e religião. Os resultados mostraram que cerca de 3% em uma amostra de 19 mil comentários mencionam termos associados a grupos de gênero e orientação sexual frequentemente atacados e que não foram corretamente identificados pelos anotadores do dataset original. O estudo tem

como objetivo levantar a discussão sobre a qualidade dos conjuntos de dados anotados manualmente e como o desconhecimento de termos específicos do discurso pode prejudicar a qualidade de um conjunto de dados.

2.2 Detecção de Toxicidade em Língua Portuguesa

Estudos sobre a detecção automática de comentários tóxicos em línguas como o português brasileiro são menos frequentes, assim como conjuntos de dados anotados manualmente e de livre acesso para uso público e estudos de acompanhamento.

Os autores em de Pelle and Moreira (2017) disponibilizam um conjunto de dados com 1,250 comentários, extraídos de sessões de comentários do site 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.

Em Fortuna et al. (2019), os autores 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 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.

Os autores em 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 em Kobellarz and 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 quanto em 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 em 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*).

Os autores em Trajano et al. (2023) apresentam o 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.

2.3 Uso de LLMs em tarefas de classificação

O trabalho em 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.

O trabalho em Belal et al. (2023) explora o uso do modelo de linguagem ChatGPT como uma ferramenta para a tarefa de análise de sentimento. Os autores compararam o modelo com abordagens baseadas em léxico, como VADER e TextBlob, em dois conjuntos de dados distintos: um composto por tweets sobre partidas de futebol e outro com avaliações de produtos da Amazon. Os resultados mostraram um ganho significativo em relação aos métodos tradicionais, com algumas descobertas-chave para o uso do GPT: (1) desempenho impressionante do ChatGPT mesmo em configurações zero-shot, (2) capacidade de lidar com emojis e sarcasmo e (3) habilidade para processar frases longas em avaliações.

Em Gilardi et al. (2023), os autores compararam o desempenho do ChatGPT com o de anotadores de crowdsourcing na tarefa de anotação de dados em quatro conjuntos de dados distintos. Os resultados indicam que o desempenho e a concordância da anotação sintética de dados superaram os dos anotadores manuais [2] em cerca de 25% em média,

[2] O processo de anotação manual foi realizado utilizando a ferramenta Amazon Mechanical Turk:

além de reduzir os custos operacionais.

Os autores em Mishra and Chatterjee (2024) investigam o uso de LLMs para a classificação de comentários tóxicos em discussões de *issues* no GitHub. O conjunto de dados utilizado é composto por *issues* fechadas, previamente curadas para identificar comportamentos tóxicos. O estudo emprega o ChatGPT, explorando previsões zero-shot, variações de parâmetros e técnicas de *prompt engineering*. Os resultados indicam um desempenho promissor, especialmente considerando que o modelo não foi treinado especificamente para identificar toxicidade em dados do Github.

2.4 Diferencial do Trabalho

Conforme discutido nesse capítulo, existem alguns trabalhos que lidam com o comportamento tóxico e ofensivo em redes sociais, direcionados para a língua portuguesa. Estes trabalhos se dividem em construir conjuntos de dados cuidadosamente curados e, posteriormente, suas aplicações em modelos de aprendizado de máquina para a tarefa de detecção de toxicidade.

Assim, a construção de um conjunto de dados de qualidade é etapa essencial para auxiliar na detecção automática de interações tóxicas nesse ambiente. Embora existam alguns trabalhos na literatura com foco na língua portuguesa, ainda há uma carência de dados rotulados manualmente para a rede social Reddit, que possui características que diferem de outras redes sociais online, tais como interações de usuários através de comunidades, anonimidade e, principalmente, quantidade ilimitada de caracteres para os comentários compartilhados. E, como consequência, modelos de aprendizado com boa acurácia para a tarefa de aprendizado de interesse.

A Tabela 2.1 apresenta uma visão comparativa de estudos de detecção de toxicidade encontrados na literatura e os resultados apresentados nesta dissertação. Resumidamente, 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. Além disso, investigamos diferentes modelos de aprendizado de máquina para a detecção de toxicidade, treinados com o conjunto de dados apresentado neste estudo, juntamente com uma metodologia para a geração de anotações sintéticas. Os modelos analisados apresentam diferentes níveis de complexidade e podem ser selecionados com base nos requisitos específicos do

https://www.mturk.com/.

Trabalho	Dados	# exemplos	# classes	Anotadores	Especialistas
de Pelle and Moreira (2017)	Comentários de notícias	1.250	2	*Crowdsourcing*	✗
Fortuna et al. (2019)	Twitter	5.668	2+	Especialistas e não especialistas	✓
Leite et al. (2020)	Twitter	21.000	6	Estudantes voluntários	✗
Vargas et al. (2022)	Instagram	7.000	2+	Não informado	✓
Trajano et al. (2023)	YouTube e Twitter	7.943	5	Anotadores contratados	✗
Nosso trabalho Lima et al. (2024a,b)	Reddit	2.500	2	Estudantes voluntários	✗

Tabela 2.1: Resumo das características dos trabalhos relacionados à construção de conjuntos de dados e detecção de toxicidade para o Português.

problema [3].

[3]Os modelos treinados para este estudo serão disponibilizados em futuras publicações.

Capítulo 3

Metodologia

Neste capítulo, descrevemos os métodos que serão aplicados para responder às Questões de Pesquisa (QPs) abordadas nesta dissertação. O capítulo está organizado da seguinte forma: primeiramente, apresentamos uma visão geral da metodologia proposta neste trabalho (Seção 3.1). Na sequência, a Seção 3.2 detalha o conjunto de dados utilizado, abordando desde os critérios para seleção das comunidades de interesse até a coleta, processamento dos dados e geração da amostra para o processo de anotação. A seguir, discutimos em detalhes o processo de anotação manual do conjunto de dados (Seção 3.3), seguido pela Seção 3.5, que aborda o estudo de caracterização linguística dos comentários anotados manualmente. Por fim, a Seção 3.6 apresenta os modelos de aprendizado de máquina que serão analisados neste trabalho.

3.1 Visão Geral da Metodologia

A Figura 3.1 apresenta a visão geral da metodologia proposta neste estudo. Para a coleta de dados, primeiramente realizamos a seleção das comunidades de interesse. A seleção de comunidades busca elencar as principais comunidades brasileiras em número de inscritos no Reddit, dessa forma, selecionando as comunidades mais relevantes em termos de geração de conteúdo no contexto nacional (**Etapa 1: Coleta dos dados**). Os dados coletados originalmente são filtrados conforme os critérios discutidos na Seção 3.2 para gerar uma amostra representativa das comunidades e do período selecionado (**Etapa 2: Amostragem dos dados**). Essa amostra é essencial para a condução do processo de anotação manual.

A partir da seleção das comunidades de interesse, os dados foram coletados para todo o ano de 2022 e para as 10 maiores comunidades brasileiras em número de inscritos. Para responder à QP1, um grupo de anotadores voluntários - estudantes de graduação e pós-graduação - participou do processo de anotação manual dessa amostra de dados. Para compreender os padrões linguísticos dos comentários que são *tóxicos* e os que não são,

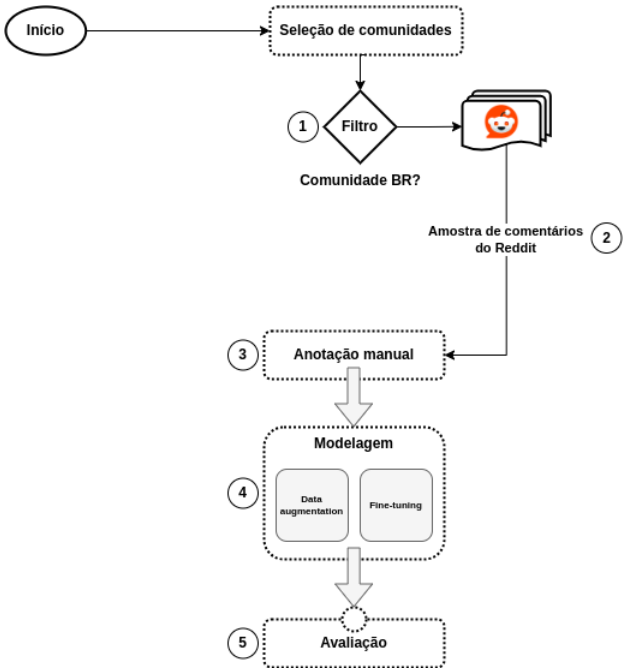

Figura 3.1: Diagrama ilustrando a visão geral dos processos e etapas da metodologia proposta.

realizamos uma rica caracterização utilizando diferentes técnicas de análise de conteúdo.

Na próxima etapa da metodologia, aplicamos técnicas de aprendizado de máquina para treinar modelos para a tarefa de classificação de toxicidade, especificamente para comentários na língua portuguesa (**Etapa 4: Modelagem**). Para endereçar a QP2, avaliamos o desempenho de diferentes categorias de modelos de aprendizado de máquina no conjunto de dados anotado. Como veremos mais adiante, o conjunto de dados anotado possui um grande desbalanceamento na proporção das classes. Por esse motivo, exploramos abordagens para lidar com essa característica e reduzir o impacto do desbalanceamento nos modelos treinados (**Data augmentation**). Na sequência, exploramos como a representação textual impacta no desempenho dos modelos analisados e como modelos pré-treinados generalizam para essa tarefa específica (**Fine-tuning**). Por fim, propomos uma análise detalhada dos diferentes métodos empregados (**Etapa 5: Avaliação**).

3.2 Conjunto de Dados

O Reddit é uma rede social online multilíngue fundada em 2005 e organizada em subcomunidades por áreas de interesse (subreddits). Diferente de outras redes sociais, como Facebook e Twitter, por exemplo, no Reddit os usuários interagem através das comunidades nas quais fazem parte. Os usuários podem publicar nas comunidades em

que participam através de postagens (*submissions*) e os participantes da comunidade podem comentar essas publicações. Os comentários são organizados em árvores (*threads*) e as respostas a um comentário são estruturadas por níveis. Comentários podem ser feitos diretamente em uma postagem (*parent*), como respostas diretas a outro comentário (nível 1), assim por diante.

Considerando as estatísticas de Reddit, números de 2025 mostram que esta plataforma possui mais de 1,5 bilhão de usuários registrados globalmente Wise (2023) e 22,5 milhões de usuários inscritos no Brasil Review (2023). Devido ao grande volume de dados gerados pelos usuários, neste trabalho definimos um conjunto de critérios para a escolha de um subconjunto de comunidades para realizar nosso estudo. Seguindo Almerekhi et al. (2022b,a, 2019), utilizamos os critérios a seguir para a escolha das comunidades de interesse:

- **Número de usuários ativos**: Comunidades populares no contexto brasileiro em número de usuários ativos até Março de 2023[1];
- **Diversidade de temas**: Comunidades que permitem a discussão de temas diversos entre os usuários, desde que esteja de acordo com as normas de moderação;
- **Moderação**: Comunidade que possui moderação ativa e com pelo menos 3 moderadores;
- **Classificação de tópicos**: Os posts em geral possuem uma classificação (*flair*) que pode ser usada para filtrar threads de assuntos específicos;
- **Tempo de existência**: Comunidades com tempo de existência de pelo menos 5 anos.

A Tabela 3.1 apresenta os subreddits selecionados e algumas estatísticas descritivas. Coletamos um total de 7.348.257 comentários e 390.924 postagens por meio do projeto 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 múltiplos 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.[2] Por fim, também excluímos comentários gerados por contas de automoderadores e bots que detectamos em

[1]Esta foi a data de início da pesquisa tema desta dissertação.

[2]Em português, textos de risadas são representados por sequências de caracteres como kkkkk, haha, hehe, etc.

Subreddit	Inscritos	Postagens	Commentários	Posts por dia	Comentários por dia
r/brasil	1.516.433	110.829	2.382.928	303,64	6.528,52
r/desabafos	490.049	115.876	1.487.076	317,46	4.074,16
r/futebol	369.925	35.826	1.272.009	98,15	3.484,93
r/saopaulo	358.681	7.308	88.894	20,02	243,53
r/eu_nvr	308.064	12.631	221.348	34,60	606,41
r/botecodoreddit	270.451	7.059	62.999	19,60	172,6
r/conversas	247.545	21.967	355.761	60,18	974,67
r/investimentos	232.485	9.756	156.695	26,7	429,2
r/tiodopave	219.926	2.371	12.106	6,58	33,16
r/brasilivre	210.582	67.301	130.844	184,38	3.584,75
	Total	390.924	7.348.257	–	–

Tabela 3.1: Subreddits selecionados, número de assinantes, postagens e comentários do ano 2022.

nossos dados. Esses filtros reduziram nosso corpus para aproximadamente 6,6 milhões de comentários. A Tabela 3.2 apresenta algumas estatísticas para os subreddits analisados após a aplicação dos filtros.

Subreddit	Postagens	Comentários
r/brasil	110.829	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 3.2: Subreddits selecionados e total de postagens e comentários (2022) após a filtragem.

