BOU-Guard: Uma Abordagem para Detecção de Conteúdo Impróprio na Internet

Guilherme Bou[1], Adriano M. Rocha[1], Vagner E. Quincozes[2], Silvio E. Quincozes[3] e Juliano F. Kazienko[4]

[1]FACOM – Universidade Federal de Uberlândia (UFU) – Monte Carmelo, Brasil

[2]Universidade Federal do Pampa (UNIPAMPA) – Alegrete, Brasil

[3]IC – Universidade Federal Fluminense (UFF) – Niterói, Brasil

[4]CTISM – Universidade Federal de Santa Maria (UFSM) – Santa Maria, Brasil

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

vequincozes@unipampa.edu.br, sequincozes@gmail.com

kazienko@redes.ufsm.br

Abstract. *Amidst the ever-expanding digital landscape, exposure to inappropriate content such as racism, homophobia, and sexism has become an increasingly pressing concern. Despite the existing literature on online hate speech, significant limitations persist, including a lack of automation and effective warning mechanisms. This article proposes an innovative approach, introducing the BOU-Guard (Behavior Observation Unit - Guard), based on GPT-3.5-Turbo technology, to detect and filter prejudiced or offensive content. Through a proof of concept, we demonstrated that the proposed mechanism applied to 30 web pages can detect offensive content with a high F1-Score, on average, to content related to homophobia (94.69%), racism (98.45%), and sexism (98.09%).*

Resumo. *Em meio ao cenário digital em constante expansão, a exposição a conteúdo impróprio, como racismo, homofobia e sexismo, tornou-se uma preocupação cada vez mais premente. Apesar da literatura existente sobre discurso de ódio online, persistem limitações significativas, incluindo a falta de automação e mecanismos de alerta eficazes. Este artigo propõe uma abordagem inovadora, apresentando o BOU-Guard (Behavior Observation Unit - Guard), baseado na tecnologia GPT-3.5-Turbo, para detectar e filtrar conteúdos preconceituosos ou ofensivos. Por meio de uma prova de conceito, demonstrou-se que a aplicação do mecanismo proposto na análise de 30 páginas web é capaz de detectar conteúdos ofensivos com alta F1-Score média para conteúdos relacionados a homofobia (94,69%), racismo (98,45%) e machismo (98,09%).*

1. Introdução

Nos últimos anos, com a popularização da Internet e das redes sociais, as pessoas de todas as idades estão frequentemente expostas a conteúdos impróprios, como racismo, homofobia e machismo. De acordo com [Bliuc et al. 2018], as formas pelas quais a Internet pode facilitar a expressão e disseminação de visões e ideologias racistas têm sido objeto de um crescente corpo de pesquisa em várias disciplinas. O

racismo pode se manifestar em diferentes níveis, incluindo interpessoal, institucional e cultural, e vivenciar o racismo pode levar a uma série de consequências negativas, como depressão, hipertensão e doenças cardíacas coronárias [Bliuc et al. 2018]. Ademais, com a utilização crescente da Internet para estabelecer novas relações sociais e buscar apoio online, especialmente entre os adolescentes, observa-se um aumento na exposição a riscos de vitimização sexual online e conteúdos prejudiciais, como homofobia e discriminação [Gámez-Guadix and Incera 2021].

Diante das preocupações geradas pela produção e conteúdo impróprio na Internet, as principais plataformas de mídia social, como Facebook, YouTube e Twitter, têm usado inteligência artificial para moderar, monitorar e remover conteúdos prejudiciais ou ilegais. No entanto, esses sistemas automatizados têm sido criticados por questões como falta de transparência, viés e possíveis danos a comunidades marginalizadas. Além dessas preocupações, há uma ênfase na moderação de conteúdo do lado das plataformas. Essa abordagem muitas vezes não é suficiente para suprimir o conteúdo prejudicial e evitar que o mesmo chegue até seus usuários [Wang et al. 2023]. Por outro lado, a literatura carece de soluções centradas no usuário que sejam eficientes para combater a exposição a tais conteúdos. A implementação de medidas de controle parental, por exemplo, requer o desenvolvimento de soluções inovadoras para evitar que crianças e adolescentes sofram tal exposição [Martins et al. 2020].

