Detecção de Conteúdo Impróprio na Internet por meio da Tecnologia Generative Pre-Trained Transformer (GPT)

Guilherme Bou[1], **Daniel Caetano**[1], **Adriano M. Rocha**[1]

[1]Faculdade da Computação – Universidade Federal de Uberlândia (UFU)

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

Abstract. *With the expansion of the digital environment, exposure to prejudiced content has become a challenge. BOU-Guard demonstrated effectiveness in detecting hateful comments, with GPT-4.0 outperforming GPT-3.5-Turbo on preprocessed data, achieving F1-Scores of up to 97.74%. On manually validated data, GPT-4.0 maintained its lead, but recall remained challenging, particularly for racism. Despite limitations in the datasets, rigorous curation and manual analysis validated the results. The study reinforces the feasibility of GPT-4.0 in real-world scenarios.*

Resumo. *Com a expansão do ambiente digital, a exposição a conteúdos preconceituosos tornou-se um desafio. O BOU-Guard demonstrou eficácia na detecção de comentários odiosos, com o GPT-4.0 superando o GPT-3.5-Turbo em dados pré-processados, atingindo F1-Scores de até 97,74%. Nos dados validados manualmente, o GPT-4.0 manteve a liderança, mas o recall permaneceu desafiador, especialmente para racismo. Apesar de limitações nos datasets, a curadoria rigorosa e a análise manual validaram os resultados. O estudo reforça a viabilidade do GPT-4.0 em cenários reais.*

1. Introdução

A Internet oferece liberdade de comunicação e expressão de opinião. No entanto, as mídias sociais atuais vêm sendo regularmente utilizadas de forma inadequada para disseminar mensagens violentas e discurso de ódio. O discurso de ódio online é definido como qualquer comunicação que menospreze uma pessoa ou um grupo com base em características como raça, cor, etnia, gênero, orientação sexual, nacionalidade, religião ou afiliação política [Zhang and Luo 2019].

Em reflexo deste 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 and dos Direitos Humanos e da Cidadania MDHC 2023].

[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 1986].

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 [Presidência da República do Brasil 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 and dos Direitos Humanos e da Cidadania MDHC 2023] indicam que essa legislação não tem surtido o efeito esperado no combate a esse problema.

Para enfrentar esse problema, propõe-se uma prova de conceito desenvolvida para identificar conteúdos racistas, homofóbicos e sexistas em comentários de redes sociais. Para isso, foi desenvolvido o BOU-Guard, uma ferramenta baseada em *Large Language Models* (LLMs), especialmente GPT-4.0 e GPT-3.5-Turbo [3]. Em um estudo anterior, a solução foi testada em cenários gerais com dados de sites diversos e conteúdos preconceituosos [Bou et al. 2023]. O BOU-Guard implementa a Unidade de Observação de Comportamento (*Behavior Observation Unit* - BOU) para classificar conteúdos impróprios, sendo uma ferramenta versátil e integrável a diferentes contextos. A ferramenta "BOU-Guard — *Extension*"está disponível no *GitHub*[4].

2. 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 realizado por [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 feito por [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* (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 comentários odiosos, ofensivos e tóxicos (*Hateful, Offensive, and Toxic* - HOT) em redes sociais. O estudo compara o desempenho do modelo com anotações humanas, realizando quatro experimentos para avaliar consistência, confiabilidade e raciocínio na classificação de comentários. Foram usados cinco tipos de *prompts*, divididos em duas categorias: binários (sim/não) e probabilísticos (pontuação entre 0 e 1 representando a probabilidade de o conteúdo ser HOT). Os resultados mostram que o *ChatGPT* alcança 80% de precisão em relação às anotações humanas e fornece respostas consistentes em 90% dos casos [Li et al. 2024]. O estudo destaca que a engenharia de *prompt* adequada é crucial para otimizar a confiabilidade e consistência, tornando o *ChatGPT* uma ferramenta promissora para moderação de grandes volumes de conteúdo.

