Análise Comparativa de LLMs para Detecção de Racismo, Sexismo e Homofobia em Redes Sociais

Guilherme Bou

Universidade Federal de Uberlândia
Faculdade de Computação
Bacharelado em Sistemas de Informação

Monte Carmelo - MG
2025

Guilherme Bou

Análise Comparativa de LLMs para Detecção de Racismo, Sexismo e Homofobia em Redes Sociais

Trabalho de Conclusão de Curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Área de concentração: Sistemas de Informação

Orientador: Prof. Dr. Adriano Mendonça Rocha

Monte Carmelo - MG
2025

Este trabalho é dedicado, primeiramente, a mim, pela dedicação e resiliência por ter passado grande parte da graduação em outro estado, longe da família e dos amigos. Dedico também ao meu pai, minha mãe, meus irmãos, minha avó, minha namorada, aos amigos de longa data e àqueles que conheci ao longo desta jornada acadêmica, por sempre estarem presente de alguma forma e me apoiarem.

Agradecimentos

Agradeço à minha família, amigos, professores, colegas e tantos outros que fizeram parte dessa minha jornada acadêmica. Não foi fácil, iniciei a graduação semanas antes do estouro da pandemia da COVID-19, onde tive apenas uma semana de aulas presenciais antes da suspensão total das atividades. Foi um período conturbado, repleto de incertezas, dúvidas e ansiedades.

Antes mesmo de ingressar na universidade, a rotina já era intensa, com trabalho, cursinho, uma jornada que me mantinha fora de casa das 7h às 22h, sempre com a cabeça cheia e a ansiedade pelas incertezas da vida. Durante a pandemia, conciliar o trabalho com o ensino a distância em período integral foi um grande desafio, o que acabou atrasando minha trajetória acadêmica. O mais importante é que, ao retornar ao ensino presencial, reencontrei o ânimo para concluir a graduação. Toda a jornada que enfrentei foi essencial para o meu amadurecimento, fortalecimento, consistência, aprendizado, aperfeiçoamento, construção de relações e compreensão das diferentes etapas da vida.

"É preciso estar atento e forte."
(Letra de Divino, maravilhoso - Gilberto Gil / Caetano Veloso)

Abstract

This study investigated the effectiveness of large language models (LLMs), including GPT-3.5-Turbo, GPT-4.0, DeepSeek-V3, and Gemini-2.0-Flash, in detecting hate speech on social media, focusing on three main categories: homophobia, sexism, and racism. The primary objective was to compare the performance of these models in real-world scenarios, analyzing both raw and pre-processed comments to identify the optimal balance between accuracy, cost, and computational efficiency. The methodology involved collecting data from real-world datasets, totaling over 2,000 manually labeled comments. Pre-processing techniques were then applied to assess their impact on detecting linguistic nuances, and standardized prompts were implemented for classification.

The results indicated that DeepSeek-V3 provided the best balance between performance and cost, achieving an average F1-score of 93.37% and excellent performance on homophobia (99.63%), with a cost of US$ 0.66, significantly lower than GPT-4.0 (US$ 26.99). Although GPT-4.0 achieved the highest overall accuracy (94.19%), its high cost makes large-scale use impractical. Gemini-2.0-Flash, while the most economical (US$ 0.12), showed inconsistent performance, particularly on pre-processed racism comments (46.15% F1-score). It was observed that traditional pre-processing generally reduced the F1-score by 6.93%, except for the racism category in GPT models, where detection improved.

Keywords: LLMs, Racism, Homophobia, Sexism, Detection.

Resumo

Este trabalho investigou a eficácia de modelos de linguagem de grande escala (LLMs), incluindo os modelos GPT-3.5-Turbo, GPT-4.0, DeepSeek-V3 e Gemini-2.0-Flash, na detecção de discursos de ódio em redes sociais, com foco em três categorias principais: homofobia, sexismo e racismo. O objetivo central foi comparar o desempenho desses modelos em cenários reais, analisando tanto comentários brutos quanto pré-processados, para identificar a melhor relação entre precisão, custo e eficiência computacional. A metodologia adotada envolveu a coleta de dados de bases reais, totalizando mais de 2.000 comentários rotulados manualmente. Em seguida, aplicou-se técnicas de pré-processamento para avaliar o impacto na detecção de nuances linguísticas e implementou-se *prompts* padronizados para classificação. Os resultados indicaram que o DeepSeek-V3 apresentou o melhor equilíbrio entre desempenho e custo, alcançando F1-*score* médio de 93,37% e excelente desempenho em homofobia (99,63%), com custo de US$ 0,66, muito inferior ao do GPT-4.0 (US$ 26,99). Embora o GPT-4.0 tenha obtido a maior precisão agregada (94,19%), seu alto custo inviabiliza o uso em larga escala. O Gemini-2.0-Flash, apesar de ser o mais econômico (US$ 0,12), apresentou desempenho inconsistente, sobretudo em racismo pré-processado (46,15% de F1-*score*). Observou-se que o pré-processamento tradicional, em geral, reduziu o F1-*score* em 6,93%, exceto para a categoria de racismo nos modelos GPTs, em que houve melhoria na detecção.

Palavras-chave: LLMs, Racismo, Homofobia, Machismo, Detecção.

Lista de ilustrações

Lista de tabelas

Lista de siglas

API *Application Programming Interface*

CSV *Comma Separated Values*

CNNs *Convolutional Neural Networks*

GPT *Generative Pretrained Transformer*

HOT *Hateful, Offensive and Toxic*

ID *Identity*

LLMs *Large Language Models*

LSTM *Long Short-Term Memory*

LGBTQIAP+ Lésbicas, Gays, Bissexuais, Transgêneros, Queer, Intersexo, Assexuais e todas as outras identidades de gênero e orientações sexuais

NLP *Natural Language Processing*

NLTK *Natural Language Toolkit*

RNNs *Recurrent Neural Networks*

Sumário

Capítulo 1

Introdução

A expansão das redes sociais ampliou o alcance das interações digitais, o que inclui também a presença de discursos de ódio, como racismo, sexismo e homofobia. O discurso de ódio *online* é definido como qualquer comunicação que menospreze uma pessoa ou um grupo com base em características como cor, etnia, gênero, orientação sexual, nacionalidade, religião ou afiliação política (ZHANG; LUO, 2019).

Em reflexo desse problema, crimes de ódio na *internet* no Brasil atingiram mais de 74 mil casos em 2022, o maior número desde 2017, segundo a Central Nacional de Denúncias de Crimes Cibernéticos, da SaferNet. Esses dados constam no Observatório Nacional dos Direitos Humanos (ObservaDH), do Ministério dos Direitos Humanos e da Cidadania (MDHC)[1]. Entre 2017 e 2022, a plataforma registrou 293,2 mil denúncias de crimes motivados por preconceito ou intolerância contra identidade, orientação sexual, gênero, etnia, nacionalidade ou religião. Esses crimes incluem ofensas, ameaças, difamações, incitações à violência e divulgação de conteúdos humilhantes. Durante o período, a misoginia[2] foi o crime de ódio que mais cresceu, de 961 denúncias em 2017 para 28.679 em 2022, um aumento de quase 30 vezes (SaferNet Brasil; Ministério dos Direitos Humanos e da Cidadania, 2023).

O discurso de ódio proliferado na *internet* sempre foi um tema amplamente debatido no Brasil. Mesmo com a regulamentação estabelecida pela Lei nº 12.965/2014 (Brasil – Presidência da República, 2014), conhecida como Marco Civil da *internet*, que assegura princípios, direitos e deveres no uso da *internet* no país, os dados alarmantes apontados por pesquisas (SaferNet Brasil; Ministério dos Direitos Humanos e da Cidadania, 2023) indicam que essa legislação não tem surtido o efeito esperado no combate a esse problema.

[1] Dados SaferNet. Disponível em: <https://experience.arcgis.com/experience/6a0303b2817f482ab550dd024019f6f5/page/Enfrentamento-ao-discurso-de-%C3%B3dio/>

[2] O Dicionário Houaiss define misoginia como "ódio ou aversão às mulheres, aversão ao contato sexual com as mulheres" (**??**).

1.1 Contextualização

Nos últimos anos, com a popularização da *internet*, as pessoas de todas as idades estão frequentemente expostas a conteúdos impróprios, como racismo, homofobia e machismo. De acordo com Bliuc et al. (2018), as formas pelas quais a *internet* pode facilitar a expressão e disseminação de visões e ideologias racistas têm sido objeto de um crescente corpo de pesquisa em várias disciplinas. O racismo pode se manifestar em diferentes níveis, incluindo interpessoal, institucional e cultural, e vivenciar o racismo pode levar a uma série de consequências negativas, como depressão, hipertensão e doenças cardíacas coronárias (BLIUC et al., 2018). Ademais, com a utilização crescente da *internet* para estabelecer novas relações sociais e buscar apoio *online*, especialmente entre os adolescentes, observa-se um aumento na exposição a riscos de vitimização sexual *online* e conteúdos prejudiciais, como homofobia e discriminação (GÁMEZ-GUADIX; INCERA, 2021).

1.2 Motivação

Diante das preocupações geradas pela produção de conteúdo impróprio na *internet*, as principais plataformas Web ligadas a redes sociais, como *Facebook*, *YouTube*, *Twitter (X)*, *Reddit* e *Instagram*, têm usado inteligência artificial para moderar, monitorar e remover conteúdos prejudiciais ou ilegais. No entanto, esses sistemas automatizados têm sido criticados por questões como falta de transparência, viés e possíveis danos a comunidades marginalizadas. Além dessas preocupações, há uma ênfase na moderação de conteúdo por parte das plataformas. Essa abordagem muitas vezes não é suficiente para suprimir o conteúdo prejudicial e evitar que o mesmo chegue até seus usuários (WANG et al., 2023).

Esta pesquisa também se motiva pela continuação do trabalho de Bou et al. (2023), mas com um olhar mais amplo e aplicado investigando o comportamento de modelos de linguagem de larga escala na detecção de discurso de ódio em comentários reais de usuários da *internet*. Este estudo busca analisar o desempenho desses modelos em cenários práticos, onde fatores como ambiguidade linguística, contexto cultural e sarcasmo podem influenciar significativamente os resultados.

1.3 Objetivo

Este estudo tem como objetivo desenvolver e avaliar uma abordagem para a detecção de discursos de ódio utilizando modelos de linguagem em larga escala *Large Language Models* (LLMs), com foco especial nos modelos GPT-4.0 e GPT-3.5-Turbo[3], DeepSeek-V3[4] e Gemini-2.0-Flash[5]. Uma proposta similar já havia sido explorada em um estudo preli-

[3] Modelos GPT-4.0 e GPT-3.5-Turbo disponíveis em: <https://platform.openai.com/docs/models>
[4] Modelo DeepSeek-V3 disponível em: <https://api-docs.deepseek.com/>
[5] Modelo Gemini-2.0-Flash disponível em: <https://ai.google.dev/gemini-api/docs>

minar, publicado anteriormente, no qual foi avaliada a viabilidade do desenvolvimento de uma extensão de navegador para detectar conteúdos preconceituosos na *web*. Embora o cenário fosse mais restrito e com menor profundidade, a análise qualitativa aplicada a 30 *sites* segmentados por temática (homofobia, racismo e sexismo) revelou resultados expressivos, com F1-*scores* de 94,69%, 98,45% e 98,09%, respectivamente (BOU et al., 2023). Esses achados iniciais indicaram não apenas a viabilidade da proposta, mas também a necessidade de expandir e aprofundar a investigação, especialmente no que se refere ao desempenho comparativo entre diferentes LLMs. Assim, o presente trabalho tem como objetivo ampliar essa análise, utilizando uma base de dados ampla e explorando de forma sistemática as diferenças de desempenho entre os modelos, a fim de compreender suas similaridades, limitações e efetividade em diferentes categorias e temáticas.

1.3.1 Objetivos Especificos

1. Realizar uma avaliação comparativa entre diferentes modelos de LLMs;
2. Conduzir uma análise por categorias temáticas (Racismo, Homofobia e Sexismo);
3. Avaliar o impacto do pré-processamento textual na avaliação das LLMs;
4. Identificar o modelo mais adequado para cada temática (Racismo, Homofobia, Sexismo e Normal);
5. Identificar o modelo mais adequado para cada tipo de conteúdo analisado (Bruto e Pré-processado);
6. Avaliar o desempenho das LLMs com dados reais de redes sociais, em volume superior ao utilizado no trabalho anterior (BOU et al., 2023).