3.2.1 Estatísticas Gerais do Conjunto de Dados

Nessa seção, apresentamos um estudo de caracterização geral do engajamento dos usuários com as principais comunidades brasileiras no Reddit. Os principais indicadores analisados são descritos abaixo:

- **Score médio:** Cada publicação no Reddit possui um indicador de engajamento denominando score. Essa métrica indica a quantidade de votos positivos que uma

Comunidade	Score médio das publicações	Número médio de comentários	Publicações deletadas (%)	Publicações removidas (%)
r/brasil	76,65	20,60	14,41%	11,62%
r/brasilivre	37,28	18,63	6,46%	10,85%
r/desabafos	11,88	12,23	19,52%	26,25%
r/conversas	7,51	15,40	40,80%	12,29%
r/futebol	66,50	34,62	8,80%	10,31%
r/saopaulo	24,10	11,71	19,54%	6,99%
r/investimentos	16,38	15,21	41,12%	7,09%
r/botecodoreddit	104,91	8,37	2,98%	2,63%
r/tiodopave	32,85	4,91	9,02%	7,42%
r/eu_nvr	306,01	16,92	5,14%	16,03%

Tabela 3.3: Estatísticas de score médio, quantidade média de comentários e publicações removidas e deletadas por comunidade analisada.

publicação teve de outros usuários. Portanto, o score médio é a média simples desse indicador.

- **Comentários controversos:** Assim como existe um indicador para votos positivos, há um outro atributo que indica a desaprovação dos usuários (*downvote*) Jasser et al. (2022). Um comentário controverso é um comentário que apresenta a razão votos positivos / votos negativos próximo de 0,5, representando um equilíbrio entre votos positivos e negativos.

- **Publicações removidas:** Publicações removidas pelo autor.

- **Publicações deletadas:** Publicações moderadas, ou seja, removidas pela plataforma.

A Tabela 3.3 descreve as estatísticas de score médio, quantidade de comentários e publicações removidas e deletadas por comunidade. Das comunidades que possuem as maiores médias de comentários, a comunidade r/futebol possui o maior score médio, o que é condizente com o tipo de publicação dessa comunidade, geralmente em torno de eventos esportivos de futebol, gerando elevado engajamento entre os usuários. As comunidades r/investimentos e r/conversas tiveram o maior percentual de comentários deletados no período, enquanto a comunidade r/desabafos teve o maior percentual de comentários removidos. Isso indica que comunidades que debatem temas mais polêmicos - como é o caso das comunidades r/conversas e r/desabafos - estão mais propensas à moderação de conteúdo, mesmo que discutir tópicos sensíveis seja o propósito dessas comunidades.

3.3 Anotação Manual de Conteúdo Tóxico

Conforme apresentado no Capítulo 1, o objetivo principal desta dissertação é a análise, caracterização e classificação automática de conteúdo tóxico compartilhado nas comunidades brasileiras no Reddit. Trabalhos na literatura discutem a importância de conjuntos de dados anotados por humanos para que a tarefa de classificação de toxicidade seja a mais acurada possível Trajano et al. (2023); Kobellarz and Silva (2022). Apesar da existência de dados anotados de outras redes sociais, dados do Reddit em português não são encontrados na literatura, conforme discutimos no Capítulo 2. Assim, uma etapa importante da nossa metodologia foi a anotação manual de conteúdo tóxico no nosso conjunto de dados.

Devido ao alto custo inerente da rotulação humana de dados Golazizian et al. (2024), extraímos uma amostra de 2.500 comentários do nosso corpus após aplicados os filtros descritos na Seção 3.2, 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 da Universidade Federal de Minas Gerais (UFMG) 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*.[3] Para fins de anotação, consideramos linguagem tóxica todo comentário *rude, desrespeitoso ou irracional com grande probabilidade de fazer com que alguém saia de uma discussão*, conforme definido pela API Perspective Perspective (2022). 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.

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 Concordância observada McHugh (2012), a estatística Cohen's Kappa Score Cohen et al. (2009), Kappa de Fleiss Fleiss (1971) e o Alfa de Krippendorf Krippendorff (2004). As métricas de avaliação são detalhadas abaixo.

Concordância Observada: A Concordância Observada é uma medida geral que mede a proporção de exemplos em que dois ou mais anotadores fornecem a mesma classificação ou anotação. Essa medida é definida por $P_o = \frac{A}{N}$, onde A representa o número de instâncias

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

em que os anotadores concordaram e N representa o número total de exemplos.

Cohen's Kappa Score: O escore de Kappa de Cohen mede o nível de concordância entre dois avaliadores ou classificadores, ajustado para a concordância que seria esperada ao acaso, definido pela Equação 3.1:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{3.1}$$

Fleiss' Kappa Score: É uma generalização do Kappa score de Cohen para mais de dois avaliadores. Mede a concordância entre n avaliadores em N itens com k categorias, onde $\bar{P}$ é a proporção média de concordância entre os avaliadores e $\bar{P}_e$ é a concordância esperada ao acaso:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}. \tag{3.2}$$

onde p_o é a proporção de casos em que os avaliadores concordam e p_e é a concordância esperada ao acaso, definida pela Equação 3.3, onde P_{A_i} e P_{B_i} são as proporções de observações em cada categoria para os avaliadores A e B, respectivamente, e k é o número de categorias.

$$p_e = \sum_{i=1}^{k} (P_{A_i} \cdot P_{B_i}) \tag{3.3}$$

Enquanto a métrica do score de Kappa de Cohen é utilizada para calcular a concordância par-a-par (entre anotadores e modelos, por exemplo), o Kappa de Fleiss é utilizado para calcular a concordância entre os anotadores, uma vez que no processo manual de anotação, cada comentário foi anotado por três anotadores distintos de forma anônima.

Alfa de Krippendorf: O alfa de Krippendorf é uma estatística que mede a concordância entre múltiplos anotadores, considerando a concordância observada e a concordância esperada ao acaso. O score é definido pela Equação 3.4, onde D_o representa a discordância observada e D_e representa a discordância ao acaso. O valor de alfa próximo de 1 representa alta concordância e próximo de 0 representa concordância aleatória.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{3.4}$$

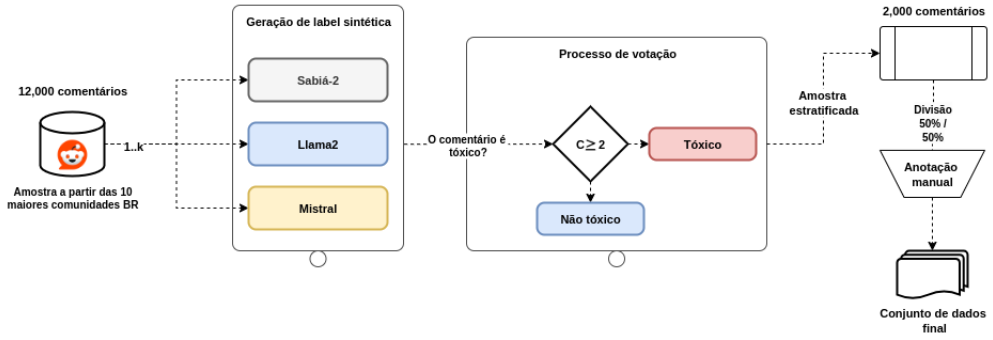

Figura 3.2: Visão geral do processo de anotação assistido por LLMs para o processo de *data augmentation* do conjunto de dados original desbalanceado. Os modelos utilizados nesta etapa são detalhados na Seção 3.6.5.

3.4 Anotação Automática de Conteúdo Tóxico

Após a etapa de classificação usando os LLMs, selecionamos uma amostra balanceada de 2.000 (50% *tóxico* - 50% *não-tóxico*) comentários para validação humana com a revisão de três anotadores independentes. Para cada exemplo, o anotador é apresentado ao comentário original do Reddit, e é solicitado a fornecer uma classificação para cada comentário (*não-tóxico*, *tóxico*, *não sei dizer* ou *falta informação*), sendo a classe final de cada comentário definida via voto majoritário. Por fim, fizemos uma amostragem aleatória selecionando apenas comentários classificados como *tóxicos* para incorporar o conjunto de dados original, desta forma aumentando a representatividade da classe minoritária. A Figura 3.2 apresenta a visão geral do processo de *data augmentation*.

3.5 Caracterização da Linguagem Tóxica

Para investigar padrões na linguagem dos comentários de conteúdo tóxico no domínio de interesse desta dissertação (comunidades do Reddit em Língua Portuguesa), selecionamos as seguintes análises linguísticas descritas a seguir que foram realizadas na amostra de 2.500 comentários anotados manualmente pelo grupo de 12 voluntários.

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. (2012)), 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% para o idioma Português do Brasil. 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.

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 de comentário; portanto, essa medida permite calcular a probabilidade de os usuários publicarem um texto curto ou longo na plataforma.

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 e se baseia na arquitetura Transformer.

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

Grafos de coocorrência: Uma análise complementar de um corpus pode ser feita por meio da observação de redes de co-ocorrência. 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 temos mencionados pelo menos 50 nos comentários analisados.

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.

3.6 Tarefa de Classificação de Toxicidade

A moderação do grande volume de dados produzidos diariamente é um enorme desafio para as plataformas de rede social online Gorwa et al. (2020); Gillespie (2020). Em alguns casos, seres humanos são responsáveis por esta moderação Jhaver et al. (2019). No entanto, modelos de aprendizado de máquina são as ferramentas mais indicadas para a execução em larga escala da tarefa de classificação de conteúdo tóxico nestas plataformas Ye et al. (2023).

Definimos o problema de distinguir um conteúdo como uma tarefa de classificação binária que, dado um comentário, classifica se o mesmo é *tóxico* ou *não-tóxico*. Formalmente, considere $\mathcal{X}$ como o conjunto de comentários e $\mathcal{Y} = \{0, 1\}$ o espaço de rotulação, onde 1 corresponde a um comentário *tóxico* e 0 um comentário *não-tóxico*. O principal objetivo do classificador binário é aprender a função $f : \mathcal{X} \rightarrow \mathcal{Y}$ utilizando um conjunto de treinamento com m itens $\{(x_i, y_i)|1 \leq i \leq m\}$. Aqui, cada item $x_i \in \mathcal{X}$ representa um comentário representado por um vetor de atributos pertencentes a um espaço de alta dimensionalidade, e $y_i \in \mathcal{Y}$ denota a classe alvo (*tóxico* ou *não-tóxico*).

O conjunto de dados apresentado nesse estudo possui desbalanceamento de classes. O desbalanceamento de classes é um problema comum em tarefas de classificação automática, caracterizado por uma distribuição desigual entre as classes do conjunto de dados, onde uma ou mais classes (minoritárias) possuem significativamente menos exemplos do que outras (majoritárias). Esse desequilíbrio pode levar os modelos de aprendizado de máquina a favorecer a classe majoritária, resultando em alta acurácia aparente, mas baixo desempenho em métricas mais sensíveis ao desbalanceamento, como precisão, recall ou F1-score para a classe minoritária.

A seguir, apresentamos as técnicas para tratamento de desbalanceamento de dados e representação vetorial dos comentários analisados, seguido pelos modelos de aprendizado cujos os desempenhos na execução da tarefa de classificação de toxicidade descrita anteriormente foram avaliados e comparados.

3.6.1 Técnicas para Balanceamento de Dados

Para lidar com o desbalanceamento do conjunto de dados original, aplicamos técnicas de reamostragem de dados que visam equilibrar a proporção das classes ao reduzir a classe majoritária ou aumentar sinteticamente a proporção dos exemplos da classe minoritária. Existem diversas abordagens para fazer reamostragem, porém, neste estudo,

iremos experimentar com duas técnicas clássicas: subamostragem aleatória e SMOTE, descritas a seguir.

Subamostragem aleatória. Essa abordagem visa reduzir a quantidade de exemplos da classe majoritária para atenuar o efeito da diferença na proporção entre as classes. É executada de forma a selecionar exemplos aleatoriamente até que determinada proporção de exemplos para cada classe seja alcançada Noor et al. (2022). Embora seja uma técnica bem simples e que pode funcionar bem em alguns cenários, a subamostragem aleatória tem algumas desvantagens, principalmente em conjuntos de dados pequenos, como perda de informação relevante e introdução de ruídos no processo de treinamento.

