Proposta recebida em Setembro 2024 e aceite para publicação em Novembro 2024.

Explorando Técnicas de Aprendizado em Modelos de Linguagem para Classificação de Discurso de Ódio e Ofensivo em Português

Exploring Learning Techniques in Language Models for Classifying Hate and Offensive Speech in Portuguese

Gabriel Assis
Universidade Federal Fluminense

Annie Amorim
Universidade Federal Fluminense

Jonnathan Carvalho
Instituto Federal Fluminense

Mariza Ferro
Universidade Federal Fluminense

Daniel de Oliveira
Universidade Federal Fluminense

Daniela Vianna
JusBrasil

Aline Paes
Universidade Federal Fluminense

Resumo

As Redes Sociais, que desempenham um papel significativo no debate e na comunicação moderna, enfrentam o desafio contemporâneo do grande volume desordenado de conteúdo nocivo, como discurso de ódio e desinformação. Este artigo aborda a detecção de discurso de ódio em português, considerando suas particularidades linguísticas e nuances culturais. Utilizando-se modelos derivados de *Transformers*, juntamente com diversas estratégias de treinamento e ativação, são investigados nove modelos com variações em arquitetura, tamanho e *corpora* de pré-treinamento. Os resultados obtidos demonstram que, apesar de grandes modelos generativos acessados via *prompts* apresentarem resultados promissores, modelos de linguagem de menor escala ajustados permanecem superiores na realização dessa delicada tarefa.
⚠ **O texto contém exemplos potencialmente nocivos e ofensivos.**

Palavras chave

transformers; classificação; discurso de ódio

Abstract

Social Media platforms, significant in modern debate and communication, face the challenge of managing a vast and disorderly volume of hateful content and disinformation. This work examines the detection of hate speech in Portuguese, contemplating its unique linguistic and cultural nuance. Leveraging Transformer-based models and different training and activation strategies, nine models with variations in architecture, size, and pre-training *corpora* are evaluated. Our findings show that, even though large generative models with enhanced prompts exhibited promising results, tuned small language models remain superior in addressing this task.
⚠ **The text contains potentially harmful and offensive examples.**

Keywords

transformers; classification; hate speech

1. Introdução

As redes sociais são um canal preeminente para disseminação de informações com uma velocidade sem precedentes, aumentando significativamente o alcance e a capacidade de comunicação e expressão de opiniões (Pelle et al., 2018). Tais plataformas evoluíram para praças públicas de debate, onde indivíduos e grupos podem compartilhar seus pontos de vista sobre uma ampla gama de tópicos (Moura, 2016; Paiva et al., 2019). No entanto, essas plataformas também amplificam problemas sociais, como a propagação de informações falsas e a proliferação de insultos e discursos de ódio (Aluru et al., 2020). Nesse contexto, comentários ofensivos são definidos como *aqueles que contêm qualquer tipo de comunicação ofensiva, transpassando desde linguagem inapropriada até insultos diretos* (Pelle et al., 2018). Por outro lado, o discurso de ódio é caracterizado como *qualquer expressão pública de ódio ou incentivo à violência contra um indivíduo, ou um grupo, baseados em características como etnia, raça, nacionalidade, orientação sexual e gênero* (Vargas et al., 2021). Essas expressões, quando endossadas, potencialmente resultam em ameaças à integridade individual e coletiva, emergindo assim como uma preocupação essencial para comunidades digitais, plataformas de mídia social, entidades governamentais e a sociedade como um todo (Saraiva et al., 2021).

Adicionalmente, momentos de impacto expressivo no debate público podem tornar esse

DOI: 10.21814/lm.16.2.446

cenário ainda mais desafiador. A título de exemplo, nas eleições dos Estados Unidos em 2016, registrou-se um aumento nos crimes de ódio (Edwards & Rushin, 2018). Um efeito semelhante foi observado nas eleições federais brasileiras de 2018, quando houve um aumento massivo nos relatos de xenofobia, homofobia, racismo e intolerância religiosa nas mídias sociais (Vargas et al., 2021). Países como Canadá[1], Dinamarca[2] e Brasil[3] possuem arcabouços legais específicos para combater conteúdo de ódio em suas legislações. Especificamente na constituição brasileira, a discriminação baseada em raça, cor, etnia, religião ou origem nacional é legalmente reconhecida como crime. Contudo, certos usuários utilizam indevidamente as plataformas digitais para disseminar tal conteúdo, alicerçando-se erroneamente na prerrogativa de liberdade de expressão. Embora a liberdade de expressão também seja um direito constitucional, ela não deve promover ódio ou intolerância (de Freitas & de Castro, 2013). Não obstante, aplicar a lei permanece um desafio, principalmente devido ao volume de postagens e à complexidade de identificar e classificar corretamente comentários abusivos (Vargas et al., 2021). A análise precisa desse tipo de conteúdo é crucial, pois enfrentar o discurso de ódio transcende identificar comentários com linguagem abusiva, por exemplo. Apesar de plataformas digitais implementarem seus próprios sistemas de prevenção, eles apresentam várias limitações. Como ilustração, filtros sobre palavras-chave[4] podem endereçar o uso de expressões chulas, mas não nuances na expressão de ódio (Yin & Zubiaga, 2021). Além disso, muitos usuários empregam táticas inventivas ao escrever comentários ofensivos, como a troca de caracteres. Dessa forma, é decisivo construir métodos automatizados e precisos para filtrar e detectar conteúdos de discurso ofensivo e de ódio.

Este artigo concentra-se no contexto brasileiro, almejando contemplar as particularidades culturais do país e da língua portuguesa. A utilização contextual de expressões e palavras pode modificar profundamente o significado de um discurso, tornando fundamental uma abordagem que considere essas variações. Ilustrativamente, as sentenças ⚠ "*Ainda bem que gay não se reproduz..*", "*@USER canalha!*" e "*Olha que lindo!*" são classificadas nas bases de dados analisadas neste estudo como discurso de ódio, ofensivo e neutro, respectivamente. Enquanto a primeira sentença apresenta um indicativo claro de homofobia, a segunda representa uma ofensa direta, e a terceira reflete uma afirmação positiva, características que minimizam possíveis ambiguidades quanto à interpretação. Contudo, essa distinção nem sempre é evidente. Por exemplo, na sentença "*comecei a lavar a louça pra bichinha [...]*", originalmente rotulada como neutra, o uso do termo "bichinha"[5] pode gerar interpretações diversas, dependendo do leitor e do contexto. Em algumas regiões, a expressão é utilizada como vocativo informal entre interlocutores com quem se tem intimidade. No entanto, a comparação implícita com um animal pode ser considerada ofensiva por determinados receptores. Além disso, o termo pode carregar conotações homofóbicas, dependendo da intenção subjacente e do contexto em que é usado. Desse modo, fica evidente que a distinção das classes não é trivial, embora seja essencial. Os erros de classificação tornam-se particularmente críticos nesse contexto. Potenciais falsos positivos de discurso de ódio podem levar à censura indevida e, igualmente relevante, falsos negativos em tais classificações podem falhar em proteger grupos vulneráveis e comprometer a execução de leis.

Nesse ínterim, a arquitetura *Transformer* (Vaswani et al., 2017) emergiu, demonstrando resultados que compreendem o estado-da-arte em vários cenários, incluindo problemas de classificação e geração de texto (Fortuna & Nunes, 2018; Fu et al., 2024). Classificar postagens com origem em redes sociais é um campo de pesquisa ativo em Processamento de Linguagem Natural (PLN) (Fortuna & Nunes, 2018; Paiva et al., 2019; Jahan & Oussalah, 2023). Nessa perspectiva, enquanto o mundo se fascina pelas habilidades notáveis de grandes modelos de linguagem (*Large Language Models*, LLMs), como o ChatGPT[6], derivados dessa proeminente arquitetura, sobrepujar tarefas especialmente desafiadoras, como a identificação de discurso de ódio, permanece.

Ainda que utilizar LLMs para essa tarefa seja uma opção, ajustá-los é por vezes impraticável devido ao seu enorme número de parâmetros (sobretudo quando na ordem de centenas de bilhões) e a consequente implicação de custos diretos e indiretos. Nos casos de modelos de código-fechado, o acesso costuma ser realizado por meio de APIs[7] que resultam em cobranças associadas ao seu uso direto. Contudo, mesmo nos mode-

[1] https://bit.ly/canadian-hate-speech-law
[2] https://bit.ly/danish-hate-speech-law
[3] https://bit.ly/planalto-lei-7716
[4] https://bit.ly/words-filter-threads-instagram
[5] Uso do termo "bichinho(a)": https://pt.m.wiktionary.org/wiki/bichim
[6] https://chat.openai.com/
[7] Interface de programação de aplicações, do inglês *Application Programming Interface*

los de código-aberto, emergem os custos indiretos decorrentes da necessidade de *hardware* capaz de comportar tais modelos. Dentre essas, uma possibilidade, ainda que não isenta de custos, porém mais viável do que ajustar os pesos do modelo, é o emprego do aprendizado contextualizado (*in-context learning*) (Brown et al., 2020). Nesse método, demonstrações são inseridas diretamente nos *prompts*—instruções textuais dadas ao modelo—para fornecer contexto relevante (Chiu et al., 2022). Por exemplo, ao instruir um modelo com um comando como *"Qual o sentimento em 'o filme é péssimo', dado que a frase 'a música é maravilhosa' tem sentimento positivo?"*, o modelo utiliza a demonstração fornecida para capturar como deve analisar e classificar a frase alvo, considerando o novo contexto dado sem a necessidade de ajustar diretamente os seus pesos.

Neste artigo, são avaliados diversos métodos de seleção de demonstrações: *one-shot* (que usa um único exemplo, independentemente da classe), *one-class-shot* (com um exemplo de cada classe) e *few-shot* (que utiliza mais de um exemplo para cada classe). Para selecionar exemplos de demonstração, é proposto escolhê-los com base em seu tamanho e proximidade de similaridade com as instâncias de teste. Essas estratégias são comparadas com a seleção de exemplos aleatórios e a não seleção de algum exemplo de demonstração (*zero-shot*). Outrossim, a adição de contexto extra no *prompt* por meio de palavras-chave selecionadas com técnicas de modelagem de tópicos (Amorim et al., 2023; Pham et al., 2024), enquanto a instrução é mantida fixa, também é experimentada.

Todavia, desponta a questão de se, mesmo com *prompts* aprimorados, LLMs estão preparados para lidar com as nuances específicas para identificar o discurso de ódio considerando-se apenas o aprendizado de seu pré-treinamento. Investiga-se também como classificadores baseados em modelos menores e ajustados se comparam aos LLMs mais recentes. Selecionaram-se modelos baseados em *encoder* e em *decoder*. Enquanto o primeiro grupo possui reconhecida eficácia em classificação (Fortuna & Nunes, 2018), o segundo demonstra capacidades em alinhar o significado semântico de rótulos com o texto de entrada (Li et al., 2023). Mesmo que ajustar esses modelos seja mais factível, outros fatores devem ser considerados, como as características dos *corpora* de pré-treinamento, contemplando estilo e comprimento do texto.

Assim, a tarefa abordada neste artigo é formulada da seguinte maneira: **Dada uma postagem de rede social P escrita em português, realize o pré-processamento retornando X e classifique-o como pertencente a uma das classes em Y = {"discurso de ódio", "ofensivo" ou "neutro"}**. Em síntese, para essa tarefa de classificação, procedeu-se sobre três conjuntos de dados rotulados — HateBR (Vargas et al., 2022), OLID-BR (Trajano et al., 2023) e ToLD-Br (Leite et al., 2020) — à análise de nove modelos distintos, os quais se diferenciam quanto a sua arquitetura, tamanho e aos *corpora* em que foram pré-treinados. De forma específica, os modelos adotados podem ser categorizados em três grupos: **(i)** Modelos *encoder* fundamentados na arquitetura BERT (Devlin et al., 2019), o que neste trabalho incluem quatro variantes. Destacam-se, nesse grupo, três modelos especializados para a língua portuguesa, nomeadamente o BERTimbau (Souza et al., 2020), o AlBERTina (Rodrigues et al., 2023) e o BERTweet.BR (Caneiro et al., 2024) — esse último, um modelo pré-treinado especificamente em um *corpus* composto por *tweets*. Ademais, uma alternativa multilíngue, também pré-treinada em *tweets*, denominada Bernice (DeLucia et al., 2022); **(ii)** Modelos *decoder* ajustados para o português baseados na arquitetura LLaMA (Touvron et al., 2023a,b), compreendendo dois modelos de 7 bilhões de parâmetros — Sabiá (Pires et al., 2023) e Gervásio (Santos et al., 2024) — pré-treinados em textos formais; e, por fim, **(iii)** Grandes modelos de linguagem de propósito geral, dentre os quais os populares GPT (Brown et al., 2020) e Gemini (Google, 2023), e a MariTalk[8], um agente conversacional brasileiro desenvolvido especificamente para o português.