1.4 Organização da Monografia

Esta monografia está estruturada em capítulos que visam apresentar de forma clara e sequencial o desenvolvimento do trabalho. Inicialmente, é apresentada a fundamentação teórica no Capítulo 2, abordando os conceitos centrais relacionados ao discurso de ódio, incluindo definições, estatísticas sobre crimes de ódio na sociedade e na *internet*, além de autores e estudos relevantes às temáticas de homofobia, racismo e sexismo. Também são discutidas as tecnologias utilizadas, com foco nos modelos de linguagem de larga escala (LLMs). Na sequência, o Capítulo 3 de metodologia descreve o processo de construção da pesquisa. A Figura 1 apresenta de forma visível o funcionamento da aplicação da metodologia, desde a coleta e preparação dos dados até a implementação dos experimentos com os diferentes modelos de LLM. Em seguida, são apresentados os resultados e análises comparativas na Seção 3.2. Na Tabela 1 estão todos os resultados obtidos, diferenciando

modelos, tipos de conteúdos, categorias, quantidade de comentários e métricas de *precision*, *recall* e F1-*score*. Na Tabela 2 é demonstrada a média do F1-*score* dos modelos, somando os conteúdos do tipo bruto e pré-processado de cada modelo com a categoria respectiva. Nessa seção, destaca-se o desempenho dos modelos avaliados e discussões relevantes a partir dos testes realizados. Por fim, a monografia é concluída com o Capítulo 4, que apresenta as considerações finais, discutindo as contribuições do trabalho, suas limitações e possíveis direções para pesquisas futuras.

Capítulo 2

Fundamentação Teórica

Este capítulo estabelece os conceitos teóricos essenciais que embasam esta pesquisa. Inicialmente, são apresentadas as áreas, conceitos, arquiteturas e modelos ligados à inteligência artificial, bem como os modelos utilizados no desenvolvimento da proposta deste trabalho, introduzidos nas Seções 2.5 a 2.7. Em seguida, discute-se o discurso de ódio em ambientes digitais, suas categorias e impactos sociais. Por fim, na Seção 2.13, é referenciado os trabalhos correlatos à abordagem proposta.

2.1 Aprendizado de Máquina

O campo do aprendizado de máquina, que faz parte da inteligência artificial, dedica-se ao desenvolvimento de algoritmos capazes de aprender e aprimorar seus desempenhos a partir de dados. Dentro dessa área, são utilizadas diversas técnicas, como as redes neurais, que ajudam os sistemas a ajustar seus parâmetros internos para detectar padrões complexos e realizar previsões de forma precisa. Um avanço significativo nesse campo foi a proposta das redes profundas, uma inovação apresentada por Hinton, Osindero e Teh (2006), que trouxe métodos mais eficazes para o treinamento de redes neurais com múltiplas camadas ocultas. A abordagem proposta inclui o uso de redes restritas de *Boltzmann* (RBMs) para realizar o pré-treinamento das camadas, seguido de um refinamento das conexões usando o algoritmo *wake-sleep*. Em resumo, a contribuição de Hinton, Osindero e Teh (2006) representa um marco no aprendizado de máquina, fornecendo uma sólida base teórica e práticas eficientes para o treinamento de redes neurais de maior complexidade.

2.2 Processamento de Linguagem Natural

Natural Language Processing (NLP), ou Processamento de Linguagem Natural em português, é uma área da inteligência artificial focada em criar métodos e ferramentas para entender e interpretar a comunicação humana, seja por meio da linguagem escrita ou da fala. A NLP permite desenvolver sistemas capazes de analisar, processar e responder

a textos e comandos de voz, o que facilita a interação entre pessoas e máquinas em várias aplicações, como assistentes virtuais, tradutores automáticos e *softwares* de análise de sentimentos.

No estudo *Distributed Representations of Words and Phrases and their Compositionality* de Mikolov et al. (2013), os autores introduzem o modelo *Skip-gram*, que cria representações vetoriais para palavras, facilitando o entendimento semântico em tarefas de processamento de linguagem natural. A ideia é transformar cada palavra em um "vetor" no espaço matemático, de modo que palavras com significados parecidos fiquem próximas. Imaginando uma proximidade no contexto das palavras próximas, representando assim palavras com significados semelhantes: por exemplo, "mãe" e "pai" estariam mais próximas entre si, e "mãe" poderia se aproximar de "filho" e "familia". Com isso, sistemas de NLP conseguem identificar relações e padrões nas palavras, o que pode ser usado em traduções, buscas inteligentes e compreensão de contexto em conversas. Esse estudo foi pioneiro no campo do NLP por proporcionar uma maneira eficiente de entender e processar a linguagem de forma mais parecida com a humana.

2.3 Arquitetura *Transformer*

O estudo *Attention Is All You Need* de Vaswani (2017), introduz a arquitetura *Transformer*, um modelo que revolucionou o campo do aprendizado de máquinas e processamento de linguagem natural. Ao contrário dos modelos anteriores que dependiam de *Recurrent Neural Networks* (RNNs) ou *Convolutional Neural Networks* (CNNs), o *Transformer* utiliza exclusivamente mecanismos de atenção para capturar relações entre *tokens*, o que permite modelar dependências complexas sem necessidade de processamento sequencial. Com isso, a arquitetura consegue explorar a paralelização de forma eficaz, o que reduz drasticamente o tempo de treinamento, especialmente em tarefas de tradução de sequências longas, como tradução automática.

O *Transformer* é composto por uma estrutura de codificador-decodificador, onde tanto o codificador quanto o decodificador são formados por múltiplas camadas de atenção própria e redes totalmente conectadas. No núcleo da arquitetura está o mecanismo de atenção multi-cabeça, que permite ao modelo prestar atenção a diferentes partes de uma sequência simultaneamente, facilitando a identificação de relações entre palavras independentemente da distância entre elas no texto. Além disso, o uso de codificação posicional permite que o *Transformer* capture a ordem das palavras, o que é crucial em tarefas de linguagem. O impacto desse trabalho foi imediato, pois a arquitetura foi rapidamente adotada em várias aplicações de inteligência artificial, estabelecendo as bases para modelos avançados que seguem o mesmo princípio de atenção, como *BERT* e *Gemini* da *Google*, GPT da *Openai*, *Copilot* da *Microsoft* e *Claude* da *Anthropic*. Empresas líderes no setor de tecnologia rapidamente reconheceram o potencial dos *Transformers*, aplicando-os em seus modelos

generativos para aprimorar sistemas de linguagem e expandir o uso da inteligência artificial para novas áreas. Esses avanços impulsionaram significativamente o desenvolvimento de soluções mais eficientes e robustas no processamento de linguagem natural e em outras áreas que envolvem sequências complexas de dados.

2.4 LLM (*Large Language Models*)

As LLMs, ou Modelos de Linguagem de Grande Escala, são modelos avançados de inteligência artificial que utilizam técnicas de *Machine Learning* (ML) para compreender e gerar linguagem humana de forma natural. Esses modelos são aplicados para processar, interpretar e criar textos. Sua aplicação auxilia em diversos setores, desde atendimento ao cliente e *chatbots* até análise de dados e criação de conteúdo.

O estudo de Devlin et al. (2019) apresenta um novo modelo que explora conceitos fundamentais das LLMs, que seria o *BERT* que significa *Bidirectional Encoder Representations from Transformers*, ou Representações Codificadoras Bidirecionais de Transformadores, em português. Esse modelo se destaca por conta da utilização de grandes quantidades de dados e arquiteturas complexas de aprendizado profundo, utilizando a arquitetura *Transformers*, já discutido na seção anteior, assim permitindo avanços significativos na compreensão de linguagem natural. As LLMs, como o BERT, foram projetadas para captar nuances e dependências contextuais em dados textuais extensos, utilizando milhões, ou até bilhões, de parâmetros. Esses modelos são capazes de "aprender" representações ricas do significado das palavras e sentimentos ao considerar seu contexto completo, ou sua característica central no processamento de linguagem.

2.5 GPT - Openai

O *Generative Pretrained Transformer* (GPT)[1], desenvolvido pela OpenAI, é um modelo de linguagem avançado que utiliza redes neurais baseadas na arquitetura *Transformer*, capaz de compreender e gerar textos com fluidez e precisão contextual. Ao ser treinado em um vasto conjunto de dados textuais, o GPT identifica padrões e estruturas linguísticas, o que o torna apto a interpretar e produzir linguagem de forma autônoma. Sua metodologia envolve um pré-treinamento não supervisionado, seguido por ajustes supervisionados, permitindo-lhe realizar tarefas complexas, como responder perguntas, redigir textos e detectar discursos ofensivos em redes sociais. Dessa forma, o GPT se destaca em diversas aplicações que exigem análise de grandes volumes de dados textuais, tornando-se uma ferramenta útil na automação e no suporte a processos comunicacionais complexos.

[1] Documentação GPT Disponível em: <https://platform.openai.com/docs/models/>

2.6 Gemini - Google

Gemini[2], uma família de modelos multimodais desenvolvidos pelo Google, se destaca por sua capacidade de entender e operar em diversos tipos de informação, incluindo texto, código, áudio, imagem e vídeo. Diferente de modelos que aprendem principalmente por reforço em ambientes interativos, o Gemini foi treinado em vastos e diversos conjuntos de dados, permitindo-lhe realizar tarefas complexas como raciocínio, compreensão de linguagem natural, e geração de conteúdo criativo.

2.7 DeepSeek - Hangzhou DeepSeek AI

DeepSeek[3], desenvolvido pela *Hangzhou DeepSeek AI*, é uma modelo de inteligência artificial generativa avançada projetada para compreender e gerar linguagem natural com alta precisão contextual. Sua arquitetura baseia-se em redes neurais profundas do tipo transformador, otimizadas para processar e interpretar grandes volumes de dados textuais. Treinada com extensos conjuntos de dados multilíngues, a tecnologia demonstra capacidade robusta de identificar padrões complexos, gerar respostas coerentes e realizar tarefas como tradução, síntese de textos e resolução de problemas lógicos. Com suporte a contextos de até 128 mil *tokens*, aplica-se a diversos domínios, incluindo educação, pesquisa científica, desenvolvimento de *software* e análise de negócios, atuando como ferramenta estratégica para potencializar produtividade e inovação.

2.8 Discurso de Ódio

Discurso de ódio abrange várias formas de expressão que incitam, promovem ou justificam o ódio, discriminação ou hostilidade contra indivíduos ou grupos, com base em características inerentes, como raça, religião, etnia, orientação sexual, gênero, entre outros (**??**). Esta seção explora dados e manifestações de discurso de ódio na sociedade, no Brasil e no mundo.

A expansão das redes sociais ampliou o alcance das interações digitais, no entanto, a mídia social atual está sendo regularmente usada indevidamente para espalhar mensagens violentas, comentários e discurso de ódio. Isso foi conceituado como discurso de ódio *online*, definido como qualquer comunicação que menospreze uma pessoa ou um grupo com base em características como cor, etnia, gênero, orientação sexual, nacionalidade, religião ou afiliação política (ZHANG; LUO, 2019).

Os crimes de ódio na *internet* chegaram a mais de 74 mil casos em 2022 – maior número desde 2017 – de acordo com dados registrados pela Central Nacional de Denúncias

[2] Documentação Gemini Disponível em: <https://ai.google.dev/gemini-api/docs?hl=pt-br>
[3] Documentação do DeepSeek. Disponível em: <https://api-docs.deepseek.com/>

de Crimes Cibernéticos, da organização SaferNet. Os números constam no Observatório Nacional dos Direitos Humanos (ObservaDH), do Ministério dos Direitos Humanos e da Cidadania (MDHC). Entre 2017 e 2022, a plataforma revela um total de 293,2 mil denúncias de crimes de ódio motivados por preconceito ou intolerância contra grupos ou indivíduos por sua identidade ou orientação sexual, gênero, etnia, nacionalidade ou religião. Estes crimes podem assumir diversas formas na *internet*, como ofensas, ameaças, injúrias, difamações, incitações à violência, apologias ao crime e divulgação de imagens ou vídeos humilhantes (SaferNet Brasil; Ministério dos Direitos Humanos e da Cidadania, 2023).

2.9 Racismo

O racismo envolve preconceito, discriminação ou antagonismo dirigido a pessoas de uma raça ou etnia diferente, sustentado pela crença na superioridade de uma raça sobre as outras. É um problema persistente que afeta diversos aspectos da sociedade, incluindo o mercado de trabalho, sistemas legais e interações sociais, perpetuando desigualdade e injustiça (FREDRICKSON, 2002).

As redes sociais oferecem aos usuários uma plataforma onde perspectivas impopulares podem ser expressas *online* (CHAUDHRY; GRUZD, 2020). Uma suposição comum é que esses pontos de vista tendem a ser racistas e discriminatórios por natureza.

Com base nesse problema, uma pesquisa coordenada pela Faculdade Baiana de Direito, Jusbrasil e pelo Programa das Nações Unidas para o Desenvolvimento (PNUD), mapeou e analisou casos julgados pelos Tribunais brasileiros envolvendo os tipos penais da Injúria Racial e/ou Racismo praticados contra vítimas negras em Redes Sociais. Esse estudo levantou que as mulheres são quase 60% das vítimas dos crimes de racismo e injúria racial julgados em segunda instância no Brasil. Homens são apenas 18,29%. Outros 23,17% não têm gênero identificado. Os insultos relacionados às mulheres negras remetem à sua sexualidade, à sua higiene e à sua estética, já os relacionados aos homens negros, os associam a inferiorização social (NICORY et al., 2022).