Synthetic Minority Over-sampling Technique (SMOTE). É uma técnica de superamostragem que gera amostras sintéticas para a classe minoritária em vez de simplesmente duplicá-las. As amostras sintéticas são criadas interpolando entre amostras reais da classe minoritária e seus vizinhos mais próximos Chawla et al. (2002). A criação de um novo ponto de dados é definida pela Equação 3.5, onde $\mathbf{x}_i$ é um ponto de dados real da classe minoritária, $\mathbf{x}_n$ é um dos vizinhos mais próximos de $\mathbf{x}_i$ dentro da mesma classe e λ é um numero aleatório no intervalo [0,1] usado para a interpolação entre $\mathbf{x}_i$ e $\mathbf{x}_n$. Por fim, o novo ponto sintético é definido por $\mathbf{x}_{new}$.

$$\mathbf{x}_{new} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_i - \mathbf{x}_n). \tag{3.5}$$

O algoritmo seleciona um ponto aleatório no espaço entre uma amostra real e um de seus vizinhos mais próximos, definido pela distância Euclidiana.

$$\text{dist}(\mathbf{x}_i, \mathbf{x}_n) = \sqrt{\sum_{j=1}^{d} (x_{i,j} - x_{n,j})^2}. \tag{3.6}$$

Embora o algoritmo SMOTE evite gerar exemplos duplicados e preserve a distribuição original da classe minoritária, ele pode introduzir ruídos e instabilidades no processo de treinamento, dado o fato de que os pontos selecionados ao acaso podem mudar em diferentes iterações, além de aumentar a complexidade computacional em relação a métodos mais simples, como amostragem aleatória. Mesmo com essas desvantagens, o procedimento SMOTE é amplamente utilizado na literatura, especialmente em tarefas onde há pouca disponibilidade de dados Rupapara et al. (2021); Sun et al. (2009); Padurariu and Breaban (2019). Por esse motivo, selecionamos esse método para a análise comparativo deste trabalho.

3.6.2 Técnicas de Representação de Dados

Para a aplicação de modelos de aprendizagem, os dados textuais devem ser traduzidos em uma representação numérica. A técnica utilizada para a obtenção desta representação impacta na acurácia do modelo utilizado Selva Birunda and Kanniga Devi (2021); Xu and Wu (2014). A seguir, apresentamos as duas técnicas de representação vetorial utilizadas nesta dissertação. Essas técnicas são amplamente utilizadas na literatura e refletem métodos clássicos e modernos de representação de características textuais Ge and Moh (2017); Wang et al. (2020).

Term Frequency-Inverse Document Frequency (TF-IDF). É uma técnica de vetorização de texto amplamente utilizada para converter documentos textuais em vetores numéricos, capturando a importância de cada termo em relação ao contexto global e ao contexto local de um conjunto de documentos Robertson (2004). A primeira parte da equação para cálculo do TF-IDF contabiliza a frequência em que um termo ocorre em um dado documento. A Equação 3.7 representa a quantidade de vezes em que o termo ocorre em um documento, onde $f_{t,d}$ é o número de vezes que o termo t ocorre no documento d, e o denominador representa o número total de termos no documento d.

$$\mathrm{TF}_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}. \tag{3.7}$$

O segundo termo do cálculo do TF-IDF visa calcular a frequência invertida dos documentos (IDF, *Inverse document frequence*), de forma a penalizar termos que são compartilhados por muitos documentos, ou seja, possuem menor capacidade de discriminar entre as classes. A Equação 3.8 ilustra o cálculo do termo IDF, onde $|D|$ é o número total de documentos na coleção, e $|\{d \in D : t \in d\}|$ representa o número de documentos nos quais o termo t aparece.

$$\mathrm{IDF}(t) = \log\left(\frac{N}{1 + \mathrm{DF}(t)}\right). \tag{3.8}$$

Por fim, para calcular o valor de TF-IDF para representar o texto de um dado documento, utilizamos a Equação 3.9, onde multiplicamos os valores encontrados nos termos da equação descritos anteriormente.

$$\text{TF-IDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t). \tag{3.9}$$

Word Embeddings. Uma abordagem mais eficiente para representação semântica de dados textuais são os vetores de palavras (*Word Embeddings*). Word embeddings são

representações densas e contínuas de palavras em um espaço vetorial, projetadas para capturar relações semânticas e sintáticas entre palavras com base nos contextos em que as palavras são utilizadas. Através do processo de treinamento de uma rede neural para gerar os vetores de representação, cada palavra w é mapeada para um vetor de dimensão fixa $\vec{w} \in \mathbb{R}^d$, onde d é a dimensionalidade do vetor de embedding. Alguns dos principais algoritmos para criação de vetores de embeddings são listados abaixo.

1. **Word2Vec** Mikolov et al. (2013). Utiliza redes rasas (*shallow networks*) para aprender embeddings de palavras. Existem duas variações de arquitetura usando esse tipo de rede: CBOW (*Continuous Bag of Words*) e Skip-gram. Na estratégia CBOW, dado o contexto (um conjunto de k palavras), o modelo faz a predição da palavra central. Já na estratégia Skip-gram, dada uma palavra de entrada, o modelo prevê as palavras de contexto (palavras adjacentes à palavra de interesse).

2. **GloVe** Pennington et al. (2014). Ao invés de prever a palavra de dado um contexto, GloVe usa contagens globais de coocorrência para modelar as relações entre palavras. A função de custo é projetada para otimizar a similaridade entre palavras que coocorrem frequentemente, levando em consideração a distribuição global de palavras no corpus.

3. **BERT** Kenton and Toutanova (2019). Diferente de abordagens anteriores, o BERT utiliza um mecanismo de atenção bidirecional que considera o contexto completo de uma palavra na sentença, tanto à esquerda quanto à direita. Isso resulta em embeddings contextuais, onde o mesmo termo pode ter diferentes representações dependendo do contexto de onde está inserido. Os vetores de embeddings gerados pelo BERT são então extraídos para serem utilizados em tarefas de classificação.

3.6.3 Modelos de Classificação Avaliados

A seguir, apresentamos os modelos de aprendizado avaliados nesta dissertação. Como *baselines*, selecionamos o modelo de Regressão Logística Kleinbaum et al. (2002) e API Perpective Lees et al. (2022), que é um modelo proprietário amplamente utilizado na literatura de detecção de toxicidade em diversos idiomas Chong and Kwak (2022); Almerekhi et al. (2022a); Salehabadi et al. (2022). O desempenho destes modelos foi comparado com um conjunto de modelos baseados na arquitetura *transformers*: BERT Kenton and Toutanova (2019), RoBERTa Zhuang et al. (2021) e os grandes modelos de linguagem (LLMs): Llama 2 Touvron et al. (2023), Sabiá-2 Almeida et al. (2024) e Mistral Jiang et al. (2023).

3.6.4 Baselines

Os seguintes modelos foram escolhidos como *baselines*:

Regressão Logística Kleinbaum et al. (2002). O modelo de Regressão Logística é um modelo linear de aprendizado de máquina utilizado em tarefas de classificação, especialmente nas tarefas de classificação binária. Em sua arquitetura, a regressão faz uma combinação linear entre os valores de entrada e os pesos da rede . A saída do modelo, ou *logit*, é definida abaixo, onde w é o vetor de pesos, x é o vetor de entradas (nesse caso, a representação vetorial, como TF-IDF ou word embeddings), $\mathbf{w}^T$ é o produto interno entre os vetores de entrada e de peso e b é o termo de viés (*bias)*.

$$z = \mathbf{w}^T\mathbf{x} + b. \tag{3.10}$$

Para transformar a saída do modelo em uma distribuição de probabilidade, a função Sigmoide é aplicada. A função $\sigma(z)$ escala o valor de saída da combinação linear para o limite $[0, 1]$, representando a probabilidade da predição pertencer a uma classe.

$$P(y = 1|\mathbf{x}) = \sigma(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}}. \tag{3.11}$$

Por fim, a predição pode ser feita usando um limiar de decisão. Em outras palavras, se o valor resultante da operação de sigmoide for acima de um limite, a classificação é atribuída a uma classe ou outra, caso contrário. O limiar de decisão pode variar a depender da aplicação, porém um valor muito utilizado é $0,5$, conforme definido abaixo. Esse valor será adotado nos experimentos de baseline nesse projeto.

$$\hat{y} = \begin{cases} 1 & \text{if } \sigma(\mathbf{w}^T\mathbf{x} + b) \geq 0.5 \\ 0 & \text{caso contrário.} \end{cases} \tag{3.12}$$

Perspective API Lees et al. (2022). Conjunto de classificadores de toxicidade proposto pela Google Jigsaw [4], amplamente usado em pesquisas anteriores Almerekhi et al. (2020); Salehabadi et al. (2022); Zannettou et al. (2020); ElSherief et al. (2018). O modelo da Perspective retorna um score entre 0 e 1 com a probabilidade de um comentário ser considerado tóxico. A seleção de um *threshold* adequado para classificação dos comentários depende da aplicação e pode ser configurado para atender critérios pré-definidos. Como o objetivo deste estudo é reduzir a incerteza na detecção de conteúdo tóxico, o limiar de 0,7 foi utilizado como critério de corte, assim como em Kumar et al. (2023).

[4] Disponível em https://perspectiveapi.com/how-it-works/.

3.6.5 Modelos Baseados na Arquitetura *Transformers*

A arquitetura Transformer Vaswani (2017) representa uma inovação no campo de aprendizado profundo para processamento de linguagem natural (NLP) e surgiu como uma alternativa às arquiteturas baseadas em redes recorrentes (Recurrent Neural Networks Medsker et al. (2001)) ou convolucionais (Convolutional Neural Networks O'shea and Nash (2015)). O sucesso desta arquitetura pode ser explicado pela utilização do mecanismo de atenção proposto em Vaswani (2017). Esse mecanismo permite que o processo de treinamento seja mais eficiente em termos de paralelismo e modelagem de longas dependências, possibilitando a representação vetorial dinâmica dos textos analisados, com base no contexto atual que está sendo analisado. A arquitetura possui dois componentes principais: Encoder e Decoder. O primeiro é responsável por processar os dados de entrada e gerar uma representação intermediária, enquanto o decodificador processa essa representação e gera a saída do modelo, que é a distribuição de probabilidades de um próximo *token*[5] em uma sequência de texto. Os modelos usados nesta dissertação são brevemente apresentados a seguir. Os modelos transformers bidirecionais como BERTimbau e XLM RoBERTa foram selecionamos por serem extensamente utilizados na literatura Almerekhi et al. (2022b); Leite et al. (2020); Trajano et al. (2023) e por possuírem variantes especializadas na língua portuguesa Souza et al. (2020); Conneau et al. (2020). Além disso, os LLMs escolhidos possuem vasto conhecimento multilíngue, em português e podem ser adaptados para tarefas específicas de classificação Oliveira et al. (2023); Almeida et al. (2024).

BERT Kenton and Toutanova (2019). O processo de treinamento do modelo BERT (Bidirectional Encoder Representations from Transformers) é baseado no uso do mecanismo de atenção bidirecional, ou seja, possibilitando o processamento das entradas de dados de forma bidirecional, capturando contexto subsequente ou precedente a um determinando *token*, possibilitando a geração de representações de dados mais ricas. O BERT é treinado em duas variações de parâmetros: $\text{BERT}_{\text{BASE}}$ (L=12, H=768, A=12, Parâmetros = 110M) e $\text{BERT}_{\text{LARGE}}$ (L=24, H=1024, A=16, Parâmetros = 340M), onde L é a quantidade de blocos de codificação, H é a quantidade de camadas ocultas das redes neurais usadas pelo modelo e A o número de módulos de mecanismo de atenção que são executados paralelamente. Neste trabalho, adotamos a versão $\text{BERTimbau}_{\text{LARGE}}$ do modelo pré-treinado para o Português Souza et al. (2020) por possuir maior quantidade de parâmetros em sua construção.

[5]Token é uma unidade fundamental de texto, podendo representar uma palavra, partes de uma palavra ou um sinal de pontuação, que é utilizada por modelos de aprendizado de máquina para processar características textuais.

RoBERTa Zhuang et al. (2021). RoBERTa (Robustly Optimized BERT Pretraining Approach) é uma variante do modelo BERT que introduz melhorias no processo de treinamento e parametrização do modelo. O modelo inicial foi inspirado na variante BERT LARGE, com modificações no conjunto de dados de treinamento, no uso de máscaras dinâmicas no processo de modelagem de linguagem (*language modeling*) [6] e no aumento do tamanho dos *batches* de treinamento, bem como no número de iterações. Neste trabalho, utilizamos a variante XLM RoBERTa, um modelo pré-treinado com dados multilínguas Conneau et al. (2020).

LLMs Zhao et al. (2023); Kukreja et al. (2024). Grandes Modelos de Linguagem (*Large Language Models -LLMs*) são modelos de aprendizado profundo que processam grande quantidade de dados na fase de treinamento. O processo de treinamento desses modelos pode variar a depender da implementação, mas os modelos de fundação (*Foundational models*) usam o framework autoregressivo e aprendizado auto-supervisionado (*self-supervised learning*) para mapear as relações semânticas do corpus de texto Brown et al. (2020). A tarefa destes modelos é predizer o próximo *token* dado uma sequência prévia de *tokens*.

Apesar da tarefa original destes modelos estar relacionada a geração de texto, o treinamento com uma grande base de dados faz com que estes modelos sejam também eficientes para diferentes tarefas de aprendizado, fenômeno conhecido como *habilidades emergentes* Tan et al. (2024); Kumar et al. (2024). Devido ao alto custo computacional de treinamento, os primeiros LLMs disponíveis foram construídos por empresas privadas e disponibilizados comercialmente. Entretanto, com o avanço e melhorias na arquitetura dos modelos de fundação, alguns modelos abertos (*open-source*) estão atualmente disponíveis na literatura. Nesse trabalho, iremos abordar modelos comerciais e abertos. A Tabela 3.4 detalha as principais diferenças entre os LLMs adotados nesse estudo.