O estudo de [Salminen et al. 2020] propõe um classificador de ódio online para plataformas sociais, aplicável em sistemas de moderação que reportam comentários, pro-

[3]Modelos GPTs disponíveis em: `https://platform.openai.com/docs/models`

[4]Ferramenta Disponível em: `https://github.com/guilhermebou/BOU-Guard-Extension`

movendo uma experiência mais harmônica para os usuários. O modelo classifica comentários como “ódio” ou “não ódio”, utilizando um conjunto de 197.566 comentários coletados de *YouTube, Reddit, Wikipedia e Twitter (X)*. A ferramenta utiliza algoritmos de aprendizado de máquina para identificar automaticamente comentários de ódio [Salminen et al. 2020].

Em contraste com as propostas existentes, a aplicação apresentada neste trabalho, utiliza uma abordagem automatizada para detectar discursos de ódio em tempo real, com foco em temas de racismo, sexismo e homofobia. A ferramenta integra LLMs, como GPT-4.0, GPT-3.5 Turbo, para identificar e listar expressões ofensivas em textos, em vez de simplesmente categorizar palavras isoladas. Além disso, nem todas as abordagens discutidas anteriormente são centradas no usuário, ou seja, não permitem instalação fácil em navegadores ou dispositivos pessoais, limitando o acesso direto pelo usuário final.

Conforme introduzido anteriormente, o presente trabalho propõe uma análise de viabilidade em ambientes reais, utilizando comentários de usuários provenientes de redes sociais que contenham discursos de ódio. O objetivo é analisar a capacidade dos modelos em interpretar a linguagem humana e a escrita real dos usuários, considerando que, em determinados contextos, variações culturais e formas de escrita podem gerar situações desafiadoras para as LLMs na tomada de decisões ao classificar o conteúdo desses comentários. Diferentemente do estudo apresentado em [Bou et al. 2023], cuja análise focou na viabilidade de uma ferramenta de extensão para detecção de conteúdos na web em geral com dados provenientes de sites diversos e conteúdos preconceituosos alocados em tags HTML, este trabalho apresenta um enfoque distinto.

Neste estudo é realizado uma análise detalhada de desempenho entre os modelos de linguagem GPT-3.5 Turbo e GPT-4.0, permitindo uma avaliação abrangente das capacidades e limitações de cada modelo na identificação de conteúdos ofensivos, reforçando a relevância e aplicabilidade prática desta abordagem.

3. Metodologia

Nesta nova aplicação do BOU-Guard, a proposta é avaliar a eficiência da ferramenta em cenários reais, utilizando os modelos de LLMs GPT-3.5 Turbo e GPT-4.0, aplicados a comentários provenientes de usuários reais, o qual contém conteúdo homofóbico, racista, sexista e também comentários neutros. Comentários neutros são aqueles que não contêm termos ofensivos, pejorativos ou discriminatórios, sendo compostos por expressões comuns e sem conotação de ódio. A Figura 1 ilustra as principais etapas da proposta, juntamente com os detalhes de implementação empregados neste estudo.

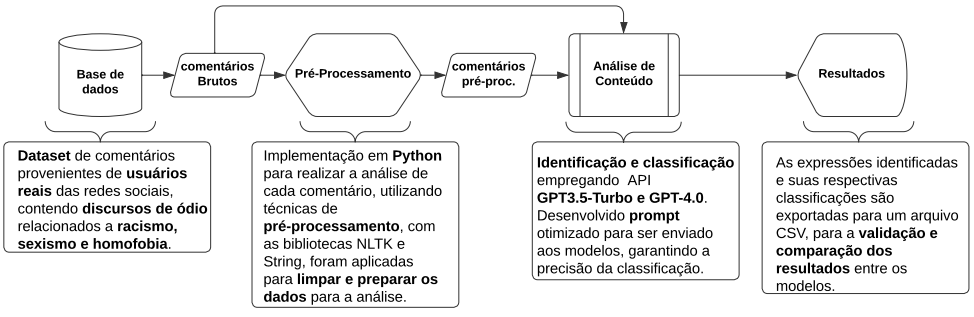

Figura 1. Detalhamento da Metodologia Empregada.

