UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Discurso de ódio nas redes sociais: um olhar pela Inteligência Artificial

Pâmela Ferreira

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Discurso de ódio nas redes sociais: uma análise pela Inteligência Artificial

Pâmela Ferreira

USP - São Carlos

2023

Pâmela Ferreira

Discurso de ódio nas redes sociais: uma análise pela Inteligência Artificial

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Profa. Dra. Solange O. Rezende
Coorientador: Dr. Brucce Neves Santos

USP - São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

P383d

Ferreira, Pâmela
Discurso de ódio nas redes sociais: uma análise pela Inteligência Artificial / Pâmela Ferreira; orientador Profa. Dra. Solange O. Rezende; coorientador Dr. Brucce Neves Santos. -- São Carlos, 2023.
31 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. discurso de ódio. 2. Twitter. 3. ChatGPT. 4. pré-processamento . I. O. Rezende, Profa. Dra. Solange , orient. II. Neves Santos, Dr. Brucce , coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

AGRADECIMENTOS

Peço licença primeiramente a agradecer a minha ancestralidade e orixás por estar aqui, materializando um projeto no qual acho grandioso e necessário, a minha mãe Iraci que desde sempre me apoiou nos meus sonhos e movimentos, se caminho hoje para a grandiosidade sou imensamente grata a ti, ao meu pai Lucival por me incentivar no desenvolvimento e conclusão da pós-graduação, a continuar com firmeza e persistência.

A mãe Oyassi Roxinha, por sempre acreditar na minha potência desde o início dos estudos em tecnologia, me fortalecer dia após dia a ter coragem de enfrentar sempre os maiores desafios, ao professor Valdinei Freire, agradeço o apoio que foi fundamental na área de Inteligência Artificial, pelos estudos semanais que me possibilitaram seguir firme nesse projeto e que me fortaleceu para continuar na pós-graduação mesmo com todos os desafios.

A minha família em especial a Simone, Isabela e Ivanilton pela paciência e acolhimento nos dias mais desafiadores.

Aos meus amigos, amigas e amigues pelo cuidado, incentivo, apoio nos dias mais intensos que fossem, pela parceria, encorajamento de continuidade do projeto que por vezes tive dúvidas se iria ser realizado mesmo sabendo que precisava ser feito, algo necessário para mim e que vejo como importância de discussão para o mundo.

Cito alguns por aqui como a Bárbara Cavalcante, Elô Nunes, Barbara Bispo, João Paulo Gomes, entre outros que mesmo distantes sei que sempre estão me apoiando de alguma forma, sou grata demais por essa rede, aprendo a cada dia a importância e sabedoria que é pedir ajuda e contar com quem amamos nos momentos mais desafiadores.

Não poderia deixar de memorar a minha grande amiga Laís, na qual eu fiz a citação de uma frase para essa monografia, onde eu estiver lembrarei de ti com carinho e amor, sou muito grata por tudo, em tudo que me incentivou e que sua memória seja sempre lembrada, você foi muito importante para mim.

Por último e não menos importante, quero agradecer imensamente pelo apoio para a conclusão desse trabalho, a professora Solange pela dedicação, incentivo e organização. Ao Brucce Neves pelo incrível apoio que tive, pela paciência, desempenho, olhar analítico e certeiro naquilo que precisa ser melhorado, agradeço demais!

“Respeito ao ser humano”
Mãe Oyassi Roxinha (2020)

“Que os ventos sombrios de ontem e de hoje não permaneçam em nosso futuro. Que nossas lutas de hoje, garantam dias de tranquilidade, amor e paz no amanhã que se aproxima. Sem jamais esquecer dos reais motivos pelos quais lutamos|”.

Laís Marçal (2022)

RESUMO

FERREIRA, Pâmela Discurso de ódio nas redes sociais: uma análise pela Inteligência Artificial. 2023. 29 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Com a limitada disponibilidade de conjuntos de dados na língua portuguesa, contrastada com a imensa quantidade de informações dos usuários em plataformas digitais, com rápida disseminação de conteúdo no Twitter em poucos caracteres, este estudo realiza a etapa de pré-processamento e análise dos *tweets* submetidos ao modelo de arquitetura do ChatGPT 3.5. Feito a análise, o modelo obteve como resultado na comparação de dois experimentos, uma F1-Score menor na análise métrica de 37%, que por ser uma classificação de multirrótulo, tem a dependência da detecção de discurso de ódio e identificação das categorias.

Palavras-chave: discurso de ódio, Twitter, ChatGPT, pré-processamento

ABSTRACT

FERREIRA, E. P. "Hate Speech on Social Media: An Analysis by Artificial Intelligence" 2023. 29 f. - Completion of Course Work (MBA in Artificial Intelligence and Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2020.

With the limited availability of datasets in the Portuguese language, contrasted with the immense amount of user information on digital platforms and the rapid dissemination of content on Twitter within a few characters, this study undertakes the preprocessing and analysis stage of *tweets* submitted to the ChatGPT 3.5 architecture model. After the analysis, the model resulted in a lower F1-Score of 37% in the comparison of two experiments. As it is a multi-label classification, it is dependent on hate speech detection and category identification.

Keywords: hate speech, Twitter, ChatGPT, preprocessing.