Uma forma padronizada de comparar os LLMs é através de benchmarks, que são tarefas bem definidas em que os modelos precisam gerar respostas válidas. Através dessa avaliação, é possível comparar o desempenho de um determinado modelo em diferentes domínios, como raciocínio (*reasoning*), capacidade de responder questões matemáticas e gerar códigos, por exemplo.

A Tabela 3.5 apresenta as principais informações sobre alguns dos benchmarks utilizados para comparar os modelos de linguagem quanto à sua capacidade de raciocínio, conhecimentos gerais, habilidades técnicas e conversacionais.

A avaliação padronizada via *benchmarks* é importante para comparar modelos diferentes em uma mesma escala. Para esse trabalho, selecionamos LLMs que possuem diferentes abordagens de pré-treinamento, coleta e curadoria de dados e, por fim, são

[6] Modelagem de linguagem é a tarefa de predizer a próxima palavra baseado na probabilidade de observar um termo dado um conjunto de palavras adjacentes como contexto.

Modelo	# Parâmetros	Contexto	Dados de treino	Capacidades	Open-source
GPT-3	175B	16.385	Livros Wikipedia CommonCrawl Artigos da Internet Redes sociais	Raciocínio de propósito geral, geração e compreensão de texto	✗
GPT-4	1T+ (estimado)	128.000	Similar ao GPT-3, mas com curadoria aprimorada	Raciocínio avançado e compreensão de tarefas complexas	✗
Llama2	7B 13B 70B	4.096	Livros CommonCrawl Fóruns online Sites de notícias	Geração de texto e compreensão competitivas; bom desempenho em tarefas multi-idiomas	✓
Mistral	123B	4.096	Conjunto de dados otimizado para tarefas multi-idiomas e geração de código	Bom desempenho em tarefas multi-idiomas e geração de código	✓
Sabiá-2	Não divulgado	8.192	Dados coletados da internet, majoritariamente em PT-BR	Desempenho focado em domínio cultural, especialmente em tarefas do Português do Brasil	✓

Tabela 3.4: Comparativo entre os modelos de linguagem (LLMs) selecionados para esse trabalho.

Benchmark	Área	Formato da tarefa	Habilidades	# Dataset
MMLU Hendrycks et al. (2021a)	Conhecimento geral	Questões de múltipla escolha	Amplitude de raciocínio e conhecimento	57 tarefas em diversos domínios, totalizando 15, 908 questões
MATH Hendrycks et al. (2021b)	Matemática	Questões com resposta de referência	Raciocínio matemático e resolução de problemas	12,500 exemplos
HumanEval Chen et al. (2021)	Programação	Geração de código com testes automatizados	Resolução de problemas através de programação	164 problemas de programação
Chatbot Arena Chiang et al. (2024)	Interação / conversação	Comparação par a par via chat	Capacidade conversacional, coerência, objetividade e raciocínio geral	Interação contínua com usuários reais

Tabela 3.5: Diferentes benchmarks utilizados para avaliar capacidades de raciocínio, conhecimentos gerais e específicos de modelos de linguagem.

competitivos em benchmarks gerais. A Tabela 3.6 apresenta o desempenho dos modelos analisados nesta dissertação, considerando diferentes *benchmarks*. Embora os modelos *open-source* sejam promissores, os LLMs proprietários ainda apresentam desempenho superior em diversas tarefas, como é o caso do GPT-4-turbo. Vale ressaltar que o modelo Sabiá-2 não possui avaliações para alguns *benchmarks*, pois este modelo é especializado no português brasileiro. Entretanto, os autores em Almeida et al. (2024) propõem uma visão comparativa em *benchmarks* nacionais, comparando os modelos em exames brasileiros, como ENEM (Exame Nacional do Ensino Médio), ENADE (Exame Nacional de Desempenho dos Estudantes) e POSCOMP (Exame Nacional para Ingresso na Pós-Graduação em Computação). Os LLMs *open-source* utilizados na etapa de anotação sintética deste trabalho foram: Sabiá-2, Llama2 e Mistral. Por fim, os LLMs comerciais analisados foram utilizados em Junho de 2024 e, portanto, representam os modelos mais avançados disponíveis à época do experimento.

Modelo	MMLU	MATH	HumanEval	Chatbot Arena
GPT-3-turbo	0.68	0.39	0.69	1107
GPT-4-turbo	0.87	0.74	0.87	1257
Llama2	0.45	0.14	0.21	1063
Mistral	0.64	0.33	0.43	1114
Sabiá-2	–	–	–	–

Tabela 3.6: Resultados em benchmarks padrão dos modelos de linguagem selecionados nesta dissertação. Esses resultados foram fornecidos pelos trabalhos listados na Tabela 3.5.

3.6.6 Métricas de avaliação

As métricas utilizadas para a comparação de desempenho entre os modelos analisados nesta dissertação são descritas a seguir. **Correlação de Pearson:** Mede a força e a direção do relacionamento linear entre duas variáveis Sedgwick (2012). O coeficiente de correlação de Pearson é definido pela Equação 3.13

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3.13}$$

onde cov(X, Y) é a covariância entre as variáveis X e Y e σ_X e σ_Y são os desvios padrão de X e Y, respectivamente.

Precisão: A precisão mede a proporção de instâncias corretamente classificadas como positivas dentre todas as instâncias previstas como positivas Buckland and Gey (1994). A Equação 3.14 define o cálculo de precisão

$$\text{Precisão} = P = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{3.14}$$

onde VP são os verdadeiros positivos (instâncias positivas que foram corretamente classificadas como positivas) e FP, os falsos positivos (instâncias negativas que foram erroneamente classificadas como positivas). Nesse trabalho, consideramos a classe de interesse como a classe positiva, ou seja, o objetivo é classificar corretamente conteúdo tóxico.
Recall: O recall (revocação ou sensibilidade) é uma métrica de recuperação que mede a proporção de instâncias positivas corretamente classificadas Buckland and Gey (1994), definido pela Equação 3.15:

$$\text{Recall} = R = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{3.15}$$

Em aprendizado de máquina, geralmente existe um compromisso (*trade-off*) entre as métricas de precisão e recall, uma vez que a precisão mede a qualidade e o recall mede

a recuperação das instâncias. Em outras palavras, aumentar o recall para recuperar mais instâncias da classe positiva pode resultar em uma redução na qualidade da precisão e vice-versa.

F1-Score: O F1-score é uma métrica de avaliação utilizada em problemas de classificação que combina precisão e recall em uma única medida de desempenho Hossin and Sulaiman (2015). Ela é definida como a média harmônica entre precisão e recall, proporcionando um equilíbrio entre elas. Quando se usa F1 score, o peso das métricas de precisão e recall contribui igualmente para a métrica combinada. O F escore é definido pela Equação 3.16.

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{3.16}$$

As métricas de avaliação de modelos de aprendizado de máquina foram apresentadas em sua forma geral. Entretanto, é possível adaptá-las para um contexto específico, favorecendo precisão ou recall, por exemplo. Em uma aplicação onde o custo do erro é muito elevado, pode-se aumentar o peso de contribuição da precisão, enquanto em outra aplicação, pode ser desejável apenas retornar o maior número possível de instâncias positivas. Como neste trabalho o objetivo é apenas traçar um comparativo geral entre os modelos, utilizaremos as métricas em sua forma padrão.

Capítulo 4

Caracterização Linguística dos Dados Anotados Manualmente

Neste capítulo, apresentamos as análises realizadas para responder à QP1: (*é possível distinguir conteúdo tóxico e não-tóxico considerando diferentes características linguísticas encontradas em seus textos?*), considerando o conjunto amostrado de 2.500 comentários. Iniciamos, na Seção 3.2.1, com a apresentação das estatísticas descritivas gerais das comunidades, incluindo o percentual de toxicidade, a média de comentários e o número de publicações removidas. Em seguida, detalhamos a avaliação das anotações manuais, abordando tanto a concordância entre os anotadores envolvidos no estudo quanto a comparação entre as rotulagens manual e automática (Seção 4.1). A Seção 4.2 apresenta a análise linguística conduzida para caracterizar o conteúdo tóxico e não tóxico, bem como discutimos os principais achados desse estudo. Por fim, sumarizamos os principais achados da nossa caracterização na Seção 4.3. Os resultados apresentados neste Capítulo foram publicados em Lima et al. (2024a,b).

4.1 Avaliação da Anotação Manual

Nesta seção, apresentamos os resultados das análises de concordância entre os anotadores na rotulagem manual de comentários do Reddit. Inicialmente, são apresentadas as métricas globais de concordância — incluindo a concordância observada, o *Kappa de Cohen*, *Kappa de Fleiss* e o *Alfa de Krippendorff* — que evidenciam níveis variados de acordo entre os anotadores, com casos tanto de concordância total quanto de discordância total Cohen et al. (2009); Fleiss (1971); Krippendorff (2004). Em seguida, a análise é estratificada por grupos e lotes de comentários, revelando que, enquanto alguns grupos alcançaram concordância razoável a moderada, outros apresentam maior incerteza na classificação. Por fim, exploramos a distribuição dos rótulos atribuídos por cada anotador, ressaltando as diferenças individuais na percepção do que constitui conteúdo tóxico

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 4.1: Concordância entre anotadores.

e confirmando a alta subjetividade inerente à tarefa de detecção de toxicidade.

4.1.1 Concordância entre Anotadores

Primeiramente, calculamos a concordância global entre os anotadores nos comentários do Reddit rotulados manualmente, cujos resultados são mostrados na Tabela 4.1.

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. A quantidade de comentários que tiveram concordância ou discordância totais foi de 1.594 e 107, 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 88,70% 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 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 **??** 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.

A Tabela 4.2 analisa a rotulagem feita por cada anotador. O grupo A rotulou a menor porcentagem de comentários como *tóxico*. Já o Grupo B apresenta a maior varia-

Rótulo	Anot. 1	Anot. 2	Anot. 3
Lote 1 (Grupo A)			
Não-tóxico	84.60%	88.96%	90.60%
Tóxico	9.40%	9.84%	7.40%
Não sei dizer	0.60%	1.00%	0.00%
Info. insuficiente	5.40%	0.20%	2.00%
Lote 2 (Grupo B)			
Não-tóxico	83.17%	69.48%	74.95%
Tóxico	7.82%	21.29%	4.81%
Não sei dizer	3.81%	3.82%	2.81%
Info. insuficiente	5.21%	5.42%	17.43%

Lotes 1 e 2

Rótulo	Anot. 1	Anot. 2	Anot. 3
Lote 3 (Grupo C)			
Não-tóxico	75.90%	68.01%	78.51%
Tóxico	19.28%	21.73%	17.87%
Não sei dizer	4.02%	7.65%	3.01%
Info. insuficiente	0.80%	2.62%	0.60%
Lote 4 (Grupo D)			
Não-tóxico	72.80%	93.59%	69.14%
Tóxico	11.60%	5.21%	9.02%
Não sei dizer	4.20%	0.80%	6.41%
Info. insuficiente	11.40%	0.40%	15.43%

Lotes 3 e 4

Rótulo	Anot. 1	Anot. 2	Anot. 3
Lote 5 (Grupo C)			
Não-tóxico	84.51%	68.60%	75.20%
Tóxico	14.49%	25.00%	19.72%
Não sei dizer	1.01%	4.20%	4.67%
Info. insuficiente	0.00%	2.20%	0.41%

Lote 5

Tabela 4.2: Distribuição dos rótulos por anotador e por grupo de lotes.

bilidade 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).

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

4.1.2 Comparação entre a Rotulagem Manual e as Rotulagens Automáticas

Esta seção apresenta uma comparação entre a classificação manual de toxicidade em comentários do Reddit e as avaliações realizadas por métodos automáticos, utilizando tanto a API da Perspective quanto LLMs. Inicialmente, analisamos o desempenho da API Perspective, evidenciando que o limiar de 0.7 proporciona um equilíbrio (*trade-off)* adequado entre precisão e revocação, alinhando a proporção de comentários tóxicos com as anotações humanas. Em seguida, detalhamos a comparação entre as predições dos LLMs e as anotações manuais, destacando as diferenças nos comportamentos dos modelos. Por fim, são discutidas as métricas de correlação e os índices de concordância (como o Kappa de Cohen) entre as anotações e as predições, ressaltando as variações de desempenho entre os diferentes lotes e grupos, bem como os desafios em interpretar o uso de gírias e expressões regionais.

4.1.2.1 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 severa** 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.

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 4.3 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 pela Perspective tem uma grande margem de melhoria para textos em português brasileiro, 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 Cohen 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 anotadores 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. As maiores diferenças observadas foram entre os Lotes 1 e 4.