Além disso, foi realizada uma segunda avaliação para verificar o impacto do pré-processamento desses comentários. Nesse processo, foram aplicadas técnicas como lematização e exclusão de palavras vazias, com o objetivo de melhorar o rendimento do processamento. A análise buscou determinar se a aplicação do pré-processamento influenciaria positiva ou negativamente a capacidade das LLMs em detectar discurso de ódio.

1ª Etapa - Seleção de *Datasets*: Foram escolhidos *datasets* contendo discursos de ódio relacionados a racismo, sexismo e homofobia, extraídos de comentários de usuários nas redes sociais *Twitter (X), Reddit e YouTube*. Os *datasets* utilizados estão disponíveis publicamente no *GitHub*[5] e no *Kaggle*[6].

2ª Etapa - Pré-processamento: Foi implementada uma solução em Python, utilizando a biblioteca Pandas, para realizar a análise dos comentários. Técnicas de pré-processamento, utilizando as bibliotecas NLTK e *String*, foram aplicadas para limpar e preparar os dados para a análise.

A preparação adequada dos dados é fundamental para garantir a precisão e a relevância nas análises subsequentes. Nesta etapa, é aplicado duas atividades distintas para o pré-processamento de dados. A primeira utilizando técnicas de análise de dados, onde é utilizado a biblioteca *Pandas*[7] para a leitura e organização dos *datasets*, enquanto a segunda se concentra em técnicas de limpeza e normalização de texto por meio das bibliotecas *NLTK (Natural Language Toolkit)*[8] e a blibioteca *String*[9] nativa da linguagem de programação *python*.

A análise de dados será realizada utilizando a biblioteca *Pandas*[3], como já foi introduzido, que permite a manipulação e análise de dados em formato tabular. Este passo é essencial para organizar as informações e facilitar a percepção de todo conteúdo.

A normalização de dados será realizada utilizando funções da biblioteca *NLTK (Natural Language Toolkit)*[4], incorporando pacotes como "*punkt*", "*stopwords*" e "*wordnet*". Estes pacotes serão empregados para realizar tokenização, remoção de palavras irrelevantes e lematização, respectivamente. Já na utilização da biblioteca *String* foi utilizado a função *"string.punctuation"*, que retorna uma sequência contendo todos os caracteres de pontuação definidos, como: " !"#$%&'()*+,-./:;<>=?@[\]^_{|}~"

3ª Etapa - Análise de conteúdo para GPT 3.5 Turbo e GPT 4.0: Tanto o conteúdo pré-processado quanto os comentários brutos retirados dos *datasets* são utilizados nesta etapa, permitindo uma avaliação mais abrangente das capacidades dos modelos de linguagem. Os comentários são passados como parâmetro 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 e testado 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

[5]Disponível na URL: https://github.com/guilhermebou/Analysis-Dataset-BOU-Guard-A-Study-in-Real-World-Scenarios/tree/main/Data

[6]Disponível na URL: https://www.kaggle.com/datasets/kw5454331/anti-lgbt-cyberbullying-texts e https://www.kaggle.com/datasets/munkialbright/classified-tweets

[7]Documentação *Pandas* Disponível em: https://pandas.pydata.org/

[8]Documentação NLTK Disponível em: https://www.nltk.org/

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

CSV com três colunas, indicando a presença (“1”) ou ausência (“0”) das temáticas (homofobia, sexismo ou racismo) no comentárop fornecido. A seguir, apresenta-se o *prompt* e sua tradução.

> ***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}”

A estrutura do *prompt* é objetiva e busca evitar ambiguidades ao especificar que o retorno deve ser exclusivamente numérico e no formato delimitado por vírgulas, sem incluir textos explicativos ou espaços. O modelo responde com “1”nas colunas correspondentes aos discursos ofensivos identificados e “0”caso contrário. Por exemplo, para a sentença analisada, a saída poderia ser algo como “0,1,0”, indicando a presença de sexismo, mas ausência de homofobia e racismo, ou até mesmo “1,1,0”, indicando presença de homofobia e sexismo no mesmo comentário, clasificando como um comentário multilabel.