Em reflexo desses dados, a escritora Grada Kilomba aprofunda a discussão sobre o conceito de "Outro"[4] ao argumentar que as mulheres negras, por não serem nem brancas nem homens, ocupam um lugar complexo na sociedade supremacista branca, pois representam uma "dupla ausência": são opostas tanto à branquitude quanto à masculinidade. Nesse sentido, ela observa que o status das mulheres brancas é ambíguo, uma vez que, embora sejam mulheres, ainda são brancas. Da mesma forma, os homens negros, mesmo sendo negros, ainda desfrutam do privilégio de gênero. Já as mulheres negras, segundo

[4] A teoria do "Outro" é idealizada pela filósofa francesa Simone de Beauvoir, dizendo que, a mulher foi constituída como o "Outro", pois é vista como um objeto. De forma simples, seria pensar na mulher como algo que possui uma função. Uma cadeira, por exemplo, serve para que a gente possa sentar, uma caneta, para que possamos escrever (RIBEIRO, 2019).

essa análise, não são nem brancas nem homens, assumindo, assim, o papel de "Outro do Outro" (KILOMBA, 2021).

O racismo é um fenômeno enraizado nas sociedades ao redor do mundo, historicamente contestado por líderes, pensadores e artistas. Por meio de suas obras, é destacado as consequências do preconceito racial e reforçaram o orgulho e a identidade da comunidade negra. Personalidades como Malcolm X, Zumbi dos Palmares, Angela Davis, Grada Kilomba, Nelson Mandela, Racionais MCs, Jorge Ben Jor, entre outros, marcaram a luta contra o racismo com suas contribuições literárias e musicais, cada uma oferecendo perspectivas poderosas e distintas que ainda reverberam na sociedade contemporânea. Além disso, grupos e ativistas, como os Panteras Negras, Black Lives Matter, Coletivo de Mulheres Negras e Movimento Negro Unificado, desempenham papéis cruciais na busca por direitos e justiça. Esses movimentos abordam questões sociais e políticas, promovendo a conscientização e mobilizando ações em prol de uma sociedade mais justa e igualitária.

Malcon-X e Haley (1973), em *The Autobiography of Malcolm X*, narra sua trajetória de vida e a transformação de sua visão sobre a identidade negra nos Estados Unidos. Suas ideias contribuíram para uma visão radical e orgulhosa da luta contra o racismo, desafiando as estruturas de opressão.

Racionais MCs, através do álbum Sobrevivendo no Inferno (1997), abordam as realidades das periferias brasileiras, expondo a violência, o racismo e a exclusão social enfrentados pela comunidade negra no Brasil. O grupo em suas letras retrata as vivências cotidianas, que ecoam nas experiências de muitos jovens negros. Racionais (1997), em sua obra "Capítulo 4, versículo 3", diz:

> **A cada quatro pessoas mortas pela polícia, três são negras**
> **Nas universidades brasileiras, apenas 2% dos alunos são negros**
> **A cada quatro horas, um jovem negro morre violentamente em São Paulo**
> **(RACIONAIS, 1997)**

Ben-Jor (1974), em sua obra "Zumbi", no álbum "Tábua de Esmeralda", canta:

> **Quando Zumbi chegar**
> **O que vai acontecer**
> **Zumbi é senhor das guerras**
> **É senhor das demandas**
> **Quando Zumbi chega**
> **É Zumbi é quem manda**
> **(BEN-JOR, 1974)**

Essa canção retrata a força de Zumbi dos Palmares como símbolo de resistência negra e lembrando a importância da luta por liberdade e igualdade. Zumbi dos Palmares foi o líder do Quilombo dos Palmares, um dos maiores quilombos do Brasil, que se tornou símbolo de resistência contra a opressão colonial e a escravidão.

2.10 Sexismo

Sexismo, ou discriminação de gênero, frequentemente tem como alvo pessoas do genêro feminino, decorrente de crenças enraizadas na superioridade de um gênero e em papéis de gênero tradicionais. Ele se manifesta em várias esferas da sociedade, como no ambiente de trabalho, na mídia e nos sistemas legais, contribuindo para disparidades significativas em oportunidades, remuneração e representação (RISMAN, 2018).

Segundo a análise feita na Central Nacional de Denúncias de Crimes Cibernéticos entre o período de 2017 e 2022, a misoginia foi o tipo de crime de ódio que mais cresceu durante esse período, iniciou com 961 denúncias em 2017 para 28679 em 2022, com um aumentando de quase 30 vezes (SaferNet Brasil; Ministério dos Direitos Humanos e da Cidadania, 2023).

A Ouvidoria Nacional de Direitos Humanos (ONDH) é outra fonte relevante sobre a ocorrência desse tipo de crime. Entre janeiro de 2021 e setembro de 2023, a ONDH registrou 29.297 denúncias de violações de direitos humanos na *internet*. Nesse período, o grupo mais afetado por essas violações foi o de mulheres vítimas de violência online, com 4.953 casos em 2021, 5.669 em 2022 e 4.551 até setembro de 2023 (SaferNet Brasil; Ministério dos Direitos Humanos e da Cidadania, 2023). Esses números correspondem a aproximadamente metade das denúncias em cada ano, evidenciando a gravidade e a recorrência desse tipo de violência nas redes.

A importância de líderes, escritoras, pensadoras e artistas que lutam contra o sexismo é fundamental para o empoderamento das mulheres e para a promoção da representatividade em diferentes esferas da sociedade. Essas figuras desempenham um papel crucial ao desafiar normas e estruturas de poder que perpetuam a desigualdade de gênero. Personalidades como Angela Davis, Grada Kilomba, Bell Hooks, Simone de Beauvoir, Djamila Ribeiro, Linda Nochlin, Malala Yousafzai, Conceição Evaristo, Elza Soares, Nina Simone, Elis Regina, Rita Lee, Gal Costa, entre outras, têm sido essenciais na construção de um movimento que busca a representatividade e fortalecer o protagonismo feminino. A ação e o pensamento dessas mulheres são essenciais para transformar a sociedade, promovendo a igualdade e o respeito entre os gêneros. Em conjunto, suas contribuições ampliam o debate sobre a equidade de gênero, tornando-se pilares de resistência e inspiração para futuras gerações.

No Brasil, um grupo que se destaca no âmbito femista é o "Grupo Mulheres do Brasil", que é uma organização suprapartidária formada por mulheres comprometidas com a promoção da igualdade de gênero e o empoderamento feminino. Seu trabalho se concentra na defesa dos direitos das mulheres, buscando aumentar sua participação em espaços de decisão e combater a violência, o racismo e as desigualdades sociais. Através de parcerias com diversos setores da sociedade, a organização contribui para a formulação e implementação de políticas públicas voltadas à melhoria da saúde, educação, e condições de vida das mulheres. Além disso, o grupo investe em ações práticas para promover a

inclusão social, a paz e o respeito aos direitos humanos. Seu objetivo é garantir a efetiva igualdade entre homens e mulheres, com ênfase na formação de uma sociedade mais justa e igualitária. Com uma abordagem colaborativa, o grupo busca transformar realidades por meio de projetos de impacto social e cultural. Dessa forma, o Grupo Mulheres do Brasil se posiciona como uma força importante na luta por mudanças estruturais no país (RODRIGUES; SEVERINO, 2020).

2.11 Feminismo Negro

Existe diferentes vertentes dentro do movimento feminista, como o feminismo negro, feminismo marxista e o feminismo radical. O feminismo negro se destaca muito pela sua história de luta, especialmente ao unir as questões de gênero e cor, e ao dar voz a mulheres que enfrentam múltiplas camadas de opressão. A autora Djamila Ribeiro comenta em sua obra "Quem tem medo do feminismo negro?" o início do feminismo negro no Brasil, enfatizando a fala da socióloga Núbia Moreira:

> **A relação das mulheres negras com o movimento feminista no Brasil se estabelece a partir do III Encontro Feminista Latino-Americano ocorrido em Bertioga em 1985, de onde emerge a organização atual de mulheres negras com expressão coletiva com o intuito de adquirir visibilidade política do campo feminista. A partir daí, surgem os primeiros coletivos de mulheres negras, época em que aconteceram alguns encontros estaduais e nacionais de mulheres negras. Em momentos anteriores, porém há vestígios de participação de mulheres negras no Encontro Nacional de Mulheres, realizado em março de 1979. No entanto, a nossa compreensão é que, a partir do encontro ocorrido em Bertioga, se consolida entre as mulheres negras um discurso feminista, uma vez que em décadas anteriores havia uma rejeição por parte de algumas mulheres negras em aceitar a identidade feminista (RIBEIRO, 2018, p. 25).**

Como discutido na seção anterior sobre racismo, a pesquisa realizada por Nicory et al. (2022) sobre casos julgados de injúria racial e/ou racismo praticados em redes sociais destaca que as mulheres negras são o principal alvo desse tipo de crime.

Nesse sentido, é essencial ressaltar novamente as autoras, já mencionadas na seção sobre sexismo, que desenvolvem a discussão com foco no feminismo negro, como Angela Davis, Bell Hooks, Grada Kilomba, Djamila Ribeiro e Conceição Evaristo. Conforme mencionado no quarto parágrafo da seção sobre o racismo, a obra de Kilomba reflete sobre a posição da mulher negra na sociedade. Nesse contexto, Angela Davis contribui com uma visão crítica sobre as intersecções entre raça, gênero e classe, ampliando o debate ao evidenciar as múltiplas opressões enfrentadas pela mulher negra e a necessidade de uma luta antirracista e feminista integrada.

Uma de suas obras que tem esse destaque, foi publicado em 1981: "Mulheres, Raça e Classe" é uma obra central do feminismo negro e dos estudos sobre raça, gênero e classe. Angela Davis apresenta uma análise histórica e filosófica das condições de vida das mulheres negras, desde o período da escravização nas Américas até os movimentos feministas e antirracistas dos séculos XIX e XX, evidenciando a omissão de suas especificidades por

parte dos movimentos feministas e negros. A contribuição desta obra destaca a importância de incluir as experiências das mulheres negras na luta por igualdade, ampliando a compreensão das opressões interligadas e promovendo um feminismo mais inclusivo e transformador (DAVIS, 1981).

2.12 Homofobia

Homofobia envolve medo, aversão ou discriminação contra homossexuais ou contra a homossexualidade em geral. Ela reflete preconceitos mais amplos da sociedade e pode causar graves danos psicológicos e físicos aos indivíduos, restringindo sua capacidade de viver de forma aberta e autêntica (HEREK, 2009).

No contexto das redes sociais, a homofobia encontra um ambiente propício para a disseminação de discursos de ódio e estereótipos, frequentemente amparada pelo anonimato da *internet*. Segundo a análise feita por SaferNet Brasil (2021), entre 1 de janeiro e 15 de junho de 2021 registraram um aumento de 106,3% nas denúncias de conteúdo LGBTfóbico na *internet*, em 2020 tiveram 1226 denúncias, contra 2529 em 2021.

A luta contra a homofobia tem sido historicamente representada por ativistas, artistas e movimentos sociais. Com obras que questionam a intolerância e celebram a diversidade, figuras como Madonna, Frida Kahlo, Lady Gaga, Ellen DeGeneres, Cazuza, Marta, Ney Matogrosso, Laerte, Pabllo Vittar, Erica Hilton, entre outros representantes. Além disso, movimentos como a parada do Orgulho Lésbicas, Gays, Bissexuais, Transgêneros, Queer, Intersexo, Assexuais e todas as outras identidades de gênero e orientações sexuais (LGBTQIAP+) e a organização Stonewall, desempenham papéis essenciais no fortalecimento da identidade e dos direitos dessa comunidade. Esses líderes e grupos buscam visibilidade, igualdade e representatividade, promovendo uma sociedade mais inclusiva e respeitosa.

2.13 Trabalhos Relacionados

Existem várias propostas recentes voltadas para minimizar a exposição das pessoas a conteúdos inapropriados. Entre essas propostas, há aquelas que exploram o potencial de modelos de linguagem de larga escala, como o GPT-3, na detecção de discurso de ódio e linguagem abusiva. Um exemplo é a proposta de (CHIU; COLLINS; ALEXANDER, 2021), na qual os autores fornecem um trecho de texto ao GPT-3 para que ele classifique esse trecho como “Racista”, “Sexista” ou “Neutro”. Outro trabalho (WANG et al., 2023) adota o uso do GPT-3 para gerar explicações sobre *tweets* que contêm trechos com e sem discurso de ódio. Em ambos os trabalhos, *prompts* de entrada são gerados para o GPT-3.

Além disso, há outras propostas para identificar e filtrar conteúdo inapropriado em discussões *online* usando (i) aprendizado profundo (YENALA et al., 2018) e (ii) apren-

dizado estatístico (SHETH; SHALIN; KURSUNCU, 2022). Enquanto a primeira utiliza uma combinação de camadas de convolução e *Long Short-Term Memory* (LSTM) bidirecional para capturar padrões sequenciais e semântica global em textos, a segunda busca incorporar conhecimento explícito em um algoritmo de aprendizado estatístico para detectar toxicidade em comunicações *online*.