4.1.2.2 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 diversas línguas (inclusive o português brasileiro), enquanto o segundo é uma versão ajustada (*fine-tuned*) para dados em português. Para ambos 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 5.1 mostra o *prompt* utilizado para a tarefa de

> 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*.
> 4. Marque somente como "Tóxico" os comentários em que há forte evidência da presença de conteúdo *tóxico*.

Figura 4.1: Instruções fornecidas ao assistente para classificação de comentários no Reddit usando a abordagem *zero-shot*.

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 modelos consideramos o tamanho máximo de tokens de entrada de cada modelo igual a 500 e temperatura[1] igual a zero.

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 pela presença de conteúdo tóxico, 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"*.

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

		Lote 1	Lote 2	Lote 3	Lote 4	Lote 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**
	FP	6	4	15	5	13	**43**
	Total	**161**	**133**	**161**	**165**	**131**	

Tabela 4.4: Falso positivos (FP) e falso negativos (FN) por grupo.

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 4.4 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 à análise anterior de que o Sabia-2 é menos conservador, enquanto o ChatGPT é mais sensível à presença de toxicidade.

Essas características, embora diferentes, podem ser aplicadas a depender do cenário em questão. Por exemplo, se o interesse for minimizar a presença de falsos positivos, 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 4.5 mostra o Kappa score para cada grupo. Observa-se, novamente, menor concordância entre o Grupo D e maior concordância com

	Lote 1	Lote 2	Lote 3	Lote 4	Lote 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 4.5: *Kappa score* entre anotações e modelos de linguagem (PT-BR) instruídos para executar a tarefa de classificação de toxicidade.

os Grupo C.

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, enquanto o ChatGPT favoreceu muito mais a métrica de precisão. Todos os modelos de classificação automática foram influenciados pela presença de expressões regionais e palavrões, como discutido anteriormente.

4.2 Caracterização da Linguagem do Conteúdo Tóxico e Não-tóxico

Nesta seção, 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.

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 [2] 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].

[2]Type se refere a tokens distintos em uma determinada sentença.

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.

Análise de etiquetas de POS: 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.

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 $p < 0{,}01$.

As duas etiquetas de POS mais comuns para comentários *tóxicos* e *não-tóxicos* são NOUN (substantivo) e ADJ (adjetivo). 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 ADJ, NOUN e PROPN. Os resultados são mostrados na Tabela 4.6. Uma descoberta relevante é o termo *mulher* em comentários *tóxicos*. De fato, realizamos um

	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 [3], Liberdade, Rússia, Paris

Tabela 4.6: 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 4.7: Tópicos e palavras-chave relevantes em comentários em que os três anotadores discordaram (desacordo total).

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

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 4.10).

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 4.7). As palavras no tópico 0 (*feedback, remoção*) revelam que alguns comentários foram moderados anteriormente pela DMCA (Digital Millennium Copyright Act, Congress (1998)).

É interessante notar que os principais tópicos dos comentários sobre os quais os anotadores concordaram totalmente também discutem os mesmos temas (Tabela 4.8). 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

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

Tabela 4.9: 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).

concordaram no rótulo *tóxico* e que a API Perspective classificou como *não-tóxico* (Tabela 4.9), 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. 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.

A Tabela 4.10 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*.

Análise de n-gramas: A Figura 4.2 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.

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

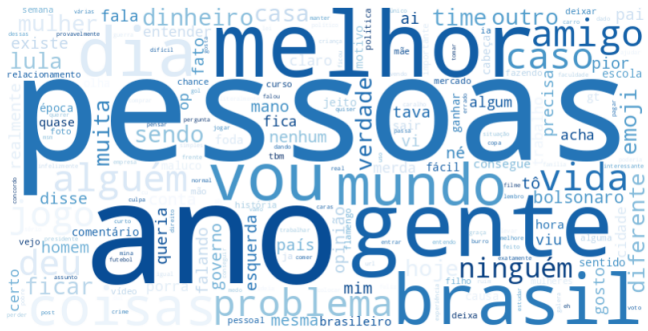

Figura 4.2: Unigramas mais frequentes encontrados nos comentários das principais comunidades brasileiras no Reddit.

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.

Já a Figura 4.3 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 presidência 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 4.11 apresenta os bigramas mais recorrentes nas comunidades analisadas neste artigo. Conforme mencionado anterior-

Figura 4.3: Termos frequentes em comentários classificados como *tóxico* (esquerda) e comentários classificados como *não-tóxicos* (direita).

mente, 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 a 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 dado 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 extracur-*

	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 4.11: 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*.

riculares, e *curso voltado*, indicando discussões focadas no desenvolvimento pessoal e financeiro. Na comunidade 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.*

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 4.4 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

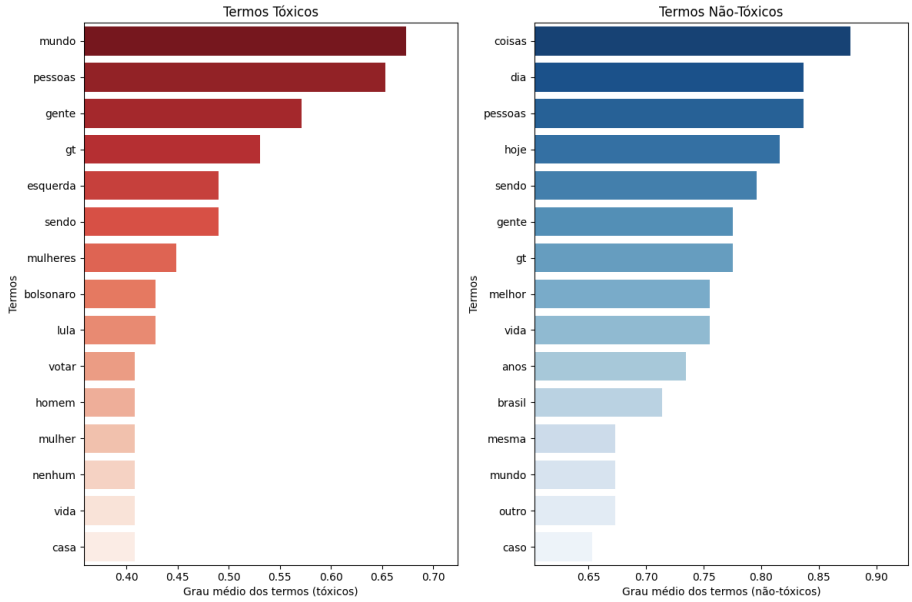

Figura 4.4: Top-15 termos *tóxicos* e *não-tóxicos* mais influentes extraídos a partir do grafo de coocorrência.

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"* e *"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"* e *"hoje"*, possuem alta centralidade de grau e representam discussões mais cotidianas e neutras, voltadas a interações 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 frequentemente são utilizadas em conjunto nestes tipos de discussão.

Reconhecimento de entidades nomeadas (NER): A Tabela 4.12 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

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

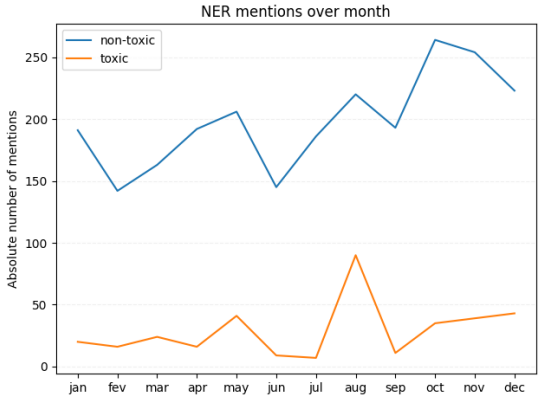

Figura 4.5: Série temporal mensal de menções de entidades nomeadas.

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"*.

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 série temporal mensal dos números de citações de entidades na Figura 4.5.[4] 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"*.

[4]MISCELÂNEA foi excluída.

4.3 Sumarização dos Resultados

Nesta seção, sumarizamos os principais achados da caracterização linguísticas dos dados anotados manualmente como *tóxicos* ou *não-tóxicos*.

Qualidade da anotação. Avaliamos a qualidade do conjunto de dados calculando a concordância entre anotadores, com resultados que corroboram trabalhos semelhantes (Perspective (2022)). 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.

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.

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*. A presença de termos repetidos para ambas as classes indica que esses termos são frequentes no geral, ou seja, eles ocorrem frequentemente tanto para comentários tóxicos quanto 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.

Capítulo 5

Avaliação da Tarefa de Classificação de Toxicidade

Neste capítulo, apresentamos os resultados dos experimentos conduzidos para a criação de modelos de aprendizado de máquina para classificação de toxicidade para os dados do Reddit em PT-BR. O principal objetivo é responder à QP2 (*Quais são os principais aspectos que devem ser considerados para a proposta de modelos acurados para a classificação automática de conteúdo tóxico no Reddit, considerando a língua portuguesa? Quais classes de modelos de classificação são as mais adequadas para esta tarefa de aprendizado?*) e avaliar como diferentes métodos de classificação automática desempenham esta tarefa, discutindo as vantagens e limitações de cada método, especialmente considerando sua aplicabilidade em larga escala, para o processamento de um grande volume de comentários.

As análises realizadas neste capítulo são divididas em três principais conjuntos:

- **Avaliação do impacto do desbalanceamento de dados na tarefa de classificação de toxicidade**. Esta análise investiga como o desbalanceamento de dados afeta a tarefa de classificação de toxicidade. Inicialmente, revisamos técnicas clássicas de balanceamento, como o *over-sampling* e o *under-sampling*, e discutimos seus impactos nos modelos de classificação. Em seguida, exploramos abordagens recentes na literatura aplicando LLMs para automatizar a anotação dos dados e promover um balanceamento mais efetivo do conjunto de dados.

- **Avaliação do impacto da representação de dados na tarefa de classificação de toxicidade**. Nesta análise, avaliamos o impacto da representação textual na tarefa final de classificação. Para isso, comparamos o desempenho de modelos que utilizam representações convencionais (por exemplo, TF-IDF) com aqueles que incorporam vetores de *embeddings* extraídos de modelos de linguagem estado-da-arte (como ChatGPT).

- **Comparação do desempenho de modelos tradicionais e modelos baseados na arquitetura de *transformers***. Este estudo apresenta a comparação entre modelos tradicionais e aqueles baseados na arquitetura de *transformers*. Em nos-

sos experimentos, o *baseline* foi definido por abordagens mais simples – como a Perspective API e Regressão Logística.

Este capítulo está organizado da seguinte forma: a Seção 5.1 descreve a configuração experimental utilizada em cada um dos cenários analisados. Já a Seção 5.2 apresenta a comparação dos desempenhos entre os modelos de aprendizado propostos neste trabalho. Por fim, consolidamos os principais resultados encontrados (Seção 5.3). Os modelos comerciais considerados neste trabalho foram disponibilizados em junho de 2024, representando, portanto, o estado da arte disponível até aquela data.

5.1 Setup Experimental

Nesta seção, descrevemos as configurações e os parâmetros empregados na condução dos experimentos de treinamento dos modelos de aprendizado de máquina para a tarefa de classificação automática de toxicidade. Inicialmente, apresentamos o conjunto de dados anotado manualmente, bem como as etapas de pré-processamento adotadas. Em seguida, detalhamos o processo de *data augmentation* usando anotações sintéticas geradas por LLMs e a divisão final do conjunto de dados para treino e avaliação dos modelos selecionados. Por fim, discutimos as configurações e os hiperparâmetros definidos para os modelos de classificação automática, além das especificações de *hardware* do ambiente computacional utilizado no processo de treinamento.

Inicialmente, utilizamos o conjunto de dados apresentado na Seção 3.3. Esse conjunto de dados contém 2.500 comentários, sendo 10.5% de comentários *tóxicos* e 89.5% de *não-tóxicos*. Para a avaliar o desempenho dos modelos selecionados, um conjunto de dados de teste com 593 exemplos foi criado a partir da amostra inicial. Os dados de teste incluem somente comentários que foram anotados manualmente para evitar que haja distorção no cálculo das métricas de avaliação. Para o pré-processamento dos dados, foram aplicadas as seguintes etapas: remoção de *stopwords*, eliminação de caracteres repetidos e exclusão de comentários compostos unicamente por risadas [1]. Além disso, realizamos a lematização, com o objetivo de reduzir as palavras à sua forma base. Para lidar com o problema de desbalanceamento das classes, aplicamos os métodos de reamostragem aleatória e SMOTE, utilizando as configurações padrão da implementação do pacote *scikit-learn*.

Os dados de entrada para o modelo são o corpo do comentário em texto e a respectiva classe. Os comentários foram representados usando duas estratégias: TF-IDF

[1] A etapa de pré-processamento dos comentários não foi aplicada no texto na geração dos vetores de embeddings para evitar a mudança de sentido do texto original.

> 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 5.1: Instruções fornecidas ao assistente para classificação de comentários no Reddit usando a abordagem *zero-shot*.

e *word embeddings*. Na abordagem usando TF-IDF, configuramos o *max_features* = 2.000 e *ngram_range* = 3. Para *word embeddings*, utilizamos vetores pré-treinados do modelo *text-embedding-ada-002* da OpenAI com 1.536 dimensões fixas Whitehouse et al. (2023).