LISTA DE TABELAS

SUMÁRIO

1) INTRODUÇÃO

O crescente fenômeno do discurso de ódio representa um desafio crucial na contemporaneidade, impulsionado em parte pelo volume sem precedentes de dados gerados no âmbito de big data. Esse avanço tecnológico, embora tenha impactado positivamente áreas como saúde e educação, trouxe consigo um aumento alarmante do discurso de ódio, gerando problemas sociais significativos.

O termo "discurso de ódio" é definido como o uso de palavras que tendem a insultar, intimidar ou assediar pessoas com base em características como raça, cor, etnia, nacionalidade, sexo ou religião (Winfried Brugger, 2007). No contexto brasileiro, a Safernet (2022) revelou um preocupante crescimento desse fenômeno em diversas categorias, como a apologia a crimes e contra a vida, LGBTFobia, misoginia, entre outros, registrando um aumento de 67% nas denúncias à Central Nacional de Denúncias (CND) em comparação ao ano anterior.

Esse aumento do discurso de ódio não apenas afeta diretamente a segurança dos usuários no ambiente online, mas também se reflete de maneira violenta no mundo real, como evidenciado pelos ataques e invasões a escolas brasileiras, correlacionados ao discurso de ódio propagado anteriormente (CARA, 2023).

Torna-se, portanto, essencial compreender como as instituições podem criar um ambiente virtual seguro, minimizando suas repercussões no mundo físico. A análise da Safernet (2022) demonstrou que os períodos eleitorais brasileiros entre 2018 e 2022 registraram um aumento significativo no discurso de ódio. Essas manifestações virtuais frequentemente extrapolam para o mundo real, apresentando desafios na identificação e responsabilização dos agressores (GLOBO, 2023).

Embora existam desafios consideráveis na identificação e moderação desses discursos, avanços significativos foram alcançados no desenvolvimento de ferramentas e métodos para combatê-los. Diversas estratégias, abrangendo desde o emprego de algoritmos de Inteligência Artificial até métodos fundamentados em abordagens comunitárias, têm sido estudadas para enfrentar esse desafio em constante crescimento (SCHÄFER, 2022).

Neste contexto, este estudo direciona seu foco para o Twitter, uma plataforma ágil na disseminação de informações e discursos. A restrição de caracteres pode tornar os discursos de ódio mais evidentes, o que somado à acessibilidade da API do Twitter, oferece oportunidades significativas para a detecção desses discursos (AYO et al, 2020).

A pesquisa sobre o discurso de ódio é vital, especialmente pela falta de regulamentação eficaz em meio ao crescimento exponencial das *Big Techs*. Este trabalho busca, portanto, desenvolver e aplicar modelos de Processamento de Linguagem Natural (PLN) para identificar discursos de ódio em *tweets* em língua portuguesa, visando contribuir para a mitigação desses discursos em plataformas online baseados na nossa língua.

A continuação deste estudo abrange uma análise mais profunda dos impactos sociais do discurso de ódio, especialmente para grupos mais afetados diretamente, como mulheres negras e população LGTBQIA+ (DATA LABE, 2023). O objetivo é compreender e identificar padrões predominantes nesses discursos, fornecendo subsídios para estratégias eficazes de mitigação.
Essa necessidade urgente de pesquisas e ações assertivas para lidar com o discurso de ódio online é acentuada pela escassez de estudos interdisciplinares na interseção entre a computação, Inteligência Artificial e ética. É fundamental direcionar essas inovações tecnológicas por princípios éticos, contribuindo efetivamente para a mitigação desse problema social. De certo os avanços da Inteligência Artificial têm ocasionado diversos incentivos para várias finalidades ligadas a agilidade e contribuições significativas, mas se torna urgente tratar na mesma proporção os debates éticos que ela já opera (UNESCO, 2023).

1.1. Objetivos:

Esse trabalho tem como objetivo geral analisar a capacidade do modelo ChatGPT na detecção de discursos de ódio na plataforma Twitter. Para alcançar esse objetivo são propostos dois objetivos específicos:

1. Realizar uma análise detalhada do conjunto de dados. Isso inclui as etapas pré-processamento e limpeza dos dados para prepará-los para o uso do ChatGPT.
2. Utilizar o ChatGPT para classificar os *tweets* e identificar discursos de ódio. Apesar de não ser um modelo de classificação convencional, será utilizado de modo a classificar os conteúdos associados a discursos de ódio dos demais *tweets* presentes no conjunto de dados.

1.2. Organização do Texto

Esse trabalho está organizado em 5 capítulos, no Capítulo 2 serão abordados os principais conceitos sobre classificação de texto e discurso de ódio. No Capítulo 3 é explicada a proposta deste TCC enquanto no Capítulo 4 são apresentados os resultados obtidos com essa abordagem. Por fim, no Capítulo 5 são feitas as considerações finais.

2) REVISÃO BIBLIOGRÁFICA

Neste capítulo serão apresentados os principais conceitos sobre o que caracteriza um discurso de ódio. Em seguida, serão apresentados os fundamentos relacionados à PLN e o ChatGPT. Por fim, será feita uma breve discussão sobre os trabalhos relacionados.

2.1) Fundamentos do Discurso de Ódio

O entendimento do conceito de discurso de ódio abarca quatro contextos essenciais: jurídico, lexical, científico e prático, cada um com suas nuances e âmbitos específicos. Na Seção 2.1.1, serão explorados detalhadamente esses conceitos, enquanto na Seção 2.1.2 será discutido a distinção entre discurso de ódio e liberdade de expressão.