Na proposta realizada por Li et al. (2024) os autores investigam o potencial do ChatGPT para detectar e discriminar comentários odiosos, ofensivos e tóxicos (*Hateful, Offensive and Toxic* (HOT)) em redes sociais. O trabalho compara o desempenho do *ChatGPT* com anotações humanas, conduzindo quatro experimentos para avaliar a consistência, confiabilidade e raciocínio do modelo ao classificar esses comentários. Foram elaborados cinco tipos de *prompts* e divididos em duas categorias principais: aqueles que solicitavam uma resposta binária (sim ou não) e aqueles que pediam uma resposta probabilística (uma pontuação entre 0 e 1 representando a probabilidade de o conteúdo ser HOT). Para assim, interagir com o *ChatGPT*, explorando o impacto que diferentes estruturas de *prompts* têm na capacidade do modelo de identificar e discriminar o conteúdo HOT. Os resultados indicaram que o *ChatGPT* atinge uma precisão aproximada de 80% em comparação com as anotações humanas e fornece a mesma resposta com 90% do tempo, garantindo a confiabilidade e consistência (LI et al., 2024). O estudo sugere que o *ChatGPT* pode ser uma ferramenta viável para moderação de grandes volumes de conteúdo, mas destaca a importância de engenharia de *prompt* adequada para otimizar a confiabilidade e consistência das classificações.

O estudo realizado por Salminen et al. (2020) propõe o desenvolvimento de um classificador de ódio *online* aplicável para plataformas sociais, contribuindo para um ambiente online mais seguro, sendo aplicável em sistemas de moderação, que reportam tais comentários para melhorar a experiência e o bem-estar dos usuários. O modelo é aplicado ao classificar comentários em categorias de "ódio" e "não ódio", usando como base um conjunto de dados composta por 197.566 comentários coletados das redes sociais *YouTube, Reddit, Wikipedia e Twitter*. A ferramenta funciona por meio de algoritmos de aprendizado de máquina (como *XGBoost* e Redes Neurais) que utilizam várias representações de características de linguagem (como BERT, TF-IDF e Word2Vec) para detectar automaticamente comentários de ódio em redes sociais (SALMINEN et al., 2020).

No estudo preliminar (BOU et al., 2023), foi realizada uma avaliação qualitativa sobre a detecção de discursos de ódio em uma amostragem segmentada e controlada, envolvendo 30 *sites* divididos igualmente entre homofobia, racismo e sexismo. Mesmo com essa escala reduzida, os resultados foram expressivos, com F1-score de 94,69% para homofobia, 98,45% para racismo e 98,09% para sexismo, indicando a viabilidade inicial do estudo e entendimento na detecção de discurso de odio com LLM. Esse estudo teve papel crucial como base conceitual, permitindo não apenas testar hipóteses iniciais, mas também consolidar fundamentação sobre discurso de ódio.

Em contraste com as propostas existentes, este trabalho diferencia-se da literatura existente ao desenvolver uma metodologia integrada de avaliação de modelos de linguagem de grande escala para detecção de discurso de ódio em cenários realistas. Enquanto abordagens convencionais frequentemente analisam categorias genéricas ou modelos isolados, nossa proposta avança em três eixos: (i) avaliação comparativa de múltiplos LLMs em categorias específicas (racismo, sexismo, homofobia e neutro); (ii) análise sistemática do impacto do pré-processamento textual na capacidade de detecção de nuances linguísticas; e (iii) validação dualística com dados reais de redes sociais, examinando tanto conteúdo bruto quanto pré-processado. Esta abordagem triangular, articulando modelos, categorias temáticas e fluxos de processamento, oferecendo um *framework* decisório para órgãos de moderação e desenvolvedores de plataformas sociais, e também, superando as limitações de abordagens fragmentadas reportadas em Li et al. (2024) e Bou et al. (2023).

Capítulo 3

Metodologia e Análise dos Resultados

Este capítulo descreve a metodologia empregada para avaliar a eficácia de LLMs na detecção de discursos de ódio em redes sociais, com foco em três categorias principais: homofobia, sexismo e racismo. A abordagem proposta utiliza modelos avançados de processamento de linguagem natural, incluindo GPT-3.5, GPT-4.0, Gemini-2.0-Flash e DeepSeek-V3, para classificar automaticamente comentários extraídos de conjuntos de dados reais. O processo metodológico abrange desde a seleção e preparação dos dados até a implementação de técnicas de pré-processamento e análise de conteúdo, garantindo uma avaliação robusta e comparativa entre esses diferentes modelos.

Além disso, são apresentados os resultados na Seção 3.2, destacando métricas de desempenho como precisão (*precision*), revocação (*recall*) e F1-*Score*, que permitem quantificar a eficácia de cada modelo na identificação de conteúdos ofensivos. A discussão dos resultados inclui uma análise detalhada das diferenças entre comentários brutos e pré-processados, bem como uma avaliação de custo-benefício entre os LLMs testados.

3.1 Metodologia

A metodologia deste trabalho visa avaliar tanto a viabilidade quanto o desempenho de LLMs na identificação de conteúdos impróprios em ambientes de rede social, com foco em manifestações de racismo, homofobia e sexismo. O estudo parte de um cenário simulado que reflete o contexto real das interações sociais *online*, incluindo a linguagem informal e as variações regionais presentes em redes sociais populares. Além disso, o estudo apresenta o detalhamento de custos gerados pelas implementações das LLMs.

Para alcançar esse objetivo, a investigação será conduzida através de uma análise experimental que utiliza LLMs amplamente conhecidas, como o *GPT* da *OpenAI*, *Gemini* da *Google DeepMind* e *DeepSeek* da *Hangzhou DeepSeek Artificial Intelligence Co., Ltd.* Cada uma desses modelos, será testada em parâmetros e configurações semelhantes, permitindo uma análise justa e comparativa, diante da tarefa de identificar conteúdos impróprios propagados em redes sociais.

Ao longo do processo, serão aplicadas métricas de desempenho como precisão (*precision*), revocação (*recall*) e F1-*Score*, que avaliam a eficácia dos modelos na classificação e no reconhecimento de conteúdo ofensivo, incluindo taxas de falsos positivos e falsos negativos. Essa abordagem nos permitirá identificar o potencial das LLMs em minimizar a exposição dos usuários a conteúdos prejudiciais de forma prática, simulando os casos de uso para os quais a aplicação será direcionada no futuro. A partir desses dados, será possível obter um panorama detalhado sobre a capacidade das LLMs de operar em cenários reais e os ajustes necessários para que esses modelos atuem com maior eficácia no combate à disseminação de linguagem ofensiva nas plataformas sociais. Sendo assim, o ambiente metodológico seguiu os passos detalhados na Figura 1:

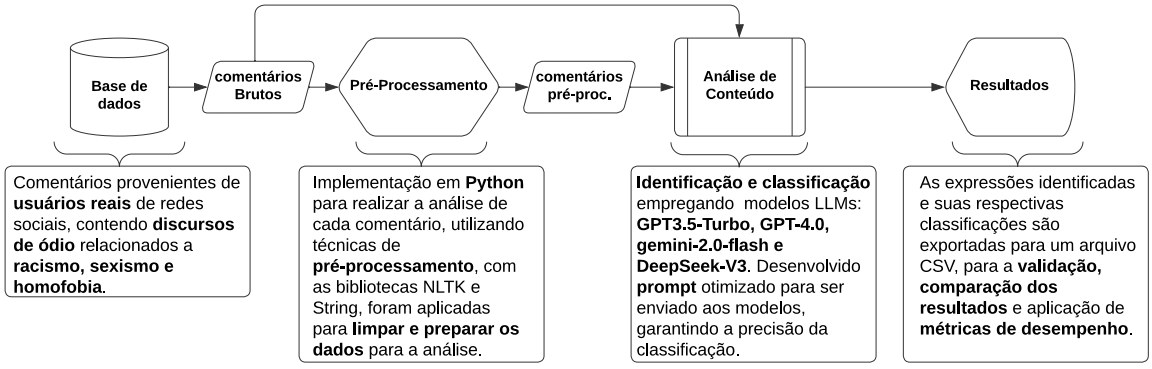

Figura 1 – Fluxograma da Metodologia Empregada.

3.1.1 Base de dados

A análise da base de dados de discursos de ódio, contendo manifestações de homofobia, sexismo e racismo, permite uma compreensão aprofundada das dinâmicas de discriminação presentes nas redes sociais. Os conjuntos de dados que foram analisados e classificados são: o "*Anti-LGBT Cyberbullying Texts*"[1] para os comentarios que possui contexto homofóbico e neutro, os conjuntos relacionados ao sexismo e ao racismo foram obtidos a partir do *"Classified Tweets"*[2], disponibilizado por Albright (2021).

3.1.1.1 Conjunto de dados homofóbico

O *corpus* "*Anti-LGBT Cyberbullying Texts*" oferece uma base sólida para a análise de discursos de ódio com foco na homofobia, integrando uma abordagem perspectivista que considera as visões dos anotadores em suas rotulagens. Esse conjunto de dados foi

[1] Dataset "*Anti-LGBT Cyberbullying Texts*" Disponível em: <https://www.kaggle.com/datasets/kw5454331/anti-lgbt-cyberbullying-texts>

[2] Base de dados "*Classified Tweets*" Disponível em: <https://www.kaggle.com/datasets/munkialbright/classified-tweets/data>

desenvolvido com o objetivo de apoiar pesquisas na área, fornecendo uma versão limpa e reanotada para classificação binária de *ciberbullying anti-LGBT*. Ele conta com mais de 50 mil comentários de plataformas como *Twitter*, *Reddit* e *YouTube*, rotulados por uma equipe diversa de 11.143 anotadores (SACHDEVA et al., 2022). Com isso, é possível uma análise detalhada da gravidade e natureza dos discursos de ódio contra a comunidade LGBTQIAP+, utilizando o espectro contínuo para identificar nuances que uma classificação binária tradicional não captaria com a mesma precisão.

3.1.1.2 Conjunto de dados sexista e racista

Já para o conjunto de dados ligado ao sexismo e racismo, foi utilizado o conjunto *"Classified Tweets"*, disponibilizado pelo autor Albright (2021). Essa versão realiza uma classificação binária inicial entre *tweets* suspeitos e não suspeitos, seguida de uma categorização mais específica entre os *tweets* considerados suspeitos, subdividindo-os em casos de racismo e sexismo.

A rotulagem dos dados foi conduzida por meio do algoritmo de Florestas Aleatórias (*Random Forest*). Utilizando 100 estimadores, ou seja, 100 árvores de decisão, o modelo alcançou uma precisão de 83%, conforme reportado pelo autor Albright (2021).

Tendo em vista a baixa precisão identificada no estudo da base de dados de sexismo e racismo, foi proposto uma análise manual dos comentários presentes nessas bases, com a finalidade de avaliar a acurácia dos rótulos previamente definidos por Albright (2021).

Na análise manual, para os comentários racistas, foi utilizados conceitos do autor Hall (1997), onde classificamos como racistas manifestações que reforçam divisões e estereótipos negativos contra grupos culturais diversos, como imigrantes, cor da pele, minorias religiosas e povos indígenas. Essa análise manual da base de dados, foi identificado que a maioria dos comentários não se referiam ao racismo relacionado à cor da pele, mas sim à diversidade cultural, a imigrantes, religiões e outros aspectos. Como Hall explica em *Representation: Cultural representations and signifying practices*:

> "O racismo é um processo cultural que cria 'outros' sociais a partir de diferenças construídas, que vão além da cor da pele, incluindo nacionalidade, religião, língua e cultura."

Essa perspectiva amplia o entendimento tradicional de racismo, permitindo identificar atitudes preconceituosas que não se restringem à questão da cor da pele, mas que também manifestam intolerância e exclusão social em função da diversidade cultural e identitária (HALL, 1997).

Para a avaliação dos comentários presentes na base de dados classificada como sexista, adotamos a perspectiva de Davis (1981), segundo a qual o sexismo deve ser compreendido como um sistema de opressão tão estruturado quanto o racismo. Essa ideologia atua

no enraizamento da desigualdade entre os sexos e na exclusão histórica das mulheres em diversas esferas sociais. Segundo a autora:

> "O sexismo é uma ideologia que legitima a opressão das mulheres, da mesma forma que o racismo legitima a opressão de negros."

Essas compreensões fundamentaram a classificação dos comentários considerados sexistas e racistas nesta avaliação manual, alinhando desde a reprodução de estereótipos de gênero e o tom agressivo de fala, até o uso de sarcasmo, brincadeiras de má-fé e termos pejorativos associados às temáticas de racismo e sexismo.

3.1.2 Pré-processamento de Dados

A preparação adequada dos dados é fundamental para garantir a precisão e a relevância nas análises subsequentes. Nesta seção, empregamos duas abordagens distintas para o pré-processamento de dados. A primeira utiliza a biblioteca *Pandas*[3] para a leitura e organização da base de dados, enquanto a segunda se concentra em técnicas de limpeza e normalização de texto por meio das bibliotecas *Natural Language Toolkit* (NLTK)[4] e *String* [5] nativas da linguagem de programação *python*. Ambas as etapas são essenciais para preparar os dados, permitindo que sejam analisados de forma eficiente e que os resultados sejam obtidos com maior confiabilidade.

3.1.2.1 Análise de dados

A análise de dados será realizada utilizando a biblioteca *Pandas*, que permite a manipulação e análise de dados em formato tabular. Este passo é essencial para organizar as informações, facilitando o manuseio e assegurando que os dados sejam lidos corretamente. O pré-processamento será aplicado na coluna que referencia o comentário na base de dados.

