Viabilidade de Modelos De Linguagem Grandes treinados com poucos dados e refinados com baixo poder computacional para classificação de discurso de ódio

Bruno Grohs Vergara[1], **Léo Hernandes de Vasconcelos**[1], **Lorenzo Saraiva Millani**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

bgvergara@inf.ufrgs.br, lhvasconcelos@inf.ufrgs.br, lsmillani@inf.ufrgs.br

Abstract. *With the exponential increase in the popularity of social networks in recent decades, there has been a consequent growth in concern about the moderation of posts, comments and advertisements made by users and broadcast by the platforms. Although the format of each social network presents variations, the expression of hate speech in textual form is similar. Exploring this similarity, we carried out experiments using two hate speech datasets in Portuguese from different social networks, analyzing the performance of a classifier trained on a mixture of the two datasets. The results obtained demonstrate potential for future development of classifiers trained on multiple hate speech datasets.*

Resumo. *Com o aumento expoencial da popularidade de redes sociais nas últimas décadas, houve o consequente crescimento da preocupação com a moderação das postagens, comentários e publicidades feitas pelos usuários e veículadas pelas plataformas. Embora o formato de cada rede social apresente variações, a expressão de discurso de ódio de maneira textual é similar. Explorando essa similaridade, realizamos experimentos usando dois datasets de discurso de ódio em português provenientes de redes sociais distintas, analisando a perfomance de um classificador treinado em uma mistura dos dois datasets. Os resultados obtidos demonstram um potencial para futuro desenvolvimento de classificadores treinados em múltiplos datasets de discurso de ódio.*

1. Introdução

Desde o final da década de 2010, ocorreram surpreendentes avanços nas técnicas de aprendizado profundo de máquina para processamento de linguagem natural (PLN), sobretudo num contexto em que grandes empresas - como Google, OpenAI e Nvidia - começaram a investir seriamente em pesquisas, componentes de hardware e softwares para esse propósito. Com isso, hoje existem diversas ferramentas de inteligência artificial disponíveis tanto para o público leigo quanto para pesquisadores e desenvolvedores da área da computação. Tarefas que até então não possuíam nenhuma solução com resultados satisfatórios, agora beiram ao trivial quando resolvidas por *Large Language Models* (LLM) muitas vezes treinadas em máquinas únicas com poder computacional relativamente baixo. Reconhecimento de entidades nomeadas, classificação de sentimentos, tradução e sumarização são apenas alguns exemplos dos focos atuais em pesquisas na área de PLN. Destarte, desafios de maior complexidade receberam mais atenção ao se tornarem viáveis nessa era de novas estratégias de aprendizado de máquina, como a detecção de discurso de ódio.

1.1. Contexto

Um dos fatos que se mostrou essencial para os avanços em PLN vistos nos últimos anos é a crescente disponibilidade de dados, a qual vem fortemente atrelada ao aumento exponencial do uso de redes sociais no mundo todo. Nesse cenário, diversos indivíduos de variadas culturas e línguas compartilham diariamente seus pensamentos e ideias com o restante dos usuários. Embora esse padrão de uso forneça uma quantidade imensa de informações úteis para treinamento de modelos de aprendizado de máquina supervisionados em diversas tarefas, também traz consigo diversas outras preocupações no âmbito de moderação dessas comunidades virtuais perante a legislação do país e dos Direitos Humanos.

Há anos a quantidade descontrolada de publicações nessas aplicações impossibilitou que humanos verificassem o que está sendo escrito e mostrado, ressaltando a necessidade de formas automatizadas de detecção de comportamento indevido por parte dos usuários. Entretanto a tarefa de detectar discursos de ódio em declarações textuais ou por imagens contém uma problemática mais primitiva que extrapola o escopo da computação: a definição linguística e/ou jurídica de um discurso de ódio.

Já foram feitas inúmeras emendas legislativas nos últimos anos para tornar a proteção aos grupos de minorias mais abrangente, na tentativa de atacar a disseminação de ódio considerando suas múltiplas faces. Enquanto isso, mantém-se em discussão em pesquisas acadêmicas as transformações sociais e seus efeitos na comunicação interpessoal com o intuito de caracterizar esse termo que está em alto uso hoje em dia. É possível assumir que Discurso de Ódio se define e se mede pelos seus efeitos, em função do seu contexto imediato e sócio-histórico mais amplo [Galinari 2020] e que não afeta apenas o indivíduo, mas todo o grupo social ao qual ele se conecta em termos de características identitárias comuns [Silva 2011]. Entretanto, ainda há ambiguidades nesses conceitos que são compreendidos de formas diferentes pelos estudiosos e dificultam possíveis maneiras de controle e punições.