Substancialmente, são apresentadas quatro questões de pesquisa principais a respeito da classificação de postagens em redes sociais como neutras, ofensivas ou de discurso de ódio, em três conjuntos de dados:

QP1. Qual o desempenho de ajustar classificadores *encoder* clássicos a partir de modelos de linguagem "pequenos" em uma tarefa de classificação ternária e em um domínio delicado como a classificação de conteúdo de ódio?

QP2. A capacidade de alinhamento semântico de classificadores baseados em *decoder* os colocam como uma alternativa viável para a classificação de conteúdo de ódio?

QP3. Adicionar contexto extra e demonstrações selecionadas de maneira acurada aprimora

[8] https://chat.maritaca.ai/

a resposta de grandes modelos de linguagem generativos acessados via *prompt*?

E, finalmente,

QP4. Modelos da arquitetura *Transformer* se mostram alternativas viáveis para a tarefa de classificação ternária de conteúdo de ódio em redes sociais?

No que concerne aos principais achados e contribuições deste trabalho, elencam-se os seguintes itens:

- Realizar o ajuste fino (*fine-tuning*) em um modelo de menor escala pré-treinado em *corpora* adequados impera neste domínio.
- Abordagens alicerçadas na capacidade de representação semântica de grandes modelos adotados como classificadores não prevalecem sobre os modelos clássicos de menor escala neste domínio. Isso sugere uma possível lacuna no pré-treinamento desse grupo de modelos sobre conteúdo odioso e ofensivo.
- Adicionar contexto e exemplos bem selecionados beneficia modelos generativos ativados por *prompt*. Desse modo, este artigo contribui com novas estratégias para o aprimoramento de *prompts* que podem ser investigadas em outros domínios e tarefas.

Destacamos que este artigo é uma versão estendida do estudo intitulado *"Exploring Portuguese Hate Speech Detection in Low-Resource Settings: Lightly Tuning Encoder Models or In-Context Learning of Large Models?"*, publicado nos anais da *16th International Conference on Computational Processing of Portuguese* (Assis et al., 2024). Nesta versão, novos experimentos foram realizados utilizando o conjunto de dados OLID-BR e novos modelos, Bernice e Gemini, foram avaliados. Além disso, aplicamos uma nova abordagem com classificadores baseados em *decoders*, Gervásio e Sabiá-1, e aprofundamos a análise qualitativa dos resultados. Também abordamos uma limitação do trabalho anterior relativa à subamostragem no conjunto de testes, proporcionando uma investigação mais abrangente e precisa do tema.

Por fim, o trabalho se encontra organizado em cinco seções além da Introdução. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve os modelos selecionados e as estratégias de aprendizado aplicadas. A Seção 4 detalha os conjuntos de dados utilizados. A Seção 5 discute os resultados experimentais. Por último, a Seção 6 conclui o presente artigo e apresenta trabalhos futuros.

O código desta investigação está disponível publicamente no GitHub[9].

2. Trabalhos Relacionados

Não obstante a identificação de discurso de ódio em redes sociais representar um tópico imperativo contemporaneamente, o número de estudos considerando as particularidades da língua portuguesa permanece limitado, sobretudo, se comparado ao Inglês (Trajano et al., 2023; Jahan & Oussalah, 2023). Todavia, alguns trabalhos aplicaram e investigaram classificadores tradicionais de aprendizado de máquina (Souza et al., 2022; Plath et al., 2022; Pelle et al., 2018; Silva et al., 2018; da Silva & Rosa, 2023; Paiva et al., 2019; Vargas et al., 2021, 2022), classificadores baseados em *Transformers* (da Silva & Rosa, 2023; Leite et al., 2020; Oliveira et al., 2023; Plath et al., 2022; Santos et al., 2022; Vargas et al., 2021) e grandes modelos generativos para abordar essa questão (Chiu et al., 2022; Das et al., 2023; Nguyen et al., 2023; Oliveira et al., 2023, 2024; Assis et al., 2024).

Em conjunturas específicas, como associadas à detecção de racismo, misoginia e homofobia, classificadores baseados em algoritmos como Naïve Bayes (NB), Máquinas de Vetores de Suporte (*Support Vector Machines*, SVMs) e Florestas Aleatórias (*Random Forests*, RFs) demonstram desempenhos preditivos expressivos (Souza et al., 2022; Plath et al., 2022; Silva et al., 2018). Outrossim, alguns trabalhos utilizam representações baseadas em *embeddings* (Pelle et al., 2018; da Silva & Rosa, 2023). Contudo, essa abordagem pode apresentar limitações na representação de palavras sensíveis a contextos dinâmicos, como em redes sociais (Seno et al., 2024).

Modelos baseados em BERT (Devlin et al., 2019) despontam como o estado-da-arte proeminente na classificação de discurso de ódio, com alguns modelos específicos para línguas superando alternativas multilíngues em contextos não anglófonos (Jahan & Oussalah, 2023). Nesse sentido, da Silva & Rosa (2023) avaliaram 11 métodos distintos de classificação, incluindo o BERTimbau (Souza et al., 2020), que alcançou os melhores resultados para a língua portuguesa. De maneira similar, outros trabalhos destacaram o desempenho superior do BERTimbau (da Silva & Rosa, 2023; Santos et al., 2022; de Souza et al., 2024) e do BERT multilíngue (Leite et al., 2020) sobre abordagens que utilizam representações construídas sobre ex-

[9] https://github.com/MeLLL-UFF/hate_speech_in_context_pt

tração de *embeddings*, entretanto considerando a detecção de discurso de ódio apenas como um problema de classificação binário.

Para mais, os LLMs e suas notáveis habilidades emergentes também têm sido avaliados na tarefa de detecção de texto ofensivo e discurso de ódio. Chiu et al. (2022) investigaram as capacidades do ChatGPT para a detecção de linguagem sexista e racista, fazendo uso de técnicas de aprendizado *zero-shot*, *one-shot* e *few-shot*. Em contraste, Oliveira et al. (2023) utilizaram-se exclusivamente da abordagem *zero-shot* para avaliar o desempenho do GPT na detecção de discurso de ódio em *tweets* em português. Nesse trabalho, a comparação com o BERTimbau ajustado demonstra a viabilidade de modelos generativos para classificar conteúdo odioso. Em (Das et al., 2023), o ChatGPT demonstrou resultados promissores na detecção de discurso de ódio em português, porém com limitações sobre a capacidade de distinção entre discurso abusivo e não odioso dirigido a indivíduos e minorias. Ainda, Nguyen et al. (2023) avaliaram modelos LLaMA-2 ajustados para a detecção de textos sexuais, predatórios e abusivos. Por fim, os trabalhos de ambos, Assis et al. (2024) e Oliveira et al. (2024), contrastaram o GPT-3.5 e o modelo brasileiro MariTalk com alternativas baseadas em BERT ajustadas para o português brasileiro, concluindo que o último grupo alcança melhor desempenho, mesmo com resultados promissores das alternativas generativas.

Nenhum dos trabalhos mencionados conduziu um estudo que incluísse, de forma comparativa, a mesma vasta quantidade de modelos em língua portuguesa ajustados para tratar de um problema de classificação ternária sobre este domínio. As pesquisas anteriores concentram-se em problemas de classificação com configurações distintas da adotada neste trabalho. Por exemplo, Trajano et al. (2023) reporta métricas gerais de F1 na ordem de 0,70 para a tarefa de classificação multiclasse no conjunto OLID-BR, abrangendo 10 rótulos, incluindo racismo, sexismo e xenofobia. Observa-se, entretanto, que os resultados variam significativamente entre as classes. As classes de xenofobia e racismo, por exemplo, que possuem um suporte inferior a 40 amostras, apresentam métrica de F1 inferiores a 0,50. No caso da configuração binária, em que se considera uma classe neutra e uma contra-classe tóxica, Oliveira et al. (2024) reporta resultados de F1 de 0,75 para a classe de ódio no conjunto ToLD-BR, enquanto de Souza et al. (2024) alcança métricas na ordem de 0,90 no conjunto HateBR. Neste estudo, optamos pela abordagem de classificação ternária, o que restringe comparações diretas com os resultados mencionados. Nesse sentido, essa escolha busca proporcionar uma análise mais granular das classes, ampliando o escopo de investigação para além das configurações binárias frequentemente exploradas. Ao mesmo tempo, busca-se equilibrar a representação dos rótulos, evitando a formação de classes extremamente subamostradas e que podem ter seus resultados sub-representados. De mais a mais, da mesma maneira, modelos baseados em *decoder* como base para classificadores e um LLM mais recente em uma abordagem de aprendizado contextualizado não foram também avaliados. Além disso, este trabalho introduz novas estratégias para a seleção de demonstrações em *prompts* e propõe um método baseado em modelagem de tópicos para enriquecer instruções.

3. Método

Esta seção descreve os modelos selecionados, suas características, os métodos de treinamento adotados e as estratégias de inferência propostas.

3.1. Modelos

Para responder às questões de pesquisa QP1 e QP2, que abordam o desempenho de classificadores baseados em arquiteturas de *encoder* e *decoder*, respectivamente, bem como a QP4, que trata do desempenho geral de modelos *transformers* no domínio da classificação de discurso de ódio, foi realizada uma seleção criteriosa de modelos. Essa seleção incluiu explicitamente tanto arquiteturas baseadas em *encoders* quanto em *decoders*, além de grandes modelos de linguagem acessados por meio de *prompts*. Primeiro, destacam-se os modelos baseados em BERT (Devlin et al., 2019), pré-treinados com *corpora* específicos da variante brasileira do português: (i.) BERTimbau (Souza et al., 2020), em sua versão "*large*", e (ii.) AlBERTina PT-BR (Rodrigues et al., 2023), na sua versão de 100 milhões de parâmetros, ambos pré-treinados com textos mais formalmente redigidos. Além desses, (iii.) BERTweet.BR (Caneiro et al., 2024), pré-treinado com um *corpus* de *tweets* brasileiros, e (iv.) Bernice (DeLucia et al., 2022), que, embora multilíngue, também foi pré-treinado com dados do Twitter[10].

No que tange aos classificadores baseados em *decoder*, selecionam-se (v.) Sabiá-7B-1 (Pires et al., 2023), ajustado sobre a arquitetura LLaMA-1 (Touvron et al., 2023a), e (vi.)

[10] A plataforma possui atualmente o nome X: `https://x.com/`

Gervásio-7B-PTBR (Santos et al., 2024), baseado na arquitetura LLaMA-2 (Touvron et al., 2023b). Ambos os modelos foram pré-treinados com textos formais e orientados para tarefas de geração de conteúdo autorregressivo.

Quanto aos LLMs acessados via *prompt*, incluem-se (vii.) GPT-3.5-turbo (Ouyang et al., 2022), da reconhecida família GPT; (viii.) Gemini-pro 1.0 (Google, 2023), modelo notável por seus resultados expressivos em *benchmarks* recentes; e (ix.) MariTalk, um agente brasileiro aplicado na sua versão construída sobre o modelo Sabiá-2-medium (Almeida et al., 2024), ajustado sobre um *corpus* em português para também realizar tarefas gerais de um *chatbot*.

Considerando os diversos tamanhos, a natureza e a disponibilidade desses modelos, são aplicadas estratégias distintas para cada grupo. No entanto, enfatiza-se que o desempenho preditivo é avaliado consistentemente sobre o mesmo conjunto de testes. Detalhamentos adicionais serão apresentados nas seções subsequentes.

3.2. Treinamento dos classificadores

O processo de ajuste de pesos de classificadores permite que o modelo preserve as representações linguísticas obtidas no pre-treinamento, todavia se adequando a novos padrões de um novo domínio ou tarefas específicas, melhorando não só o seu desempenho nas previsões, mas também a sua capacidade de generalização (Yosinski et al., 2014). Desse modo, as técnicas empregadas para o ajuste dos modelos classificadores usados neste trabalho são especificadas a seguir.

3.2.1. Classificadores Baseados em Encoder

Os modelos de *encoder* foram ajustados utilizando duas abordagens principais: extração de características (*feature extraction*) e ajuste fino (*fine-tuning*). Apesar de o ajuste fino geralmente oferecer um melhor desempenho preditivo, a extração de características também foi explorada neste trabalho, considerando a sua adoção em diversos trabalhos anteriores no domínio da detecção de discurso de ódio (Fortuna et al., 2019; Plath et al., 2022).

Neste estudo, a abordagem de extração de características utiliza o *token* [CLS] como base para as representações vetoriais de entrada para treinar classificadores SVM. Esse *token* é um marcador especial inserido no início das sequências de entrada em modelos de linguagem. Ele desempenha a função de capturar a essência da sequência inteira, uma vez que agrega um resumo contextual de todas as entradas posteriores. Dessa forma, sua representação vetorial é frequentemente utilizada em tarefas de classificação que envolvem sentenças. Os vetores de características são então extraídos dos modelos de linguagem, originando a matriz $\mathbf{X} \in \mathrm{R}^{n \times d}$, onde $\mathbf{X}$ representa a matriz de exemplos, n o número de exemplos, e d as dimensões correspondentes ao *token* [CLS]. Assim, apenas os parâmetros do novo classificador são ajustados com base no conjunto de dados de treinamento, mantendo-se os pesos do modelo de linguagem pré-treinado inalterados. A outra estratégia adotada é a do ajuste fino, que consiste em acoplar uma camada classificatória diretamente ao modelo de linguagem. Por fim, os modelos são então inteira e diretamente ajustados conforme os exemplos do conjunto de treinamento.