3.1.2.2 Limpeza dos Dados

A limpeza dos dados será realizada utilizando funções da biblioteca NLTK, incorporando pacotes como "*punkt*","*stopwords*" e "*wordnet*". Esses pacotes serão empregados para realizar tokenização, remoção de palavras irrelevantes e lematização, respectivamente. Já na utilização da biblioteca *String*, foi utilizada a função *"string.punctuation"*, que retorna uma sequência contendo todos os caracteres de pontuação definidos, como: "`!"#$%&'()*+,-./:;<>=?@[\]^_{|}~`". A limpeza ligada ao pré-processamento tem como

[3] *Pandas* Disponível em: <https://pandas.pydata.org/>
[4] NLTK Disponível em: <https://www.nltk.org/>
[5] Documentação String Disponível em <https://docs.python.org/3/library/string.html>

objetivo reduzir a redundância textual e o número de *tokens* enviados aos modelos, isso impacta diretamente no custo de requisições e no tempo de processamento. Além disso, sua aplicação é fundamental para permitir uma análise comparativa entre o comportamento das LLMs ao classificarem os comentários brutos e pré-processados. Essa comparação é essencial para compreender como o pré-processamento afeta a capacidade dos modelos de identificar possiveis nuances linguísticas, ironias, gírias ou termos pejorativos e ambíguos, sobretudo em temáticas sensíveis como racismo, homofobia e sexismo. Avaliar essas diferenças permite identificar peculiaridades na forma como as LLMs respondem a cada temática, contribuindo para o refinamento da metodologia.

Como resultado do pré-processamento textual, é gerada uma nova coluna denominada "*Text_Lem*". Cada instância dessa coluna é então encaminhada como entrada para a função de análise de conteúdo, que é integrada em cada modelo de LLM, permitindo uma avaliação sistemática e precisa de cada comentário, já padronizados e otimizados para serem processados em cada um dos modelos.

3.1.3 Análise de Conteúdo

Nesta etapa, foi realizada a análise dos comentários contidos nas bases de dados, por meio das *Application Programming Interface* (API) das *LLMs*, *GPT* utilizando os modelos *GPT-3.5-Turbo* e *GPT-4.0*, *Gemini* sendo utilizado o modelo *gemini-2.0-flash* e *DeepSeek* utilizando o modelo *DeepSeek-V3 (deepseek-chat)*. A implementação dessas *LLMs* permitirá uma abordagem comparativa robusta na classificação desses tipos de conteúdos, visando selecionar o melhor modelo de classificação entre diferentes temáticas abordadas nos discursos de ódio.

Tanto o conteúdo pré-processado quanto os comentários brutos retirados das bases de dados são utilizados nesta etapa, permitindo uma avaliação mais abrangente das capacidades dos modelos de linguagem. Os comentários são passados como parâmetro, acompanhando com *prompt* de contexto aos modelos de linguagem, para assim identificar e classificar expressões impróprias relacionadas aos termos específicos referentes às temáticas abordadas, garantindo assim uma maior consistência na detecção de padrões linguísticos.

O *prompt* que foi desenvolvido para maximizar a precisão da classificação, foi redigido em inglês, tanto pela clareza de interpretação pelas LLMs quanto pela língua dos comentários analisados. Ele instrui o modelo a gerar uma linha de dados em formato *Comma Separated Values* (CSV) com três colunas, indicando a presença ("1") ou ausência ("0") das temáticas de homofobia, sexismo ou racismo no comentário fornecido. A estrutura desenvolvida no *prompt* busca evitar ambiguidades ao especificar que o retorno deve ser exclusivamente numérico, delimitado por vírgulas, sem incluir textos adicionais ou espaços. Por exemplo, para a sentença analisada, a saída poderia ser algo como "0,1,0" (Homofobia, Sexismo, Racismo), indicando a presença de sexismo, mas ausência de homo-

fobia e racismo, ou até mesmo "1,1,0", múltiplas categorias ofensivas presentes no mesmo comentário, classificando como um comentário homofóbico e sexista. Essa uniformização no formato de entrada e saída permitiu maior controle sobre o processo de análise e viabilizou comparações quantitativas entre os diferentes modelos. A seguir, apresenta-se o *prompt* utilizado, juntamente com sua tradução para o português - Brasil, que serviu como base padronizada para os testes com os modelos *GPT-3.5 Turbo*, *GPT-4.0* e *DeepSeek-V3*, a fim de garantir uniformidade na tarefa de classificação.

> ***Prompt*:** "*Generate a CSV row with three columns, based on the following sentence, where: the first column should contain the value 1 if the sentence contains any indication of homophobia; the second column should contain the value 1 if the sentence contains any indication of sexism; the third column should contain the value 1 if the sentence contains any indication of racism. (Please only write according to the CSV template. I don't want text, just numbers for the tags and commas to separate them and no spaces as well.) The sentence is: {data}*"

> **Tradução do *prompt* - português (BR):** "Gere uma linha CSV com três colunas, baseada na seguinte sentença, onde: a primeira coluna deve conter o valor 1 se a sentença contiver qualquer indicação de homofobia; a segunda coluna deve conter o valor 1 se a sentença contiver qualquer indicação de sexismo; e a terceira coluna deve conter o valor 1 se a sentença contiver qualquer indicação de racismo. (Por favor, escreva apenas de acordo com o formato do CSV. Não quero texto, apenas números para as tags e vírgulas para separá-los, sem espaços.) A sentença é: {dados}"

Apesar da intenção metodológica de manter o mesmo *prompt* entre os modelos, visando assegurar padronização, imparcialidade no contexto inserido aos modelos e isonomia nos testes, o modelo da *Google* apresentou restrições ao processar diretamente comentários com conteúdo sensível, o modelo não retornava resposta ou encerrava a geração, quando utilizado o mesmo *prompt* dos demais modelos. Acredita-se que essa limitação esteja associada às políticas internas do *Gemini*, que restringem respostas a conteúdos sensíveis. Para contornar esse obstáculo, foi desenvolvida uma nova versão do *prompt* do *gemini-2.0-flash*, adotando uma abordagem mais descritiva e ética. O novo *prompt* contextualiza explicitamente que os comentários analisados continham discurso de ódio e que a tarefa fazia parte de uma pesquisa científica voltada à detecção desse tipo de conteúdo em redes sociais. Essa reformulação foi fundamental para que o Gemini executasse corretamente a classificação. Segue abaixo, o texto do *prompt* reformulado desenvolvido exclusivamente para o modelo *gemini-2.0-flash*, juntamente com sua tradução para o português - Brasil.

> ***Prompt*:** "*For research purposes on hate speech detection in social media comments, generate a CSV row with three columns, based on the following sentence: The first column should contain the value 1 if the sentence contains any indication of homophobia; the second column should contain the value 1 if the sentence contains any indication of sexism; the third column should contain the value 1 if the sentence contains any indication of racism. (Output only the CSV row: three comma-separated numbers without spaces.) The sentence is: {data}* "
>
> **Tradução do *prompt* - português (BR):** "Para fins de pesquisa sobre a detecção de discurso de ódio em comentários de redes sociais, gere uma linha CSV com três colunas, com base na seguinte frase: A primeira coluna deve conter o valor 1 se a frase contiver qualquer indicação de homofobia; a segunda coluna deve conter o valor 1 se a frase contiver qualquer indicação de sexismo; a terceira coluna deve conter o valor 1 se a frase contiver qualquer indicação de racismo. (Forneça apenas a linha CSV: três números separados por vírgulas, sem espaços.) A frase é: {dados}"

Apesar dessa necessidade de reformulação textual para o modelo *gemini-2.0-flash*, é importante destacar que a modificação realizada não alterou o objetivo semântico da tarefa de classificação. Ambas as versões do *prompt* — a utilizada nos modelos *GPT-3.5 Turbo*, *GPT-4.0* e *DeepSeek*, e a adaptada especificamente para o Gemini — compartilham a mesma estrutura lógica e finalidade: identificar a presença de homofobia, sexismo e racismo em comentários extraídos de redes sociais, retornando os resultados em formato CSV com três colunas numéricas. A diferença central reside na inclusão, no *prompt* do *Gemini*, de uma justificativa explícita de caráter acadêmico, que contextualiza o uso dos dados como parte de uma pesquisa sobre discurso de ódio, medida necessária para contornar os filtros de conteúdo sensível da ferramenta da Google. Assim, a alteração pode ser considerada uma adequação estratégica de linguagem, sem impacto na interpretação ou execução da tarefa pelos modelos, garantindo a consistência da análise comparativa entre os diferentes sistemas de LLMs.

O ocorrido evidenciou a importância de alinhar a engenharia de *prompt* não apenas às capacidades técnicas dos modelos, mas também às diretrizes éticas e operacionais impostas por cada fornecedor de LLMs.

Além do *prompt*, todos os modelos foram executados utilizando os parâmetros padrão recomendados em suas respectivas documentações oficiais, de forma a manter a imparcialidade experimental e garantir uma base comparável entre os resultados obtidos.

3.1.4 Métricas para a Análise dos Resultados

As expressões identificadas e suas respectivas classificações são exportadas para um arquivo CSV, para a validação e comparação dos resultados entre os modelos da *OpenAI*, *Google DeepMind* e *Hangzhou DeepSeek Artificial Intelligence Co., Ltd.*

O conteúdo identificado pela análise é classificado e anexado em um novo arquivo CSV, levando em conta sua respectiva temática. A validação do mecanismo proposto baseia-se no cálculo das métricas: Verdadeiros Positivos (VP), representando conteúdo impróprio corretamente identificado; Falsos Negativos (FN), conteúdo impróprio existente, mas não detectado; e Falsos Positivos (FP), conteúdo erroneamente apontado como impróprio. A partir desses indicadores, foram computadas as métricas precisão (*precision*), revocação (*recall*) e F1-*Score*.

A precisão é a proporção de identificações positivas corretas (conteúdo identificado como ofensivo que realmente é ofensivo). A revocação mede a proporção de positivos reais corretamente identificados (percentual de conteúdo ofensivo detectado). Já a F1-*Score* é a média harmônica entre precisão e revocação, fornecendo uma visão geral do desempenho do modelo. Essas métricas são definidas pelas equações:

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{1}$$

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{2}$$

$$F1\text{-Score} = \frac{2 \cdot \text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3}$$

Esta prova de conceito encontra-se disponível no *GitHub*[6]. Informações sobre os detalhes para reprodutividade e resultados podem ser encontrados no arquivo `README.MD`.

3.2 Resultados e Discussão

Esta seção apresenta e discute os resultados obtidos a partir da metodologia descrita na seção 3.1. Serão detalhados os achados da abordagem, analisando os dados coletados e suas implicações para a eficácia e aplicabilidade da ferramenta desenvolvida. Além disso, os resultados serão contextualizados em relação aos objetivos propostos, destacando as contribuições, limitações e desafios encontrados durante o processo de avaliação.

Como mencionado anteriormente, uma aplicação semelhante e preliminar da metodologia empregada nesta pesquisa, realizada em escala reduzida, já havia apresentado resultados promissores em estudo anterior (BOU et al., 2023). Na análise de 30 *sites* segmentados (10 de homofobia, 10 de racismo e 10 de sexismo), obtendo F1-*Score* de

[6] Disponível publicamente em: <https://github.com/guilhermebou/Analysis-Dataset-BOU-Guard-A-Study-in-Real-World-Scenarios>

94,69%, 98,45% e 98,09%, respectivamente . No presente trabalho, com uma base de dados significativamente mais robusta, composta por centenas de comentários reais coletados diretamente de redes sociais, os resultados mantiveram sua relevância, demonstrando a consistência e a aplicabilidade da metodologia em cenários mais complexos. Como demonstrado nas Tabelas 1, 2 e Figura 2, os modelos de LLMs, especialmente o GPT-4.0 e o DeepSeek-V3, mantiveram e em muitos casos superaram o desempenho do estudo preliminar, mesmo diante de desafios mais complexos, como maior variabilidade linguística e volume de dados. Isso evidencia não apenas a consistência metodológica, mas também o avanço tecnológico desses modelos mais recentes em relação a versões anteriores, como o GPT-3.5-Turbo, utilizado no estudo anterior.

Tabela 1 – Comparativo de desempenho dos modelos de LLMs para detecção de discursos de ódio.