De modo a resolver as lacunas da literatura, neste trabalho é proposto o BOU-Guard, um mecanismo baseado na tecnologia *Generative Pre-Trained Transformer* (GPT-3) que implementa Unidade de Observação de Comportamento, do inglês, *Behavior Observation Unit* (BOU) para monitorar e detectar conteúdo impróprio na Internet. Ao passo que o BOU-Guard é um mecanismo que pode ser integrado à diversas aplicações e contextos, neste trabalho assume-se um cenário onde tal mecanismo seja acoplado a uma extensão para o navegador *Google Chrome* para aplicações tais como o controle parental de conteúdo acessado. Diante desse cenário, foi desenvolvida uma prova de conceito visando a detecção de conteúdo de cunho racista, homofóbico e machista. Nas demais seções são apresentados os trabalhos relacionados (Seção 2), a abordagem proposta (Seção 3), uma prova de conceito (Seção 4) e os resultados preliminares obtidos (Seção 5).

2. Trabalhos Relacionados

Há diversas propostas recentes que objetivam minimizar a exposição das pessoas a conteúdos impróprios. Dentre tais propostas, existem aquelas que exploram o potencial dos modelos de linguagem de grande escala, como o GPT-3, na detecção de ódio e linguagem abusiva. Um exemplo consiste na proposta de [Chiu et al. 2021], na qual os autores fornecem um trecho de texto para o GPT-3 e ele classifica esse trecho como “Racista”, “Sexista” ou “Neutro”. Outro trabalho [Wang et al. 2023] adota o uso do GPT-3 a fim de gerar explicações para *tweets* que contém trechos com e sem ódio. Em ambos os trabalhos, são gerados *prompts* de entrada para o GPT-3.

Também, há outras propostas que visam identificar e filtrar conteúdo inadequado em discussões online usando (i) aprendizado profundo [Yenala et al. 2018] e (ii) aprendizado estatístico [Sheth et al. 2022]. Enquanto a primeira usa uma combinação de camadas convolucionais e bidirecionais LSTM para capturar padrões sequenciais e semântica global em textos, a segunda visa incorporar conhecimento explícito em um algoritmo de

aprendizado estatístico para detectar a toxidade em comunicações online.

Em contraponto com tais propostas, o mecanismo BOU-Guard, proposto neste trabalho, adota o modelo GPT-3.5-Turbo para listar expressões impróprias em textos, ao invés de apenas categorizar as palavras. Além disso, nenhuma das abordagens mencionadas anteriormente é centrada no usuário, ou seja, elas não podem ser facilmente instaladas em navegadores ou dispositivos do lado do usuário.

3. BOU-Guard

Neste trabalho, é proposto o BOU-Guard, uma abordagem baseada na tecnologia GPT 3.5 para minimizar a exposição das pessoas a conteúdo racista, homofóbico e/ou machista. A abordagem consiste em utilizar o GPT-3.5-Turbo[1] com o objetivo de detectar expressões relacionadas a conteúdos impróprios associados a tais temáticas.

A abordagem adotada consiste na realização de uma raspagem de dados utilizando a linguagem de programação *Python* e a biblioteca *Beautiful Soup* para extrair o conteúdo dentro das *tags HyperText Markup Language* (HTML) de uma página. Esse conteúdo é convertido em uma *string*, que é passada como parâmetro para uma função responsável por analisar, identificar e listar expressões impróprias relacionadas a termos específicos, como homofobia, racismo ou machismo. Essa função é responsável por realizar a integração do BOU-Guard com o modelo GPT-3.5-Turbo, popularmente conhecida por integrar o *ChatGPT*[2], através de sua Interface de Programação de Aplicação, do inglês, *Application Programming Interface* (API).