3.2.2. Classificadores Baseados em Decoder

Modelos baseados em *decoders* também são adaptáveis a tarefas específicas, incluindo contextos que envolvem conteúdo sensível (Nguyen et al., 2023). No entanto, a estratégia de ajuste comumente utilizada aproveita a tarefa original de pré-treinamento desses modelos, que, geralmente, é a geração de texto de forma autorregressiva. Esse método de ajuste, embora permita a adaptação dos modelos a novos domínios e instruções, muitas vezes não utiliza diretamente conjuntos de dados rotulados, mesmo quando esses estão disponíveis. Para abordar esse quesito, uma estratégia semelhante à utilizada em classificadores baseados em *encoders* pode ser aplicada: incorporar uma camada classificatória diretamente ao modelo de linguagem, ao encontro do processo de ajuste fino. Essa abordagem se baseia no conceito de que as saídas dos modelos baseados em *decoders* podem alinhar o significado semântico da entrada com os rótulos, podendo, pois, funcionar como representações textuais para tarefas de classificação, inclusive alcançando resultados notáveis (Li et al., 2023).

3.3. Ativação dos Modelos Generativos via Prompts

As respostas obtidas dos agentes GPT-3.5-turbo[11] e MariTalk[12] são coletadas das suas respectivas APIs. Já o Gemini-pro é acessado por meio do serviço da Google Cloud Platform, Vertex AI[13]. Esses agentes recebem como entrada

[11] https://platform.openai.com/
[12] https://www.maritaca.ai/
[13] https://cloud.google.com/vertex-ai/

uma instrução que define a tarefa a ser realizada, um contexto que incorpora informações adicionais relevantes, e um ou mais exemplos de demonstração, que são pares de entrada e saída (X, Y_i) utilizados como referência. Neste estudo, são abordadas três classes, portanto, Y_i pode ser neutro, ofensivo ou discurso de ódio. São propostas múltiplas maneiras de selecionar exemplos de demonstração. Adicionalmente, também são experimentados diferentes contextos. A instrução permanece inalterada ao longo dos experimentos. Embora se entenda a sensibilidade desses agentes em relação as suas entradas (Liu et al., 2023), estudos anteriores que exploraram previamente instruções para a detecção de discurso de ódio em português são considerados como alicerce (Oliveira et al., 2023). Complementarmente, almeja-se, sobretudo, investigar o papel do contexto e das demonstrações na composição dos *prompts*, considerando a questão de pesquisa QP3, que aborda os efeitos da inclusão de informações contextuais adicionais e exemplos sistematicamente selecionados.

3.3.1. Construção do Prompt

Duas principais fontes direcionam a instrução aplicada neste trabalho. A primeira é o PromptHub[14], um repositório de código aberto de *prompts* categorizados por tarefa. Os *prompts* dessa coleção relacionados a tarefas semelhantes, como análise de sentimentos, auxiliaram a moldar a formulação da instrução. Em contrapartida, Pires et al. (2023) embasaram a integração de demonstrações dentro dos *prompts*. Assim, define-se a seguinte instrução[15]: Classifique o texto de rede social como "discurso de ódio" ou "ofensivo" ou "neutro".\n Texto: ≪ *alvo* ≫ \n Classe:.

Especificamente para o MariTalk, a instrução para ativação do modelo foi complementada com a sentença "Responda apenas com a classe". Isso se deve ao fato de que, em testes preliminares, observou-se que esse modelo tende a ser verboso em suas respostas, por vezes elaborando justificativas para suas classificações. Embora essa característica possa ser vantajosa em certos contextos, como na busca por explicabilidade, a crescente no número de *tokens* e o consequente aumento nos custos podem ser limitantes, considerando-se, sobretudo, o grande volume de experimentos realizados neste trabalho. Dessa forma, a adição dessa diretriz permitiu que o modelo realizasse classificações de maneira concisa e com respostas autocontidas no número de *tokens* estabelecido.

3.3.2. Estratégias para Seleção de Demonstrações

Sobre o número de exemplos de demonstração, foram aplicadas quatro maneiras de compor os *prompts*: ***(a.)* zero-shot**, caracterizado pela ausência de exemplos no comando; ***(b.)* one-shot**, que inclui um único exemplo, sem distinção de classe; ***(c.)* one-class-shot**, que engloba um exemplo para cada classe no comando; e ***(d.)* few-shot**, que apresenta vários exemplos para cada classe. Esses exemplos são selecionados a partir do conjunto de dados de treinamento.

Para a seleção de demonstrações do conjunto de treinamento, foram propostas três estratégias visando reduzir a variabilidade e aumentar a eficácia. A primeira estratégia consiste em ***(e.)* selecionar exemplos aleatoriamente**, considerando, naturalmente, a quantidade estabelecida de exemplos de demonstração. Por exemplo, essa abordagem, combinada com a modalidade (c.), implica a escolha aleatória de um exemplo de cada classe; já na modalidade (b.), envolve a seleção de um único exemplo de todo o conjunto de treinamento. Outras duas estratégias levam em conta ***(f.)* a similaridade semântica**, baseando-se nas representações de *embeddings*, ou ***(g.)* o tamanho em número de *tokens***, para a escolha dos exemplos. Intuitivamente, pretende-se fornecer informações adicionais, porém relevantes, para orientar melhor a capacidade de aprendizado contextualizado.

Ambas as últimas estratégias iniciam-se com a criação de agrupamentos

$$C = \{C_{1,1}, \ldots, C_{1,k_1}, C_{2,1}, \ldots, C_{2,k_2}, C_{3,1}, \ldots, C_{3,k_3}\},$$

formados separadamente para cada uma das três classes, visando assegurar uma clara distinção. Pressupõe-se que todas as instâncias de teste se enquadrarão no mesmo agrupamento C_t, uma vez que sua classe é desconhecida em tempo de execução. O passo seguinte é identificar, um para cada classe, os agrupamentos mais próximos $C_{1,p}, C_{2,q}$ e $C_{3,r} \in C$ e os agrupamentos $C_{1,s}, C_{2,t}$ e $C_{3,u} \in C$ mais distantes das representações de *embeddings* médias das instâncias de teste C_t, visando compreender como esses extremos podem influenciar positiva ou negativamente o aprendizado. Destaca-se que, embora o conjunto de testes seja utilizado para formar C_t, somente sua informação textual é adotada, o que viabiliza a aplicabilidade da estratégia em tempo de execução.

[14] https://github.com/deepset-ai/prompthub/

[15] Um exemplo completo de um *prompt one-shot* encontra-se no Apêndice A.

Na análise do impacto das informações extremas, a **estratégia baseada na similaridade semântica (f.)** seleciona (f.1.) os exemplos

$$\text{Ex} = \{ex_{x_1} \in C_{1,p},\ ex_{x_2} \in C_{2,q},\ ex_{x_3} \in C_{3,r}\}$$

que estão mais próximos dos *embeddings* médios de C_t, conforme a similaridade de cosseno, ou, alternativamente, (f.2.) os exemplos

$$\text{Ex} = \{ex_{z_1} \in C_{1,s},\ ex_{z_2} \in C_{2,t},\ ex_{z_3} \in C_{3,u}\}$$

que se distanciam mais. Novamente, essa escolha deve alinhar-se às configurações de a-d, adaptando-se ao número de exemplos requeridos em cada caso. Por exemplo, a estratégia (d.) realiza o processo múltiplas vezes até selecionar N exemplos, enquanto a (c.), apenas uma. Já para a estratégia (b.), seleciona-se apenas um exemplo de cada classe em Ex por vez. A **estratégia baseada no tamanho (g.)** é construída sobre (f.). Todavia, prioriza os exemplos que, além de estarem semanticamente próximos ou distantes, apresentam um tamanho em número de *tokens* mais próximo à moda das instâncias de teste. Essa abordagem é fundamentada na percepção de que conteúdos semanticamente relacionados podem compartilhar uma quantidade de *tokens* similar, facilitando a transmissão de mensagens análogas.

3.3.3. Contexto Adicional

Foram abordadas duas estratégias distintas para investigar se há benefício em incluir mais contexto aos *prompts*: a primeira absteve-se do uso de contexto adicional, ao passo em que a segunda incorporou palavras-chave e *emojis*[16] para dotar os modelos de exemplos representativos do tipo de discurso em foco. A seleção dessas palavras-chave pode desempenhar um papel significativo em um contexto em que a anotação de dados manual é realizada sob orientação de classificação com base na existência de termos específicos (Vargas et al., 2022). O método proposto é articulado em quatro etapas. Inicialmente, procede-se à remoção de pronomes possessivos, nomes próprios, verbos, caracteres especiais, numerais e termos com menos de duas letras. Em um segundo momento, para cada classe definida, são gerados dez tópicos a partir do conjunto de dados de treinamento, utilizando-se o BERTopic (Grootendorst, 2022) em integração com o BERTweet.BR. Subsequentemente, computa-se o cálculo da frequência de palavras por classe, identificando-se as dez mais prevalentes. No último passo, selecionam-se as dez palavras mais pertinentes dos tópicos gerados de acordo com seus pesos, desde que não constem nos tópicos ou no conjunto de palavras frequentes de outras classes.

As palavras-chave são incorporadas ao *prompt*, situando-se entre a instrução descrita e as demonstrações selecionadas, conforme o seguinte formato[17]: Considerando que os assuntos da classe "Classe A" estão associados com as palavras e emojis ≪*10 termos mais relevantes para Classe A* ≫. \n Da classe "Classe B" estão associados com as palavras e emojis ≪*10 termos mais relevantes para Classe B* ≫. \n Da classe "Classe C" estão associados com as palavras e emojis ≪*10 termos mais relevantes para Classe C* ≫.

4. Conjuntos de Dados

Os modelos foram avaliados em três conjuntos de dados distintos: HateBR (Vargas et al., 2022), OLID-BR (Trajano et al., 2023) e ToLD-Br (Leite et al., 2020). O conjunto HateBR consiste em 7.000 comentários do Instagram[18], recolhidos de perfis de políticos brasileiros durante o segundo semestre de 2019. Esse conjunto foi dividido em três classes para esta pesquisa: discurso de ódio (abrangendo textos rotulados com ao menos uma das classes dentre misoginia, gordofobia, xenofobia), textos ofensivos (que não se encaixam em algum conteúdo de ódio relacionado à construção identitária) e textos neutros.

Por outro lado, o conjunto de dados OLID-BR contém 7.000 registros que remontam a 2019, provenientes de comentários no Youtube[19] e postagens no Twitter. De maneira semelhante, textos identificados com ao menos um dos rótulos associados a manifestações de LGBTQ+fobia, racismo, sexismo, intolerância religiosa, ou outras formas de insultos dirigidos a grupos identitários, foram agrupados na classe de discurso de ódio. Textos que, não estando diretamente ligados ao discurso de ódio, porém marcados em rótulos como insultos ou profanidades, constituem a classe de textos ofensivos neste conjunto. Textos que não se ajustam a qualquer uma das classes de ofensividade ou discurso de ódio são rotulados como neutros.

O conjunto ToLD-Br, por sua vez, inclui

[16] https://www.significados.com.br/emoji/

[17] Um exemplo completo de um *prompt* com contexto encontra-se no Apêndice B.

[18] https://www.instagram.com/

[19] https://www.youtube.com/

21.000 *tweets* coletados entre julho e agosto de 2019, rotulados como não tóxico, LGBTQ+fobia, obsceno, insulto, racismo, misoginia e xenofobia. É notável que apenas cerca de 300 *tweets* foram estritamente rotulados em um dos rótulos de discurso de ódio. Neste estudo, as mensagens categorizadas como obscenas ou insultuosas foram agrupadas na classe ofensiva, enquanto aquelas consideradas não tóxicas foram classificadas como neutras, e as demais foram incluídas na classe de discurso de ódio.

Outrossim, a classificação do HateBR foi conduzida por anotadores voluntários com formação de doutorado e especialização em linguística, discurso de ódio ou ciência da computação. Já o processo para o OLID-BR começou com um filtro inicial feito por modelos automáticos de detecção de toxicidade[20] e prosseguiu com anotação humana. Para serem considerados minimamente qualificados para a tarefa, os anotadores precisavam ser nativos ou fluentes em português, além de atenderem a critérios de formação específicos. O ToLD-Br, por sua vez, contou com pelo menos três anotadores por instância, sem impor restrições quanto à formação anterior desses. Em todos os conjuntos de dados, foi dada ênfase à diversidade de gênero, à orientação política e à diversidade racial entre os selecionados. Vale ressaltar que o OLID-BR foi o único conjunto de dados que remunerou seus anotadores.