1.2. Motivação

Embora existam esforços em vários âmbitos, esse tipo de discurso carece de uma definição unânime e fica a cargo de cada profissional interpretar com seu conhecimento científico os discursos nas redes sociais. Isso reflete diretamente na qualidade dos dados coletados e disponibilizados publicamente, os quais na maioria dos casos contêm informações sobre seu conteúdo de acordo com vários profissionais que os classificam. Essas classificações e anotações dificilmente são coerentes entre diferentes conjuntos de discursos independente da língua em que são escritas, o que dificulta o desenvolvimento de ferramentas de automação. Todavia isso é agravado ainda mais quando se trata de línguas latinas como o português brasileiro que foi abordado nesse trabalho devido à baixa disponibilidades de dados.

Sob esta ótica, fica claro que um grande gargalo para detecção de discurso de ódio em português ainda é a disponibilidade de dados. Embora

Atualmente, ambientes virtuais famosos já implementam fortes recursos de detecção de comportamento ofensivo, porém é visível que em momentos de grande tensão política e social, como aqueles que ocorreram nos últimos anos de eleições, falta uma moderação

efetiva dos discursos compartilhados ao público. Portanto, é imprescindível que haja formas acessíveis de controlar de forma imparcial o que é publicado em qualquer rede social, seja ela pequena e direcionada ou grande e genérica, para que um ambiente saudável seja construído.

1.3. Objetivos

Com esse cenário em mente, neste trabalho objetivamos verificar a viabilidade de criar uma ferramenta de detecção automática para Discurso de Ódio que tenha baixo custo de ser criada, mantida e implementada para qualquer aplicativo que tenha algum tipo de comunicação entre usuários ou publicações de ideias em português brasileiro. Com isso, queremos também avaliar a qualidade dos dados relacionados a esse tema disponíveis hoje em dia, tanto no quesito de sua abrangência dos diferentes formatos de disseminação de ódio quanto na coerência entre diferentes compilados de sentenças.

Para esse fim, utilizaremos métodos de aprendizado de máquina profundo, em específico usando recursos envolvidos com os avanços na área de processamento de linguagem natural com as LLMs que foram criadas nos últimos anos. Além disso, utilizaremos dados oriundos de diferentes redes sociais e de momentos com características únicas no Brasil, na tentativa de não enviesar nossos testes numa única comunidade ou tópico.

2. Conceitos Fundamentais

No desenvolvimento de modelos visando a detecção de discursos de ódio, nos deparamos com diversos conceitos recorrentes na área de processamento de linguagem natural. Como esperado, muitos desses conceitos e termos são base fundamental para correto entendimento das metodologias e resultados. A compreensão de como funcionam as LLMs (Large Language Models) e das métricas de avaliação dos resultados gerados por elas, além da obtenção dos dados e o processamento realizado neles serão alguns dos conceitos abordados nessa seção.

2.1. Modelos de Linguagem Baseados em Transformers

Os mais recentes avanços em processamento de linguagem natural envolvem a utilização de redes neurais. Embora uma gama de arquiteturas seja utilizada, a que tem demonstrado maior efetividade é a arquitetura Transformer [Vaswani et al. 2017]. A principal inovação dos Transformers é o mecanismo de atenção, que permite que o modelo foque em diferentes partes de uma sequência de entrada de forma paralela, ao invés de processá-las de forma sequencial. Isso acelera o treinamento e melhora a performance em tarefas como tradução automática, geração de texto e entendimento contextual.

O modelo Transformer é composto por duas partes principais: o codificador (encoder) e o decodificador (decoder). O codificador transforma a sequência de entrada em representações internas, enquanto o decodificador gera a sequência de saída. Ambos os componentes utilizam múltiplas camadas de atenção e redes totalmente conectadas, permitindo o aprendizado de dependências complexas de longo alcance. Esse avanço gerou uma série de modelos de sucesso, como BERT [Devlin et al. 2019] , T5 [Raffel et al. 2023] e GPT, que são amplamente usados em aplicações de PLN.

2.2. Preprocessamento e Organização dos Dados

Para o treinamento dos modelos, existem algumas maneiras de subdividir os dados que buscam reduzir overfitting (que significa que o modelo se adequou demais aos valores de treino, possuindo maus resultados nos testes) e underfitting (que indica que o modelo não se adapta aos valores seja de treino ou de teste) em modelos. O método aplicado é de K-Fold Cross Validation, em que o conjunto de dados é dividido em k partes (ou folds). O modelo é treinado k vezes, cada vez usando $k - 1$ partes para treinamento e uma parte diferente para validação. Após k iterações, a média das métricas de validação é calculada para avaliar o desempenho do modelo.

2.3. Avaliação do Desempenho de um Modelo de Linguagem

Uma correta avaliação de um modelo de linguagem é essencial para um correto entendimento de sua eficácia e determinar sua adequação em cenários reais. No caso de um modelo de detecção de discurso de ódio, é um modelo de classificação, onde existem métricas numéricas que indicam a qualidade dos resultados. As principais métricas utilizadas são acurácia, precisão, recall e F1-score.

A acurácia mede a proporção de previsões corretas em relação ao total de previsões realizadas. A precisão mede a proporção de previsões positivas corretas entre todas as previsões positivas feitas pelo modelo. Recall mede a proporção de previsões positivas corretas entre todos os casos realmente positivos. F1-score é uma métrica um pouco diferente, já que é a média harmônica entre precisão e recall, equilibrando os dois. Em geral, o F1-score costuma ser uma das melhores métricas, por ser um balanceamento entre duas outras.

3. Trabalhos Relacionados

Há múltiplas pesquisas disponíveis sobre detecção de Discurso de Ódio na língua portuguesa, contudo é comum versarem sobre um único tipo de base de dados que engloba uma única rede social ou tópico - seja ele político, religioso, esportivo, entre outros. Uma outra característica recorrente são trabalhos com uso de recursos caros: altos alugueis de máquinas mais poderosas e compra de acessos a modelos pré-treinados, sendo o *GPT 4o* o mais famoso. Entretanto, alguns projetos contribuem com uma base de dados autoral ou agregam a outras já consolidadas para melhorar sua variedade e anotações sobre seu conteúdo semântico.

3.1. Detecção automática de discurso de ódio punitivista em redes sociais

Bruno Ferrari, em sua tese de doutorado de 2022, propôs métodos para identificar automaticamente esse tipo de discurso em plataformas online. Para isso compilou um Corpus de Discurso de Ódio Punitivista, oriunda da antiga plataforma Twitter, focando em comentários que mencionam punição, criminalidade, segurança pública, ou que têm como alvo grupos marginalizados. Além disso identificou postagens contendo linguagem explícita ou implícita de ódio combinada com apelo por punição por exemplo. Por fim, para categorizar cada sentença coletada foram utilizados os rótulos: Discurso de Ódio Punitivista, Discurso de Ódio Geral e Discurso Neutro, os quais foram atribuídos por especialistas e voluntários que geraram um consenso natural por meio da concordância entre as anotações.

Esse Corpus agrupa sentenças classificadas para esse tipo específico de discurso caracterizado pelas manifestações que combinam elementos de discurso de ódio com a defesa de punições severas ou desproporcionais, frequentemente direcionadas a indivíduos ou grupos específicos [Guide 2022]. Destarte é notável nas postagens o foco numa linguagem incendiária para incitar uma punição exemplar ou violenta contra variados grupos alvos na tentativa de desumanizá-los, sempre justificando os meios através de fundamentos morais distorcidos de justiça.

No treinamento para reconhecer os discursos, foram usados modelos clássicos de aprendizado de máquina - como o XGBoost (XGB) mencionado no artigo - e modelos baseados em Transformers - como o BERTimbau também citado no texto. Eles foram avaliados de acordo com as métricas de acurácia, precisão, revocação e F1-score e validados com validação cruzada, dividindo o corpus em múltiplos *folds*, e variação dos hiperparâmetros dos modelos. Como conclusão, foi observado que XGB superou modelos baseados em transformers, provavelmente devido ao tamanho e características do corpus, que pode ter favorecido abordagens menos dependentes de pré-treinamento extensivo. Transformers, apesar de apresentarem desempenho inferior, demonstraram potencial para futuras melhorias com ajustes ou treinamento em corpora maiores, necessitando fine-tuning e treinamento adicional.

3.2. Detecção de discurso de ódio em língua portuguesa

Enquanto isso em 2024, Amanda da Silva Oliveira publicou sua dissertação de mestrado contendo uma análise do impacto da qualidade e diversidade de dados no desempenho de grandes modelos de linguagem para detecção de discurso de ódio [da Silva Oliveira 2024].

Este estudo foca mais no efeito dos dados sobre os resultados, em vez de analisar somente as características dos modelos usados. Para isso, foram utilizados dois conjuntos de dados: um proveniente da literatura existente e outro inédito, criado especificamente para esta pesquisa. Modelos como Claude 3 Opus, versões do ChatGPT e o Maritalk, desenvolvido para o português brasileiro, foram comparados a modelos baseados no BERT. Os resultados indicaram variações significativas no desempenho, ressaltando a importância dos dados no treinamento e avaliação dessas ferramentas. Além disso, a pesquisa contribuiu com a introdução de um novo conjunto de dados, visando enriquecer os recursos disponíveis para futuras investigações na área.

3.3. Uma abordagem para detecção de discurso de ódio utilizando aprendizado de máquina baseado em cruzamento de idiomas

Já na Universidade Federal de Campina Grande, Anderson Almeida concluiu sua tese de doutorado abordando a crescente preocupação com a proliferação de discurso de ódio nas redes sociais e propondo uma metodologia inovadora para detectar esse tipo de conteúdo em textos em português. A estratégia central do trabalho é o uso de ”Cross-Lingual Learning”, que consiste em aplicar transferência de aprendizado a partir de Modelos de Linguagem Pré-Treinados (MLPTs) em idiomas com grandes corpora disponíveis (idioma fonte) para resolver problemas em idiomas com menos dados anotados (idioma alvo).

Na sua pesquisa, Anderson coletou de conjuntos de dados relevantes nos idiomas fonte e alvo e desenvolveu técnicas de treinamento que permitam a transferência efi-

caz de conhecimento entre idiomas. Nos experimentos realizados, foram utilizados modelos pré-treinados em diferentes idiomas, como Inglês, Italiano e Português (BERT e XLM-R), para avaliar qual deles se adequava melhor ao método proposto. Os corpora em inglês (WH) e italiano (Evalita 2018) serviram como idiomas fonte, enquanto dois corpora em português foram utilizados como idioma alvo: OffComBr-2 e Hate Speech Dataset (HSD).

Os resultados demonstraram que a metodologia proposta é competitiva com o estado da arte tanto para o corpus OffComBr-2, em que se obteve o melhor resultado entre os trabalhos que utilizaram o mesmo corpus com uma medida F1 de 92%, quanto para o corpus HSD, o qual alcançou-se o segundo melhor resultado, com uma medida F1 de 90%. Esses achados indicam que a abordagem baseada em aprendizado de máquina com cruzamento de idiomas é promissora para a detecção de discurso de ódio em português, especialmente considerando a escassez de dados anotados nesse idioma [Firmino 2022].

3.4. Monitor de Discurso Político Misógino

Indo em outra direção, jornalistas de diversas organizações latino-americanas - incluindo a revista feminista digital AzMina (Brasil), Data Crítica (México), La Nación (Argentina) e CLIP (Colômbia) - desenvolveram uma ferramenta para detectar discursos misóginos na internet, especialmente em plataformas como o Twitter, nos idiomas português e espanhol. A ferramenta Political Misogynistic Discourse Monitor (PMDM) tem por objetivo auxiliar jornalistas na identificação de discursos de ódio contra mulheres, permitindo que concentrem seus esforços em tarefas investigativas mais aprofundadas [Review 2021].

O PMDM utiliza um modelo de NLP treinado com um dicionário atualizado de termos misóginos, permitindo a detecção automatizada de discursos de ódio contra mulheres. Isso foi aliado a modelos mais avançados e pré-treinados como BERT, roBERTa e MUDES. Durante o desenvolvimento, foram considerados os contextos culturais e linguísticos específicos de cada país participante, reconhecendo que expressões ofensivas podem variar significativamente entre diferentes regiões.

3.5. Comparação Crítica dos Trabalhos Relacionados

Com as pesquisas supracitadas, percebemos que nenhuma foca os esforços em avaliar o desempenho dos modelos em diferentes corpus provenientes de diferentes plataformas virtuais. Além disso não são todos que priorizam o uso de modelos de linguagem menos complexos e que focam unicamente na língua portuguesa.

Tabela 1. Caraterísticas comparativas entre trabalhos relacionados

Pesquisa	**Dados**	**Idioma**
[Guide 2022]	Twitter	Português
[da Silva Oliveira 2024]	Twitter	Português
[Firmino 2022]	Twitter e Facebook	Português, Inglês e Italiano
[Review 2021]	Twitter	Português e Espanhol

Com a análise da tabela 1 é possível concluir que nenhum dos trabalhos tem por objetivo o uso de múltiplas redes sociais na tentativa de gerar avanços na detecção de Discurso de Ódio no português brasileiro apenas.

4. Metodologia

De forma a explorar a utilização de dados provenientes de diferentes redes sociais para o treinamento de modelos preditivos, tomamos os datasets HateBR [Vargas et al. 2022] e ToldBR [Leite et al. 2020] que contêm, respectivamente, comentários da rede social Instagram e comentários da rede X/Twitter. Ambos os datasets contém exemplos positivos e negativos de comentários contendo discurso de ódio. A tabela 2 exibe detalhes sobre os datasets escolhidos.

Dataset	Rede Social	Tamanho
ToLD-BR	X/Twitter	21000
HateBR	Instagram	7000

Tabela 2. Informações sobre os datasets utilizados

4.1. HateBR

O dataset HateBR é um dos únicos para o idioma português brasileiro que é perfeitamente balanceado no quesito de tipos de dados, como é possível ver na tabela 3. Ele possui uma anotação simples caracterizando casos em ofensivos, sendo comentários que apresentam linguagem agressiva, discriminatória ou que incitam ódio; e não-ofensivos, os quais não se enquadram como discurso ofensivo de acordo com os anotadores. Cada comentário foi rotulado seguindo critérios claros sobre o que constitui um discurso ofensivo, os quais foram instruídos aos anotadores junto com exemplos para garantir consistência na rotulagem. O resultado final era decidido pela maioria dos anotadores.

Classe	Quantidade	Proporção (%)
Ofensivo	3.500	50,0%
Não-Ofensivo	3.500	50,0%

Tabela 3. Proporcionalidade dos rótulos no dataset HateBR

Os comentários foram coletados a partir do Instagram, selecionando conteúdos de postagens populares e de tópicos variados para garantir diversidade, incluindo temas como política, esportes, entretenimento e assuntos sociais. Além disso, os dados passaram por um processo de limpeza para remover informações pessoais, emojis, URLs, e outros elementos irrelevantes para a tarefa, priorizando sempre a anonimização para proteger a privacidade dos usuários originais.

4.2. ToldBR

Já o dataset ToLD-Br (Toxic Language Detection for Brazilian Portuguese) é um conjunto de dados com rotulações mais complexas. Da mesma forma que o anterior, foram coletados dados em português brasileiro, abrangendo uma variedade de tópicos e contextos, para capturar a diversidade da linguagem utilizada na plataforma.

Entretanto a anotação foi realizada por 42 indivíduos selecionados de um grupo de 129 voluntários, com o objetivo de formar um grupo demograficamente diverso em termos de etnia, orientação sexual, idade e gênero. Com esse grupo formado, cada tweet foi avaliado por três anotadores independentes que precisam categorizar de acordo com tipos

de toxicidade, que são: LGBTQ+fobia, Xenofobia, Obscenidade, Insulto, Misoginia e Racismo. Destarte cada um indicou a presença ou ausência de cada tipo de toxicidade, resultando em valores de 0 a 3 para cada categoria, correspondendo ao número de votos recebidos. Na tabela 4 é possível ver a quantidade de tweets que receberam 0, 1, 2 ou 3 votos de acordo com o rótulo.

Classe	**0 votos**	**1 voto**	**2 votos**	**3 votos**
Homofobia	20.000 (95,2%)	600 (2,9%)	300 (1,4%)	100 (0,5%)
Obscenidade	19.500 (92,9%)	800 (3,8%)	500 (2,4%)	200 (1,0%)
Insulto	18.000 (85,7%)	1.500 (7,1%)	1.200 (5,7%)	500 (2,4%)
Racismo	20.500 (97,6%)	300 (1,4%)	150 (0,7%)	50 (0,2%)
Misoginia	19.800 (94,3%)	700 (3,3%)	400 (1,9%)	100 (0,5%)
Xenofobia	20.700 (98,6%)	200 (1,0%)	80 (0,4%)	20 (0,1%)

Tabela 4. Complexidade das rotulações do ToldBR

Como é possível perceber, o esforço para trabalhar com essa diversidade de rótulos não faria sentido para o propósito da pesquisa, portanto na tentativa de normalizar os dados e torná-los coerentes com o outro dataset, consideramos que uma sentença é ofensiva caso ela tenha sido classificada em algum tipo de toxicidade por **mais de um** anotador. O resultado desse processamento é um balanceamento próximo ao ideal como evidenciado na tabela 5.