Modelo	**Tipo de Conteúdo**	**Categoria**	**Qtd Comentários**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
DeepSeek-V3	Bruto	Homofobia	680	100,00	99,26	99,63
DeepSeek-V3	Bruto	Sexismo	456	100,00	96,71	98,33
DeepSeek-V3	Bruto	Racismo	450	100,00	88,22	93,74
DeepSeek-V3	Bruto	Normal	680	100,00	83,82	91,20
DeepSeek-V3	Pré-processado	Homofobia	680	100,00	98,97	99,48
DeepSeek-V3	Pré-processado	Sexismo	456	100,00	85,53	92,20
DeepSeek-V3	Pré-processado	Racismo	450	100,00	70,67	82,81
DeepSeek-V3	Pré-processado	Normal	680	100,00	81,18	89,61
Gemini-2.0-Flash	Bruto	Homofobia	680	100,00	88,53	93,92
Gemini-2.0-Flash	Bruto	Sexismo	456	100,00	86,40	92,71
Gemini-2.0-Flash	Bruto	Racismo	450	100,00	54,44	70,50
Gemini-2.0-Flash	Bruto	Normal	680	100,00	90,88	95,22
Gemini-2.0-Flash	Pré-processado	Homofobia	680	100,00	90,59	95,06
Gemini-2.0-Flash	Pré-processado	Sexismo	456	100,00	89,04	94,20
Gemini-2.0-Flash	Pré-processado	Racismo	450	100,00	30,00	46,15
Gemini-2.0-Flash	Pré-processado	Normal	680	100,00	91,03	95,30
GPT-4.0	Bruto	Homofobia	680	100,00	96,03	97,97
GPT-4.0	Bruto	Sexismo	456	100,00	92,76	96,25
GPT-4.0	Bruto	Racismo	450	100,00	73,11	84,47
GPT-4.0	Bruto	Normal	680	100,00	92,21	95,94
GPT-4.0	Pré-processado	Homofobia	680	100,00	95,59	97,74
GPT-4.0	Pré-processado	Sexismo	456	100,00	91,67	95,65
GPT-4.0	Pré-processado	Racismo	450	100,00	81,78	89,98
GPT-4.0	Pré-processado	Normal	680	100,00	91,47	95,55
GPT-3.5-Turbo	Bruto	Homofobia	680	100,00	73,09	84,45
GPT-3.5-Turbo	Bruto	Sexismo	456	100,00	54,82	70,82
GPT-3.5-Turbo	Bruto	Racismo	450	100,00	72,82	84,27
GPT-3.5-Turbo	Bruto	Normal	680	100,00	92,79	96,26
GPT-3.5-Turbo	Pré-processado	Homofobia	680	100,00	73,97	85,04
GPT-3.5-Turbo	Pré-processado	Sexismo	456	100,00	40,35	57,50
GPT-3.5-Turbo	Pré-processado	Racismo	450	100,00	76,66	86,79
GPT-3.5-Turbo	Pré-processado	Normal	680	100,00	89,56	94,49

Destaca-se a precisão de 100% apresentada em todas as tabelas, o resultado atribuído à idealização da veracidade dos conteúdos presentes nas bases de dados.

Em relação à comparação entre os modelos DeepSeek-V3, Gemini-2.0-Flash, GPT-4.0 e GPT-3.5-Turbo, os resultados revelam variações expressivas em diferentes categorias, com destaque para a temática homofóbica com conteúdo bruto, onde o modelo DeepSeek-V3 atingiu um desempenho de 99,63% no F1-*Score*, superior ao obtido por Gemini-2.0-Flash (93,92%), GPT-4.0 (97,97%) e significativamente acima do GPT-3.5-Turbo (84,45%).

Na temática de sexismo, o DeepSeek-V3 também apresentou desempenho competitivo com conteúdo bruto (98,33% no F1-*Score*), seguido por GPT-4.0 (96,25%) e Gemini-2.0-Flash (92,71%). Entretanto, o GPT-3.5-Turbo mostrou novamente dificuldades acentuadas, com 70,82% nessa mesma métrica, reforçando sua menor eficácia frente aos outros modelos.

A categoria racismo se destacou por ser uma categoria bem desafiadora para todos os modelos, pois apresentou os menores valores de *recall* e F1-*Score* entre todas as categorias. Mesmo no conteúdo bruto, o modelo Gemini-2.0-Flash apresentou apenas 54,44% de *recall*, enquanto o DeepSeek-V3 alcançou 88,22%, e o GPT-4.0 registrou 73,11%. O modelo GPT-3.5-Turbo, por sua vez, manteve desempenho semelhante ao obtido em experimentos anteriores, com F1-*Score* de 84,27%.

Quanto aos conteúdos normais, que não apresentam discurso de ódio, todos os modelos apresentaram excelentes desempenhos em F1-*Score*, todos acima de 89%, com destaque para o Gemini-2.0-Flash (95,22% nos dados brutos e 95,30% nos pré-processados) e o GPT-3.5-Turbo (96,26% nos dados brutos). A surpresa ficou por conta dos modelos da linha GPT, já que o modelo mais recente, GPT-4.0, apresentou desempenho inferior ao seu antecessor, o GPT-3.5-Turbo. Esse resultado pode estar relacionado à sensibilidade na interpretação da linguagem: alguns comentários classificados como "normais" continham xingamentos ou expressões mais agressivas, o que pode ter levado o GPT-4.0 a interpretá-los como ofensivos, resultando em uma classificação mais rígida. Isso será apronfundado na Seção 3.2.2.

Realizando um comparativo entre os resultados obtidos neste estudo e os apresentados por Li et al. (2024), observa-se que os resultados atuais superam todas as médias de F1-*Score* dos *prompts* avaliados na análise HOT, que, no estudo de Li et al. (2024), alcançaram um máximo de 56,2% de F1-*Score* nas respostas do *ChatGPT*, comparadas às anotações humanas. No presente trabalho, as bases de dados utilizadas foram analisadas manualmente, e os resultados obtidos apresentaram desempenho significativamente superior. Considerando o mesmo modelo avaliado por Li o GPT-3.5-Turbo, a média de F1-*Score* em todas as categorias, no presente trabalho, atingiu 82%. Apesar de possíveis diferenças entre as bases de dados utilizadas, ambas apresentam uma similaridade em relação a análise humana nos comentários, o que confere certa equivalência nos critérios de avaliação. Além disso, os conteúdos analisados neste trabalho estão diretamente relacionados ao discurso de ódio, com foco em temáticas como racismo, sexismo e homofobia.

3.2.1 Comparação entre comentários brutos e pré-processados

A comparação entre os conteúdos brutos e pré-processados revela percepções significativas, incluindo possíveis peculiaridades entre os modelos. Na Figura 2, é possível visualizar graficamente os resultados de F1-*Score* em comparação entre os modelos, categorias e conteúdos analisados. O modelo DeepSeek-V3 demonstrou maior estabilidade

entre os dois contextos. Embora o desempenho nas bases de dados com conteúdos brutos tenha sido superior em todas as categorias, os altos níveis de desempenho foram preservados mesmo após o pré-processamento. Por outro lado, o modelo Gemini-2.0-Flash, por exemplo, apresentou uma queda acentuada na temática de racismo, reduzindo seu *recall* de 54,44% (bruto) para apenas 30,00% (pré-processado), resultando em um F1-*Score* de 46,15%. Esse comportamento reforça a hipótese de que a remoção de elementos contextuais durante o pré-processamento pode, em alguns casos, comprometer significativamente a sensibilidade do modelo na identificação de discursos preconceituosos. Isso ocorre porque essa prática, ao eliminar ou lematizar certos termos, pode alterar o contexto original da mensagem, afetando a interpretação dos modelos, especialmente em categorias mais sutis ou ambíguas.

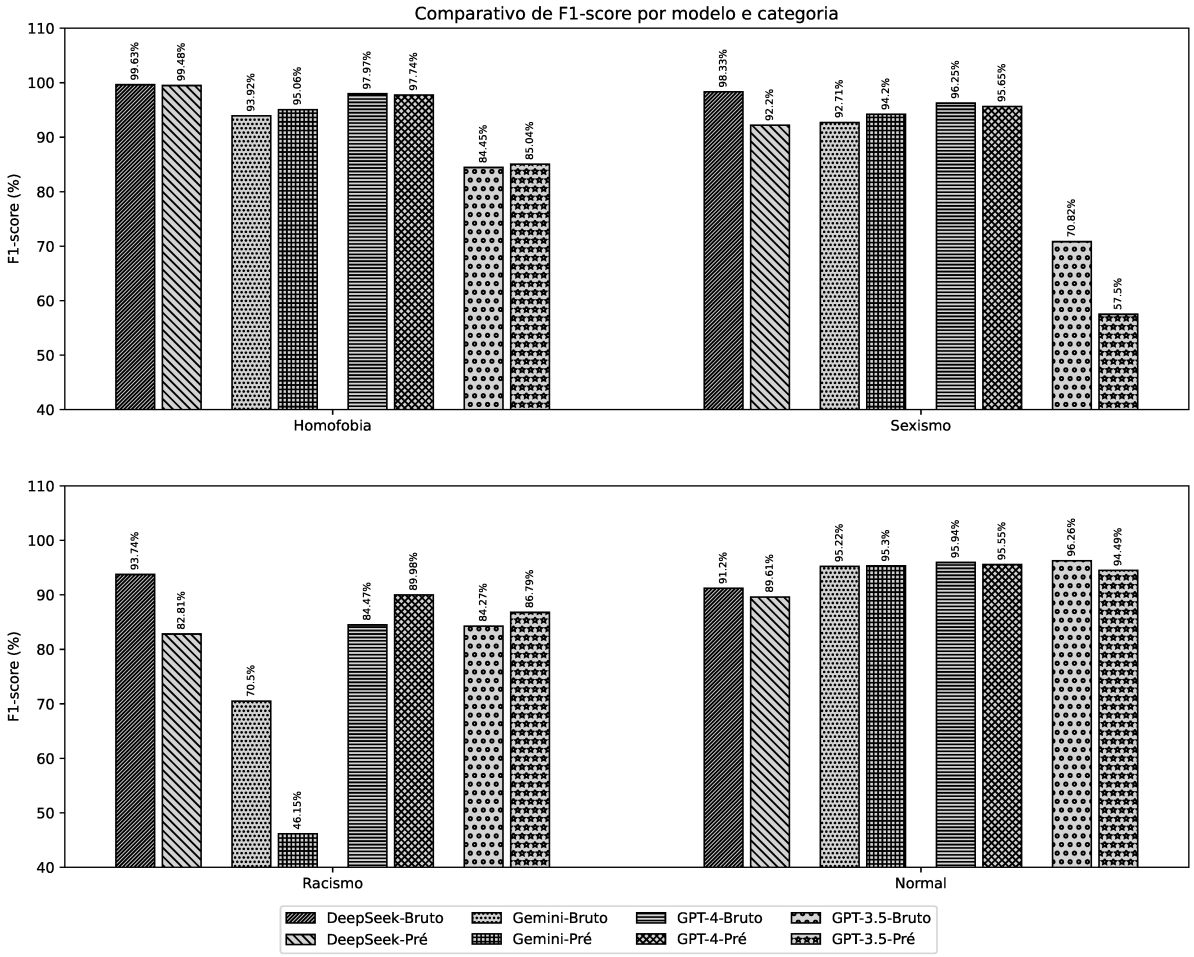

Figura 2 – Gráfico geral F1-Score.

Na comparação entre os tipos de conteúdos analisados pelo modelo GPT-4.0, os comentários racistas que passaram pelo pré-processamento se destacas, os resultados apresentaram um aumento de 5,51% no F1-*Score* em relação aos dados brutos. Esse resultado contraria a tendência observada no modelo DeepSeek e até mesmo nas demais categorias analisadas pelo próprio GPT-4.0, que geralmente obteve resultados superiores ao analisar

os comentários brutos. A exceção foi justamente na categoria de racismo, na qual o pré-processamento parece ter contribuído significativamente para a melhoria do desempenho.

O modelo GPT-3.5-Turbo seguiu a mesma tendência observada no GPT-4.0 na categoria de racismo, apresentando um aumento de desempenho de 84,27% nos dados brutos para 86,79% nos pré-processados. Embora o aumento não seja expressivo, ele evidencia um comportamento semelhante entre os modelos da OpenAI nessa categoria. Já na base de dadosco com conteúdo homofóbica, o GPT-3.5-Turbo apresentou um crescimento ainda menor, de apenas 0,59%, não indicando uma diferença significativa no desempenho.

Tabela 2 – Média geral de F1-*Score* por modelo LLM.

Modelo	**F1-*Score* (Bruto)**	**F1-*Score* (Pré)**	**Média Geral**
GPT-4.0	93,65%	94,73%	94,19%
DeepSeek-V3	95,73%	91,02%	93,37%
Gemini-2.0-Flash	88,08%	82,67%	85,26%
GPT-3.5-Turbo	83,95%	80,95%	82,45%

As médias de F1-*Score* por modelo e tipo de conteúdo estão apresentadas na Tabela 2. O DeepSeek-V3 obteve o melhor desempenho na análise de comentários brutos, com 95,73%, superando o GPT-4.0 (93,65%). Por outro lado, o GPT-4.0 destacou-se na análise de conteúdos pré-processados, alcançando 94,73% e registrando a maior média geral entre todos os modelos, com 94,19%. Já o DeepSeek-V3 apresentou 91,02% para conteúdos pré-processados, resultando em uma média geral de 93,37%, ficando abaixo do GPT-4.0, que manteve maior consistência entre os dois tipos de análise. Quanto aos demais modelos, ambos mantiveram médias na faixa dos 80%. O Gemini-2.0-Flash registrou 88,08% nos comentários brutos e 82,67% nos pré-processados, com média geral de 85,26%, enquanto o GPT-3.5-Turbo apresentou 83,95% e 80,95%, respectivamente, resultando em média geral de 82,45%.

3.2.2 Avaliação dos Modelos Baseada pelo tipo de Conteúdo

Nesta subseção, será apresentada a comparação entre os resultados obtidos para determinados comentários que foram classificados como positivos em alguma categoria de conteúdo. Serão discutidas possíveis causas para essas ocorrências, especialmente nos casos em que os comentários foram identificados como ofensivos no conteúdo bruto, mas não após o pré-processamento, ou vice-versa.