Dados	N	O	D	#tokens
HateBR	3410	2670	695	15.83
OLID-BR	1016	3896	2029	28.39
ToLD-Br	11634	8241	290	18.78

Tabela 1: Número de instâncias por classe, denotados como **N**eutro, **O**fensivo e **D**iscurso de Ódio, e contagem média de *tokens* por instância nos conjunto de dados.

A Tabela 1 apresenta a proporção das classes, acompanhada do número médio de *tokens* por instância, calculado com base no modelo BERTweet.BR, dado o contexto de redes sociais. A contagem média reflete possíveis diferenças no comportamento dos usuários nas plataformas que originaram os conjuntos de dados (French & Bazarova, 2017). Por exemplo, o conjunto HateBR, apresenta a menor contagem média de *tokens*, o que pode estar relacionado ao padrão de interação no Instagram, em que as postagens textuais geralmente consistem em comentários breves e reativos às fotografias publicadas. Em contraste, o conjunto ToLD-BR possui uma contagem ligeiramente maior, potencialmente devido a sua origem no Twitter, uma rede predominantemente textual, porém com limitações no número de caracteres por postagem. Já o conjunto OLID-BR, por combinar dados de múltiplas fontes, dentre elas o YouTube, talvez represente um contexto no qual os usuários dispõem de mais espaço para elaborar comentários, opiniões e análises relacionadas a vídeos.

Enfim, todos os conjuntos de dados foram particionados em aproximadamente 60% para treinamento e 20% para validação e testes cada, mantendo a proporção das classes. Para alcançar o balanceamento, a subamostragem aleatória das classes majoritárias foi aplicada no conjunto de treinamento.

5. Resultados Experimentais

Esta seção apresenta as configurações usadas para os experimentos e discute o desempenho dos modelos e estratégias adotados, com foco na classe de discurso de ódio, conjecturada por nós como a mais crítica no domínio analisado. Em seguida, realiza-se uma análise comparativa dos desempenhos proeminentes nas demais classes, prosseguindo-se então para uma inspeção qualitativa sobre os resultados.

5.1. Configurações Experimentais

Realizou-se um procedimento de pré-processamento simples sobre os dados, envolvendo a eliminação de duplicidades, a substituição de menções a usuários pelo *token* @USER, de links por HTTPURL e de *emojis* por sua representação textual, por meio da biblioteca Emoji[21]. A determinação dos agrupamentos C, como parte das estratégias (f.) e (g.), referidas na Seção 3.3.2, fundamenta-se no método tradicional KMeans (Jin & Han, 2011). Estabelece-se o número de grupos consoante o Critério do Cotovelo (Thorndike, 1953), sendo $k = 4$ para todas as classes.

O ajuste de classificadores baseados em *encoder*, abordado na Seção 3.2.1, recorre a hiperparâmetros tipicamente aplicáveis a modelos desse tipo. Em um cenário de recursos limitados, os modelos foram calibrados adotando uma taxa de aprendizado de $2e-5$, um tamanho de lote de 16 e com uso do critério de parada antecipada igual a 3, almejando evitar o sobreajuste.

[20] https://www.perspectiveapi.com/

[21] https://pypi.org/project/emoji/

Dados	Modelo	acurácia	precisão	revocação	F1
HateBR	**BERTimbau (extração de características)**	0.756	0.361	0.633	0.460
	BERTimbau (ajuste fino)	**0.862**	0.632	**0.705**	**0.667**
	AlBERTina PT-BR (extração de características)	0.430	0.000	0.000	0.000
	AlBERTina PT-BR (ajuste fino)	0.800	**0.785**	0.446	0.569
	BERTweet.BR (extração de características)	0.292	0.115	0.612	0.194
	BERTweet.BR (ajuste fino)	0.846	0.529	0.647	0.583
	Bernice (extração de características)	0.725	0.312	0.698	0.431
	Bernice (ajuste fino)	0.805	0.750	0.583	0.656
OLID-BR	**BERTimbau (extração de características)**	0.581	0.600	0.533	0.565
	BERTimbau (ajuste fino)	0.663	0.729	0.504	0.596
	AlBERTina PT-BR (extração de características)	0.253	0.387	0.257	0.309
	AlBERTina PT-BR (ajuste fino)	0.613	**0.749**	0.427	0.544
	BERTweet.BR (extração de características)	0.525	0.475	**0.649**	0.548
	BERTweet.BR (ajuste fino)	**0.672**	0.625	0.588	**0.606**
	Bernice (extração de características)	0.558	0.646	0.501	0.565
	Bernice (ajuste fino)	0.666	0.737	0.477	0.579
ToLD-Br	**BERTimbau (extração de características)**	0.615	0.046	0.534	0.084
	BERTimbau (ajuste fino)	0.599	0.035	**0.569**	0.065
	AlBERTina PT-BR (extração de características)	0.546	0.000	0.000	0.000
	AlBERTina PT-BR (ajuste fino)	0.538	0.033	0.241	0.059
	BERTweet.BR (extração de características)	0.535	0.000	0.000	0.000
	BERTweet.BR (ajuste fino)	**0.708**	**0.105**	**0.569**	**0.178**
	Bernice (extração de características)	0.616	0.046	0.552	0.085
	Bernice (ajuste fino)	0.704	0.082	0.500	0.141

Tabela 2: Resultados preditivos de classificadores baseados em *encoders* nas estratégias de extração de características e ajuste fino. Exceto pela acurácia, os valores são computados para a classe de discurso de ódio. Valores em **negrito** representam as melhores métricas para cada conjunto de dados.

Já o ajuste para os classificadores baseados em *decoder*, tratado na Seção 3.2.2, ocorreu sobre a estratégia LoRA (Hu et al., 2022), com $r = 16$, $lora\_alpha = 32$, e taxa de aprendizado $1e-4$. O tamanho de lote de 16 e o critério de parada antecipada também foram adotados.

Quanto ao grupo de experimentos realizados com modelos generativos acessados via *prompts*, fixou-se o limite máximo de *tokens* por resposta em 20, e a amostragem foi desabilitada. O ajuste do parâmetro de temperatura foi de 0.1 para todos os modelos, *i.e.*, GPT, Gemini-pro e MariTalk. Particularmente, é importante mencionar que o parâmetro $BLOCK\_NONE$ foi utilizado nas configurações de segurança da API do Gemini-pro para atenuar os bloqueios decorrentes das políticas éticas do Google, devido à sensibilidade dos dados. Além disso, optou-se por configurar o parâmetro *chat_mode* do MariTalk como $True$, de encontro à recomendação da documentação oficial para desativá-lo[22]. Essa decisão foi tomada para evitar respostas insubstanciais observadas nos testes realizados no contexto deste trabalho, incluindo respostas compostas unicamente de caracteres vazios (“ ”) ou contendo somente o caractere indicador de quebra de linha (\n). Já para a estratégia (d.) *few-shot*, dois exemplos por classe foram adotados.

Por fim, implementaram-se os classificadores baseados em *encoder* e *decoder* utilizando como arcabouço o *framework transformers* da Hugging Face (Wolf et al., 2020) e com o uso de recursos em nuvem da Google Cloud Platform (GCP)[23]. Foram utilizadas instâncias com quatro GPUs T4. O Scikit-learn (Pedregosa et al., 2011) foi adotado para o treinamento das SVMs, no esquema *one-vs-one* (OvO).

5.2. Resultados dos Classificadores

5.2.1. Resultados dos Classificadores Baseados em Encoder

A fim de responder à primeira questão de pesquisa— *Qual o desempenho de ajustar classificadores encoder clássicos a partir de modelos de linguagem “pequenos” em uma tarefa de classificação ternária e em um domínio delicado como a classificação de conteúdo de ódio?* —, a Tabela 2 detalha os resultados em termos de precisão, revocação e medida F1 para a classe de discurso de ódio, além de apresentar o valor macro de acurácia.

BERTimbau e Bernice dividem os melhores resultados para a estratégia de extração de características. Esperava-se que o modelo BERTweet.BR, dada sua fase de pré-treinamento em um *corpus* de redes sociais e especificamente no português do Brasil, apresentasse desempenhos superiores. No entanto, uma presunção considerada é a de que esse modelo possa apresen-

[22] https://maritaca-ai.github.io/maritalk-api/maritalk.html

[23] https://cloud.google.com

Dados	Modelo	acurácia	precisão	revocação	F1
HateBR	**Sabiá-1-7B**	0.526	0.206	0.094	0.129
	Gervásio-7B-PTBR	**0.672**	**0.270**	**0.475**	**0.345**
OLID-BR	**Sabiá-1-7B**	**0.532**	0.442	0.385	0.412
	Gervásio-7B-PTBR	0.495	**0.465**	**0.486**	**0.475**
ToLD-Br	**Sabiá-1-7B**	**0.531**	0.000	0.000	0.000
	Gervásio-7B-PTBR	0.446	**0.023**	**0.259**	**0.042**

Tabela 3: Resultados preditivos de classificadores baseados em *decoders* na estratégia de ajuste fino. Exceto pela acurácia, os valores são computados para a classe de discurso de ódio. Valores em **negrito** representam as melhores métricas para cada conjunto de dados.

tar uma representação de vocabulário excessivamente especializada. Tal especificidade, sem algum ajuste, pode não ter sido eficaz em um procedimento de classificação de múltiplas classes. BERTimbau, por sua vez, foi pré-treinado em um *corpus* em português, mas não de redes sociais, enquanto Bernice foi pré-treinado em um *corpus* de *tweets*, mas multilíngue. Essas características podem conferir a ambos os modelos uma representação menos restritiva, que, sem a superespecialização, facilitou possivelmente a atuação da SVM na melhor distinção das diferentes classes.

Por outro lado, ao se considerar a estratégia de ajuste fino, o BERTweet.BR apresenta superioridade geral ao alcançar os melhores resultados na metade das métricas. Contudo, esses resultados são concentrados nos conjuntos OLID-BR e ToLD-BR, que incluem *tweets* em seus dados. Em contrapartida, o modelo BERTimbau, na configuração de ajuste fino, apresenta um desempenho superior no conjunto de dados HateBR, composto por comentários provenientes do Instagram. Essa observação pode estar associada à existência de comportamentos distintos entre os usuários dessas plataformas. Isso porque usuários tendem a escrever de maneira mais espontânea e informal em microblogs como o Twitter, enquanto adotam uma postura mais formal nos comentários do Instagram (French & Bazarova, 2017). Essa variação pode ter favorecido o BERTimbau nesse contexto.

5.2.2. Resultados dos Classificadores Baseados em Decoder

Associada à segunda questão de pesquisa— *A capacidade de alinhamento semântico de classificadores baseados em decoder os colocam como uma alternativa viável para a classificação de conteúdo de ódio?* —, a Tabela 3 dispõe os resultados dos modelos *decoders* adaptados como classificadores por meio do ajuste fino.

Apesar de possuírem mais pesos, os modelos classificadores baseados em *decoders* de 7 bilhões de parâmetros demonstraram ser menos eficazes quando comparados aos modelos fundamentados em *encoders*, consoante contraste da Tabela 2 e Tabela 3. Isso sugere uma possível lacuna no processo de pré-treinamento desses modelos no que tange ao conteúdo de ódio, uma limitação conjecturável dado os plausíveis filtros implementados para prevenir que esses modelos, tipicamente empregados como generativos, propaguem tal conteúdo. Adicionalmente, a configuração usual da pilha *decoder* implica um processamento unidirecional dos *tokens*, orientado para a emissão de símbolos. Tal configuração pode dificultar o seu uso direto, sem modificações, como classificadores. Por fim, a relação entre o tamanho dos conjuntos de dados e o número de parâmetros desses modelos, bem como a viabilidade de adaptá-los utilizando exclusivamente a estratégia LoRA, pode representar um percalço durante a fase de ajuste. De modo geral, o modelo Gervásio sobressaiu-se, apresentando os melhores desempenhos, com exceção da acurácia nos conjuntos OLID-BR e ToLD-BR. Nesse sentido, é relevante destacar que o Gervásio é um modelo de construção mais recente, baseado em uma arquitetura mais atual, LLaMA-2, em contraste com a versão avaliada do Sabiá, que utiliza a arquitetura LLaMA-1.