Essa abordagem permitiu a criação de uma estrutura de processamento eficiente para classificação automática, combinando o conteúdo dos comentários brutos e os pré-processado. Além disso, a uniformização do formato dos resultados viabilizou análises quantitativas e comparações entre o desempenho dos modelos GPT 3.5 Turbo e GPT 4.0 em diferentes contextos (conteúdo será discutido na seção 4 que centraliza toda coleta do estudo). A construção do *prompt* foi idealizada para assegurar com maior consistência os dados processados, para assim inibir possíveis ruídos gerados por interpretações ambíguas ou inconsistências no retorno do modelo.

4ª Etapa - Análise do resultado: As expressões identificadas e suas respectivas classificações são exportadas para um arquivo CSV, para a validação e comparação dos resultados entre os modelos GPT-3.5 Turbo e GPT-4.0.

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 positi-

vos 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*[10]. Informações sobre os detalhes para reprodutibilidade e resultados podem ser encontrados no arquivo `README.MD`.

4. Resultados e Discussões

Esta seção apresenta e discute os resultados obtidos a partir da metodologia descrita na seção 3. 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.

Inicialmente, a aplicação da metodologia apresentou resultados abaixo do esperado, especialmente em comparação com os dados do estudo anterior [Bou et al. 2023]. Os novos resultados, apresentados nas Tabelas 1, 2 e 3, evidenciam aspectos que geraram preocupações sobre a eficácia e a aplicabilidade da metodologia proposta, exigindo uma análise mais aprofundada dos fatores que influenciaram esse desempenho.

Tabela 1. Análise Conteúdo Bruto GPT-3.5-Turbo

Temática	Precision	Recall	F1-Score	Comentários
Homofobia	100%	73,09%	84,45%	680
Racismo	100%	50,15%	66,80%	680
Sexismo	100%	40,15%	57,29%	680
Neutro	100%	92,79%	96,26%	680

Tabela 2. Análise pré-processada GPT-3.5-Turbo

Temática	Precision	Recall	F1-Score	Comentários
Homofobia	100%	73,97%	85,04%	680
Racismo	100%	56,91%	72,54%	680
Sexismo	100%	30,59%	46,85%	680
Neutro	100%	89,56%	94,49%	680

Tabela 3. Análise pré-processada GPT-4.0

Temática	Precision	Recall	F1-Score	Comentários
Homofobia	100%	95,59%	97,74%	680
Racismo	100%	57,35%	72,90%	680
Sexismo	100%	69,12%	81,74%	680
Neutro	100%	91,47%	95,55%	680

Inicialmente destaca-se a precisão de 100% apresentada em todas as tabelas, o resultado atribuído à idealização da veracidade dos conteúdos presentes nos *datasets* utilizados. Contudo, análises posteriores revelaram que a precisão desses dados, especialmente nos conjuntos relacionados a racismo e sexismo, não é completamente confiável, e onde será discutido ao decorrer do texto. Em relação aos resultados obtidos na análise com o conteúdo pré-processado expostos nas Tabelas 2 e 3, a temática de homofobia, apresentou

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

na métrica F1-*Score* um desempenho de 85,04% para o modelo GPT-3.5-Turbo e 97,74% para o modelo GPT-4.0, evidenciando um aumento superior a 12% entre os modelos. Esses dados destacam a eficácia aprimorada do modelo GPT-4.0.

Para os conteúdos neutros, as métricas de *recall* e F1-*Score* no modelo GPT-3.5-Turbo alcançaram 89,56% e 94,49%, respectivamente, como apresentado na Tabela 2. De forma semelhante, o modelo GPT-4.0 apresentado na Tabela 3 demonstrou um desempenho superior, com um aumento aproximadamente 2% na métrica de *recall*, resultando em 91,47% e 1% no F1-*Score*, finalizando com 95,55%, desempenhando uma sutil melhora da precisão na análise.