Para essa análise, selecionou-se um comentário de cada base de dados. Cada comentário foi associado a um modelo específico, escolhido com base na diferença dos resultados entre os comentários brutos e pré-processados, observada entre os modelos. O comentário da base homofóbica foi atribuído ao modelo DeepSeek-V3, que apresentou os melhores resultados nesta categoria e na sua análise os conteúdos brutos, mostrou superioridade comparado ao conteúdo pré-processado. Para a categoria sexismo, utilizou-se o modelo

Gemini-2.0-Flash, no qual os comentários pré-processados apresentaram melhor desempenho em sua análise. Já na base de racismo, o modelo selecionado foi o GPT-4.0, seguindo a mesma tendência do modelo Gemini, com resultados superiores a partir do conteúdo pré-processado. Por fim, o comentário da base de dados neutra foi avaliado pelo GPT-3.5-Turbo, que obteve melhor desempenho na análise do conteúdo bruto.

A seguir, são apresentados os comentários selecionados, juntamente com seu *Identity* (ID) e tradução para o português - Brasil e discussões.

> **comentário bruto ID 338 Base de Dados Homofobia:** "*this movie is so gay it could be the AIDS quilt.*"
>
> **Tradução do comentário bruto ID 338 Base de Dados Homofobia - português (BR):** "Esse filme é tão gay que poderia ser o cobertor da AIDS."
>
> **Comentário Pré-processado ID 338 Base de Dados Homofóbica:** "*movie gay could aid quilt*"
>
> **Tradução do comentário Pré-processado ID 338 Base de Dados Homofóbica - português (BR):** "filme gay poderia ajudar colcha"

O modelo DeepSeek-V3 apresentou resultados positivos para o conteúdo bruto analisado nesta categoria, identificando a associação homofóbica entre a homossexualidade e a AIDS (*Acquired Immunodeficiency Syndrome* - Síndrome da Imunodeficiência Adquirida). No entanto, ao examinarmos o conteúdo pré-processado, observa-se que o termo "AIDS" foi removido, por ter sido considerado uma palavra vazia pelo pacote "*stopwords*" da biblioteca NLTK. Com a exclusão desse termo, perde-se a associação explícita entre o vírus e a homossexualidade, o que contribui para a mudança na classificação do comentário. Nesse caso, o problema é ligado diretamente ao dicionário de palavras vazias do pacote "*stopwords*" da biblioteca NLTK.

> **Comentário bruto ID 129 Base de Dados Sexista:** "*Must put words in dead man;s mouth and claim all his work as #feminism 's #adviceforyoungfeminists ex:HuffPostCanada*"
>
> **Tradução do comentário bruto ID 129 Base de Dados Sexista - português (BR):** "Precisam colocar palavras na boca de um homem morto e atribuir todo o trabalho dele ao #feminismo. #conselhoparafeministas ex:HuffPostCanada"
>
> **Comentário Pré-processado ID 129 Base de Dados Sexista:** "*must put word dead man mouth claim work feminism 's adviceforyoungfeminists ex*

huffpostcanada"

Tradução do comentário Pré-processado ID 129 Base de Dados Sexista - português (BR): "deve colocar palavra de homem morto na boca afirmar trabalho feminismo conselhoparafeministas ex huffpostcanada"

Este comentário foi considerado sexista após o pré-processamento, no conteúdo bruto foi clasificado como normal. Levando em consideração o contexto geral, contém sinais sexistas para ambos os gêneros, em relação às mulheres ao sugerir que o feminismo se apropria indevidamente dos méritos de um homem falecido, como se as mulheres não fossem capazes de alcançar reconhecimento por seus próprios esforços. E, em relação aos homens, o comentário também implica que um homem não poderia, por vontade própria, apoiar o feminismo ou reconhecer sua importância. Para o conteúdo bruto, o Gemini pode ter interpretado de forma literal e isolada, sem levar em consideração o contexto histórico ou social, considerando uma crítica genérica ao comentário. Já analisando o conteúdo pré-processado, pode ter levado em consideração o peso dos termos para a sua interpretação sexista ("deve colocar palavra de homem morto na boca", "afirmar trabalho feminismo conselho para feministas"), esses trechos identifica a desvalorização do movimento feminista e a implicação de que homens não poderiam apoiá-lo genuinamente.

comentário bruto ID 55 Base de Dados Racista: "*She is crying, "Boohoo, you are terrorizing our Islamic terrorists*"

Tradução do comentário bruto ID 55 Base de Dados Racista - português (BR): "Ela está chorando: 'Buááá, vocês estão aterrorizando nossos terroristas islâmicos"

comentário Pré-processado ID 55 Base de Dados Racista: "*cry " boohoo terrorizing islamic terrorist*"

Tradução do comentário Pré-processado ID 55 Base de Dados Racista - português (BR): "choram 'buá buá aterrorizando terrorista islâmico"

Para o comentário acima, foi analisado pelo modelo GPT-4.0, onde pode ter sido inicialmente classificado como "normal" erroniamente, com uma análise superficial, achando uma provocação direcionada a um grupo específico, sem uso explícito de termos racializados, porém o comentário contém sarcasmo e reduz muçulmanos a "terroristas islâmicos" de forma generalizada, associando toda uma identidade religiosa à violência. No entanto, quando a análise do conteúdo pré-processado, o viés racista foi evidenciado na classificação do modelo, acusando positivo ao racismo, provavelmente considerando o peso dos termos para a sua interpretação, associando uma abordagem pejorativa ao grupo atacado.

comentário bruto ID 382 Base de Dados Normal: "*you forgot to precede that tweet with the word FUCK.*"

Tradução do comentário bruto ID 382 Base de Dados Normal - português (BR): "Você esqueceu de começar esse tweet com a palavra FODA."

comentário Pré-processado ID 382 Base de Dados Normal: "*forgot precede tweet word fuck*"

Tradução do comentário Pré-processado ID 382 Base de Dados Normal - português (BR): "esqueceu preceder tweet palavra foda"

Neste comentário o modelo GPT-3.5-Turbo classificou o conteúdo pré-processado como homofóbico, a redução para "*forgot precede tweet word fuck*" remove elementos de coesão, deixando a palavra "*fuck*" mais destacada e potencialmente ligada a um contexto homofóbicos. Já analisando o comentário da sua forma bruta, pode ser interpretado como um contexto de frustação e o termo "*fuck*" foi levado apenas como um palavrão, reforçando a indignação por ter esquecido de começar o comentário com a palavra mencionada.

A análise dos resultados para as categorias de sexismo (Gemini-2.0-Flash), racismo (GPT-4.0) e normal (GPT-3.5-Turbo) destaca que o processamento de linguagem natural (NLP), nesses casos, manteve o viés na classificação de conteúdo como odioso, sem considerar o contexto do comentário original. Embora tenha acertado nas categorias de racismo e sexismo, na categoria normal, o conteúdo foi classificado erroneamente como homofóbico. Isso ocorreu porque o modelo levou em conta apenas o termo "*fuck*", associado mais a uma expressão do que a um xingamento direto. Essa palavra, no contexto analisado, funcionava como uma gíria do vocabulário cotidiano para expressar indignação, frustração, raiva, entre outros sentimentos, e não como um discurso de ódio.

Quanto ao conteúdo pré-processado da base de dados homofóbica, analisado pelo modelo DeepSeek-V3, houve um equívoco: a biblioteca NLTK não incluía em seu dicionário do pacote "*stopwords*" a sigla AIDS (Síndrome da Imunodeficiência Adquirida).

3.2.3 Análise de Tempo e Custo Operacional

Nesta seção, será apresentada uma análise que compara o custo financeiro e o tempo de processamento entre os modelos DeepSeek-V3, Gemini-2.0-Flash, GPT-4.0 e GPT-3.5-Turbo, avaliando eficiência e custo-benefício para diferentes aplicações. O objetivo é identificar o melhor equilíbrio entre desempenho e gastos em processamento de linguagem natural.

Vale destacar que os valores apresentados nesta seção são estimativas aproximadas, obtidas a partir da execução da aplicação em ambiente de testes, com foco na análise de

Tabela 3 – Resumo comparativo entre modelos LLMs quanto a desempenho e custo (valores aproximados).

Indicador	**DeepSeek-V3**	**GPT-4.0**	**GPT-3.5 Turbo**	**Gemini 2.0 Flash**
Quantidade de Requisições	8.449	11.688	17.954	4.117
Tokens Processados (total)	1.081.571	1.309.650	2.065.000	540.356
Média de Tokens (por 680 req.)	87.048	76.195	78.211	89.250
Tempo Médio por 680 req. (min)	13,06	12,31	16,25	13,19
Custo por 680 req. (US$)	0,05	1,57	0,25	0,06
Custo Total (US$)	0,66	26,99	3,32	0,12

viabilidade técnica e econômica do uso de LLMs na detecção de discursos ofensivos. As métricas de tempo foram obtidas diretamente durante a execução do código, por meio da soma dos tempos de resposta a cada requisição e posterior cálculo da média ponderada. Já os dados de custo foram extraídos a partir dos painéis de uso disponibilizados pelas próprias plataformas responsáveis por cada modelo.

A análise do tempo médio, apresentado na quarta linha da Tabela 3, indica que o GPT-4.0 foi o mais rápido, com média de 12,31 minutos para processar 680 requisições, ao custo de US$1,57. Contudo, essa vantagem em desempenho foi acompanhada de um custo significativamente elevado, totalizando US$ 26,99, o maior entre todos os modelos avaliados. Em contrapartida, o Gemini 2.0 Flash apresentou o menor custo operacional, com um total de apenas US$ 0,12, embora com um tempo médio de resposta um pouco superior, 13,19 minutos.

O modelo GPT-3.5 Turbo apresentou o pior desempenho em tempo, com uma média de 16,25 minutos, e custo intermediário de US$ 3,32. Já o DeepSeek-V3 demonstrou ser o modelo com o melhor equilíbrio entre tempo de execução e custo, obtendo 13,06 minutos de tempo médio, com um custo total de apenas US$ 0,66 e média por 680 requisições US$ 0,05. Esses valores, aliados a um F1-*score* médio de 93,37%, reforçam a escolha do modelo DeepSeek-V3 como a opção mais eficiente e econômica.

Ao comparar a média de custo por 680 requisições, apresentada na quinta linha da Tabela 3, o DeepSeek-V3 lidera como o modelo mais barato, com US$0,05. Logo atrás, com apenas US$0,01 de diferença, está o Gemini-2.0-Flash, que apresentou o menor custo total nesta análise, embora tenha processado apenas 4.117 requisições. Nesse caso, a projeção para 680 requisições foi feita considerando sua maior média de *tokens*, de 89.250. Já os modelos da família GPT apresentaram os maiores custos, tanto no total quanto na média por 680 requisições, sendo o GPT-4.0 com US$1,57 e o GPT-3.5-Turbo com US$0,25, conforme já mencionado anteriormente.

Quando combinamos os dados de desempenho, tempo e custo operacional, observa-se que o GPT-4.0 lidera na média geral de F1-*score* com 94,19% visivel na Tabela 2, mas seu alto custo pode representar uma limitação em aplicações com restrições orçamentárias. Já o Gemini 2.0 Flash, apesar de seu excelente custo, apresentou desempenho inferior (85,26%), o que pode comprometer a precisão da classificação. Nesse contexto, o

Tabela 4 – Comparativo de preços e recursos entre os modelos).

Categoria	**DeepSeek-V3**	**Gemini 2.0 Flash**	**GPT-4.0**	**GPT-3.5 Turbo**
Comprimento de Contexto	64K	Não especificado	128K (padrão)	16K (padrão)
Saída Máxima	4K (padrão) / 8K (máx)	Não especificado	4K–8K+	4K
Preço Entrada (1M Tokens)	US$ 0,07 (cache hit) US$ 0,27 (cache miss)	US$ 0,30	US$ 30,00	US$ 0,50
Preço Entrada com Desconto (1M Tokens)	US$ 0,035 (cache hit) US$ 0,135 (cache miss) (entre 16:30 e 00:30 UTC)	Não disponível	Não disponível	Não disponível
Preço Saída (1M Tokens)	US$ 1,10 (com desconto: US$ 0,55)	US$ 2,50	US$ 60,00	US$ 1,50
Entrada em Cache	Sim	Não disponível	Não disponível	Não disponível
Preço de Armazenamento em Cache	Incluso	US$ 0,075 (acesso) US$ 1,00 por milhão tokens/hora	Não disponível	Não disponível
Disponível gratuitamente	Não	Sim (versão gratuita)	Não	Não
Observações	Descontos por horário Cache *hit/miss* influencia no custo *Cache Hit* → Dados em cache (acesso rápido) *Cache Miss* → Dados não armazenados em cache (processamento completo)	Inclui *tokens* de pensamento Nível gratuito disponível	Modelo mais caro Ideal para tarefas complexas	Mais acessível Equilíbrio custo-benefício

DeepSeek-V3 se destaca como a alternativa mais vantajosa em termos de custo-benefício, equilibrando alto desempenho, tempo de execução competitivo e baixo custo, sendo especialmente recomendável para projetos que demandam escalabilidade e contenção de gastos.