2.1.1) Contextualização

Segundo a ONU, o discurso de ódio é compreendido como:
"Qualquer tipo de comunicação, seja oral, escrita, ou também comportamento, que ataca e utiliza uma linguagem pejorativa ou discriminatória em referência a uma pessoa ou grupo com base em quem ela é, ou seja, com base em sua religião, etnia, nacionalidade, raça, cor, ascendência, gênero ou outras formas de identidade (ONU, 2023)".

O significado de odiar alguém transcende a mera aversão ou desagrado; é um sentimento profundamente negativo que alberga o desejo de mal ao sujeito ou objeto odiado. O ódio, intrinsecamente vinculado à inimizade e à repulsa, muitas vezes motiva ações voltadas para evitar ou aniquilar aquilo que é objeto desse sentimento. Quando esse sentimento se volta a outro ser humano, suas manifestações podem ser evidentes por meio de insultos, agressões físicas e até mesmo pela busca ativa de prejudicar emocionalmente ou socialmente a pessoa alvo. A pesquisa conduzida pela professora da Faculdade de Direito de Harvard, Martha Minow (1986), revela que grupos discriminados e vítimas de discurso de ódio frequentemente enfrentam o risco de perder sua identidade e adquirir os valores dos agressores.

Para compreender claramente o conceito de discurso de ódio, este é definido como o uso de palavras que têm o propósito de insultar, intimidar ou assediar indivíduos com base em sua raça, cor, etnia, nacionalidade, gênero ou religião. Além disso, pode ser associado à sua

capacidade de incitar violência, promover o ódio ou fomentar a discriminação contra esses grupos específicos (WINFRIED BRUGGER, 2007).

Segundo a Cartilha de Orientação para Vítimas de Discurso de Ódio, da Defensoria Pública do Rio de Janeiro os discursos de ódio podem ser categorizados como: (I) Insulto e/ou ofensa a uma pessoa, incluindo um grupo socialmente vulnerável ao qual ela pertence; (II) Fala, gesto, expressão que instiga a violência, seja ela explícita ou implícita na fala do agressor.

2.1.2) Diferenças entre Discurso de Ódio e Liberdade de Expressão

Para compreender plenamente o conceito de discurso de ódio, suas implicações sociais, jurídicas e comportamentais, é crucial discernir suas diferenças em relação a outros conceitos frequentemente confundidos. A liberdade de expressão, muitas vezes mal interpretada, é um desses conceitos, erroneamente associada ao discurso de ódio, que, por sua vez, caracteriza-se pelo tratamento que viola os direitos fundamentais de outrem, podendo ser interpretado como um abuso da liberdade de expressão. Essa distinção precisa é essencial para uma análise precisa e contextualizada do tema.

“A liberdade de expressão é o direito de manifestar opiniões e ideias praticamente sem obstáculos (BBC, 2023)”. No século 20, a liberdade de expressão foi reconhecida como direito universal pelas Nações Unidas, no que se refere ao artigo 19, “Todo ser humano tem à liberdade de opinião e expressão; este direito inclui a liberdade de, sem interferência, ter opiniões e de procurar, receber e transmitir informações e ideias por quaisquer meios e independentemente de fronteiras (ONU, 2023)”.

O Secretário-Geral das Nações Unidas, António Guterres, explicita que: “Combater o discurso de ódio não significa limitar ou proibir a liberdade de expressão. Trata-se de prevenir o incitamento ao ódio para algo mais perigoso, em particular instigando a discriminação, a hostilidade e a violência, o que é proibido pelo direito internacional (ONU, 2019)”.
A urgência deste estudo em ser considerado como uma questão emergente na sociedade reside na necessidade premente de abordar e enfrentar o crescimento alarmante do discurso de ódio. A proposta de estabelecer um gabinete dedicado à análise e combate do discurso de ódio surge de uma série de fatores, um dos quais é o contexto histórico que evidencia a prevalência global desses discursos de ódio e ameaça à democracia social. Esse fenômeno não é novo, remontando

a períodos anteriores à era *big data* e da Inteligência Artificial. No entanto, a atual consideração reside na compreensão de que este fenômeno representa um risco significativo para a sociedade.

Em uma entrevista concedida à ONU *News*, Adama Dieng, que é o ex-conselheiro Especial da ONU para a Prevenção do Genocídio, ex-membro do conselho do Instituto Internacional para a Democracia e Assistência Eleitoral e ex-registrador do Tribunal Criminal Internacional para Ruanda, destacou a correlação direta entre discursos de ódio e crimes de ódio. Ele salientou que tais discursos são frequentemente precursores de genocídios, muitas vezes incitando a violência nas redes sociais e desencadeando eventos catastróficos para a sociedade (ONU News, 2019).
O exemplo emblemático do genocídio em Ruanda [4], onde cerca de 800 mil pessoas da etnia tutsis foram mortas em apenas 100 dias por extremistas étnicos hutus (BBC, 2014), ressalta a gravidade e a brutalidade dos resultados que o discurso de ódio pode provocar. Da mesma forma, o Holocausto durante a Segunda Guerra Mundial é outro exemplo contundente de como o discurso de ódio desencadeou uma violência devastadora contra os judeus, resultando em uma das páginas mais sombrias da história.