Para gerar uma nova amostra de dados com rotulações sintéticas, amostramos 12.000 comentários a partir dos dados apresentados no Capítulo 3 (Seção 3.2), excluindo comentários que já haviam sido selecionados anteriormente para o processo de rotulação manual. O *prompt* utilizado para instruir os LLMs a classificar cada comentário foi inspirado nas instruções fornecidas aos anotadores humanos e é apresentado na Figura 5.1. Os LLMs utilizados nesta etapa foram configurados com os seguintes parâmetros: *temperature*=0 e *max_tokens*=800.

Na condução do processo de treinamento e avaliação, aplicamos a técnica de validação cruzada estratificada com k=5, ou seja, o conjunto de treinamento foi dividido em 5 partições, de forma que, a cada iteração, 4 partições foram utilizadas para o treinamento e 1 para a validação Berrar (2019). Esse procedimento foi repetido 5 vezes, garantindo que cada partição atuasse, ao menos uma vez, como conjunto de validação. Para os modelos comerciais e pré-treinados, adotamos a mesma estratégia, porém, a cada iteração, uma partição distinta foi selecionada para a rotulação automática.

Os modelos usados para a análise de desempenho na tarefa de aprendizado de máquina são descritos a seguir. Os modelos *baseline* são a Regressão Logística e API da Perspective. Para a Regressão Logística, foram utilizados os seguintes parâmetros: (C=1.0, *class_weight*="balanced", *solver*="lbfgs", *max_iter*=100). Como a API da Perspective é um modelo comercial fechado, não temos acesso aos seus parâmetros internos. Para classificar cada comentário, enviamos o texto do corpo do comentário e extraímos o atributo *TOXICITY*, que retorna o escore de probabilidade do comentário possuir conteúdo tóxico. Os LLMs fechados foram configurados usando os mesmos parâmetros dos

Modelo	Taxa de Aprendizado	Tamanho do Lote	Épocas	Decaimento do peso	Passos de aquecimento
BERT	1e-6	2	5	1e-4	50
RoBERTa	2e-6	8	3	1e-3	10

Tabela 5.1: Hiperparâmetros dos modelos baseados em transformers: BERT e RoBERTa.

LLMs abertos apresentados anteriormente nesta seção (*temperature*=0 e *max_tokens*=800). Por fim, para representar os modelos de arquiteturas baseadas em *transformers* de código aberto, selecionamos os modelos BERT e RoBERTa. Para o BERT, ajustamos um modelo pré-treinado em português Souza et al. (2020). O modelo RoBERTa também foi ajustado especificamente para a tarefa de classificação de toxicidade em PT-BR. Os hiperparâmetros foram escolhidos a partir de uma lista de possíveis valores (*Gridsearch*) e usando um conjunto de validação separado (*hold-out*), onde a proporção original das classes não foi alterada para evitar sobreajuste (*overfitting*). Os modelos baseados em *transformers* foram treinados e ajustados utilizando *hardware* especializado com múltiplas GPUs P100, com os hiperparâmetros apresentados na Tabela 5.1.

Para comparar e avaliar os diferentes modelos estudados, os modelos foram avaliados com base na média e no desvio padrão do F1-score obtido em cada partição (*fold*) da validação cruzada. Adicionalmente, foram calculadas métricas intra-classe, com o objetivo de analisar a capacidade dos diferentes modelos em identificar corretamente a classe de comentários *tóxicos*.

5.2 Comparação entre os Modelos de Classificação

Nesta seção, apresentamos o estudo comparativo entre as técnicas adotadas na metodologia proposta e modelos selecionados. Iniciamos a análise com abordagens mais simples aplicadas ao conjunto de dados anotado, e, progressivamente, outros componentes para avaliar o impacto de técnicas avançadas no desempenho dos modelos.

Para investigar a hipótese de que o processo de aumento de dados (*data augmentation*) pode melhorar o desempenho na tarefa de classificação, experimentamos duas técnicas amplamente utilizadas de amostragem: subamostragem aleatória (*random undersampling*) e SMOTE. Estes métodos foram empregados visando equilibrar o conjunto de dados de treino, resultando em uma proporção de comentários de 50% para cada classe. A subamostragem aleatória foi aplicada para reduzir o tamanho da classe majoritária, igualando-a aos 279 comentários da classe *tóxica*, resultando em um conjunto de treino final de aproximadamente 558 exemplos. O SMOTE foi utilizado para aumentar a proporção de exemplos na classe minoritária, gerando exemplos sintéticos com o objetivo de

Modelo	Método de amostragem	Proporção (tóxico/não-tóxico)	Validação cruzada (F1-Macro)	F1-Macro	F1 (tóxico=0)	F1 (tóxico=1)
Regressão Logística	Anotação manual	10.5%:89.5%	58.20% ± 2.227	62.00%	93.00%	31.00%
Regressão Logística	Undersampling	50%:50%	60.99% ± 1.878	61.00%	83.00%	39.00%
Regressão Logística	Oversampling	50%:50%	57.18% ± 1.4392	60.00%	92.00%	27.00%
Regressão Logística	SMOTE	50%:50%	54.12% ± 1.6780	61.00%	92.00%	27.00%
Perspective API	–	–	74.77% ± 1.4209	74.00%	93.00%	56.00%

Tabela 5.2: Métricas de avaliação para os modelos *baseline*.

Modelo	Método de amostragem	Proporção (tóxico/não-tóxico)	Validação cruzada (F1-Macro)	F1-Macro	F1 (tóxico=0)	F1 (tóxico=1)
Regressão Logística	Dados sintéticos - LLMs	22.4%:77.59%	74.53% ± 1.7713	70.00%	91.00%	49.00%
Regressão Logística	Dados sintéticos - LLMs	34.6%:65.33%	75.40% ± 1.0773	74.00%	94.00%	53.00%
Regressão Logística	Dados sintéticos - LLMs + GPT Embeddings	34.6%:65.33%	80.23% ± 1.4234	79.00%	94.00%	64.00%

Tabela 5.3: Métricas de avaliação para o modelo linear com diferentes abordagens de data augmentation e representação semântica a partir de transferência de aprendizado.

alcançar o mesmo número de comentários da classe majoritária, totalizando 4.036 exemplos.

Desbalanceamento dos dados. Para comparar o desempenho dos modelos *baselines* (Regressão Logística e API da Perspective) quanto ao desbalanceamento dos dados, avaliamos o desempenho da Regressão Logística com diferentes configurações de reamostragem e a Perspective API no conjunto de dados de teste. A Tabela 5.2 apresenta os resultados das métricas de avaliação para esses modelos. Como pode ser observado, o modelo treinado no conjunto de dados original alcançou F1-macro de 62%, enquanto o F1-macro para a Perspective API foi 74%. Esse resultado evidencia o impacto negativo do desbalanceamento de classes sobre modelos mais simples, além de destacar o desempenho competitivo de soluções especializadas como a Perspective, mesmo sem ajustes prévios ao nosso conjunto de dados.

Para explorar abordagens clássicas de tratamento do desbalanceamento, realizamos uma subamostragem aleatória da classe majoritária para construir um conjunto de treino balanceado. A Tabela 5.2 apresenta os resultados comparativos. Com a subamostragem da classe *não-tóxica*, obtivemos um F1-macro de 57.18% ± 1.439 no treino e 60.00% no teste. Também avaliamos o uso do SMOTE, técnica que gera exemplos sintéticos da classe minoritária até igualar a proporção com a classe majoritária (*não-tóxico* 50%:50% *tóxico*), resultando em um conjunto de dados aumentado. No entanto, essa abordagem não trouxe ganhos de desempenho: o F1-macro foi de 54.12% ± 1.678 no treino e 61.00% no teste. Esses resultados mostram que métodos tradicionais de balanceamento não foram eficazes neste domínio específico, uma vez que não superaram o desempenho obtido com o conjunto de dados original desbalanceado.

Para expandir as técnicas de aumento de dados, exploramos o uso de LLMs na geração de rótulos sintéticos em PT-BR. Com o objetivo de avaliar o impacto das diferentes estratégias de reamostragem no desempenho dos modelos, conduzimos um experimento

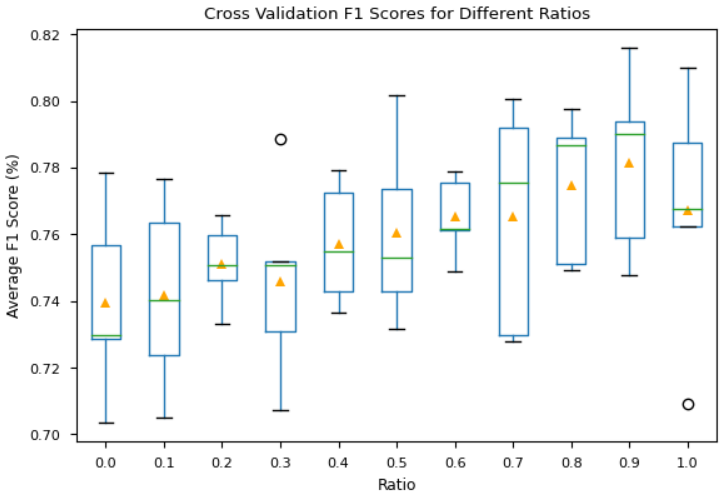

Figura 5.2: F1-Macro em diferentes limiares de decisão para o modelo de melhor desempenho na tarefa de classificação de toxicidade.

em que um novo conjunto de dados foi rotulado com auxílio de LLMs *open-source*, conforme descrito na Seção 5.1. Em seguida, balanceamos o conjunto final aumentando a proporção de comentários tóxicos. Para isso, definimos, via validação cruzada, uma porcentagem da classe majoritária a ser descartada, mantendo todos os exemplos da classe minoritária. Por exemplo, ao fixar a proporção em 0.5, apenas 50% dos comentários classificados como *não-tóxicos* foram incluídos, enquanto todos os tóxicos foram mantidos. Essa abordagem introduz variabilidade em ambas as classes, ao mesmo tempo que reduz o impacto do desbalanceamento.

A Tabela 5.3 apresenta os resultados para diferentes proporções das duas classes de toxicidade. As duas primeiras linhas da tabela apresentam o resultado utilizando a representação TF-IDF, enquanto a terceira linha emprega o uso de vetores de *embeddings*. Na primeira configuração, o modelo alcançou um F1-macro médio de 74.53% $\pm$ 1.773 no conjunto de treino e 70.00% no conjunto de teste. Com uma proporção ainda maior da classe tóxica, o modelo obteve 75.40% $\pm$ 1.0773 no conjunto de treino e 74.00% no conjunto de teste, evidenciando que o aumento da representatividade da classe minoritária contribuiu para um melhor desempenho. Embora o aumento da classe tóxica tenha contribuído para uma melhora no desempenho do modelo, os ganhos se estabilizam a partir de certo ponto, convergindo para os resultados apresentados na segunda linha da Tabela 5.3. As diferentes configurações de proporção foram avaliadas usando validação cruzada e o resultado é mostrado na Figura 5.2. Entre as configurações avaliadas, a proporção 0.6 (65.33% *não-tóxico*: 34.6% *tóxico*) apresentou o melhor desempenho nos experimentos.

Representação dos dados. Para investigar essa questão, examinamos a importância da representação dos dados e seu impacto no desempenho da tarefa em PT-BR. Para testar essa hipótese, utilizamos o modelo *text-embeddings* para extrair embeddings via a

Modelo	Método de amostragem	Proporção (tóxico/não-tóxico)	Validação cruzada (F1-Macro)	F1-Macro	F1 (tóxico=0)	F1 (tóxico=1)
BERT-large (*fine-tuned*)	Anotação manual	10.5%:89.5%	68.50% ± 0.658	69.47%	75.00%	64.00%
BERT-large (*fine-tuned*)	Dados sintéticos - LLMs	34.6%:65.33%	76.40% ± 0.463	77.00%	92.00%	62.00%
RoBERTa (*fine-tuned*)	Dados sintéticos - LLMs	34.6%:65.33%	80.64% ± 1.7316	77.00%	94.00%	60.00%

Tabela 5.4: Métricas de avaliação para os modelos transformers pré-treinados e ajustados para o PT-BR.

API do ChatGPT para cada comentário em nossos conjuntos de treino e teste.[2] Os vetores pré-treinados extraídos a partir desse modelo possuem 1536 dimensões e não houve nenhuma etapa de pós-processamento após a geração dos vetores de *embeddings*. Os resultados, apresentados na terceira linha da Tabela 5.3, indicam que o modelo que utilizou os vetores de *embeddings* de uma modelo pré-treinado como representação dos dados de entrada teve o melhor desempenho geral, com um macro F1-score médio no conjunto de treinamento de 80.23% ± 1.423 e um macro F1-score no conjunto de teste de 79.00%. Em outras palavras, o modelo se beneficiou significativamente da transferência de conhecimento para melhorar a classificação dos comentários.