5.3. Resultados da Inferência com LLMs

As Tabelas 4, 5, 6 e 7 apresentam os resultados das inferências feitas com LLMs generativos acessados via *prompt*, variando as diversas estratégias de seleção de demonstração introduzidas na Seção 3.3.2. Dessa maneira, pretende-

Modelo	#exemplos	seleção	acurácia	(±%)	precisão	(±%)	revocação	(±%)	F1	(±%)
GPT-3.5-turbo	**zero-shot**	**-**	0.649	ref.	0.237	ref.	0.770	ref.	0.362	ref.
	one-shot	**aleatória**	0.697	+7%	0.287	+21%	0.604	-22%	0.389	+7%
		similaridade	0.691	+6%	0.256	+8%	0.748	-3%	0.381	+5%
		tamanho	0.694	+7%	0.291	+23%	0.676	-12%	0.407	+12%
	one-class-shot	**aleatória**	0.535	-18%	0.210	+11%	**0.906**	**+18%**	0.341	-6%
		similaridade	0.641	-1%	0.237	0%	0.856	+11%	0.371	+2%
		tamanho	0.697	+7%	0.278	+17%	0.763	-1%	**0.408**	**+13%**
	few-shot	**aleatória**	0.672	+4%	0.223	-6%	0.640	-17%	0.330	-9%
		similaridade	**0.751**	**+16%**	**0.306**	**+29%**	0.424	-45%	0.355	-2%
		tamanho	0.683	+5%	0.269	+14%	0.748	-3%	0.396	+9%
*Gemini-pro**	**zero-shot**	**-**	0.656	ref.	0.247	ref.	0.446	ref.	0.318	ref.
	one-shot	**aleatória**	0.574	-13%	0.254	+3%	0.655	+47%	0.366	+15%
		similaridade	0.601	-8%	0.277	+12%	0.691	+55%	0.395	+24%
		tamanho	0.601	-8%	0.305	+23%	0.609	+37%	**0.407**	**+28%**
	one-class-shot	**aleatória**	0.616	-6%	0.220	-11%	**0.848**	**+90%**	0.349	+10%
		similaridade	0.677	+3%	0.240	-3%	0.691	+55%	0.357	+12%
		tamanho	0.640	-2%	0.245	-1%	0.820	+84%	0.377	+19%
	few-shot	**aleatória**	0.697	+6%	0.285	+15%	0.633	+42%	0.393	+24%
		similaridade	**0.760**	**+16%**	**0.322**	**+30%**	0.396	-11%	0.355	+12%
		tamanho	0.690	+5%	0.296	+20%	0.583	+31%	0.392	+23%
MariTalk	**zero-shot**	**-**	**0.765**	**ref.**	0.298	ref.	0.122	ref.	0.173	ref.
	one-shot	**aleatória**	0.625	-18%	**0.320**	**+7%**	0.223	+83%	0.263	+52%
		similaridade	0.603	-21%	0.253	-15%	0.317	+160%	0.281	+62%
		tamanho	0.635	-17%	0.317	+6%	0.317	+160%	0.317	+83%
	one-class-shot	**aleatória**	0.625	-18%	0.247	-17%	0.511	+319%	**0.333**	**+92%**
		similaridade	0.536	-30%	0.201	-33%	0.446	+266%	0.277	+60%
		tamanho	0.334	-56%	0.154	-48%	0.489	+301%	0.234	+35%
	few-shot	**aleatória**	0.587	-23%	0.225	-24%	0.129	+6%	0.164	-5%
		similaridade	0.286	-63%	0.112	-62%	**0.647**	**+430%**	0.191	+10%
		tamanho	0.431	-44%	0.134	-55%	0.604	+395%	0.219	+27%

Tabela 4: Resultados dos LLMs no conjunto HateBR. Exceto pela acurácia, eles são computados na classe de discurso de ódio. Os valores em **negrito** são os melhores para cada modelo, enquanto os melhores para o conjunto estão sublinhados. As porcentagens indicam a comparação com a respectiva referência de *zero-shot*. Os resultados do Gemini* podem apresentar ligeiras flutuações devido à taxa de respostas bloqueadas pelos filtros da API do Google. Nesse conjunto, esse valor foi de 0.15%.

Modelo	#exemplos	seleção	acurácia	(±%)	precisão	(±%)	revocação	(±%)	F1	(±%)
GPT-3.5-turbo	**zero-shot**	**-**	0.528	ref.	0.452	ref.	0.679	ref.	0.542	ref.
	one-shot	**aleatória**	0.592	+12%	0.532	+18%	0.551	-19%	0.541	0%
		similaridade	**0.597**	**+13%**	0.504	+12%	0.560	-18%	0.531	-2%
		tamanho	0.586	+11%	0.495	+10%	0.642	-5%	0.559	+3%
	one-class-shot	**aleatória**	0.592	+12%	0.532	+18%	0.551	-19%	0.541	0%
		similaridade	0.553	+5%	0.504	+12%	0.640	-6%	**0.564**	**+4%**
		tamanho	0.548	+4%	0.461	+2%	**0.691**	**+2%**	0.553	+2%
	few-shot	**aleatória**	0.545	+3%	**0.599**	**+33%**	0.405	-40%	0.483	-11%
		similaridade	0.559	+6%	0.587	+30%	0.425	-37%	0.493	-9%
		tamanho	**0.597**	**+13%**	0.590	+31%	0.536	-21%	0.561	+4%
*Gemini-pro**	**zero-shot**	**-**	0.588	ref.	0.509	ref.	0.486	ref.	0.497	ref.
	one-shot	**aleatória**	0.590	0%	0.481	-6%	0.603	+24%	0.535	+8%
		similaridade	**0.607**	**+3%**	0.502	-1%	0.702	+44%	**0.554**	**+11%**
		tamanho	0.458	-22%	0.424	-17%	**0.779**	**+60%**	0.549	+10%
	one-class-shot	**aleatória**	0.545	-7%	0.494	-3%	0.525	+8%	0.509	+2%
		similaridade	0.456	-22%	0.467	-8%	0.570	+17%	0.513	+3%
		tamanho	0.478	-19%	0.442	-13%	0.671	+38%	0.533	+7%
	few-shot	**aleatória**	0.551	-6%	**0.555**	**+9%**	0.415	-15%	0.475	-4%
		similaridade	0.515	-12%	0.468	-8%	0.514	+6%	0.490	-1%
		tamanho	0.550	-6%	0.470	-8%	0.616	+27%	0.533	+7%
MariTalk	**zero-shot**	**-**	0.421	ref.	0.429	ref.	0.096	ref.	0.157	ref.
	one-shot	**aleatória**	0.434	+3%	0.591	+38%	0.128	+33%	0.211	+34%
		similaridade	0.416	-1%	**0.660**	**+54%**	0.077	-20%	0.137	-13%
		tamanho	0.429	+2%	0.602	+40%	0.131	+36%	0.215	+37%
	one-class-shot	**aleatória**	0.457	+9%	0.393	-8%	0.422	+340%	0.407	+159%
		similaridade	0.397	-6%	0.421	-2%	0.368	+283%	0.393	+150%
		tamanho	**0.482**	**+14%**	0.447	+4%	0.548	+471%	**0.492**	**+213%**
	few-shot	**aleatória**	0.381	-10%	0.587	+37%	0.274	+185%	0.374	+138%
		similaridade	0.387	-8%	0.486	+13%	0.086	-10%	0.147	-6%
		tamanho	0.478	+14%	0.441	+3%	**0.484**	**+404%**	0.462	+194%

Tabela 5: Resultados dos LLMs no conjunto OLID-BR. Exceto pela acurácia, eles são computados na classe de discurso de ódio. Os valores em **negrito** são os melhores para cada modelo, enquanto os melhores para o conjunto estão sublinhados. As porcentagens indicam a comparação com a respectiva referência de *zero-shot*. Os resultados do Gemini* podem apresentar ligeiras flutuações devido à taxa de respostas bloqueadas pelos filtros da API do Google. Nesse conjunto, esse valor foi de 0.79%.

Modelo	#exemplos	seleção	acurácia	(±%)	precisão	(±%)	revocação	(±%)	F1	(±%)
GPT-3.5-turbo	**zero-shot**	**-**	0.561	ref.	0.030	ref.	0.379	ref.	0.056	ref.
	one-shot	**aleatória**	0.555	-1%	0.029	-3%	0.500	+32%	0.055	-2%
		similaridade	0.608	+8%	0.028	-7%	0.310	-18%	0.052	-7%
		tamanho	0.610	+9%	0.042	+40%	0.293	-23%	0.074	+32%
	one-class-shot	**aleatória**	0.639	+14%	0.029	-3%	0.241	-36%	0.051	-9%
		similaridade	**0.654**	**+17%**	0.042	+40%	0.293	-23%	0.073	+30%
		tamanho	0.577	+3%	0.034	+13%	0.500	+32%	0.063	+12%
	few-shot	**aleatória**	0.647	+15%	0.017	-43%	0.121	-68%	0.029	-48%
		similaridade	0.621	+11%	**0.044**	**+47%**	0.448	+18%	**0.081**	**+45%**
		tamanho	0.572	+2%	0.040	+33%	**0.638**	**+68%**	0.075	+34%
Gemini-pro*	**zero-shot**	**-**	0.560	ref.	0.034	ref.	0.190	ref.	0.057	ref.
	one-shot	**aleatória**	0.457	-18%	0.045	+32%	0.310	+63%	0.078	+37%
		similaridade	0.619	+11%	0.040	+18%	0.421	+122%	0.073	+28%
		tamanho	0.636	+14%	0.046	+35%	0.414	+118%	0.082	+44%
	one-class-shot	**aleatória**	0.584	+4%	0.052	+53%	0.397	+109%	0.092	+61%
		similaridade	0.588	+5%	0.039	+15%	**0.586**	**+208%**	0.074	+30%
		tamanho	**0.676**	**+21%**	0.057	+68%	0.276	+45%	0.095	+67%
	few-shot	**aleatória**	0.609	+9%	**0.059**	**+74%**	0.310	+63%	**0.100**	**+75%**
		similaridade	0.641	+14%	0.048	+41%	0.466	+145%	0.086	+51%
		tamanho	0.630	+12%	0.058	+71%	0.333	+75%	0.099	+74%
MariTalk	**zero-shot**	**-**	**0.501**	ref.	0.028	ref.	0.052	ref.	0.036	ref.
	one-shot	**aleatória**	0.493	-2%	0.050	+79%	0.034	-35%	0.041	+14%
		similaridade	0.496	-1%	**0.087**	**+211%**	0.069	+33%	**0.077**	**+114%**
		tamanho	0.492	-2%	0.055	+96%	0.052	+0%	0.053	+47%
	one-class-shot	**aleatória**	0.474	-5%	0.026	-7%	0.121	+133%	0.043	+19%
		similaridade	0.495	-1%	0.043	+54%	0.224	+331%	0.072	+100%
		tamanho	0.316	-37%	0.012	-57%	0.397	+663%	0.025	-31%
	few-shot	**aleatória**	0.447	-11%	0.014	-50%	0.121	+133%	0.025	-31%
		similaridade	0.223	-55%	0.037	+32%	0.069	+33%	0.048	+33%
		tamanho	0.029	-94%	0.014	-50%	**0.931**	**+1690%**	0.027	-25%

Tabela 6: Resultados dos LLMs no conjunto ToLD-Br. Exceto pela acurácia, eles são computados na classe de discurso de ódio. Os valores em **negrito** são os melhores para cada modelo, enquanto os melhores para o conjunto estão sublinhados. As porcentagens indicam a comparação com a respectiva referência de *zero-shot*. Os resultados do Gemini* podem apresentar ligeiras flutuações devido à taxa de respostas bloqueadas pelos filtros da API do Google. Nesse conjunto, esse valor foi de 0.12%.

se responder à terceira questão de pesquisa elaborada — *Adicionar contexto extra e demonstrações selecionadas de maneira acurada aprimora a resposta de grandes modelos de linguagem generativos acessados via prompt?*.

Especialmente, um aspecto do modelo Gemini-pro demanda atenção. Durante a realização de inferências por meio da API, verificou-se que algumas solicitações foram bloqueadas devido à política de filtros a conteúdos sensíveis do Google, mesmo quando as configurações de bloqueio estavam minimizadas. Inicialmente, poder-se-ia presumir que a instância em avaliação contivesse ao menos um elemento ofensivo, dada essa ocorrência; no entanto, essa suposição não é necessariamente acurada. O bloqueio pode ter sido ocasionado pelo conjunto de exemplos fornecidos no *prompt*. Contudo, visto que as taxas de bloqueio foram extremamente baixas, optou-se por manter os cálculos das métricas apenas desconsiderando as instâncias bloqueadas. Ainda que isso cause uma flutuação mínima nos resultados métricos, uma comparação com os outros modelos não deve ser relevantemente prejudicada.

Outrossim, mesmo que o MariTalk seja fundamentado em um modelo de linguagem pré-treinado em português, os modelos GPT-3.5-turbo e Gemini-pro demonstram desempenhos superiores na maioria dos casos. Entretanto, os detalhes específicos acerca dos conjuntos de dados utilizados no pré-treinamento desses dois últimos modelos não são publicamente disponíveis, o que representa uma limitação significativa para a compreensão plena das origens desse fato.

Além disso, em um cenário em que o desbalanceamento de classes constitui uma preocupação e a minimização tanto de falsos positivos quanto de falsos negativos é imperativa, a métrica F1, particularmente para a classe de discurso de ódio, requer um enfoque particular. Para essa métrica, nos conjuntos de dados HateBR e OLID-BR, tanto o GPT-3.5-turbo quanto o MariTalk alcançaram desempenhos superiores na configuração de *one-class-shot*, enquanto o Gemini-pro exibiu melhores resultados na configuração de *one-shot*. No conjunto ToLD-Br, os modelos GPT-3.5-turbo e Gemini-pro alcançaram os seus melhores resultados utilizando a abordagem *few-shot*, ao passo que o MariTalk sobressaiu-se na abordagem *one-shot*. Essa variabilidade pode ser atribuída às diferenças na capacidade de cada modelo de integrar e responder ao número de informações fornecidas via *prompt*.