Em contextualização a aplicação do modelo GPT-4.0, conforme os resultados da Tabela 3, optou-se pelo uso apenas do conteúdo pré-processado para otimizar os custos, dado que esse modelo, embora mais preciso, possui custos de processamento significativamente elevados. Assim, a análise detalhada entre o desempenho com conteúdo bruto e pré-processado foi limitada, priorizando uma abordagem financeiramente viável e representativa de cenários práticos que demandam escalabilidade e custo-benefício.

Observou-se uma preocupação com os resultados baixos nas métricas de *recall* para todas as temáticas de discurso de ódio analisadas, principalmente no modelo GPT-3.5-Turbo em ambos os casos. As temáticas de racismo e sexismo apresentaram valores abaixo de 57%, com 56,91% e 30,59%, respectivamente. Já no modelo GPT-4.0, observou-se uma variação inesperada nas métricas de *recall* entre as temáticas de sexismo e racismo. Para sexismo, o *recall* atingiu 69,12%, enquanto para racismo o valor foi de 57,35%, onde no modelo anterior os resultados era o inverso.

As métricas de F1-*Score* também chamaram atenção pelos resultados inferiores ao esperado em relação ao estudo inicial [Bou et al. 2023], especialmente para as temáticas de racismo e sexismo, cuja expectativa era de valores próximos a 90%. No GPT-3.5-Turbo, o F1-*Score* foi de 72,54% para racismo e 46,85% para sexismo. O GPT-4.0 apresentou uma melhoria significativa para sexismo, com um aumento de 34%, alcançando 81,74%. No entanto, para racismo, o aumento foi marginal, resultando em 72,90%.

Apesar dos indicativos em relação aos comparativo dos modelos terem uma melhora nos resultados, entre GPT-3.5-Turbo e GPT-4.0, algumas inconsistências foram claras em relação às expectativas esperadas, sugerindo a presença de pequenos ruídos na análise. Diante disso, foi realizada uma análise manual detalhada nos *datasets* que apresentaram maior discrepância nos resultados, especificamente nas temáticas de sexismo e racismo, com o objetivo de validar minuciosamente cada comentário.

4.1. Análise Manual

A análise manual[11] visível nas Tabelas 4 e 5, revelou a presença de diversos comentários que não se enquadravam nas categorias de discurso de ódio avaliadas, como racistas ou sexistas, e também comentários repetidos especialmente no *dataset* de sexismo. Essa descoberta foi crucial, pois evidenciou que, na maioria dos casos, os modelos aplicados realizaram classificações corretas ao não rotularem esses comentários como preconceituosos. Esses resultados ressaltam tanto a eficácia dos modelos quanto a importância de uma curadoria rigorosa na construção e utilização de *datasets*, a fim de garantir a confiabilidade das análises.

Na análise manual, foi considerado o contexto dos comentários com base em sua

[11]Análise manual Disponível em: `https://github.com/guilhermebou/Analysis-Dataset-BOU-Guard-A-Study-in-Real-World-Scenarios/blob/main/Data/Results/manual_analysis_comments_racist_sexist.xlsx`

agressividade nas ofensas relacionadas às temáticas abordadas. Foram analisados elementos como sarcasmo, levando em conta brincadeiras de má-fé, e o uso de termos pejorativos em determinadas palavras associadas às temáticas racista e sexista.