Modelos de Arquitetura Transformers com Ajuste de Parâmetros. Ao comparar o desempenho do *baseline* com os dados anotados apresentados na Tabela 5.2, podemos notar que tanto o BERT quanto o RoBERTa apresentam resultados de F1-macro superiores mesmo com os dados desbalanceados (69.47% vs 62.00%). Além disso, ao incorporar dados sintéticos ao processo de ajuste fino dos modelos BERT e RoBERTa, pode-se observar uma melhoria em termos de F1-macro de mais de 7 pontos percentuais (77% vs 69.47%). Esse resultado é apresentado na Tabela 5.4 e indica que o processo de ajuste fino de um modelo pré-treinado traz benefícios para essa tarefa em PT-BR. Entretanto, vale ressaltar que o processo de ajuste fino é essencial para adaptar os pesos do modelo para à tarefa específica, uma vez que os dados de comentários do Reddit podem possuir características distintas de outras fontes de dados que são normalmente utilizadas no processo de treinamento de modelos de linguagens abertos Souza et al. (2020); Devlin et al. (2019).

Grandes Modelos de Linguagem. Para avaliar os modelos de linguagem (LLMs) na tarefa de classificação de toxicidade, conduzimos um experimento de aprendizado *zero-shot*, ou seja, usamos apenas instruções via *prompt* para guiar os modelos na tarefa desejada. Os resultados são apresentados na Tabela 5.5. Como podemos observar, o Chat-GPT-3.5-turbo teve desempenho semelhante ao *baseline* usando Regressão Logística (F1-macro de 62.00%), enquanto o Gemini-1.0-pro (modelo proprietário do Google) teve resultado superior, com 70.00% de F1-macro, porém ainda com desempenho inferior à API da Perspective. Por fim, o modelo mais complexo avaliado teve o melhor desempenho no conjunto

[2]Modelos de *embeddings* disponíveis em: https://platform.openai.com/docs/guides/embeddings

Modelo	Precisão	Recall	F1-Macro	F1 (tóxico=0)	F1 (tóxico=1)
Chat-GPT-3.5-turbo	64.00%	78.00%	62.00%	78.00%	46.00%
Chat-GPT-4-turbo	80.00%	77.00%	78.00%	94.00%	62.00%
Gemini-1.0-pro	69.00%	74.00%	70.00%	90.00%	49.00%

Tabela 5.5: Métricas de avaliação para os modelos de linguagem (LLMs) comparados.

Modelo	Método de amostragem	Proporção (tóxico/não-tóxico)	Validação cruzada (F1-macro)	F1-macro	F1 (tóxico=0)	F1 (tóxico=1)
Perspective API	–	–	74.77% ± 1.4209	74.00	93.00	56.00
Regressão Logística	Dados sintéticos - LLMs + GPT Embeddings	34.6%:65.33%	80.23% ± 1.4234	79.00%	94.00%	**64.00%**
RoBERTa (*fine-tuned*)	Dados sintéticos - LLMs	34.6%:65.33%	**80.64% ± 1.7316**	77.00%	94.00%	60.00%
Chat-GPT-4-turbo	–	–	–	78.00%	94.00%	62.00%

Tabela 5.6: Comparação entre os melhores modelos de cada experimento.

de teste. O Chat-GPT-4-turbo é o modelo mais avançado da família ChatGPT [3] e teve F1-macro de 78.00%, o melhor resultado dentre os LLMs avaliados, resultado superior aos *baselines* definidos previamente para este experimento. Esse resultado indica que treinar modelos cada vez maiores - em número de parâmetros e dados de treinamento - traz benefícios para o desempenho dos LLMs em na tarefa de classificação de toxicidade.

5.3 Discussão dos Resultados

Nessa seção, discutimos os principais resultados do processo de modelagem para a tarefa de classificação de toxicidade com dados do Reddit em PT-BR.

A Tabela 5.6 apresenta os melhores resultados obtidos em cada experimento descrito neste capítulo. Em linhas gerais, considerando a métrica F1-macro (conjunto de teste), os resultados de todos os modelos propostos são mais precisos do que os resultados da API Perspective (*baseline*). Além disso, ao considerar especificamente a capacidade de identificar corretamente a classe de interesse (*tóxico*=1), o modelo de Regressão Logística treinado com dados sintéticos e representação vetorial destacou-se como o mais eficaz dentre os métodos avaliados, evidenciando o potencial da metodologia proposta.

Em uma análise mais detalhada, observa-se que o problema de desbalanceamento dos dados impacta fortemente modelos de aprendizado de máquina treinados apenas com o conjunto de dados anotados manualmente. Uma das possíveis explicações é o fato do conjunto de dados ser relativamente pequeno para fins de generalização para esta tarefa, já que possui apenas 270 comentários classificados previamente como *tóxicos* pelos anotadores, resultando em uma razão próxima de 9:1 (*não-tóxico*:*tóxico*).

[3] À época do experimento, este era o modelo mais avançado disponibilizado pela OpenAI.

Dentre as abordagens analisadas para lidar com esse problema, a utilização de modelos especializados se mostrou uma alternativa eficaz. Os resultados mostram que o modelo se adapta relativamente bem para dados em PT-BR, uma vez que teve desempenho muito superior ao *baseline* usando Regressão Logística treinado com os dados desbalanceados (74.00% vs 62.00%). Esse resultado indica um desempenho competitivo do modelo da Perspective para essa tarefa, mesmo sem ser treinado com dados específicos do Reddit em Português.

Embora o treinamento do modelo *baseline* usando Regressão Logística no conjunto de dados desbalanceado não tenha produzido resultados satisfatórios, os modelos pré-treinados, como BERT e RoBERTa, demonstraram desempenho competitivo, evidenciando maior robustez ao desbalanceamento. Outra abordagem comumente utilizada para lidar com a disparidade entre as classes é a aplicação de estratégias de reamostragem, como subamostragem e SMOTE. No entanto, enquanto a subamostragem pode resultar na perda significativa de dados reais (essenciais para o processo de aprendizado dos modelos), o SMOTE cria amostras artificiais que não necessariamente representam exemplos factíveis. Por esses motivos, os modelos lineares treinados somente com dados reamostrados tiveram desempenho inferior no conjunto de teste e não foram capazes de superar a Regressão Logística. Esse resultado indica que i) a perda de dados ao realizar subamostragem é prejudicial ao modelo, uma vez que o tamanho do conjunto de dados reamostrado é significativamente inferior e ii) a criação de exemplos sintéticos de forma aleatória também não influencia positivamente na capacidade de classificação dos modelos.

Embora as técnicas tradicionais de reamostragem não tenham demonstrado desempenho satisfatório, gerar anotações sintéticas usando LLMs melhorou o desempenho no geral, mesmo nos modelos mais simples (Regressão Logística). Isso indica que modelos de linguagem pré-treinados podem ser aplicados para gerar rótulos de qualidade, reduzindo o custo e tempo para anotar um conjunto de dados manualmente.

A representação dos dados se mostrou influente no processo de treinamento e na melhoria geral do desempenho do modelo. Ao comparar técnicas clássicas, como TF-IDF, com vetores de *embeddings* pré-treinados, o modelo de classificação teve resultado superior (70.00% vs 79.00%). Esse resultado evidencia que treinar modelos híbridos a partir da transferência de representação é um caminho promissor para treinar modelos que sejam eficientes em termos de latência e desempenho.

Como vimos nessa seção, os LLMs fechados apresentam bom desempenho para essa tarefa sem precisarem de adaptação para esse conjunto de dados específico. O melhor modelo fechado alcançou F1-macro de 78% (ChatGPT-4-turbo), enquanto os outros modelos fechados tiveram desempenho competitivo. Embora os LLMs representem uma abordagem promissora para a classificação de toxicidade, esses modelos enfrentam limitações quanto à escalabilidade, tanto em termos de custo de requisição quanto à quantidade

de acessos. Em contraste, a metodologia proposta neste trabalho utiliza modelos abertos, que permitem classificações em larga escala — seja em lote ou em tempo real — com desempenho comparável ao dos melhores modelos fechados. Esse resultado reforça a vantagem do uso de modelos abertos especializados para tarefas específicas, sendo uma alternativa viável para auxiliar na moderação automática de conteúdo gerado em redes sociais online.

Capítulo 6

Conclusão e Trabalhos Futuros

As redes sociais online (RSOs) desempenham um papel cada vez mais central na vida dos usuários. Com a crescente conectividade, essas plataformas não apenas facilitam a interação entre pessoas, mas também exercem influência significativa sobre suas opiniões e comportamentos. No entanto, o papel da rede social em integrar e conectar pessoas tem sido desafiado pela intensificação de interações tóxicas e prejudiciais, frequentemente impulsionadas por divergências em temas diversos como política, religião e questões pessoais. Adicionalmente, a crescente interconexão entre o ambiente virtual e eventos do mundo real contribui para a rápida polarização das redes, amplificando discursos tóxicos e agravando tensões nas comunidades de discussão online.

O principal objetivo dessa dissertação é analisar e caracterizar o conteúdo tóxico nas comunidades brasileiras no Reddit. Neste contexto, esta dissertação focou principalmente nas seguintes perspectivas: i) coleta e anotações de um conjunto de dados de toxicidade em PT-BR ii) estudo de caracterização para entender as diferenças linguísticas no discurso tóxico e iii) a proposta de uma metodologia para o treinamento de modelos de aprendizado de máquina escaláveis para detecção automática de conteúdo tóxico. Ao viabilizar o treinamento de novos modelos de detecção de toxicidade em PT-BR, buscamos contribuir para a moderação proativa em comunidades online, promovendo a liberdade de expressão dos usuários e mitigando os impactos de comportamentos tóxicos.

Para endereçar a QP1 e entender as diferenças de linguagem entre os comentários tóxicos e não tóxicos, realizamos a caracterização linguística dos comentários para as comunidades selecionadas. A caracterização envolveu diversas análises específicas de processamento de linguagem natural (PLN), como *POS tagging*, modelagem de tópicos, extração de termos frequentes e reconhecimento de entidades nomeadas (REN). Ao analisar as etiquetas de POS, verificamos que os comentários classificados como tóxicos usam mais adjetivos e substantivos em média, além de terem muitas ocorrências do termo *mulher* e expressões ofensivas, o que pode indicar ataques direcionados a grupos específicos. Além disso, termos relacionados à guerra Rússia-Ucrânia e aos candidatos e partidos envolvidos na eleição de 2022 são muito frequentes para as duas classes de toxicidade. Em relação às entidades nomeadas, observamos que os comentários tóxicos em média mencionam maior quantidade de termos associados com *pessoas* e *localização*, enquanto *organizações* são em

média mencionadas mais frequentemente por comentários não tóxicos. A caracterização textual nos permitiu observar as diferenças linguísticas entre comentários tóxicos e não tóxicos. Por fim, foi verificado que eventos externos impactam significativamente as discussões mantidas no ambiente online, uma vez que grandes eventos que ocorrem no Brasil e no mundo são frequentemente mencionados nas discussões da plataforma.

Para avaliar se o conjunto de dados proposto por este trabalho é útil para treinar modelos de detecção de toxicidade em larga escala, comparamos tanto modelos comerciais quanto de código aberto. Entre os modelos proprietários, usamos a API Perspective, o Gemini e o ChatGPT. Entre os modelos abertos, testamos desde modelos simples (Regressão Logística) até *transformers* (como BERT) e LLMs (Llama, Sabiá e Mistral). O objetivo foi comparar modelos específicos para essa tarefa com modelos pré-treinados, avaliar o desempenho de modelos simples em contraste com modelos complexos e entender o impacto do desbalanceamento de classes e da representação dos dados. Nossos resultados indicam que gerar exemplos sintéticos com LLMs para equilibrar as classes melhora o desempenho dos modelos. Essa redução no desbalanceamento das classes contribui para melhorar a capacidade de generalização tanto de modelos simples quanto de modelos *transformers*. Além disso, o uso de vetores densos na representação textual teve um impacto positivo no desempenho dos classificadores. Nossos resultados indicam que o uso de modelos abertos para a tarefa de classificação é eficaz, especialmente considerando seu baixo custo operacional e a capacidade de escalabilidade para grandes volumes de dados. Essas características são fundamentais para viabilizar a moderação automática de conteúdo em redes sociais.

Reconhecemos algumas limitações deste estudo, especialmente quanto à subjetividade envolvida na rotulação de conteúdo *tóxico* em um ambiente contextualmente limitado, como redes sociais online. Para atenuar esse problema, pretendemos repetir o experimento de anotação incluindo informações contextuais adicionais em casos específicos. Também é importante considerar o possível viés introduzido pelo processo de amostragem e pela diversidade dos anotadores. Além disso, embora tenhamos priorizado o uso de LLMs open-source na geração de anotações sintéticas visando reprodutibilidade e baixo custo, a etapa de transferência de representação utilizou *embeddings* de um modelo comercial (ChatGPT). Futuramente, pretendemos explorar *embeddings* de modelos abertos, especialmente aqueles compatíveis com múltiplas línguas ou específicos para o português.