Dados	Modelo	acurácia	(±%)	precisão	(±%)	revocação	(±%)	F1	(±%)
HateBR	**GPT-3.5 one-class-shot/tam.**	0.697	-	0.278	-	**0.763**	-	0.408	-
	GPT-3.5 one-class-shot/tam. + ctx.	**0.765**	**+10%**	**0.328**	**+18%**	0.633	-17%	**0.432**	**+6%**
	Gemini-pro one-shot/tam.	0.601	-	0.305	-	**0.609**	-	0.407	-
	Gemini-pro one-shot/tam. + ctx.	**0.783**	**+30%**	**0.393**	**+29%**	0.424	-30%	**0.408**	**0%**
	MariTalk one-class-shot/aleat.	**0.625**	-	**0.247**	-	0.511	-	**0.333**	-
	MariTalk one-class-shot/aleat. + ctx.	0.454	-27%	0.152	-38%	**0.748**	**+46%**	0.253	-24%
OLID-BR	**GPT-3.5 one-class-shot/sim.**	0.533	-	0.504	-	**0.640**	-	**0.564**	-
	GPT-3.5 one-class-shot/sim. + ctx.	**0.626**	**+17%**	**0.560**	**+11%**	0.563	-12%	0.562	0%
	Gemini-pro one-shot/sim.	**0.607**	-	0.502	-	**0.702**	-	**0.554**	-
	Gemini-pro one-shot/sim. + ctx.	0.605	0%	**0.600**	**+20%**	0.267	-62%	0.369	-33%
	MariTalk one-class-shot/tam.	**0.482**	-	**0.447**	-	0.548	-	**0.492**	-
	MariTalk one-class-shot/tam. + ctx.	0.312	-35%	0.317	-29%	**0.884**	**+61%**	0.466	-5%
ToLD-Br	**GPT-3.5 few-shot/sim.**	0.621	-	0.044	-	0.448	-	0.081	-
	GPT-3.5 few-shot/sim. + ctx.	**0.671**	**+8%**	**0.070**	**+59%**	**0.483**	**+8%**	**0.123**	**+52%**
	Gemini-pro few-shot/aleat.	0.609	-	**0.059**	-	**0.310**	-	**0.100**	-
	Gemini-pro few-shot/aleat. + ctx.	**0.680**	**+12%**	0.042	-29%	0.190	-39%	0.069	-31%
	MariTalk one-shot/sim.	**0.496**	-	**0.087**	-	0.069	-	**0.077**	-
	MariTalk one-shot/sim. + ctx.	0.065	-87%	0.015	-83%	**0.966**	**+1.3k%**	0.029	-62%

Tabela 7: Resultados da adição de contexto extra às melhores configurações dos LLMs ativados via *prompt* com base no F1. Exceto pela acurácia, eles são computados na classe de discurso de ódio. Os valores em **negrito** representam os melhores resultados quanto à estratégia de adição de contexto, enquanto os melhores para todo o conjunto de dados estão sublinhados. O valor de porcentagem indica a comparação com a respectiva referência, nesse caso, a configuração sem contexto adicional.

Por exemplo, no conjunto ToLD-Br, dois dos modelos exigiram um número maior de exemplos, constatado pelo melhor desempenho ao empregar a estratégia *few-shot*. Isso se dá em um contexto no qual, novamente, os textos provêm exclusivamente de uma rede cujos usuários se expressam primordialmente de maneira textual, o que pode implicar a presença de uma maior diversidade de recursos linguísticos. Em contraste, para os conjuntos de dados HateBR e OLID-BR, nenhum dos melhores desempenhos dos modelos foi nessa configuração.

No que concerne à seleção de demonstrações para os *prompts*, de modo geral, todos os modelos beneficiaram-se de estratégias baseadas na similaridade e tamanho dos exemplos, com apenas dois casos em que a seleção aleatória proporcionou os melhores resultados para a medida F1. Essa constatação reitera que as estratégias de seleção de elementos que compõem o *prompt* impactam na otimização do desempenho dos modelos de linguagem em tarefas específicas. Desse modo, também foram acrescentadas informações contextuais aos melhores resultados das estratégias de seleção de demonstrações. Isso é feito para verificar se é possível aprimorar ainda mais a capacidade de resposta dos modelos ao fornecer palavras-chave adicionais em conjunto dos exemplos demonstrativos. A Tabela 7 sintetiza esses resultados.

Em termos gerais, o modelo GPT-3.5-turbo, que foi pré-treinado predominantemente em inglês, apresentou uma melhoria significativa com a incorporação de contexto adicional em seus *prompts*. O modelo estabeleceu, inclusive, novos melhores resultados para a métrica F1 nos conjuntos de dados HateBR e ToLD-Br. Por sua vez, o Gemini-pro obteve um aumento na acurácia e na precisão no conjunto HateBR, melhorias em precisão para o OLID-BR, e um incremento na acurácia para o ToLD-BR. Já o MariTalk exibiu maiores resultados sobre a revocação em todos os conjuntos. Assim, a utilização de contextos adicionais pode constituir uma estratégia complementar para mitigar a incidência de falsos positivos e falsos negativos, dependendo do contexto específico almejado.

5.4. Resultados Comparativos por Classe

Esta seção apresenta uma análise comparativa centrada nos resultados da métrica F1 para os melhores representantes de cada grupo de modelos: um representante dos classificadores ba-

Classes →	**Neutro**		**Ofensivo**		**Discurso de Ódio**	
Dados ↓	**Modelo**	**F1**	**Modelo**	**F1**	**Modelo**	**F1**
HateBR	Bernice *(ajuste fino)*	**0.906**	BERTimbau *(ajuste fino)*	**0.861**	BERTimbau *(ajuste fino)*	**0.667**
	MariTalk *(one-shot/sim.)*	0.899	MariTalk *(one-shot/sim.)*	0.793	GPT-3.5-turbo *(one-class-shot/tam. + ctx.)*	0.432
	Gervásio-7B-PTBR *(ajuste fino)*	0.770	Gervásio-7B-PTBR *(ajuste fino)*	0.672	Gervásio-7B-PTBR *(ajuste fino)*	0.345
OLID-BR	BERTweet.BR *(ajuste fino)*	0.547	Bernice *(ajuste fino)*	**0.765**	BERTweet.BR *(ajuste fino)*	**0.606**
	MariTalk *(one-shot/sim.)*	**0.605**	GPT-3.5-turbo *(one-shot/sim.)*	0.719	GPT-3.5-turbo *(one-class-shot/sim.)*	0.564
	Gervásio-7B-PTBR *(ajuste fino)*	0.304	Sabiá-1-7B *(ajuste fino)*	0.657	Gervásio-7B-PTBR *(ajuste fino)*	0.475
ToLD-Br	Bernice *(ajuste fino)*	0.747	BERTweet.BR *(ajuste fino)*	**0.731**	BERTweet.BR *(ajuste fino)*	**0.178**
	MariTalk *(few-shot/sim.)*	**0.798**	MariTalk *(one-shot/tam.)*	0.701	GPT-3.5-turbo *(few-shot/sim. + ctx.)*	0.123
	Gervásio-7B-PTBR *(ajuste fino)*	0.504	Sabiá-1-7B *(ajuste fino)*	0.572	Gervásio-7B-PTBR *(ajuste fino)*	0.041

Tabela 8: Resultados dos melhores modelos para cada classe com base na métrica de F1. Foi selecionado um modelo para cada grupo avaliado, *i.e.*, um representante dos classificadores baseados em *encoder*, um dos classificadores baseados em *decoder* e um dos LLMs ativados via *prompts*. Os valores em **negrito** são os melhores para cada conjunto.

seados em *encoder*, um dos classificadores baseados em *decoder* e um dos LLMs ativados via *prompts*. São reportados valores para as classes neutro, ofensivo e discurso de ódio. As análises anteriores não incluíam as outras classes, principalmente, devido à relevância e aos desafios intrínsecos associados à categoria de discurso de ódio neste contexto, haja vista também o entrave do desbalanceamento de dados, o que dificulta a apresentação de uma média geral. Os resultados detalhados são apresentados na Tabela 8.

Entre os nove casos analisados, o modelo BERTweet.BR se destacou em três, enquanto Bernice e BERTimbau lideraram em dois casos cada. Notavelmente, na classe de discursos de ódio, o modelo pertencente à família GPT apresentou os melhores desempenhos para seu grupo. Por outro lado, nas categorias que não o discurso de ódio, o MariTalk prevaleceu como o modelo representante dos LLMs ativados via *prompt*, com exceção da classe ofensivo para o OLID-BR. Isso pode indicar que, devido ao seu *corpus* de pré-treinamento em português, o MariTalk possua uma melhor compreensão de conteúdos cotidianos não odiosos, enquanto o modelo GPT, com sua reconhecida capacidade de entendimento contextual, possa discernir de forma mais eficaz as nuances do discurso de ódio. É importante salientar que o Gemini-pro não se destacou em nenhum dos casos analisados. Quanto aos classificadores baseados em *decoder*, não foi observada alguma situação em que esse grupo superasse os demais. Tal ausência sugere que a abordagem adotada, que considera sua capacidade de alinhamento semântico, pode ser insuficiente para suas aplicações como classificadores convencionais, exigindo alterações arquiteturais conforme observado em trabalhos mais recentes (BehnamGhader et al., 2024). Porém, destaca-se que o Gervásio-7B superou o Sabiá-1-7B, que liderou apenas em dois dos nove casos dentre os classificadores baseados em *decoder*.

Os resultados corroboram a maior eficácia dos modelos baseados em *encoder* aliados à estratégia de ajuste fino. No entanto, a Tabela 9 evidencia que, mesmo entre os modelos de melhor desempenho, ainda há ocorrência de confusão entre as classes analisadas. No caso do conjunto HateBR, observa-se uma maior taxa de predições corretas, embora a principal dificuldade resida na classificação da classe "discurso de ódio". Nesse contexto, os erros de predição encontram-se equilibrados entre as classes "neutro" e "ofensivo". Em relação ao conjunto OLID-BR, a dificuldade do modelo torna-se mais evidente nas classes limítrofes, consideradas em termos de severidade, especificamente de "neutro" para "ofensivo", assim como entre "ofensivo" e "discurso de ódio". Esse cenário destaca a dificuldade do modelo em estabelecer limites claros entre essas classes. Já no conjunto ToLD-BR, verifica-se uma dificuldade significativa na separação da classe majoritária ("neutro") em relação às demais.

(a) BERTimbau – HateBR

Categoria Verdadeira	Categoria Predita		
	Neutro	Ofensivo	Ódio
Neutro	601	52	29
Ofensivo	37	469	28
Ódio	7	34	98

(b) BERTweet.BR – OLID-BR

Categoria Verdadeira	Categoria Predita		
	Neutro	Ofensivo	Ódio
Neutro	143	38	23
Ofensivo	107	550	120
Ódio	69	98	238

(c) BERTweet.BR – ToLD-BR

Categoria Verdadeira	Categoria Predita		
	Neutro	Ofensivo	Ódio
Neutro	1513	604	210
Ofensivo	265	1295	70
Ódio	10	15	33

Tabela 9: Matrizes de confusão dos melhores modelos *encoder*, selecionados com base na frequência de maiores valores de F1 entre as classes (Tabela 8), para cada conjunto de dados.

Para mais, é proeminente a discrepância nos valores observados para a categoria de discurso de ódio no conjunto de dados ToLD-Br em comparação com os conjuntos HateBR e OLID-BR. Essa diferença pode estar associada, para além do desequilíbrio de classes, à origem dos dados. Como supracitado, o ToLD-Br provém exclusivamente do Twitter, ao passo que o OLID-BR inclui também outra fonte, o YouTube, e o HateBR contempla comentários do Instagram. Todavia, essa variação também pode ser atribuída ao processo de construção dos conjuntos e rotulação dos dados, influenciada, entre outros aspectos, pela subjetividade dos anotadores. Para investigar esse fenômeno, além de elucidar ainda mais o comportamento dos modelos, a próxima seção contempla uma inspeção qualitativa de algumas instâncias classificadas.

5.5. Inspeção Qualitativa

A Tabela 10 apresenta instâncias selecionadas dos conjuntos de dados, incluindo seus respectivos rótulos e as predições realizadas pelos modelos, especificamente aqueles que obtiveram os melhores resultados para a métrica F1 para a classe discurso de ódio, consoante exibido na Tabela 8. O propósito dessa análise é proporcionar uma perspectiva qualitativa sobre como os modelos se comportam em comparação com a rotulação realizada por humanos. ⚠ **É importante salientar novamente a possibilidade de presença de conteúdo ofensivo e potencialmente nocivo nas instâncias examinadas.**

Um registro específico do conjunto HateBR destaca-se ao exibir o uso do termo "ordinária". Embora em seu sentido literal o termo denote algo "habitual, comum e corriqueiro", em um contexto figurado pode significar "mau caráter"[24]. Na rotulação original, esse caso foi identificado como discurso de ódio, possivelmente por acompanhar a palavra "mulher". No entanto, os modelos atribuíram a essa ocorrência a classificação de ofensivo. Essa decisão dos modelos parece plausível, visto que, mesmo com a inclusão de "mulher", não há evidência explícita de discriminação baseada em gênero, o que caracterizaria sexismo[25] e, por extensão, discurso de ódio. De forma semelhante, outros registros do HateBR foram classificadas como discurso de ódio pelos rotuladores, possivelmente devido à presença de expressões como "gay enrustido", enquanto os modelos as categorizaram como ofensivas apenas.