Tabela 4. *Datasets* Validados Manualmente Análise GPT-3.5-Turbo

Temática	Precision	Recall	F1-Score	Comentários
Racismo	100%	76,66%	86,79%	450
Sexismo	100%	40,35%	57,50%	456

Tabela 5. *Datasets* Validados Manualmente Análise GPT4.0

Temática	Precision	Recall	F1-Score	Comentários
Racismo	100%	84,59%	91,65%	450
Sexismo	100%	91,67%	95,65%	456

Em relação aos resultados obtidos na análise pré-processada utilizando o modelo GPT-3.5-Turbo, considerando os 680 comentários apresentados na Tabela 2, a revisão manual reduziu esse número para 456. Apesar da redução de 224 comentários, os 456 analisados manualmente foram considerados confiáveis para a avaliação. Inicialmente, o sexismo apresentou um F1-*Score* de 57,29% e um *recall* de 40,15%. Já na análise manual, houve um aumento discreto de 0,31% no F1-*Score* e de 0,20% no *recall*, alcançando valores de 57,50% e 40,35%, respectivamente. Esses resultados indicam que o modelo GPT-3.5-Turbo não está otimizado para a detecção de conteúdos relacionados ao sexismo, apresentando desempenho limitado nesse aspecto.

Em comparação com o modelo GPT-4.0, os resultados da análise manual apresentaram um progresso significativo. Conforme demonstrado anteriormente, ao comparar a análise inicial com o conteúdo bruto, este novo comparativo revelou que as métricas de *recall* e F1-*Score* superaram 91%, atingindo 91,67% e 95,65%, respectivamente. Esses resultados destacam a viabilidade de utilizar o modelo GPT-4.0 na avaliação de comentários sexistas, evidenciando sua superioridade.

Levando em consideração a análise de conteúdos racistas na revisão manual, os modelos GPT-3.5-Turbo e GPT-4.0 apresentaram uma elevação significativa em seus resultados. Inicialmente, o *recall* para o GPT-3.5-Turbo foi de 56,91%, conforme demonstrado na Tabela 2. Entretanto, nesta nova análise, visualizada na Tabela 4, com os comentários revisados, observou-se um aumento de 19,75%, atingindo 76,66%. Esse resultado superou o desempenho do GPT-4.0, que, nos 680 comentários avaliados na primeira avaliação, obteve 72,54%, como também indicado na Tabela 2. Esses dados evidenciam que a revisão e o refinamento do *dataset* foram fundamentais para melhorar os resultados obtidos.

Em comparação com o modelo GPT-4.0, os resultados obtidos em relação ao conteúdo revisado no *dataset* racista refletem sua superioridade. Anteriormente, o modelo apresentava um *recall* de 57,35% e um F1-*Score* de 72,90%, conforme representado na Tabela 3. No entanto, neste novo cenário, com os comentários revisados, o *recall* teve um aumento significativo de 27,24%, enquanto o F1-*Score* ultrapassou a marca de 91%, alcançando 91,65%. Em termos comparativos, isso representa uma elevação de 18,75%, novamente destacando a importância da revisão do *dataset* para a melhoria dos resultados.

4.2. Análise Comparativa entre comentários brutos e pré-processados

Na comparação entre comentários brutos e pré-processados, os resultados do modelo GPT-3.5-Turbo, apresentados nas Tabelas 1 e 2, indicam que, para a temática homofóbica,

o *recall* aumentou de 73,09% para 73,97%, e o F1-*Score* de 84,45% para 85,04%, embora sem grande impacto. Já para a temática racista, o *recall* subiu de 50,15% para 56,91%, e o F1-*Score* de 66,80% para 72,54%. Esses avanços mostram que, apesar da contextualização limitada, a identificação de palavras-chave, como termos pejorativos, foi decisiva para melhorar a classificação.

Para a temática neutra, as métricas de *recall* e F1-*Score* foram superiores na análise de comentários brutos, com *recall* de 92,79% e F1-*Score* de 96,26%, enquanto na análise pré-processada esses valores reduziram para 89,56% e 94,49%, respectivamente. Isso evidencia a importância da contextualização na avaliação de discursos de ódio, já que o conteúdo bruto preserva nuances essenciais para a análise.