A Tabela 4 apresenta informações adicionais sobre os custos dos modelos analisados, obtidas diretamente da documentação oficial de cada um: DeepSeek-V3[7], Gemini 2.0 Flash[8] e GPTs[9]. Ressalta-se que os valores podem ter sofrido alterações desde o início do experimento, em janeiro de 2024.

[7] Documentação de preços do DeepSeek disponível em: <https://api-docs.deepseek.com/quick_start/pricing>

[8] Documentação de preços do Gemini disponível em: <https://ai.google.dev/gemini-api/docs/pricing?hl=pt-br>

[9] Documentação de preços dos GPTs disponível em: <https://platform.openai.com/docs/pricing>

Capítulo 4

Considerações Finais e Trabalhos Futuros

Os experimentos realizados neste trabalho demonstram que os Modelos de Linguagem de Grande Escala (LLMs) possuem potencial significativo para a detecção de discursos de ódio em redes sociais, especialmente nas categorias de homofobia, sexismo e racismo. O DeepSeek-V3 se destacou como o modelo mais equilibrado, alcançando um F1-*Score* médio de 93,37% com custo operacional 41 vezes menor que o GPT-4.0 (apenas US$0,66). Esse desempenho robusto, aliado à eficiência econômica, posiciona como uma solução viável para aplicações em larga escala, onde custo e precisão são fatores críticos. Contudo, observou-se que o pré-processamento tradicional de texto, com remoção de palavras vazias e lematização, comprometeu em todas as categorias (uma queda média de 6,93% de F1-*Score* do DeepSeek-V3), pois eliminou nuances contextuais essenciais para identificar sarcasmo, ironia e referências culturais sutis.

A análise por categoria revelou disparidades importantes: enquanto a homofobia foi consistentemente bem identificada (F1-*Score* de 99,63% pelo DeepSeek-V3 em dados brutos), o racismo apresentou os maiores desafios. O Gemini-2.0-Flash, por exemplo, atingiu apenas 54,44% de *recall* nessa categoria, refletindo dificuldades dos modelos em capturar estereótipos estruturais e microagressões. Curiosamente, o pré-processamento beneficiou os modelos da OpenAI (GPT-4.0 e GPT-3.5-Turbo) na detecção de racismo (aumento de 5,51% e 2,52% no F1-*Score*, respectivamente), sugerindo que a redução de ruído pode auxiliar em contextos onde o viés é mais implícito. Essa exceção ressalta a necessidade de abordagens personalizadas por categoria, em vez de soluções generalizadas.

A avaliação de custo-desempenho evidenciou *trade-offs* marcantes. O GPT-4.0 obteve a melhor métrica agregada (94,19% de F1-*Score*), mas seu custo proibitivo (US$26,99) limita sua aplicação em cenários reais com restrições orçamentárias. Já o Gemini-2.0-Flash mostrou-se a opção mais econômica (US$0,12), porém com desempenho irregular, especialmente em racismo pré-processado (F1-*Score* de 46,15%). O GPT-3.5-Turbo, por sua vez, apresentou a pior relação custo-benefício (custo US$3,32 para F1-*Score* de 82,45%),

reforçando que versões anteriores de LLMs têm eficácia limitada em tarefas sensíveis.

As limitações do estudo também merecem destaque. A dependência de base de dados com rotulação automatizada de Albright (2021) (base de dados "Classified Tweets", com precisão reportada de 83%), introduz potenciais vieses, mesmo após validação manual. Além disso, políticas éticas de LLMs, como os filtros do *Gemini* para conteúdo sensível, exigem adaptações de *prompt* que podem influenciar resultados. Casos de falsos positivos (ex.: classificação de "*fuck*" como homofóbico pelo GPT-3.5-Turbo) e falsos negativos (ex.: perda do termo "AIDS" no pré-processamento) ilustram como ambiguidades linguísticas e limitações técnicas ainda desafiam a confiabilidade.

Os resultados desta pesquisa abrem espaço para diversas investigações complementares. Primeiramente, recomenda-se o desenvolvimento de outras soluções de pré-processamento, capazes de preservar termos culturalmente sensíveis (como "AIDS" em contextos LGBTQIAP+) e variações linguísticas regionais. Estratégias de limpeza seletiva, baseadas em dicionários dinâmicos de *stopwords* ajustados por domínio, podem minimizar perdas semânticas importantes, especialmente em categorias como racismo, onde as nuances são cruciais.

Adicionalmente, propõe-se a integração da arquitetura RAG (Retrieval-Augmented Generation) como forma de prover contexto normativo aos modelos. Ao utilizar mecanismos de recuperação de trechos embasados em literatura acadêmica, legislações e diretrizes de direitos humanos, o sistema pode enriquecer sua tomada de decisão com regras claras sobre o que constitui discurso de ódio, racismo, homofobia, sexismo ou neutralidade. Essa abordagem contextualizada pode mitigar ambiguidades e reduzir falsos positivos, principalmente em conteúdos sensíveis, como manifestações sociais legítimas.

Além das melhorias já apontadas, é fundamental avançar na validação ética e intercultural das soluções, elaborando um dicionário de referência linguística com dados multilingues (como o português brasileiro com gírias regionais, ou até mesmo outros idiomas), incluir especialistas em direitos humanos no processo de avaliação desses comentários e desenvolver métricas de justiça (*fairness*) para mensurar vieses em subgrupos marginalizados. A combinação de LLMs com modelos simbólicos baseados em regras pode ser particularmente eficaz em cenários ambíguos, reforçando a responsabilidade algorítmica.

Por fim, é importante direcionamento para trabalhos futuros é a experimentação e avaliação de modelos de linguagem mais recentes e avançados disponíveis no mercado, tais como o GPT-5.0, DeepSeek R1 e Gemini-2.5-Pro. Esses modelos mais atuais prometem avanços significativos em compreensão contextual, capacidade de generalização e eficiência computacional, podendo potencializar ainda mais a precisão na detecção de discursos de ódio e suas nuances culturais e regionais.

Referências

ALBRIGHT, M. **Monitoring system for detecting suspicious users on social network application**. 2021. Citado 3 vezes nas páginas 29, 30 e 46.

BEN-JOR, J. **A Tabua de Esmeralda**. Philips Records, 1974. Zumbi. Disponível em: <https://open.spotify.com/intl-pt/track/4SASmOXEaX8c4mK8RSHFJS?si=78694f1cb9ca44fc>. Citado na página 22.

BLIUC, A.-M. et al. Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. **Computers in Human Behavior**, Elsevier, v. 87, p. 75–86, 2018. Citado na página 14.

BOU, G. et al. Bou-guard: Uma abordagem para detecçao de conteudo impróprio na internet. In: SBC. **Anais Estendidos do XXIII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais**. Juiz de Fora, Brasil, 2023. p. 285–290. Citado 5 vezes nas páginas 14, 15, 26, 27 e 35.

Brasil – Presidência da República. **Lei n.º 12.965, de 23 de abril de 2014 (Marco Civil da Internet)**. 2014. <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm>. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Citado na página 13.

CHAUDHRY, I.; GRUZD, A. Expressing and challenging racist discourse on facebook: How social media weaken the "spiral of silence" theory. **Policy & Internet**, Wiley Online Library, v. 12, n. 1, p. 88–108, 2020. Citado na página 21.

CHIU, K.-L.; COLLINS, A.; ALEXANDER, R. Detecting hate speech with gpt-3. **arXiv preprint arXiv:2103.12407**, 2021. Citado na página 25.

DAVIS, A. Y. **Women, Race Class**. New York: Random House, 1981. Citado 2 vezes nas páginas 25 e 30.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). **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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <https://aclanthology.org/N19-1423>. Citado na página 19.

FREDRICKSON, G. M. **Racism: A Short History**. Princeton: Princeton University Press, 2002. Citado na página 21.

GÁMEZ-GUADIX, M.; INCERA, D. Homophobia is online: Sexual victimization and risks on the internet and mental health among bisexual, homosexual, pansexual, asexual, and queer adolescents. **Computers in human behavior**, Elsevier, v. 119, p. 106728, 2021. Citado na página 14.

HALL, S. **Representation: Cultural representations and signifying practices**. London: Sage Publications, 1997. "O racismo é um processo cultural que cria 'outros' sociais a partir de diferenças construídas, que vão além da cor da pele, incluindo nacionalidade, religião, língua e cultura.". Citado na página 30.

HEREK, G. M. Hate crimes and stigma-related experiences among sexual minority adults in the united states: Prevalence estimates from a national probability sample. **Journal of Interpersonal Violence**, SAGE Publications Sage CA: Thousand Oaks, CA, v. 24, n. 1, p. 54–74, 2009. Citado na página 25.

HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A fast learning algorithm for deep belief nets. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 18, n. 7, p. 1527–1554, 2006. Citado na página 17.

KILOMBA, G. **Plantation Memories: Episodes of Everyday Racism**. Toronto: Between the Lines, 2021. Citado na página 22.

LI, L. et al. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. **ACM Transactions on the Web**, v. 18, n. 2, p. 1–36, 2024. Citado 3 vezes nas páginas 26, 27 e 37.

MALCON-X, A. H. M. A.-S. .; HALEY, A. **The autobiography of Malcolm X**. Ballantine Books, 1973. Disponível em: <https://archive.org/details/autobiographyofm00malc/page/n3/mode/2up>. Citado na página 22.

MIKOLOV, T. et al. **Distributed Representations of Words and Phrases and their Compositionality**. 2013. Disponível em: <https://arxiv.org/abs/1310.4546>. Citado na página 18.

NICORY, D. et al. **Racismo e Injúria Racial Praticados nas Redes Sociais: Relatório do Observatório das Condenações Judiciais em 2ª Instância até o ano de 2022**. Salvador, BA, 2022. Apoio do Programa das Nações Unidas para o Desenvolvimento (PNUD). Disponível em: <https://drive.google.com/file/d/1pEpXmyy4g-J147-l1IrmRYe1NYERxB0W/view>. Citado 2 vezes nas páginas 21 e 24.

RACIONAIS, M. **Sobrevivendo no Inferno**. São Paulo: Cosa Nostra Fonográfica, 1997. Capítulo 4, Versículo 3. Disponível em: <https://open.spotify.com/intl-pt/track/6Wt61AZLG0bN2KasopE2sj?si=16cc63e5af244ddb>. Citado na página 22.

RIBEIRO, D. **Quem tem medo do feminismo negro?** São Paulo: Companhia das Letras, 2018. Citado na página 24.

_____. **Lugar de fala**. São Paulo: Pólen Produção Editorial, 2019. Citado na página 21.

RISMAN, B. J. Gender as a social structure. **Gender & Society**, SAGE Publications Sage CA: Los Angeles, CA, v. 32, n. 4, p. 465–480, 2018. Citado na página 23.

RODRIGUES, L. H. T. I.; SEVERINO, G. C. B. **ESTATUTO SOCIAL DO GRUPO MULHERES DO BRASIL**. 2020. Estatuto Social do Grupo Mulheres do Brasil, São Paulo, 19 de Junho de 2020. Estatuto disponivel em: <https://www.grupomulheresdobrasil.org.br/estatuto-social-do-grupo-mulheres-do-brasil/>. Acesso em: 7 nov. 2024. Citado na página 24.

SACHDEVA, P. et al. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In: **Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @ LREC2022**. Marseille, France: European Language Resources Association (ELRA), 2022. p. 83–94. Citado na página 30.

SaferNet Brasil. **Hotline SaferNet LGBTQIA+**. São Paulo, Brasil, 2021. Publicado em 18/06/2021, atualizado em 03/03/2024. Disponível em: <https://new.safernet.org.br/content/moderacao-da-safernet-elimina-ameacas-de-morte-na-live-da-parada-lgbt-de-sp#>. Citado na página 25.

SaferNet Brasil; Ministério dos Direitos Humanos e da Cidadania. **Enfrentamento ao discurso de ódio**. São Paulo, Brasil, 2023. Publicado em 12/12/2023, atualizado em 27/02/2024. Disponível em: <https://experience.arcgis.com/experience/6a0303b2817f482ab550dd024019f6f5/page/Enfrentamento-ao-discurso-de-%C3%B3dio/>. Citado 3 vezes nas páginas 13, 21 e 23.

SALMINEN, J. et al. Developing an online hate classifier for multiple social media platforms. **Human-centric Computing and Information Sciences**, Springer, v. 10, p. 1–34, 2020. Citado na página 26.

SHETH, A.; SHALIN, V. L.; KURSUNCU, U. Defining and detecting toxicity on social media: context and knowledge are key. **Neurocomputing**, Elsevier, v. 490, p. 312–318, 2022. Citado na página 26.

VASWANI, A. Attention is all you need. **Advances in Neural Information Processing Systems**, 2017. Citado na página 18.

WANG, H. et al. Evaluating gpt-3 generated explanations for hateful content moderation. **arXiv preprint arXiv:2305.17680**, 2023. Citado 2 vezes nas páginas 14 e 25.

YENALA, H. et al. Deep learning for detecting inappropriate content in text. **International Journal of Data Science and Analytics**, Springer, v. 6, p. 273–286, 2018. Citado na página 25.

ZHANG, Z.; LUO, L. Hate speech detection: A solved problem? the challenging case of long tail on twitter. **Semantic Web**, IOS Press, v. 10, n. 5, p. 925–945, 2019. Citado 2 vezes nas páginas 13 e 20.