Este trabalho explorou a caracterização e a detecção de conteúdo *tóxico* em comunidades brasileiras no Reddit, com foco no período de 2022. No entanto, alguns aspectos ficaram fora do escopo desta análise e podem ser abordados em pesquisas futuras. As principais direções para trabalhos posteriores estão listadas a seguir:

- **Incorporar atributos das interações em modelos de detecção automática:** Utilizar informações contextuais das interações, como tempo de resposta, número

de votos e estrutura da conversa, para enriquecer os modelos de classificação. Os dados coletados do Reddit são ricos em metadados sobre cada publicação e esses atributos podem ser úteis para enriquecer o treinamento de modelos de aprendizado de máquina.

- **Reproduzir o estudo de caracterização para dados recentes:** Atualizar a análise com dados mais atuais das comunidades brasileiras no Reddit, avaliando se padrões de toxicidade se mantêm ou evoluíram ao longo do tempo.

- **Anotação de novos conjuntos de dados com base em perspectivas:** Construir conjunto de dados com múltiplos pontos de vista, além de caracterizar os anotadores que participarem do processo. Esse estudo visa analisar e propor medidas para redução da subjetividade da tarefa de anotação manual.

- **Representação vetorial a partir de modelos abertos:** Como uma limitação da atual pesquisa, utilizamos os vetores pré-treinados extraídos de modelos comerciais da OpenAI. Como trabalho futuro, é interessante explorar *embeddings* gerados por modelos de linguagem abertos, como BERT ou RoBERTa, além de LLMs de código aberto (como Llama e Mistral) para representar textos de forma mais escalável e com baixo custo.

- **Explorar alternativas para classificação de toxicidade em larga escala:** Investigar abordagens eficientes para aplicação em grandes volumes de dados, como distilação de modelos (*knowledge distillation*, quantização ou uso de classificadores leves). Além disso, o uso de LLMs menores pode ser uma alternativa viável para detecção em tempo real.

- **Detecção de gatilhos de toxicidade:** Identificar eventos, temas ou padrões linguísticos que atuam como gatilhos para o surgimento de comentários tóxicos em discussões online. A partir de um modelo de classificação de toxicidade, podemos identificar comentários que possivelmente podem desencadear outras interações tóxicas. Agir proativamente para moderar essas interações é essencial para garantir um ambiente online mais seguro para todos os usuários.

Referências Bibliográficas

Thales Sales Almeida, Hugo Queiroz Abonizio, Rodrigo Frassetto Nogueira, and Ramon Pires. Sabiá-2: A new generation of portuguese large language models. *ArXiv*, abs/2403.09887, 2024. URL https://api.semanticscholar.org/CorpusID:268509975.

Hind Almerekhi, Haewoon Kwak, Bernard J Jansen, and Joni Salminen. Detecting toxicity triggers in online discussions. In *Proceedings of the 30th ACM conference on hypertext and social media*, pages 291–292, 2019.

Hind Almerekhi, Haewoon Kwak, Joni O. Salminen, and Bernard Jim Jansen. Are these comments triggering? predicting triggers of toxicity in online discussions. *Proceedings of The Web Conference 2020*, 2020. URL https://api.semanticscholar.org/CorpusID:215870543.

Hind Almerekhi, Haewoon Kwak, and Bernard J Jansen. Investigating toxicity changes of cross-community redditors from 2 billion posts and comments. *PeerJ Computer Science*, 8:e1059, 2022a.

Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J Jansen. Provoke: Toxicity trigger detection in conversations from the top 100 subreddits. *Data and Information Management*, 6(4):100019, 2022b.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

Mohammad Belal, James She, and Simon Wong. Leveraging chatgpt as text annotation tool for sentiment analysis. *arXiv preprint arXiv:2306.17177*, 2023.

Daniel Berrar. Cross-validation. In Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 542–545. Academic Press, Oxford, 2019. ISBN 978-0-12-811432-2. doi: https://doi.org/10.1016/B978-0-12-809633-8.20349-X. URL https://www.sciencedirect.com/science/article/pii/B978012809633820349X.

Shelley Boulianne, Christian P Hoffmann, and Michael Bossetta. Social media platforms for politics: A comparison of facebook, instagram, twitter, youtube, reddit, snapchat, and whatsapp. *New Media & Society*, page 14614448241262415, 2024.

Tom Brown, 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, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Michael Buckland and Fredric Gey. The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19, 1994.

Navoneel Chakrabarty. A machine learning approach to comment toxicity classification. In *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019*, pages 183–193. Springer, 2020.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. 2021.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yun Yu Chong and Haewoon Kwak. Understanding toxicity triggers on reddit in the context of singapore. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1383–1387, 2022.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

Rogers Prates de Pelle and Viviane P Moreira. Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Ashwin Geet d'Sa, Irina Illina, and Dominique Fohr. Bert and fasttext embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*, pages 1–5. IEEE, 2020.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104, 2019.

Lihao Ge and Teng-Sheng Moh. Improving text classification with word embedding. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1796–1805. IEEE, 2017.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023.

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

Preni Golazizian, Alireza Salkhordeh Ziabari, Ali Omrani, and Morteza Dehghani. Cost-efficient subjective task annotation and modeling through few-shot annotator adaptation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3474–3491, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.199. URL https://aclanthology.org/2024.findings-emnlp.199/.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945, 2020.

Anke Görzig and Kjartan Ólafsson. What makes a bully a cyberbully? unravelling the characteristics of cyberbullies across twenty-five european countries. *Journal of Children and Media*, 7(1):9–27, 2013.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Samuel S. Guimarães, Julio C. S. Reis, Filipe N. Ribeiro, and Fabrício Benevenuto. Characterizing toxicity on facebook comments in brazil. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '20, page 253–260, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381963. doi: 10.1145/3428658.3430974. URL https://doi.org/10.1145/3428658.3430974.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.

Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

Jasser Jasser, Ivan Garibay, Steve Scheinert, and Alexander V Mantzaris. Controversial information spreads faster and further than non-controversial information in reddit. *Journal of Computational Social Science*, 5(1):111–122, 2022.

Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35, 2019.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. Understanding international perceptions of the severity of harmful content online. *PloS one*, 16(8):e0256762, 2021.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

Jordan Kobellarz and Thiago Silva. Should we translate? evaluating toxicity in online comments when translating from portuguese to english. In *Anais do XXVIII Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 95–104, Porto Alegre, RS, Brasil, 2022. SBC. URL https://sol.sbc.org.br/index.php/webmedia/article/view/22110.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications, 2004.

Sanjay Kukreja, Tarun Kumar, Amit Purohit, Abhijit Dasgupta, and Debashis Guha. A literature survey on open source large language models. In *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, pages 133–143, 2024.

Deepak Kumar, T. Hancock, Jeff Kurt Thomas, and Zakir Durumeric. Understanding longitudinal behaviors of toxic accounts on reddit. In *Proceedings of the Web Conference 2023*, WWW '23. Association for Computing Machinery, 2023. doi: 10.1145/3543507.3583522. Published: 30 April 2023.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878, 2024.

Marc-André Larochelle and Richard Khoury. Generalisation of cyberbullying detection. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 296–300. IEEE, 2020.

Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3197–3207, 2022.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Kam-Fai Wong, Kevin Knight, and Hua Wu, 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, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.aacl-main.91. URL https://aclanthology.org/2020.aacl-main.91/.

Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. Toxic content detection in online social networks: a new dataset from brazilian reddit communities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 472–482, 2024a.

Luiz Henrique Quevedo Lima, Ana Clara Souza Pagano, Adriana Silvina Pagano, and Ana Paula Couto da Silva. Rotulaçao e caracterizaçao de conteúdo tóxico de comunidades do reddit no brasil. *Linguamática*, 16(2):201–218, 2024b.

Paula Reyero Lobo, Enrico Daga, and Harith Alani. Supporting online toxicity detection with knowledge graphs. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1414–1418, 2022.

Kristína Machová, Marián Mach, and Kamil Adamišín. Machine learning and lexicon approach to texts processing in the detection of degrees of toxicity in online discussions. *Sensors*, 22(17):6468, 2022.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282, 2012.

Larry R Medsker, Lakhmi Jain, et al. Recurrent neural networks. *Design and Applications*, 5(64-67):2, 2001.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (ICLR 2013), Workshop Track*, pages 1–13, 2013. URL https://research.google/pubs/efficient-estimation-of-word-representations-in-vector-space/.

Shyamal Mishra and Preetha Chatterjee. Exploring chatgpt for toxicity detection in github. In *Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*, pages 6–10, 2024.

Shagofah Noor, Omid Tajik, and Jawad Golzar. Simple random sampling. *International Journal of Education & Language Studies*, 1(2):78–82, 2022.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194: 151–175, 2013.

Amanda S Oliveira, Thiago C Cecote, Pedro HL Silva, Jadson C Gertrudes, Vander LS Freitas, and Eduardo JS Luz. How good is chatgpt for detecting hate speech in portuguese? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 94–103. SBC, 2023.

Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

Cristian Padurariu and Mihaela Elena Breaban. Dealing with data imbalance in text classification. *Procedia Computer Science*, 159:736–745, 2019.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Perspective. Using machine learning to reduce toxicity online. https://perspectiveapi.com/how-it-works/, 2022. Acessado em: 08/04/2023.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL https://aclanthology.org/L12-1115/.

Giovana Piorino, Vitor Moreira, Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. Análise de sentimentos de conteúdo compartilhado em comunidades brasileiras do reddit: Avaliação de um conjunto de dados rotulados por humanos. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 54–62. SBC, 2024.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy, September 2017. URL http://aclweb.org/anthology/W17-6523.

World Population Review. Reddit users by country. https://worldpopulationreview.com/country-rankings/reddit-users-by-country, 2023. Acessado em: 14/03/2025.

Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

Vaibhav Rupapara, Furqan Rustam, Hina Fatima Shahzad, Arif Mehmood, Imran Ashraf, and Gyu Sang Choi. Impact of smote on imbalanced text features for toxic comments classification using rvvc model. *IEEE Access*, 9:78621–78634, 2021.

Nazanin Salehabadi, Anne Groggel, Mohit Singhal, Sayak Saha Roy, and Shirin Nilizadeh. User engagement and the toxicity of tweets. *arXiv preprint arXiv:2211.03856*, 2022.

Philip Sedgwick. Pearson's correlation coefficient. *Bmj*, 345, 2012.

S Selva Birunda and R Kanniga Devi. A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*, pages 267–281, 2021.

Oyesh Mann Singh, Sandesh Timilsina, Bal Krishna Bal, and Anupam Joshi. Aspect based abusive sentiment detection in nepali social media texts. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 301–308. IEEE, 2020.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer, 2020.

spaCy. Portuguese models. https://spacy.io/models/pt#pt_core_news_lg, 2022. Acessado em: 11/04/2023.

Statista. Number of social media users worldwide from 2017 to 2027. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, 2022. Acessado em: 08/04/2023.

Aixin Sun, Ee-Peng Lim, and Ying Liu. On strategies for imbalanced text classification using svm: A comparative study. *Decision Support Systems*, 48(1):191–201, 2009.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.54. URL https://aclanthology.org/2024.emnlp-main.54/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Douglas Trajano, Rafael Bordini, and Renata Vieira. Olid-br: offensive language identification dataset for brazilian portuguese. *Lang Resources & Evaluation*, 2023.

Christopher Tuckwood. Hatebase: Online database of hate speech. *The Sentinal Project. Available at: https://www. hatebase. org*, 2017.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.777.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Emily A Vogels. The state of online harassment. *Pew Research Center*, 13:625, 2021.

Congcong Wang, Paul Nulty, and David Lillis. A comparative study on word embeddings in deep learning for text classification. In *Proceedings of the 4th international conference on natural language processing and information retrieval*, pages 37–46, 2020.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. LLM-powered data augmentation for enhanced cross-lingual performance. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.44. URL https://aclanthology.org/2023.emnlp-main.44/.

Maximilian Wich, Adrian Gorniak, Tobias Eder, Daniel Bartmann, Burak Enes Cakici, and Georg Groh. Introducing an abusive language classification framework for telegram to investigate the german hater community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1133–1144, 2022.

J. Wise. Reddit users: How many people use reddit in 2023? https://earthweb.com/how-many-people-use-reddit/, 2023. Acessado em: 08/04/2023.

Dong Dong Xu and Shao Bo Wu. An improved tfidf algorithm in text classification. *Applied Mechanics and Materials*, 651:2258–2261, 2014.

Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. Multilingual content moderation: A case study on Reddit. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3828–3844, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.276. URL https://aclanthology.org/2023.eacl-main.276/.

Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. Measuring and characterizing hate speech on news websites. In *Proceedings of the 12th ACM Conference on Web Science*, pages 125–134, 2020.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. URL http://arxiv.org/abs/2303.18223.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In Sheng Li, Maosong Sun, Yang Liu, Hua

Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China. URL https://aclanthology.org/2021.ccl-1.108/.