A proposta do BOU-Guard assume que o conteúdo da HTML de uma página seja obtido através de uma requisição por meio do protocolo *Hypertext Transfer Protocol* (HTTP). Para tanto, a *Uniform Resource Locator* (URL) da página deve ser recuperado por meio de uma extensão do navegador *Google Chrome*[3]. A resposta dessa requisição é processado pela API do modelo GPT-3.5-Turbo, que analisa o conteúdo e enumera termos e expressões impróprias, permitindo que o BOU-Guard execute as ações apropriadas. Exemplos de ações incluem bloqueio de conteúdo ofensivo ou notificação de pais ou responsáveis acerca do conteúdo acessado, quando o usuário é uma criança ou adolescente.

4. Prova de Conceito

De modo a validar a abordagem proposta, uma prova de conceito foi implementada seguindo a seguinte metodologia. Primeiramente, buscou-se no motor de pesquisa Google por páginas da internet que continham amostras de conteúdos associados ao racismo, machismo ou homofobia. Para tanto, as seguintes frases foram usadas para a realização da busca: (i) “Exemplos de frases machistas”, (ii) “Exemplos de frases homofóbicas” e (iii) “Exemplos de frases racistas”. Em seguida, para cada frase buscada, foram consideradas as 10 primeiras páginas retornadas pelo motor de busca, descartando-se aquelas que exigiam login. Então, tais páginas foram acessadas no navegador Google Chrome e processadas pela extensão que implementa o BOU-Guard. Nesta prova de conceito, foram considerados os conteúdos contidos nas *tags* `<p>`, `<span>`, `<strong>` e `<h1>`

[1]Modelo GPT-3.5-Turbo. Disponível em: `https://platform.openai.com/docs/models/`
[2]ChatGPT. Disponível em: `https://chat.openai.com/`
[3]Google Chrome. Disponível em: `https://www.google.com/intl/pt-BR/chrome/`

à <h6>. A fim de aferir a eficiência do BOU-Guard, para cada página, os termos e expressões identificados foram classificados de acordo com os seguintes indicadores:

- Verdadeiros Positivos (VP): conteúdo impróprio devidamente identificado;
- Falsos Negativos (FN): conteúdo impróprio existente na página não detectado;
- Falso Positivo (FP): o conteúdo apontado como impróprio pelo BOU-Guard, na verdade, não contém expressões ou termos ofensivos e/ou preconceituosos.

A partir desses três indicadores, foram computadas as métricas precisão, revocação e F1-Score. A precisão (*precision*) é a proporção de identificações positivas feitas corretamente (isto é, a porcentagem de conteúdo identificado como ofensivo que realmente é ofensivo). Já a revocação (*recall*) representa a proporção de positivos reais que foram identificados corretamente (isto é, a porcentagem de todo o conteúdo ofensivo que foi corretamente identificado). Por fim, a F1-Score consiste em uma média harmônica entre precisão e revocação, dando uma visão geral do desempenho do modelo. Matematicamente, tais métricas podem ser representadas pelas seguintes 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[4]. Os detalhes para reprodutibilidade dos resultados podem ser encontrados no arquivo `README.MD`.

5. Resultados e Discussões

Nesta seção serão discutidos os resultados obtidos pelo BOU-Guard na análise dos 30 sites considerados na prova de conceito introduzida na Seção 4. Em geral, os resultados apresentados na Tabela 1 demonstraram que o BOU-Guard é eficiente na identificação de palavras e frases impróprias, com todas as médias acima de 93% para as métricas referentes a todas as temáticas analisadas.

Tabela 1. Desempenho médio por temática (10 sites por temática).

Temática	Precision	Recall	F1-Score
Homofobia	93,33%	98,07%	94,69%
Racismo	100%	97,13%	98,45%
Machismo	100%	96,48%	98,09%

Em primeiro lugar, destaca-se a métrica *recall*, que apresentou uma média acima de 96,48% para os três temas, mostrando-se eficaz na recuperação dos casos positivos. Em relação a métrica *precision*, observou-se que o modelo tem uma baixa taxa de falsos positivos, pois tal métrica foi superior a 93,33% para os três temas. Por fim, os resultados da métrica *F1-Score* mostram que BOU-Guard é relevante para identificar corretamente os casos positivos, pois os resultados foram superiores a 94,69% nos três casos estudados.