2.2. Processamento de Linguagem Natural (PLN) e ChatGPT

O Processamento de Linguagem Natural (PLN) representa um ramo da Inteligência Artificial (IA) voltado para a compreensão e geração de linguagem humana por meio de computadores. Sua abrangência se estende tanto à linguagem escrita quanto à linguagem falada. Mas o que torna mais interessante é que o estudo da PLN também envolve outras áreas para além da computação, que conforme descrito por Gonzalez e Lima (2003) é a compreensão de que o estudo da PLN é a fim de fazer com que máquinas e humanos consigam ter uma comunicação cada vez mais fluída e com maior naturalidade.
Este campo multidisciplinar incorpora conhecimentos de linguística, ciência da computação e estatística, concentrando-se na interpretação, análise e produção de linguagem, frequentemente direcionando seu foco para aspectos específicos, como compreensão de texto, geração de resumos automáticos, tradução entre idiomas e até mesmo a detecção de emoções e intenções por trás da linguagem utilizada. Essa busca por uma interação mais natural e fluida entre humanos e máquinas está em constante evolução, impulsionada pelo avanço das técnicas de aprendizado de máquina e da capacidade computacional, ampliando cada vez mais as

possibilidades e aplicações do PLN em diversos campos, desde assistentes virtuais até análises de dados em larga escala (Francielle Vargas, Thiago A.S. Pardo, et e al, 2022).

Além disso, a geração de texto é uma função poderosa da PLN, capacitando modelos de linguagem a criar conteúdo autonomamente, seja redigindo histórias, respondendo perguntas ou até mesmo redigindo notícias. Essa tecnologia desempenha um papel multifacetado em diversas áreas, desde assistência virtual e análise de dados até suporte ao cliente, pesquisa acadêmica e muito mais. Dentre os modelos existentes para geração de texto, podemos destacar o ChatGPT.

O ChatGPT surgiu em novembro de 2022 como uma ferramenta de Inteligência Artificial notável, desempenhando um papel crucial na ampliação do entendimento da sociedade sobre a IA. Agora na versão 4, esse modelo aprimorado apresenta melhorias significativas em relação à sua iteração anterior, incorporando a capacidade de processar tanto imagens quanto textos em suas interações. Neste estudo, será abordado a versão anterior, o ChatGPT 3.5, explorando sua arquitetura e compreendendo por que foi escolhido para a importante tarefa de detecção e classificação de discursos de ódio.

O GPT-3.5 (*Generative Pre-trained Transformer*) é uma arquitetura de linguagem neural desenvolvida pela OpenAI. Baseado em uma extensa rede neural de *Transformers*, o GPT-3.5 passou por um treinamento prévio com vastos conjuntos de dados textuais. Esta versão representa um salto significativo em relação às suas antecessoras, oferecendo maior capacidade de processamento, habilidades linguísticas mais refinadas e uma compreensão contextual mais profunda (Santos, 2023).

A base do GPT-3.5 é uma rede neural de transformadores, uma arquitetura reconhecida por sua eficácia em lidar com tarefas de processamento de linguagem natural. O diferencial desta versão é sua escala massiva, com 175 bilhões de parâmetros treináveis. Isso se traduz em uma capacidade impressionante de capturar nuances linguísticas, gerar texto coeso e lidar com uma ampla variedade de tarefas linguísticas, desde tradução até a compreensão textual.

Seu treinamento massivo com grandes volumes de dados textuais permitiu ao GPT-3.5 adquirir um conhecimento abrangente da linguagem, capacitando-o a gerar textos mais precisos e contextualmente relevantes, além de executar uma gama diversificada de tarefas. Como

resultado, tornou-se um dos modelos mais versáteis e poderosos no campo da Inteligência Artificial voltada para a linguagem.

2.3) Detecção de Discurso de Ódio

Dentro dos estudos e pesquisas recentes sobre discurso de ódio no Twitter, destacam-se exemplos relevantes, tanto a nível mundial quanto no Brasil, que se tornaram uma base significativa para análise e estudo neste contexto. Além disso, foi identificado um modelo de dados abordando discurso de ódio multirrótulo e detecção de linguagem abusiva no Twitter Indonésio (A. Marpaung, R. Rismala and H. Nurrahmi, 2021).

O BERT (*Bidirectional Encoder Representations from Transformers*) é um dos modelos utilizados para identificar discursos de ódio (Santos, 2023). Desenvolvido por pesquisadores do Google, o BERT é uma representação de linguagem bidirecional pré-treinada, elaborada a partir de grandes volumes de texto. Esse modelo exibe desempenho notável em várias áreas, incluindo a formação de frases gramaticalmente corretas, análise de sentimentos, avaliação da relação entre sentenças (se são concordantes, discordantes ou não relacionadas), identificação das entidades referenciadas por pronomes e verificação da equivalência semântica. Sua capacidade abrangente o torna valioso na detecção e compreensão de nuances na linguagem, sendo uma ferramenta significativa na identificação de discurso de ódio e outras formas de linguagem prejudicial (Devlin et al. 2019). Nesse mesmo modelo, temos como exemplo a implementação de um projeto que visa a detecção de possíveis discursos de ódio nas redes sociais, por meio da arquitetura BERT, o modelo é pré-treinado na linguagem PT-BR, rotulado em 100 classes hierárquicas com os potenciais alvos pelo discurso de ódio (Fortuna et al. 2019).