Outro caso intrigante envolve uma referência à ex-presidente do Chile, Michelle Bachelet, utilizando o termo "senhora". Embora essa instância tenha sido rotulada como ofensiva, potencialmente sugerindo um viés de etarismo, os modelos a classificaram como neutra. Esta divergência é particularmente controversa, uma vez que o termo em questão pode ser empregado simplesmente como uma forma de tratamento respeitoso. Nesse contexto, os modelos podem não capturar a sutileza e a gravidade atribuída ao uso de certos termos da mesma maneira que um anotador humano. Não obstante, esse exemplo também pode mostrar um caso de subjetividade particular entre indivíduos.

Adicionalmente, duas sentenças que mencionam religião em um teor negativo foram classificadas como discurso de ódio pelos modelos, enquanto originalmente uma é considerada neutra no HateBR e a outra ofensiva no OLID-BR. Esse cenário destaca a necessidade de investigações sobre vieses sociais relacionados à religião, tanto em conjuntos de dados quanto eventualmente nos modelos preditivos. Além disso, múltiplos casos, sobretudo no ToLD-Br, empregam o termo "nego(a)" para referir-se a

[24] Significado de ordinário: https://www.significados.com.br/ordinario/

[25] Definição de sexismo: https://www.significados.com.br/sexismo/

Dados	Texto	Rótulo	Predição
HateBR	Q mulher ordinária	discurso de ódio	ofensivo
	esses caras dão muito é CÚ	discurso de ódio	ofensivo
	Bandido gay enrustido ! Pedófilo de merda !	discurso de ódio	ofensivo
	Senhora Michele Bachelet !	ofensivo	neutro
	Quando alguém fala que é cristão , crispa até o cabelo da orelha .	neutro	discurso de ódio
OLID-BR	KKKKKKKKKK falo tudo cara ... a mulher tenta fuder toda a porra da história dos lobisomens e vampiros !	discurso de ódio	ofensivo
	O nega besta mds	discurso de ódio	ofensivo
	chorando igual uma vagabunda porra eu sabia q ia ser triste mas nao esperava essa cenakkkkkkkk	discurso de ódio	ofensivo
	Cada dia gosto respeito menos as religiões , todas . Bando de cínicos aproveitadores . URL	ofensivo	discurso de ódio
	USER Tinham sim , nos campos de concentração os gays usavam um triângulo rosa na roupa inclusive .	neutro	discurso de ódio
ToLD-Br	eu tô só adiando pra ir na casa de vovó , a bichinha já nem acredita quando falo que vou lá	discurso de ódio	neutro
	se ngm for cmg nessa feirinha vou ficar mt puta pq todo dia nego me cobra pra ir lá	discurso de ódio	neutro
	@USER Sim , tlg que foi uma brincadeira ... Mas eu imagino se fosse ao contrário , slc , nego n ia querer saber se o vídeo tá cortado ou não , já ia falar q o ney eh lixo	discurso de ódio	ofensivo
	alguém , além de mim , consegue morder a língua dormindo ? ? ? pqp !	ofensivo	neutro
	pai comprou bud e pqp q coisa boa	ofensivo	neutro

Tabela 10: Exemplos de predições em que os modelos — os mais eficazes entre os classificadores baseados em *encoder*, *decoder* e LLM ativado via *prompt* — concordam entre si nas predições, mas divergem do rótulo verdadeiro. A escolha dos modelos foi baseada na métrica F1 para a classe de discurso de ódio em cada conjunto de dados.

alguém, e foram identificados pelos anotadores humanos como discurso de ódio. Apesar de o termo não ser necessariamente usado para expressar racismo, alguns estudos sugerem que seu uso deveria ser evitado devido a sua conotação histórica (Guimarães Nascimento & Ribeiro, 2018). As instâncias que utilizaram esse termo foram percebidas como ofensivas ou neutras pelos modelos.

Uma instância específica do OLID-BR suscita preocupações. Embora trate de um assunto altamente sensível, referindo-se às vestimentas usadas por homossexuais em campos de concentração nazistas, seu conteúdo é, aparentemente, apresentado de forma explicativa e factual. Originalmente, essa instância recebeu o rótulo neutro, mas os modelos a classificaram como discurso de ódio. Esse é um caso que poderia acarretar uma possível censura equivocada devido à falha dos modelos em discernir o contexto adequadamente. Por outro lado, outro exemplo do OLID-BR apresenta o emissor se referindo a sua avó com o termo "bichinha". Como discutido anteriormente, embora esse termo possa ser associado a discursos homofóbicos, ele também é empregado com uma conotação carinhosa em certas regiões[26], o que parece ser o caso. Essas ocorrências demonstram que a tarefa de interpretar corretamente o contexto e a intenção imputada no uso de certos termos é complexa não só para os modelos, como também para os humanos.

Ademais, frequentemente sentenças que contêm a expressão "pqp" têm seus rótulos oficiais como ofensivo, porém são atribuídas para a classe neutro pelos modelos. Essa é uma expressão considerada vulgar que, embora possa ser utilizada como interjeição, pode ser ofensiva para determinados públicos. Neste contexto, é possível que os modelos tenham falhado em captar nuances semânticas associadas ao termo, com obstáculos associados ao seu pequeno

[26]Uso do termo "bichinho": https://pt.m.wiktionary.org/wiki/bichim

comprimento —apenas três letras— e ao fato de ser uma sigla cujo significado completo não é claramente manifestado[27].

Assim, os exemplos analisados ilustram a complexidade desta tarefa, que envolve o reconhecimento e a inferência adequada de uma variedade de aspectos culturais, sociais e históricos. Eles também revelam que os modelos podem não captar nuances extremamente específicas e indicam que existem desafios até mesmo na elaboração dos conjuntos de dados.

6. Conclusão

Este trabalho investigou o uso de classificadores fundamentados em *encoders* e *decoders*, especialmente ajustados para identificar discursos de ódio em português do Brasil, além de explorar diversas adaptações de *prompts* para ativar LLMs. Visando enriquecer a semântica na ativação de modelos generativos, propomos duas abordagens para o aprimoramento dos *prompts*: a incorporação de palavras-chave por meio de modelagem de tópicos e a seleção de exemplos demonstrativos de forma criteriosa. Observamos que, apesar da recente popularidade dos LLMs e suas proclamadas habilidades contextuais que supostamente dispensam treinamento extensivo, a adaptação de classificadores baseados em *encoder* de menor escala alcança os melhores resultados, especialmente com a estratégia de ajuste fino. Avaliamos também a eficácia de modelos baseados em *decoder* como classificadores tradicionais, considerando suas capacidades de alinhamento semântico. Constatamos que essa abordagem produziu os piores resultados no geral, indicando que ela pode não ser suficiente para contextos que exigem a captura de nuances delicadas. Além disso, as estratégias de escolha de exemplos demonstrativos e a adição de contexto beneficiam os modelos generativos ativados via *prompt*, com as abordagens sem demonstrações ou com escolhas aleatórias sendo superadas na maioria dos casos. Por fim, a detecção de discurso de ódio utilizando *transformers* mostra-se promissora, mas ainda enfrenta desafios, evidentes sobretudo na inspeção qualitativa realizada.

Em vista do exposto, reforçamos a literatura recente que defende uma investigação aprofundada sobre a capacidade dos modelos de linguagem de lidar com padrões sociais delicados, como o discurso de ódio, especialmente no idioma português. Modelos menores ainda têm um papel importante para evitar a perpetuação de questões sociais em ferramentas de PLN. Estudos futuros poderiam aprofundar-se em analisar o impacto das camadas, do *corpus* de treinamento e das variações arquiteturais nos modelos BERTimbau, BERTweet.BR, AlBERTina e Bernice, buscando elucidar as razões pelas quais esses modelos conseguem processar de maneira mais eficaz esse tipo de conteúdo. O desempenho variável de diferentes modelos em distintos conjuntos de dados, aliado à divergência observada nos resultados entre conjuntos de origens variadas, destaca a importância de se aprofundar no estudo da variabilidade de comportamentos e estratégias nas diversas redes sociais. Ademais, embora se tenha avaliado muitas abordagens, novos trabalhos podem incluir estratégias que considerem comitês e fusão de modelos, assim como a combinação de conjuntos de dados para treinamento (Leonardelli et al., 2023). Por fim, trabalhos subsequentes também incluem a investigação de novas adaptações arquiteturais que aprimoram a habilidade de modelos baseados em *decoder* em atuar como classificadores, bem como das nossas propostas para estratégias de aperfeiçoamento de *prompts*, visando verificar sua aplicabilidade em outros contextos.

6.1. Limitações

Este estudo possui limitações quanto à divisão de seus conjuntos de treinamento, validação e teste, pois adota apenas uma partição única. Essa restrição decorre principalmente dos custos associados ao uso da API do GPT, da API Gemini e dos recursos computacionais no ambiente Google Cloud. Ademais, essa limitação também restringiu a variação dos hiperparâmetros para nossos modelos, como o ajuste do número de épocas ou a modificação dos parâmetros de aprendizagem e inferência. Estamos cientes de que esses fatores podem influenciar direta ou indiretamente a interpretação do comportamento dos modelos em cenários mais abrangentes e generalistas. Todavia, tais decisões viabilizaram a análise e comparação de diversas abordagens, cada uma com suas características particulares. Outro aspecto relevante é que LLMs, como os avaliados neste trabalho, utilizam *corpora* amplamente baseados em dados extraídos da *web* em sua etapa de pré-treinamento. Ainda que não tenhamos analisado a contaminação potencial desses dados em relação aos conjuntos experimentados, os resultados indicam que esses modelos não incorporaram os rótulos para os textos, dado o desempenho registrado. Dessa forma, contribuímos com uma análise comparativa entre uma vasta quantidade de modelos sobre uma tarefa de classificação extremamente sensível.

[27] Uso do termo "pqp": https://pt.wiktionary.org/wiki/puta_que_pariu

Declaração Ética

Tratando-se de conjuntos de dados provenientes de redes sociais, procedeu-se à anonimização adequada de usuários e *links*. Além disso, considerando a sensibilidade associada ao tema de discurso de ódio, todas as pessoas que, de alguma forma, acessaram os dados foram previamente cientificadas sobre a sua natureza delicada. Para mais, o domínio exige considerações criteriosas. Primeiro, conjuntos de dados podem abrigar vieses culturais e históricos, falhando em capturar a plenitude da diversidade linguística e cultural. Nesse aspecto, o Brasil é um exímio exemplo de diversidade cultural; meramente examinar diferentes perspectivas dentro de regiões do mesmo país pode revelar discrepâncias na ofensividade percebida de um termo. Ademais, ao considerar-se perspectivas inter-países, tais diferenças podem se tornar ainda mais acentuadas, mesmo entre falantes do mesmo idioma. Por exemplo, "rapariga" em Portugal significa primariamente "jovem mulher", enquanto, no Brasil, o termo pode carregar conotações depreciativas[28].

Além disso, a classificação incorreta de conteúdos ofensivos e de discurso de ódio podem acarretar implicações significativas — falsos positivos podem levar a censura indevida e falsos negativos podem falhar em proteger grupos vulneráveis. Ao se classificar um conteúdo neutro como discurso de ódio pode-se incorrer em impactos para a sociedade, tais como, o desencorajamento dos usuários de se expressarem livremente, temendo a censura e a perda de confiança nas plataformas como espaço de debate público. Muitas vezes as redes sociais são a única ferramenta que as minorias encontram para lutar contra as opressões sofridas. Ao suprimir discussões sobre tópicos sensíveis, pode-se sufocar este debate público e impedir o avanço social em pautas importantes para estes grupos, tais como racismo, sexismo e violências enfrentadas em seu meio. Socialmente o impacto deste tipo de falso positivo pode aumentar desigualdades, suprimir o direito à liberdade e ao espaço democrático, com tendência de afetar desproporcionalmente os grupos sub-representados. Por outro lado, ao se classificar o discurso de ódio como neutro, pode-se perpetuar a atuação de grupos extremistas, a incitação à violência e o assédio, amplificando o impacto social e o sofrimento psicológico dos indivíduos. Neste caso, novamente, as comunidades vulneráveis sofreriam com maiores consequências.