Em alguns casos específicos, foram registrados resultados abaixo da média devido à estruturação em *HTML* da página, o que afetou a coleta do conteúdo usado como parâmetro. Esses dados estavam localizados em outras dependências. A Figura 1 mostra os resultados das métricas *recall*, *precision* e *F1-Score*, para cada site avaliado.

A Figura 1 mostra que os resultados sobre a detecção de conteúdo ofensivo relacionado à homofobia, racismo e machismo apresenta um desempenho geralmente positivo

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

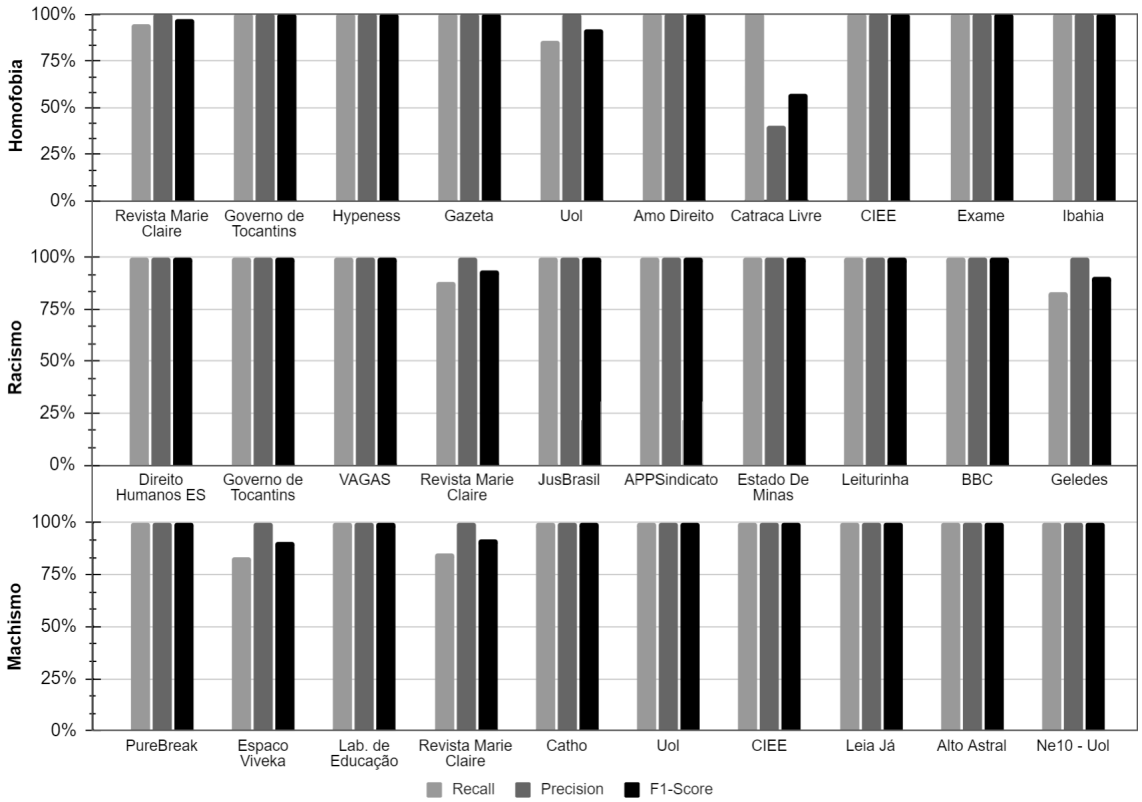

Figura 1. Desempenho do BOU-Guard na detecção de textos ofensivos.