Na temática sexista, os resultados seguiram a mesma tendência. O *recall* apresentou uma queda significativa de mais de 10%, passando de 40,15% no conteúdo bruto para 30,59% no pré-processado. O F1-*Score* também diminuiu, de 57,29% para 46,85%. Esses dados reforçam que a remoção de informações contextuais no pré-processamento pode comprometer a detecção de discursos de ódio, especialmente em temáticas onde os detalhes são determinantes para uma classificação precisa.

5. Conclusão e Limitações

O presente trabalho abordou o problema da identificação de palavras e expressões impróprias relacionadas a termos ofensivos de cunho homofóbico, racista e sexista. Para mitigar esse problema, foi proposto o mecanismo BOU-Guard, que, nesta abordagem, realiza a detecção multilabel de comentários odiosos de usuários da redes sociais. A implementação utilizou a linguagem de programação *Python*, as bibliotecas *NLTK*, *Pandas*, *String* e integra a API fornecida pelos modelos GPT-4.0 e GPT-3.5-Turbo para realizar a classificação e listagem das expressões detectadas.

Os resultados obtidos por meio de uma prova de conceito revelam que o BOU-Guard apresentou um bom desempenho na detecção de comentários odiosos associados à homofobia, racismo e sexismo em redes sociais. Em destaque, concluímos a superioridade do GPT-4.0 em relação ao GPT-3.5-Turbo, em relação ao conteúdo pré-processado e com os comentários validados manualmente, especialmente na análise de conteúdos relacionados à homofobia e ao sexismo. O modelo GPT-4.0 atingiu um F1-*Score* de 97,74% para homofobia e 81,74% para sexismo em dados pré-processados, evidenciando melhorias significativas em comparação ao GPT-3.5-Turbo, que alcançou 85,04% e 46,85%, respectivamente. Utilizando os dados validados manualmente o salto do F1-*Score* ao sexismo foi para 95,65% e racismo indo a 91,65%. No entanto, a temática de racismo apresentou uma evolução mais discreta ao conteúdo pré-processado, com F1-*Scores* de 72,54% no GPT-3.5-Turbo e 72,90% no GPT-4.0.

A análise feita com os *datasets* validados manualmente, mostrou que o *Recall* foi um ponto crítico, especialmente para racismo e sexismo, onde ambos os modelos apresentaram resultados abaixo das expectativas, reforçando a importância de curadoria rigorosa nos dados. Apesar disso, o GPT-4.0 demonstrou desempenho consistente, alcançando um F1-*Score* superior a 95% para sexismo na análise manual, confirmando sua eficácia em cenários práticos.

As limitações dos resultados estão ligadas à qualidade dos *datasets*, que apresentaram ruídos, comentários repetidos e classificações inconsistentes, dificultando a avaliação dos modelos. A validação da veracidade dos conteúdos mostrou-se essencial, pois erros ou ambiguidades comprometem o desempenho e geram resultados incoerentes. O pré-processamento, embora complexo, destacou-se nas temáticas de homofobia e racismo.

A análise manual foi crucial para corrigir inconsistências, aprimorar a interpretação dos modelos e validar os resultados, garantindo maior precisão na comparação entre o GPT-3.5-Turbo e o GPT-4.0.

Como trabalhos futuros, pretende-se desenvolver funcionalidades adicionais na extensão proposta como prova de conceito para o mecanismo BOU-Guard. Entre as futuras implementações, destacam-se um sistema de denúncia integrado ao suporte das redes sociais em que o usuário está navegando e uma funcionalidade que forneça ao usuário uma explicação sobre os motivos que levaram à classificação de um comentário como discurso de ódio.

Referências

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. SBC.

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

Houaiss, A. (1986). *Dicionário Houaiss da língua portuguesa*. Objetiva Instituto Antônio Houaiss de Lexicografia.

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.

Presidência da República do Brasil (2014). Lei n° 12.965, de 23 de abril de 2014 - marco civil da internet.

SaferNet and dos Direitos Humanos e da Cidadania MDHC, M. (2023). Enfrentamento ao discurso de ódio. Technical report, Observatório Nacional dos Direitos Humanos ObservaDH.

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.