Considerando as informações passadas nesta seção, na seção 3 é apresentado a abordagem desse trabalho que consiste na aplicação metodológica, com as ferramentas utilizadas e detalhamento do banco de dados com as informações apresentadas.

3) METODOLOGIA - PROPOSTA E DESENVOLVIMENTO

Neste capítulo será abordado a metodologia para o desenvolvimento do trabalho proposto. No qual foi utilizado um conjunto de dados para o pré-processamento, com a integração da API do ChatGPT, a análise e classificação dos *tweets* com suas categorias de identificação de discurso de ódio. Foram coletados de um banco de dados disponível na plataforma GitHub, onde os dados não estavam limpos, contendo URLs e caracteres especiais.

3.1) Dataset

A coleta de dados desempenha um papel essencial na análise de discursos de ódio nas redes sociais. Para este estudo, foi utilizada uma base de dados do Twitter com as categorias de discurso de ódio já divididas, após a análise de diversas fontes de informação. Entre essas fontes, foi selecionado o conjunto de dados "ToLD-BR"[1], um corpus elaborado para o português brasileiro, especificamente para a detecção de discursos de ódio através do Processamento de Linguagem Natural (PLN) com 21101 *tweets*, em seu estado bruto. O "ToLD-BR" é acessível via GitHub, oferecendo uma versão web para simplificar o acesso aos dados.

É relevante ressaltar que um dos principais desafios enfrentados durante este estudo está na escassez de bases de dados em língua portuguesa para a análise de discursos odiosos. O levantamento mais recente de artigos publicados nos anos de 2019 e 2020 revelou a carência desses recursos, o que impacta diretamente a condução deste tipo de pesquisa.

O conjunto de dados utilizado neste estudo foi selecionado com o intuito de compreender e analisar discursos de ódio em plataformas de redes sociais. Este conjunto de dados apresenta diversas características fundamentais para a investigação desses discursos prejudiciais. Abaixo estão as informações relevantes sobre o dataset:

Origem do Dataset: https://arxiv.org/abs/2010.04543

Fonte: ToLD-BR: https://github.com/JAugusto97/ToLD-Br

Data de Coleta: julho a agosto de 2019

Estrutura e Composição: Dataset em tabela, formato CSV

[1] ToLD-BR: https://github.com/JAugusto97/ToLD-Br/tree/main?tab=readme-ov-file6

Tamanho Total do Dataset: 21.000

Informações Disponíveis: é dividido em conjuntos padrão de treinamento (80%), desenvolvimento (10%) e teste (10%) usando uma estratégia estratificada. Além disso, o corpus é divulgado com todas as anotações. Assim, futuros usuários do ToLD-Br poderão usá-lo com todas as etiquetas e com vários níveis de concordância entre os anotadores.

Anotações ou Classificações: Categorizados por multiclasse (LGBTQIAP+[2] fobia, Racismo, Insulto, Xenofobia, Misoginia e Obsceno)

Este conjunto de dados será fundamental para a análise e classificação de discursos de ódio nas redes sociais realizada neste trabalho. Foram selecionados 2.028 exemplos aleatórios, dentre o total disponível, esse subconjunto será o foco principal da nossa investigação, permitindo uma análise mais detalhada e representativa dos discursos presentes.

Este subconjunto possui 432 exemplos que são discurso de ódio, 1550 exemplos que não são e 46 classificados como talvez. Dos exemplos que são discurso de ódio, 0.65 como insulto, 0.24 como homofobia, 0.91 como obsceno, 0.24 como misoginia, 0.16 como racismo, e 0.10 como xenofobia.

3.2) Processo de Classificação com a Utilização do ChatGPT e Configuração do Prompt

Para a realização da classificação de discursos de ódio nos *tweets*, foi adotada a API do ChatGPT, uma poderosa ferramenta de Processamento de Linguagem Natural (PLN)

desenvolvida pela OpenAI. Esta API foi configurada para analisar e classificar os *tweets*, oferecendo insights sobre a presença de discursos nocivos e categorizando-os conforme diferentes tipos de discurso de ódio.

A configuração do prompt desempenhou um papel crucial no direcionamento da análise realizada pela API do ChatGPT. O prompt foi estruturado de forma a instruir o sistema sobre as tarefas específicas a serem executadas para cada *tweet* analisado.

[2] Sigla LGBTQIAP+: Lésbicas, Gays, Bissexuais, Transexuais, Transgêneros, Travestis, Queer, Intersexo, Assexual, Pansexualidade e + (demais orientações sexuais e identidades de gênero). Disponível em: https://www.trt4.jus.br/portais/trt4/modulos/noticias/465934

O prompt incluiu uma série de instruções direcionadas ao sistema, divididas em três partes distintas: o sistema, o usuário e o assistente. O sistema foi orientado a "agir como um cientista de dados analisando *tweets* tóxicos". O usuário definiu as tarefas a serem realizadas para cada *tweet*, instruindo o sistema a prever se o *tweet* continha discurso de ódio e, caso positivo, classificá-lo em uma ou mais categorias, como homofobia, obscenidade, insulto, racismo, misoginia ou xenofobia.

Por fim, para o assistente, foi fornecido um exemplo prático para ilustrar o formato esperado de entrada e saída para a análise dos *tweets*. No exemplo dado, o sistema recebeu como entrada um *tweet* específico contendo um discurso potencialmente odioso. A saída esperada foi um formato JSON com a indicação se o *tweet* continha discurso de ódio e a categorização desse discurso em uma ou mais das categorias predefinidas.