Por fim, como forma de minimizar as preocupações éticas sobre falsos positivos e negativos, enfatizamos que mecanismos de Inteligência Artificial (IA) são viáveis enquanto auxílios na moderação de conteúdo, mas não devem ser substitutos definitivos para ela. Destaca-se a necessidade da observância dos princípios éticos, com destaque para o desenvolvimento de abordagens explicáveis e transparentes. Modelos que fornecem explicações para suas decisões podem ajudar os moderadores a entender por que o conteúdo foi sinalizado e fazer julgamentos mais bem informados. Além disso, os usuários das plataformas precisam estar cientes de que seu conteúdo foi moderado com o auxílio de IA. Alcançar o equilíbrio certo é complexo, exigindo ainda avanços na técnica e robustez dos modelos e na supervisão ética de equipes interdisciplinares.

Agradecimentos

Os autores agradecem o financiamento do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), bolsas 307088/2023-5 e 315750/2021-9, da FAPERJ - Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, processos SEI-260003/002930/2024, SEI-260003/000614/2023, SEI-260003/006057/2024, e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código Financeiro 001. Também contou-se com o apoio do programa *Google Cloud Research Credits*, sob o código GCP19980904, e do programa de créditos da Maritaca AI. Por fim, utilizamos o ChatGPT-4 para agilizar a elaboração de códigos em Python e Latex. Cada sugestão da IA foi cuidadosamente examinada, testada e, muitas vezes, ajustada por nós, imputando-nos a plena responsabilidade pela forma e conteúdo deste artigo.

Os autores agradecem ainda aos revisores e ao editor pelas observações e sugestões construtivas, que ajudaram a melhorar o artigo.

Referências

Almeida, Thales Sales, Hugo Abonizio, Rodrigo Nogueira & Ramon Pires. 2024. Sabiá-2: A new generation of portuguese large language models. arXiv [cs.CL]. 10.48550/arXiv.2403.0988

Aluru, Sai Saketh, Binny Mathew, Punyajoy Saha & Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. arXiv [cs.SL/cs.CL]. 10.48550/arXiv.2004.06465

[28] https://bit.ly/rapariga-Brasil-Portugal

Amorim, Annie, Nils Murrugarra-Llerena, Vítor Silva, Daniel de Oliveira & Aline Paes. 2023. ALTES: uma ferramenta de rotulação automática de tópicos por meio de fontes externas. Em *38th Simpósio Brasileiro de Bancos de Dados*, 120–125. 10.5753/sbbd_estendido.2023.233252

Assis, Gabriel, Annie Amorim, Jonnathan Carvalho, Daniel de Oliveira, Daniela Vianna & Aline Paes. 2024. Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? Em *16th International Conference on Computational Processing of Portuguese (PROPOR)*, 301–311.

BehnamGhader, Parishad, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados & Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. arXiv [cs.CL]. 10.48550/arXiv.2404.05961

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. Em *Neural Information Processing Systems*, 1877–1901.

Caneiro, Fernando, Daniela Viana, Jonathan Carvalho, Alexandre Plastino & Aline Paes. 2024. BERTweet.BR: A pretrained language model for tweets in Portuguese. *Neural Computing and Applications*. 10.1007/s00521-024-10711-3

Chiu, Ke-Li, Annie Collins & Rohan Alexander. 2022. Detecting hate speech with GPT-3. arXiv [cs.CL]. 10.48550/arXiv.2103.12407

Das, Mithun, Saurabh Kumar Pandey & Animesh Mukherjee. 2023. Evaluating ChatGPT's performance for multilingual and emoji-based hate speech detection. arXiv [cs.CL]. 10.48550/arXiv.2305.13276

DeLucia, Alexandra, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik & Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for twitter. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6191–6205. 10.18653/v1/2022.emnlp-main.415

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics (AACL)*, 4171–4186. 10.18653/v1/N19-1423

Edwards, Griffin Sims & Stephen Rushin. 2018. The effect of president Trump's election on hate crimes. *SSRN Electronic Journal* 10.2139/ssrn.3102652

Fortuna, Paula & Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys* 51(4). 1–30. 10.1145/3232676

Fortuna, Paula, João Rocha da Silva, Juan Soler-Company, Leo Wanner & Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. Em *3rd Workshop on Abusive Language Online*, 94–104. 10.18653/v1/W19-3510

de Freitas, Riva Sobrado & Matheus Felipe de Castro. 2013. Liberdade de expressão e discurso do Ódio: um exame sobre as possíveis limitações à liberdade de expressão. *Seqüência Estudos Jurídicos e Políticos* 34(66). 327–355. 10.5007/2177-7055.2013v34n66p327

French, Megan & Natalya N. Bazarova. 2017. Is Anybody out There?: Understanding Masspersonal Communication through Expectations for Response across Social Media Platforms. *Journal of Computer-Mediated Communication* 22(6). 303–319. 10.1111/jcc4.12197

Fu, Jinlan, See-Kiong Ng, Zhengbao Jiang & Pengfei Liu. 2024. GPTScore: Evaluate as you desire. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 6556–6576. 10.18653/v1/2024.naacl-long.365

Google, Gemini Team. 2023. Gemini: A family of highly capable multimodal models. arXiv [cs.CL]. 10.48550/arXiv.2312.11805

Grootendorst, Maarten. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv [cs.CL]. 10.48550/arXiv.2203.05794

Guimarães Nascimento, Raquel Costa & Erislane Rodrigues Ribeiro. 2018. Uma análise discursiva dos memes "nego isso, nego aquilo". *Revista do Sell* 7(1).

Hu, Edward J, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang & Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. Em *International Conference on Learning Representations*,

Jahan, Md Saroar & Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing* 546. 126232. 10.1016/j.neucom.2023.126232

Jin, Xin & Jiawei Han. 2011. K-Means clustering. *Encyclopedia of Machine Learning* 563–564. 10.1007/978-0-387-30164-8_425

Leite, João Augusto, Diego Silva, Kalina Bontcheva & Carolina Scarton. 2020. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. Em *1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 914–924. 10.18653/v1/2020.aacl-main.91

Leonardelli, Elisa, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma & Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). Em *17th International Workshop on Semantic Evaluation (SemEval)*, 2304–2318. 10.18653/v1/2023.semeval-1.314

Li, Zongxi, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu Lee Wang, Qing Li & Xiaoqin Zhong. 2023. Label supervised LLaMA finetuning. arXiv [cs.CL]. 10.48550/arXiv.2310.01208

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi & Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* 55(9). 1–35. 10.1145/3560815

Moura, Marco Aurelio. 2016. *O discurso do ódio em redes sociais*. Lura Editorial

Nguyen, Thanh Thi, Campbell Wilson & Janis Dalins. 2023. Fine-tuning Llama 2 large language models for detecting online sexual predatory chats and abusive texts. arXiv [cs.CL]. 10.48550/arXiv.2308.14683

Oliveira, Amanda, Thiago de Carvalho Cecote, João Paulo Reis Alvarenga, Vander Luis de Souza Freitas & Eduardo José da Silva Luz. 2024. Toxic speech detection in Portuguese: A comparative study of large language models. Em *16th International Conference on Computational Processing of Portuguese (PROPOR)*, 108–116.

Oliveira, Amanda, Thiago Cecote, Pedro Silva, Jadson Gertrudes, Vander Freitas & Eduardo Luz. 2023. How good is ChatGPT for detecting hate speech in Portuguese? Em *14th Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 94–103. 10.5753/stil.2023.233943

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike & Ryan Lowe. 2022. Training language models to follow instructions with human feedback. Em *36th Conference on Neural Information Processing Systems (NeurIPS)*, 27730–27744.

Paiva, Peter, Vanecy da Silva & Raimundo Moura. 2019. Detecção automática de discurso de ódio em comentários online. Em *7th Escola Regional de Computação Aplicada à Saúde*, 157–162.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Pelle, Rogers, Cleber Alcântara & Viviane P. Moreira. 2018. A classifier ensemble for offensive text detection. Em *24th Brazilian Symposium on Multimedia and the Web*, 237–243. 10.1145/3243082.3243111

Pham, Chau, Alexander Hoyle, Simeng Sun, Philip Resnik & Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2956–2984. 10.18653/v1/2024.naacl-long.164

Pires, Ramon, Hugo Queiroz Abonizio, Thales Sales Almeida & Rodrigo Frassetto Nogueira. 2023. Sabiá: Portuguese large language models. Em *Intelligent Systems*, 226–240. 10.1007/978-3-031-45392-2_15

Plath, Hannah O., Maria Estela O. Paiva, Danielle L. Pinto & Paula D. P. Costa. 2022.

Detecção de discurso de Ódio contra mulheres em textos em português brasileiro: Construção da base MINA-BR e modelo de classificação. *Revista Eletrônica de Iniciação Científica em Computação* 20(3).

Rodrigues, João, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso & Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer Albertina PT-*. arXiv [cs.CL]. 10.48550/arXiv.2305.06721

Santos, Raquel Bento, Bernardo Cunha Matos, Paula Carvalho, Fernando Batista & Ricardo Ribeiro. 2022. Semi-supervised annotation of Portuguese hate speech across social media domains. Em *11th Symposium on Languages, Applications and Technologies (SLATE)*, 11:1–11:14. 10.4230/OASIcs.SLATE.2022.11

Santos, Rodrigo, João Silva, Luís Gomes, João Rodrigues & António Branco. 2024. Advancing generative ai for Portuguese with open decoder Gervásio PT-*. arXiv [cs.CL]. 10.48550/arXiv.2402.18766

Saraiva, Ghivvago Damas, Rafael Anchiêta, Francisco Assis Ricarte Neto & Raimundo Moura. 2021. A semi-supervised approach to detect toxic comments. Em *International Conference on Recent Advances in Natural Language Processing (RANLP)*, 1261–1267.

Seno, Eloize Rossi Marques, Daniela Claro, Laila Mota & Jessica Rodrigues. 2024. Semântica distribucional. Em *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, BPLN 2nd edn.

da Silva, Rodolfo Costa Cezar & Thierson Couto Rosa. 2023. Combining data transformation and classification approaches for hate speech detection: A comparative study. SSRN Preprint. 10.2139/ssrn.4477182

Silva, Rodolvo, Deborah Fernandes & Márcio Fernandes. 2018. Caracterização de mensagens em língua portuguesa com traços de racismo no Twitter. Em *6th Escola Regional de Informática de Goiás*, 205–214.

Souza, Andrey, Eduardo Nakamura & Fabíola Nakamura. 2022. Detecção de discurso de Ódio: Homofobia. Em *16th Brazilian e-Science Workshop*, 73–80. 10.5753/bresci.2022.223222

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *Intelligent Systems*, 403–417. 10.1007/978-3-030-61377-8_28

de Souza, Lucas, Franciele Beal, André Ortoncelli & Marlon Marcon. 2024. Detection and censorship of offensive language in extended texts in Portuguese. Em *15th XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 202–211. 10.5753/stil.2024.245421

Thorndike, Robert L. 1953. Who belongs in the family? *Psychometrika* 18(4). 267–276. 10.1007/BF02289263

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave & Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models. arXiv [cs.CL]. 10.48550/arXiv.2302.13971

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov & Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. arXiv [cs.CL]. 10.48550/arXiv.2307.09288

Trajano, Douglas, Rafael H. Bordini & Renata Vieira. 2023. OLID-BR: offensive language identification dataset for Brazilian Portuguese. *Language Resources and Evaluation* 58. 1263–1289. 10.1007/s10579-023-09657-0

Vargas, Francielle, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo & Fabrício Benevenuto. 2022. HateBR: A large expert

annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. Em *Language Resources and Evaluation Conference (LREC)*, 7174–7183.

Vargas, Francielle, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto & Thiago Pardo. 2021. Contextual-lexicon approach for abusive language detection. Em *Conference on Recent Advances in Natural Language Processing (RANLP)*, 1438–1447.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. Em *31st Conference on Neural Information Processing Systems (NIPS)*,

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 38–45. 10.18653/v1/2020.emnlp-demos.6

Yin, Wenjie & Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computing Science* 7. e598. 10.7717/PEERJ-CS.598

Yosinski, Jason, Jeff Clune, Yoshua Bengio & Hod Lipson. 2014. How transferable are features in deep neural networks? Em *27th International Conference on Neural Information Processing Systems (NIPS)*, 3320–3328.

Apêndices

A. *Prompt one-shot*

Classifique o texto de rede social como "discurso de ódio" ou "ofensivo" ou "neutro".

Texto: Eu quero que ele vá pra puta que pariu , tô de saco cheio já .
Classe: ofensivo

Texto: {texto_alvo}
Classe:

B. *Prompt* com contexto adicional

Classifique o texto de rede social como "discurso de ódio", "ofensivo" ou "neutro".

Considerando que os assuntos da classe "discurso de ódio" estão associados com as palavras e emojis: mulher, bonito, bixa, viado, amo, ninguém, flor, koe, doido, boiola.
Da classe "ofensivo" estão associados com as palavras e emojis: pessoa, aborto, macho, presidente, mito, xis, podre, caneta, :rosto_sorridente_com_chifres:, pau.
Da classe "neutro" estão associados com as palavras e emojis: ano, kakakaakk, mesmo, alguém, garota, pro, conheço, óbvio, versão, kkkk.

Texto: @user eu tambem, olha a jules falando que ama a rue
Classe: neutro

Texto: {texto_alvo}
Classe: