Comparative Analysis of LLMs for Detecting Racism, Sexism, and Homophobia on Social Media

Guilherme Bou[1], **Adriano Mendonça Rocha**[1]

[1]Faculdade de Computação (FACOM) – Universidade Federal de Uberlândia (UFU)
Av. João Naves de Ávila, 2121 - Santa Mônica, Uberlândia - MG, 38408-100

{guilherme.bou, adriano.rocha}@ufu.br

Structured Abstract - Research Context: *Social media expansion has increased digital interactions, including hate speech such as racism, sexism, and homophobia. This challenges society and platforms to develop strategies for identifying and moderating harmful content;* ***Scientific and/or Practical Problem:*** *Despite AI advances, automatic detection faces limitations due to linguistic nuances, cultural context, and implementation costs. Scientifically, the challenge is evaluating model effectiveness; practically, it is developing economical, reliable large-scale solutions;* ***Proposed Solution and/or Analysis:*** *We conducted a comparative analysis of LLMs (GPT-3.5-Turbo, GPT-4.0, DeepSeek-V3 and Gemini-2.0-Flash) in detecting offensive social media comments. Tests on raw and preprocessed data using standardized prompts measured precision, cost, and execution time;* ***Related IS Theory:*** *The study draws on Information Systems theories, emphasizing socio-technical, ethical, and cost–benefit aspects (Socio-technical Theory, Actor-Network Theory, Resource-Based View, Dynamic Capabilities);* ***Research Method:*** *Over 2,000 comments were analyzed by LLMs using precision, recall, F1-score, operational cost, and processing time metrics;* ***Summary of Results:*** *GPT-4.0 achieved the highest F1-score (94.19%) but at high cost (US$ 26.99). DeepSeek-V3 balanced performance and cost (F1-score 93.37%, US$ 0.66). Gemini-2.0-Flash was the cheapest (US$ 0.12) but showed inconsistent results;* ***Contributions and Impact to IS area:*** *This work offers a practical framework for selecting LLMs for hate-speech detection based on accuracy, cost, and performance. It advances IS research by evaluating state-of-the-art models in real scenarios and providing guidance for ethical and efficient content moderation.*

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 and 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 Direi-

tos 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 and 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 and Ministério dos Direitos Humanos e da Cidadania 2023] indicam que essa legislação não tem surtido o efeito esperado no combate a esse problema.

Diante das preocupações geradas pela produção de conteúdo impróprio na *internet*, este estudo tem como objetivo desenvolver e avaliar uma abordagem para a detecção de discursos de ódio utilizando modelos de linguagem em larga escala, com foco especial nos modelos GPT-4.0 e GPT-3.5-Turbo[3], DeepSeek-V3[4] e Gemini-2.0-Flash[5].

2. Fundamentação Teórica

Esta seção estabelece a base conceitual necessária para a compreensão do estudo ao contextualizar o discurso de ódio, abordando as dimensões do racismo, sexismo e homofobia.

2.1. Discurso de Ódio

Segundo [Heinze 2016], o discurso de ódio é geralmente compreendido como um conjunto de expressões que incitam, promovem ou legitimam o ódio, a discriminação ou a hostilidade contra indivíduos ou grupos, em razão de características como raça, religião, etnia, gênero ou orientação sexual.

2.2. 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].

Com base nesse problema, uma pesquisa coordenada pela Faculdade Baiana de Direito, Jusbrasil e pelo Programa das Nações Unidas para o Desenvolvimento (PNUD),

[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" [Houaiss 2001].

[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`

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"[6] 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 essa análise, não são nem brancas nem homens, assumindo, assim, o papel de "Outro do Outro" [Kilomba 2021].

2.3. 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 aumento de quase 30 vezes [SaferNet Brasil and Ministério dos Direitos Humanos e da Cidadania 2023].

2.4. 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 foi registrado um aumento de 106,3% nas denúncias de conteúdo LGBTfóbico na *internet*, em 2020 tiveram 1.226 denúncias, contra 2529 em 2021.

[6] 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].

3. 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 et al. 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) aprendizado estatístico [Sheth et al. 2022]. Enquanto a primeira utiliza uma combinação de camadas de convolução e *Long Short-Term Memory* 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 de [Li et al. 2024], os autores investigam o uso do *ChatGPT* para detectar comentários odiosos, ofensivos e tóxicos (HOT) em redes sociais. Compararam o modelo com anotações humanas em quatro experimentos, testando cinco tipos de *prompts* categorizados em respostas binárias (sim/não) e probabilísticas (pontuação 0–1). O *ChatGPT* obteve aproximadamente 80% de precisão frente às anotações humanas e repetiu a mesma resposta em 90% dos casos, indicando boa consistência; porém, os autores destacam a necessidade de engenharia de *prompt* adequada para melhorar a confiabilidade.

O estudo de [Salminen et al. 2020] propõe o desenvolvimento de um classificador de ódio *online* aplicável a plataformas sociais, contribuindo para um ambiente virtual mais seguro. O sistema pode ser integrado a mecanismos de moderação que reportam comentários ofensivos, visando melhorar a experiência e o bem-estar dos usuários. Os autores apresentam um modelo híbrido que combina algoritmos clássicos e representações textuais, utilizando o modelo BERT como extrator de *features* devido ao seu poder de contextualização. Foram analisadas as categorias "ódio" e "não ódio", com base em um conjunto de dados composto por 197.566 comentários coletados de *YouTube*, *Reddit*, *Wikipedia* e *Twitter*. A ferramenta emprega algoritmos de aprendizado de máquina como *XGBoost* e redes neurais, para classificar conteúdos odiosos, explorando diferentes representações de características linguísticas (BERT, TF–IDF e Word2Vec) para a detecção automática de discursos 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 ódio 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 litera-

tura 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], [Salminen et al. 2020] e [Bou et al. 2023].

4. 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 um desses modelos, será testado 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:

4.1. 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*"[7] para os comentários que possuem

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

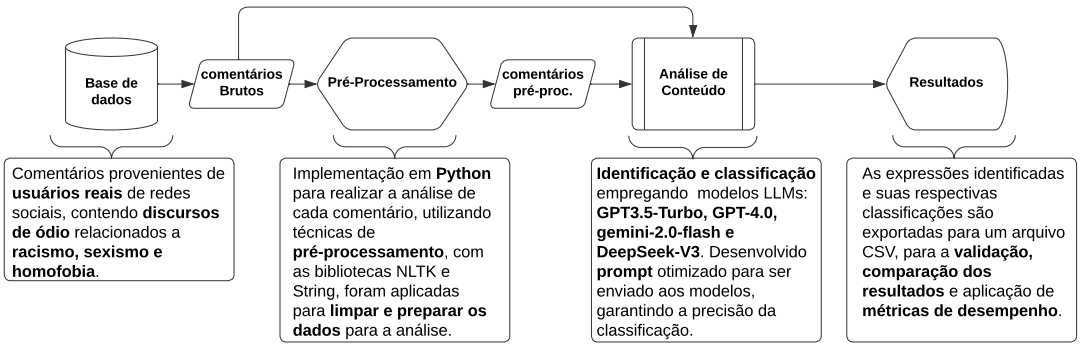

Figura 1. Fluxograma da Metodologia Empregada.

contexto homofóbico e neutro, os conjuntos relacionados ao sexismo e ao racismo foram obtidos a partir do *"Classified Tweets"*[8], disponibilizado por [Albright 2021].

4.1.1. 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 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.

4.1.2. 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

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

[Albright 2021].

Na análise manual dos comentários racistas, adotaram-se as diretrizes da Lei nº 7.716/1989, que define os crimes resultantes de preconceito de raça, cor, etnia, religião ou procedência nacional [Brasil - Presidência da República 1989], bem como os conceitos de [Hall 1997], que compreende o racismo como um processo cultural de produção da diferença. Assim, foram classificados como racistas comentários que reforçam estereótipos negativos e divisões contra grupos culturais, como imigrantes, minorias religiosas e povos indígenas. Observou-se que a maioria dos comentários não estava associada exclusivamente à cor da pele, mas a aspectos de diversidade cultural, imigração e religião.

Para a avaliação dos comentários classificados como sexistas, adotamos as previsões e diretrizes da Lei nº 11.340/2006 (Lei Maria da Penha) no que tange à proteção contra formas de violência física, psicológica e simbólica dirigidas às mulheres [Brasil - Presidência da República 2006]. Seguindo a perspectiva de [Davis 1981], 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.

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.

4.2. 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*[9] 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 NLTK[10] e *String* [11] 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.

4.2.1. 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.

[9] *Pandas* Disponível em: `https://pandas.pydata.org/`

[10] NLTK Disponível em: `https://www.nltk.org/`

[11] Documentação String Disponível em `https://docs.python.org/3/library/string.html`

4.2.2. 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 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 possíveis 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.

4.3. Análise de Conteúdo

Nesta etapa, foi realizada a análise dos comentários contidos nas bases de dados, por meio das APIs 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*, desenvolvido para maximizar a precisão e a consistência da classificação, foi redigido em inglês, uma vez que análises comparativas indicam que os LLMs apresentam desempenho mais robusto nesse idioma. Conforme demonstrado por [Khoshtab et al. 2025], o inglês obtém os melhores resultados médios em tarefas que envolvem linguagem figurativa, como expressões comparativas e expressões idiomáticas, independentemente do modelo ou da estratégia de *prompting*. Ele instrui o modelo a gerar uma linha de dados em formato CSV com três colunas, indicando a presença ("1") 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 ex-

clusivamente 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 homofobia 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, 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*}”

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*.

> ***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*} ”

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 de *prompts* 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 ocorrência de homofobia, sexismo e racismo em comentários que são submetidos ao modelo para análise, 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 destacou a necessidade de alinhar a engenharia de *prompt* não só às atividades atribuídas aos modelos, mas também às diretrizes éticas e operacionais definidas pelos fornecedores dos modelos.

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.

4.4. 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 analisados.

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: $precision = \frac{VP}{VP+FP}$, $recall = \frac{VP}{VP+FN}$ e $F1Score = \frac{2 \cdot precision \cdot recall}{precision+recall}$

Esta prova de conceito encontra-se disponível no *GitHub*[12]. Informações sobre os detalhes para reprodutividade e resultados podem ser encontrados no arquivo `README.MD`.

5. Resultados e Discussão

Esta seção apresenta e discute os resultados obtidos a partir da metodologia descrita na Seção 4. 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

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

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 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 5.2.

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.

5.1. 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.

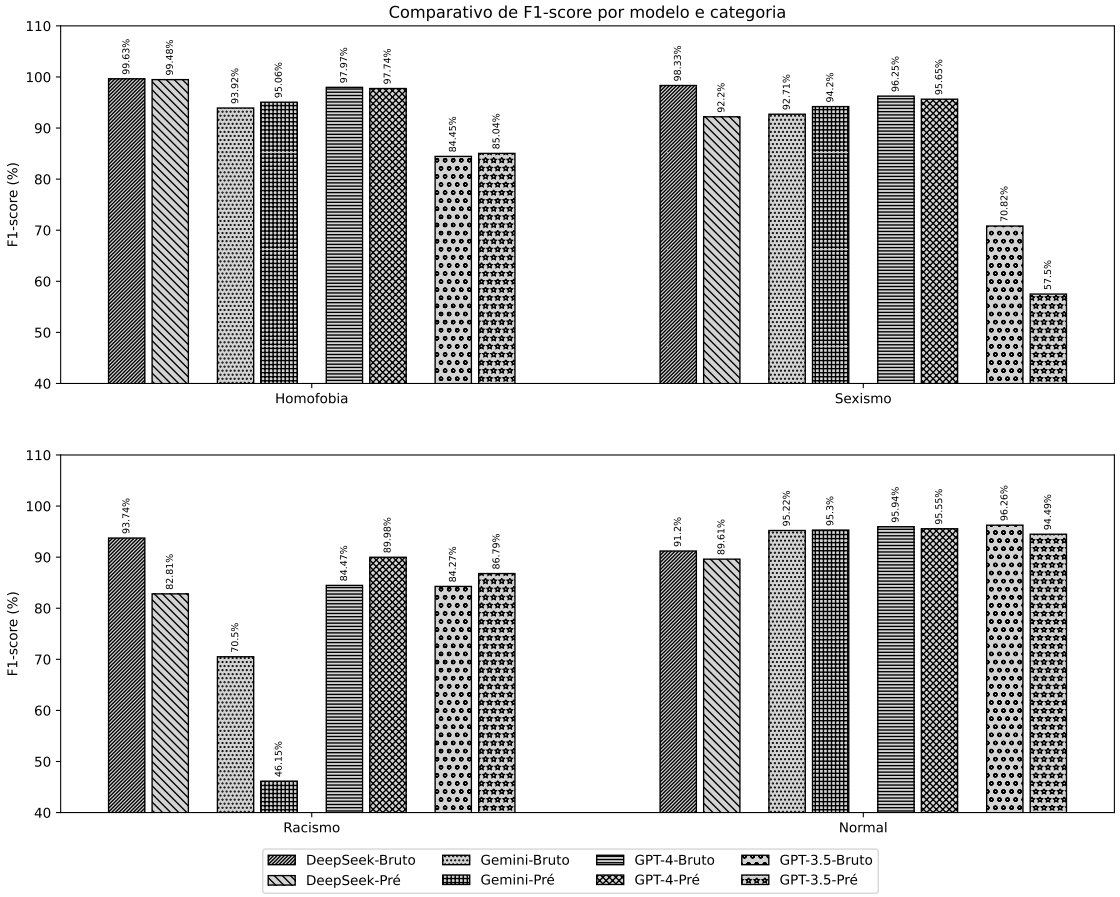

Figura 2. Gráfico geral F1-Score.