Essa estrutura de prompt foi essencial para orientar a análise e classificação dos *tweets*, permitindo que a API do ChatGPT identificasse e categorizasse os discursos de ódio presentes nos *tweets* analisados.

4) Análise de resultados e conclusões

Neste capítulo, serão explorados os resultados obtidos ao longo da pesquisa, focando na detecção e categorização de discursos de ódio nos *tweets* analisados. O processo iniciou com um *dataset* composto por 21.000 *tweets*, sendo reduzido para 2.028 após a seleção de uma amostra aleatória desses dados.

Foram conduzidos dois experimentos com o intuito de aperfeiçoar a detecção de discursos de ódio presentes nos *tweets*:

1. Primeiro experimento: tem como objetivo inicial categorizar os *tweets* em três possíveis classificações: "Sim" ou "Não", indicando a presença ou ausência de discursos de ódio. Essa análise foi realizada com o ChatGPT e a classificação permitiu uma visão inicial sobre a prevalência desses discursos no conjunto de dados analisados.

2. Segundo experimento: o foco foi a utilização do ChatGPT para reconhecer e classificar com maior detalhamento os diferentes tipos de discursos de ódio. Foram definidas seis categorias distintas para esse propósito: homofobia, xenofobia, racismo, misoginia, insulto e conteúdo obsceno. Cada categoria representou uma faceta específica do discurso de ódio, possibilitando

uma análise mais aprofundada sobre a natureza e tipos de expressões consideradas prejudiciais ou ofensivas.

4.1 Resultados Obtidos

O primeiro experimento teve como objetivo inicial a classificação dos *tweets* quanto à presença de discursos de ódio. Para esta tarefa de classificação binária - identificar se um *tweet* é ou não um discurso de ódio. Na Tabela 1 está ilustrado os resultados obtidos pela API do ChatGPT:

Tabela 1 – Resultados obtidos pela API do ChatGPT

	Precisão	Revocação	F1 score
Não	0.76	0.97	0.85
Sim	0.90	0.50	0.64
Média Macro	0.83	0.73	0.75

Ao avaliar o F1 score, que é uma métrica que leva em conta tanto a precisão quanto a revocação, observamos que o modelo alcançou um valor global de 0.75. Esse resultado indica que o modelo teve um desempenho razoável na classificação dos *tweets*, sendo melhor na identificação dos textos que não são discursos de ódio (classe negativa).

Os resultados desse primeiro experimento destacam que o modelo teve um desempenho superior na identificação de *tweets* que não continham discursos de ódio, apresentando uma alta precisão e revocação para essa categoria. No entanto, o desempenho na identificação de discursos de ódio foi relativamente inferior, mostrando uma precisão mais elevada, porém com uma revocação menor, indicando que o modelo teve dificuldades em identificar corretamente todos os casos de discursos de ódio.

O segundo experimento teve como objetivo realizar uma classificação multirrótulo para identificar os diferentes tipos de discursos de ódio presentes nos *tweets* analisados. Na Tabela 2 é apresentado as métricas de precisão, revocação e F1 score para cada categoria:

Tabela 2 – Resultados da classificação multirrótulo

Categoria	precisão	revocação	F1 score
homofobia	0.53	0.68	0.59
Insulto	0.65	0.52	0.57
Misoginia	0.24	0.18	0.21
Obsceno	0.91	0.33	0.48
racismo	0.16	0.43	0.24
xenofobia	0.10	0.26	0.14
Média Macro	0.43	0.40	0.37

Ao comparar com o primeiro experimento, este apresentou uma F1 score mais baixa, alcançando 37% de F1 score. Essa redução na métrica de avaliação indica a maior complexidade do problema enfrentado, uma vez que se trata de uma classificação multirrótulo e depende tanto da detecção do discurso de ódio quanto da identificação específica de cada categoria associada.

É relevante ressaltar que, ao observar individualmente as métricas de precisão, revocação e F1 score para cada categoria, foi identificado que as classes de homofobia, insulto e obsceno obtiveram os melhores resultados. Essas categorias apresentaram níveis razoáveis de precisão e revocação, indicando um desempenho relativamente melhor na identificação desses tipos específicos de discursos de ódio.

Os resultados desse segundo experimento refletem a maior dificuldade associada à classificação multirrótulo. Embora apresente desafios, a análise individual das categorias destacou áreas em que o modelo teve um desempenho mais promissor, indicando caminhos para refinamentos futuros visando uma melhor detecção e classificação de discursos de ódio.

5. Conclusão

Com base nos resultados apresentados na Seção 4, é evidente que a detecção de discursos de ódio nas redes sociais utilizando o modelo ChatGPT é uma tarefa desafiadora. No primeiro experimento, ao classificar se um *tweet* é um discurso de ódio ou não, observou-se uma precisão razoável na identificação de *tweets* que não continham discursos de ódio. No entanto, o desempenho na identificação de discursos de ódio foi relativamente inferior, sugerindo dificuldades em identificar corretamente todos os casos desse tipo de conteúdo.

Já no segundo experimento, ao realizar uma classificação multirrótulo para identificar diferentes tipos de discursos de ódio, percebeu-se um desafio maior, refletido na menor precisão geral e na dificuldade em identificar diversas categorias, especialmente aquelas relacionadas à misoginia, racismo e xenofobia. Isso ressalta a complexidade da tarefa, visto que a identificação precisa desses tipos específicos de discursos de ódio é crucial.