do BOU-Guard. No que diz respeito a homofobia (gráfico superior da Figura 1), a maioria dos sites apresentou um desempenho positivo, com altos valores de *recall*, *precision* e *F1-Score* (*i.e.,* próximos de 100%). No entanto, destaca-se o site "Catraca Livre", que apresentou resultados inferiores (*i.e.*, 40% e 57,14% para *precision* e *F1-Score*, respectivamente). Tais resultados se devem à presença de 6 falsos positivos em meio às 10 frases homofóbicas anunciadas no título da página. Uma vez que o título da página não estava consistente com o seu conteúdo, a API GPT-3.5-Turbo, que é também uma ferramenta generativa, foi induzida ao erro e produziu e reportou frases que não existiam na página analisada. De acordo com experimentos isolados, essa limitação pode ser mitigada através da escrita de instruções explícitas para o motor do mecanismo evitar a geração de conteúdo inexistente na página.

Quanto ao racismo (gráfico intermediário), a maioria dos sites obtiveram bons resultados, evidenciando a eficácia do BOU-Guard. Dos 10 sites avaliados, 8 apresentaram 100% de F1-Score, enquanto que somente 2 apresentam F1-Score de 93,62% e 90,91%, respectivamente. Por fim, no contexto de machismo (gráfico inferior), a detecção também apresentou um bom desempenho geral, com a maioria dos sites atingindo valores elevados de *recall*, *precision* e *F1-Score*.

Em resumo, os resultados obtidos demonstram que o BOU-Guard é capaz de detectar efetivamente conteúdos ofensivos relacionados a homofobia, racismo e machismo. Assim, a ferramenta proposta é promissora para a detecção automática de conteúdo ofensivo, especialmente considerando o quão desafiador pode ser a moderação automática de conteúdo devido às nuances da linguagem e ao risco de falsos positivos. No entanto, a

melhoria contínua do sistema será essencial para manter a precisão e minimizar possíveis falsos positivos, maximizando assim as métricas *recall* e F1-Score.

6. Conclusão, Limitações e Trabalhos Futuros

O presente trabalho em andamento abordou o problema sobre a identificação, em tempo real, de palavras e expressões impróprias relacionadas a termos ofensivos de cunho homofóbico, racista e machista. Para mitigar esse problema, foi proposto o mecanismo BOU-Guard, o qual realiza a raspagem de dados utilizando a linguagem de programação *Python*, a biblioteca *Beautiful Soup*, e possui integração com a API fornecida pelo modelo GPT-3.5-Turbo para realizar a análise, identificação e listagem das expressões desejadas.

Os resultados preliminares obtidos através de uma prova de conceito revelam que o BOU-Guard é um mecanismo promissor na detecção de termos associados à homofobia, ao racismo e ao machismo, alcançando, respectivamente, em média, uma F1-Score de 94,69%, 98,45% e 98,09% para os sites analisados.

Como ainda se trata de um trabalho em andamento o BOU-Guard possui algumas limitações. Uma limitação atual consiste no limite de 4.096 *tokens* por requisição, que é imposto pela API empregada. Tal limitação será mitigada em trabalhos futuros através de filtros para o pré-processamento de dados e eliminação de redundâncias. Uma outra abordagem simplificada seria dividir o conteúdo em múltiplas requisições. Ademais, como trabalhos futuros, pretende-se finalizar o desenvolvimento da extensão que foi proposta como prova de conceito para o mecanismo BOU-Guard. Futuras funcionalidades para a extensão incluem alertas de conteúdo impróprio em tempo real para o controle parental.

Referências

Bliuc, A.-M., Faulkner, N., Jakubowicz, A., and McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, 87:75–86.

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

Gámez-Guadix, M. and Incera, D. (2021). Homophobia is online: Sexual victimization and risks on the internet and mental health among bisexual, homosexual, pansexual, asexual, and queer adolescents. *Computers in human behavior*, 119:106728.

Martins, M. V., Formiga, A., Santos, C., Sousa, D., Resende, C., Campos, R., Nogueira, N., Carvalho, P., and Ferreira, S. (2020). Adolescent internet addiction–role of parental control and adolescent behaviours. *International Journal of Pediatrics and Adolescent Medicine*, 7(3):116–120.

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.