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 dados com conteúdo homofóbico, 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%.

5.2. 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 sua identificação (ID) e em seguida suas discussões.

> **comentário bruto ID 338 Base de Dados Homofobia:** "*this movie is so gay it could be the AIDS quilt.*"
>
> **Comentário Pré-processado ID 338 Base de Dados Homofóbica:** "*movie gay could aid quilt*"

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 lemantizado, por ter sido considerado uma palavra no plural pelo pacote "*wordnet*" da biblioteca NLTK. Com a lemantização desse termo, perde-se o contexto do comentário e 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*"
>
> **Comentário Pré-processado ID 129 Base de Dados Sexista:** "*must put word dead man mouth claim work feminism 's adviceforyoungfeminists ex huffpostcanada*"

O comentário acima analisado pelo modelo *Gemini*, ao ser submetido na forma de conteúdo pré-processado, foi classificado como sexista; entretanto, quando analisado em seu texto bruto, o mesmo comentário foi classificado como normal. A análise contextual do comentário revela traços de sexismo direcionados ao movimento feminista, enquanto representação de um grupo social associado ao gênero feminino, ao desvalorizá-lo por meio da alegação de apropriação indevida dos méritos de um "homem falecido". O modelo *Gemini* possivelmente interpretou o texto bruto de maneira mais literal e isolada, enquanto, no pré-processamento, a ênfase em termos-chave (por exemplo, "colocar palavras na boca de um homem morto") intensificou a leitura com viés sexista.

> **comentário bruto ID 55 Base de Dados Racista:** "*She is crying, Boohoo, you are terrorizing our Islamic terrorists*"
>
> **comentário Pré-processado ID 55 Base de Dados Racista:** "*cry boohoo terrorizing islamic terrorist*"

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

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 mostra que os modelos mantiveram vieses na classificação de conteúdo odioso sem considerar o contexto. Embora *Gemini-2.0-Flash* e GPT-4.0 tenham acertado nas categorias de sexismo e racismo, GPT-3.5-Turbo classificou erroneamente conteúdo normal como homofóbico, ao interpretar a palavra "*fuck*" como xingamento, quando na verdade funcionava como gíria expressando indignação ou frustração. No conteúdo homofóbico pré-processado pelo DeepSeek-V3, houve erro: a biblioteca NLTK não reconheceu a sigla AIDS no dicionário "*wordnet*", não sendo necessário lematizá-la.

5.3. 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 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

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

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 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.

Para mais informações detalhadas, os custos dos modelos analisados podem ser consultados diretamente na documentação oficial: DeepSeek-V3[13] , Gemini 2.0 Flash[14] e GPTs[15] . Ressalta-se que os valores podem ter sofrido alterações desde o início do experimento, realizado em janeiro de 2024, até a sua finalização, em agosto de 2025.

6. Ameaças à Validade

Esta seção discute as principais ameaças à validade do estudo. Os resultados podem ser influenciados pelo caráter evolutivo das LLMs, uma vez que avanços tecnológicos e atualizações nos modelos podem alterar seu comportamento ao longo do tempo. Além disso, a base de dados utilizada pode não representar todos os contextos linguísticos e sociais, e bases com rotulação automatizada, como a de [Albright 2021], utilizada nas temáticas racista e sexista e com precisão reportada de 83%, podem introduzir vieses, mesmo após validação manual. Políticas éticas das LLMs, como filtros para conteúdo sensível (ex.: *Gemini*), exigem adaptações de *prompt*, o que também pode impactar os resultados. Casos de falsos positivos e falsos negativos evidenciam desafios relacionados a ambiguidades linguísticas, pré-processamento e sensibilidade à formulação dos *prompts* e

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

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

[15]Documentação de preços dos GPTs disponível em: `https://platform.openai.com/docs/pricing`

ao idioma analisado. Apesar dessas limitações, a padronização dos *prompts* e dos critérios de avaliação contribui para mitigar tais ameaças, mantendo a consistência das análises.