É importante ressaltar que, embora o modelo ChatGPT tenha sido eficaz na identificação de alguns tipos de discursos de ódio, não conseguiu abranger toda a gama e a complexidade do problema. Esses resultados indicam a necessidade de estratégias mais refinadas e um olhar mais abrangente para compreender e lidar com o discurso de ódio nas redes sociais.

As limitações encontradas, como a complexidade da classificação multirrótulo e a necessidade de prompts mais precisos e abrangentes, apontam para a importância de pesquisas futuras. Tais estudos devem focar em abordagens mais sofisticadas, combinando métodos quantitativos e qualitativos para uma detecção mais precisa e uma compreensão abrangente dos diferentes tipos de discursos de ódio presentes nas interações online.

Além disso, a falta de *datasets* (conjunto de dados) diversificados e em quantidade adequada na língua portuguesa representa um desafio significativo para futuras pesquisas nessa área. Trabalhos futuros devem concentrar-se na obtenção de conjuntos de dados mais representativos e diversificados para treinar modelos mais precisos na identificação e compreensão do discurso de ódio.

Portanto, é fundamental adotar uma abordagem multidisciplinar, englobando diversas áreas do conhecimento, e investir em pesquisas que visem à compreensão mais profunda e à mitigação do discurso de ódio nas plataformas online. A busca por soluções mais eficazes requer um esforço contínuo na adaptação e aprimoramento dos modelos para melhor refletir a dinâmica e a diversidade do discurso nas redes sociais.

Assim, a análise dos resultados obtidos ressalta a necessidade premente de estudos mais aprofundados e a utilização de abordagens mais amplas e diversificadas para compreender, detectar e combater o discurso de ódio nas redes sociais, visando a criação de ambientes online mais seguros e inclusivos para todos os usuários.

REFERÊNCIAS

AGÊNCIA BRASIL, 2023. **“Discurso de ódio na internet tem mulheres negras como principal”**. Disponível em: https://agenciabrasil.ebc.com.br/direitos-humanos/noticia/2018-08/discurso-de-odio-na-intern et-tem-mulheres-negras-como-principal. Acesso em: 14 Jul. 2023.

AYO, Femi Emmanuel et al**. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions**. Elsevier, [S. l.], ano 38, 13 out. 2020. Computer Science Review, p. 2-31. Doi: https://doi.org/10.1016/j.cosrev.2020.100311. Disponível em: https://www.sciencedirect.com/journal/computer-science-review. Acesso em: 7 set. 2023.

BBC, 2014. Entenda o genocídio de Ruanda de 1994: 800 mil mortes em cem dias. BBC News Brasil, [S. l.], p. 1-2, 7 abr. 2014. Disponível em: https://www.bbc.com/portuguese/noticias/2014/04/140407_ruanda_genocidio_ms. Acesso em: 8 set. 2023.

BRUGGER, W. Proibição ou Proteção do Discurso de Ódio? Algumas Observações Sobre o Direito Alemão e o Americano no Direito Público. V.4. N.15, 2009. Disponível em: https://repositorio.idp.edu.br/bitstream/123456789/541/1/Direito%20Publico%20n152007_Winfried%20Brugger.pdf. Acesso em: 15 Set. 2023.

CARA, Daniel Tojeira. Combate à violência nas escolas passa pelo controle dos discursos de ódio na internet. Jornal da USP, [S. l.], p. 1-2, 13 nov. 2023. Disponível em: https://jornal.usp.br/radio-usp/combate-a-violencia-nas-escolas-passa-pelo-controle-dos-discursos-de-odio-na-internet/. Acesso em: 23 nov. 2023.

DATALAB, 2023. 8 a cada 10 pessoas negras e LGBTQIAP+ já foram vítimas de discurso de ódio. DataLab, [S. l.], p. 1-3, 29 mar. 2023. Disponível em: https://datalabe.org/iea_negres_lgbtqiap/. Acesso em: 23 ago. 2023.

DEVLIN, J.; Chang, M.W.; Lee, K.; Toutanova, K. (2019). **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** Disponível em: https://doi.org/10.48550/arXiv.1810.04805. Acesso em: 12 Set. 2023.

FORTUNA, Paula; SILVA, João Rocha da; SOLER-COMPANY, Juan; WANNER, Leo; NUNES, Sérgio. A Hierarchically-Labeled Portuguese Hate Speech Dataset. ACL Anthology,

[S. l.], p. 94-104, 1 ago. 2019. Disponível em: https://aclanthology.org/W19-3510/. Acesso em: 29 nov. 2023.

G1, 2023. **“Denuncia de crimes envolvendo discurso de ódio nas redes sociais triplicam nos últimos 6 anos”**. Disponível em: https://g1.globo.com/jornal-nacional/noticia/2023/05/01/denuncias-de-crimes-envolvendo-discurso-de-odio-nas-redes-sociais-triplicaram-nos-ultimos-6-anos-aponta-levantamento.ghtml. Acesso em: 12 Out. 2023.

GONZALEZ, M.; LIMA, V. L. S. (2003). Recuperação de informação e processamento da linguagem natural. In: XXIII Congresso da Sociedade Brasileira de Computação. [S.l.: s.n.], v. 3, p. 347-395. Acesso em: 3 Mai. 2023

LEITE, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). **Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis**. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 914–924. Disponível em: https://aclanthology.org/2020.aacl-main.91/. Acesso em: 18 Ago. 2023.