7. Conclusão

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.

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 LGBTQIAPN+) 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 eficaz em cenários ambíguos, reforçando a responsabilidade algorítmica.

Por fim, um importante direcionamento para trabalhos futuros é a experimentação e a avaliação de modelos de linguagem mais recentes e avançados disponíveis no mercado, tais como GPT-5.2, DeepSeek R1 e Gemini-3-Pro-Flash, bem como a comparação com modelos *open source* especializados na detecção de discurso de ódio, como XLM-RoBERTa e *Detoxify*. A inclusão desses modelos mais atuais, juntamente com modelos especialistas em detecção de discurso de ódio, enriquece significativamente a análise comparativa entre múltiplos modelos de LLMs, promovendo avanços na compreensão contextual, na capacidade de generalização e na eficiência computacional, e podendo potencializar ainda mais a precisão na detecção de discursos de ódio e de suas nuances culturais e regionais.

Referências

Albright, M. (2021). Monitoring system for detecting suspicious users on social network application.

Bou, G., Rocha, A. M., Quincozes, V. E., Quincozes, S. E., and Kazienko, J. F. (2023). Bou-guard: Uma abordagem para detecçao de conteudo impróprio na internet. In *Anais Estendidos do XXIII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*, pages 285–290, Juiz de Fora, Brasil. SBC.

Brasil - Presidência da República (1989). Lei nº 7.716, De 5 De Janeiro De 1989 (Lei Caó).

Brasil - Presidência da República (2006). Lei nº 11.340, De 7 De Agosto De 2006 (Lei Maria da Penha).

Brasil - Presidência da República (2014). Lei n.º 12.965, De 23 De Abril De 2014 (Marco Civil Da Internet).

Chiu, K.-L., Collins, A., and Alexander, R. (2021). Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.

Davis, A. Y. (1981). *Women, Race Class*. Random House, New York.

Fredrickson, G. M. (2002). *Racism: A Short History*. Princeton University Press, Princeton.

Hall, S. (1997). *Representation: Cultural representations and signifying practices*. Sage Publications, London.

Heinze, E. (2016). *Hate Speech and Democratic Citizenship*. Oxford University Press, Oxford.

Herek, G. M. (2009). 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*, 24(1):54–74.

Houaiss, A. (2001). *Dicionário Houaiss da língua portuguesa*. Objetiva, Rio de Janeiro.

Khoshtab, P., Namazifard, D., Masoudi, M., Akhgary, A., Sani, S. M., and Yaghoobzadeh, Y. (2025). Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698.

Kilomba, G. (2021). *Plantation Memories: Episodes of Everyday Racism*. Between the Lines, Toronto.

Li, L., Fan, L., Atreja, S., and Hemphill, L. (2024). "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.

Nicory, D., Guanabara, D., Assumpção, V., Viana, D., Leal, F., Afonso, I., Pinho, L. P., Freire, M., and Colombini, P. (2022). Racismo e injúria racial praticados nas redes sociais: Relatório do observatório das condenações judiciais em 2ª instância até o ano de 2022. Technical report, Faculdade Baiana de Direito e Jusbrasil, Salvador, BA.

Ribeiro, D. (2019). *Lugar de fala*. Pólen Produção Editorial, São Paulo.

Risman, B. J. (2018). Gender as a social structure. *Gender & Society*, 32(4):465–480.

Sachdeva, P., Barreto, R., Bacon, G., Sahn, A., Vacano, C. V., and Kennedy, C. (2022). The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @ LREC2022*, pages 83–94, Marseille, France. European Language Resources Association (ELRA).

SaferNet Brasil (2021). Hotline safernet lgbtqia+. Technical report, São Paulo, Brasil.

SaferNet Brasil and Ministério dos Direitos Humanos e da Cidadania (2023). Enfrentamento ao discurso de ódio. Technical report, São Paulo, Brasil.

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S.-g., Almerekhi, H., and Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10:1–34.

Sheth, A., Shalin, V. L., and Kursuncu, U. (2022). Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

Wang, H., Hee, M. S., Awal, M. R., Choo, K. T. W., and Lee, R. K.-W. (2023). Evaluating gpt-3 generated explanations for hateful content moderation. *arXiv preprint arXiv:2305.17680*.

Yenala, H., Jhanwar, A., Chinnakotla, M. K., and Goyal, J. (2018). Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics*, 6:273–286.

Zhang, Z. and Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.