MARPAUNG, Angela; RISMALA, Rita; NURRAHMI, Hani. Hate Speech Detection in Indonesian Twitter Texts using Bidirectional Gated Recurrent Unit. Hate Speech Detection in Indonesian Twitter Texts using Bidirectional Gated Recurrent Unit, [S. l.], p. 1-3, 21 jan. 2021. Disponível em: https://www.semanticscholar.org/paper/Hate-Speech-Detection-in-Indonesian-Twitter-Texts-Marpaung-Rismala/5711ede13fe0e163160b17cdac57e877b2c2c48c. Acesso em: 20 out. 2023.

MINOW, Martha. The Supreme Court 1986 Term: Justice Engendered. In Harvard Law Review, vol. 101, 1987. Acesso em: 04 Jul. 2023.

OLIVEIRA, Amanda S.; CECOTE, Thiago C.; SILVA, Pedro H. L.; GERTRUDES, Jadson C.; FREITAS, Vander L. S.; LUZ, Eduardo J. S. **How Good Is ChatGPT For Detecting Hate Speech In Portuguese?.** In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 14. , 2023, Belo Horizonte/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 94-103. Disponível em: https://www.researchgate.net/publication/374164741_How_Good_Is_ChatGPT_For_Detecting_Hate_Speech_In_Portuguese. Acesso em: 16 Set. 2023.

ONU; DIENG, Adama. **Temos de lembrar que crimes de ódio são precedidos por discurso de ódio.** ONU News, [S. l.], p. 1-2, 28 jun. 2019. Disponível em: https://news.un.org/pt/story/2019/06/1678221. Acesso em: 12 out. 2023.

PEREIRA, Jane Reis Gonçalves; OLIVEIRA, Renan Medeiros de; COUTINHO, Carolina Saud. REGULAÇÃO DO DISCURSO DE ÓDIO: ANÁLISE COMPARADA EM PAÍSES DO SUL GLOBAL. Revista de Direito Internacional, [S. l.], p. 1-3, 1 maio 2023. Disponível em: https://www.publicacoes.uniceub.br/rdi/article/view/6533. Acesso em: 8 nov. 2023.

RAVALLI, Maria José. Como proteger a integridade da informação nas plataformas digitais?: ONU publica orientações do secretário-geral.. In: RAVALLI, Maria José. Como proteger a integridade da informação nas plataformas digitais? : ONU publica orientações do secretário-geral.. [S. l.], 20 out. 2023. Disponível em: https://brasil.un.org/pt-br/249995-como-proteger-integridade-da-informa%C3%A7%C3%A3o-nas-plataformas-digitais-onu-publica-orienta%C3%A7%C3%B5es-do. Acesso em: 11 nov. 2023.

ROCHA, Juliana Lívia Antunes da; MENDES, André Pacheco Teixeira. **CARTILHA DE ORIENTAÇÃO PARA VÍTIMAS DE DISCURSO DE ÓDIO**. FGV: [s. n.], 2020. Disponível em: https://hdl.handle.net/10438/29490. Acesso em: 10 jul. 2023

SANTOS, Cecilia. Bert vs GPT: Principais diferenças entre esses dois modelos de IA. UPDF, [S. l.], p. 1-2, 27 set. 2023. Disponível em: https://updf.com/br/chatgpt/bert-vs-gpt/. Acesso em: 4 nov. 2023.

SAFERNET, 2022. “Denúncias de crimes de discurso de ódio e de imagens de abuso sexual infantil na internet têm crescimento em 2022”. Disponível em: https://new.safernet.org.br/content/denuncias-de-crimes-de-discurso-de-odio-e-de-imagens-de-abuso-sexual-infantil-na-internet. Acesso em: 03 Out. 2023

SCHÄFER, Johannes. Bias Mitigation for Capturing Potentially Illegal Hate Speech. Springer, [S. l.], p. 41-51, 29 mar. 2023. DOI https://doi.org/10.1007/s13222-023-00439-0. Disponível em: https://link.springer.com/article/10.1007/s13222-023-00439-0. Acesso em: 18 out. 2023.

UNESCO, 2023. Ética da Inteligência Artificial (IA) no Brasil. [S. l.], 15 mar. 2023. Disponível em: https://www.unesco.org/pt/fieldoffice/brasilia/expertise/artificial-intelligence-brazil. Acesso em: 11 ago. 2023.

UNITED, 2023. **“Detecção do discurso de ódio para combater a desinformação esteriotipada”**. Disponível em: https://www.unite.ai/pt/detec%C3%A7%C3%A3o-de-discurso-de-%C3%B3dio-ai-para-comb ater-a-desinforma%C3%A7%C3%A3o-estereotipada/. Acesso em: 06 Out. 2023.

VARGAS, Francielle; CARVALHO, Isabelle; GÓES, Fabiane; PARDO, Thiago A. S.; BENEVENUTO, Fabrício. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection, [S. l.], p. 20-25, 1 jun. 2022. Disponível em: https://www.semanticscholar.org/paper/HateBR%3A-A-Large-Expert-Annotated-Corpus-of-Comments-Vargas-Carvalho/c58e10adfc33a7660f8ef88b30ceb066224dfbef. Acesso em: 1 nov. 2023.