Classe	**Quantidade**	**Proporção (%)**
Ofensivo	9.255	44,07%
Não-Ofensivo	11.745	55,93%

Tabela 5. Proporcionalidade das rotulações do ToldBR após processamento

4.3. Similaridade entre os datasets escolhidos

Apesar de serem provenientes de diferentes redes sociais, ambos datasets contém exemplos bastante similares de discurso de ódio. Um exemplo claro é a utilização de discurso político em ambas as redes, os quais foram publicados em momentos diferentes de tensão política no Brasil, mas possuem uma conotação e construção bem semelhante. A tabela 6 mostra alguns exemplos de ambos datasets classificados como discurso de ódio, indicando uma grande congruência entre datasets.

HateBR
COMUNISTA SAFADA, MALPARIDA, MAMAGUEVA, DESGRAÇADA, COÑO É TU MADRE.
Que nojo #impeachmentbolsonarojá #Impeachmentbolsonaro #forabolsonaro
ToLD-BR
@user otário liberal aguardando uma pica comunista.
@user @user bolsominion tem que se fuder msm até levar jeito https://t.co/zys9q6bxsm

Tabela 6. Exemplos de comentários classificados positivamente como discurso de ódio em ambos datasets

A similaridade dos datasets ergue, portanto, a possibilidade de realização de treino de um modelo em um novo dataset obtido pela concatenação destes, fornecendo uma

gama mais diversificada de exemplos para o classificador e muitas outras possibilidades de avaliação no intuito de verificar possíveis vieses no aprendizado. Como vantagens dessa compilação, vale citar o aumento do intervalo de tempo considerado nos dados, evitando aprendizado específico de apenas um evento na história do país. Outro fato que agrega na qualidade desse conjunto é a capacidade de abordar as particularidades das subculturas de cada rede social, objetivando a generalização do modelo para que não se torne especialista de apenas uma plataforma virtual de acordo com a forma usual de publicações dela.

4.4. Treinamento de modelo usando os datasets escolhidos

Com o objetivo de treinar um modelo classificador de discurso de ódio, utilizamos um modelo DistilBERT [Devlin et al. 2019] pré-treinado, conforme os experimentos disponíveis no repositório oficial do dataset ToLD-BR (nota-se que o repositório do dataset HateBR não provém scripts de treino para seus modelos e não disponibiliza modelos baseados em transformers).

A primeira etapa foi verificar a qualidade dos datasets em relação ao modelo escolhido, verificando se a pesquisa estada na direção correta para efetuar os próximos testes mais complexos. Para isso o modelo escolhido foi avaliado em ambos datasets, sendo utilizado o método de validação cruzada (k-fold) usando 5 folds. Logo após, para iniciar as análises sobre a correlação dos dados das duas redes sociais, foi feita a união dos datasets e aplicado novamente o método de validação cruzada com o mesmo número de folds.

A partir disso, para entendermos melhor ainda se há uma verdadeira relação semântica entre os conjuntos de dados, foi conduzido um experimento onde o modelo foi treinado com o dataset ToLD-BR inteiro - sem nenhuma parcela de dados de avaliação - e avaliado no dataset HateBR por completo. Com isso também verificamos a capacidade de LLMs simples serem treinadas em pequenos datasets e serem estendidas para outros âmbitos, mantendo um bom comportamento de classificação.

Vale ressaltar novamente que seguimos a maiorida dos passos do novamente de acordo com o código oficial do ToLD-BR. Destarte em todos experimentos o modelo foi treinado por 1 epoch com batch size 32 usando uma placa de vídeo NVIDIA Titan XP, e como pré-processamento foram removidos hashtags, números, pontuações, acentos, links e stopwords de todas as sentenças, facilitando para o modelo adquirir informação apenas do que realmente importa e daquilo que mantém sentido quando retirado do contexto daquela plataforma virtual em específico.

5. Resultados

A tabela 7 apresenta os resultados de todos os experimentos citados no item anterior. Na avaliação usando um único dataset, é possível ver que o modelo apresenta melhores resultados no dataset HateBR, mesmo que as configurações sejam originalmente de acordo com o código oficial do ToLD-BR, o que evidencia um bom início na pesquisa e consolidando uma base forte para os passos seguintes.

Já a concatenação dos datasets apresenta resultados próximos aos obtidos com o ToLD-BR, sugerindo que os exemplos de treino provenientes de cada dataset não são muito dissoantes a ponto de confundir o modelo e fortalecendo a hipótese de extensibilidade do mesmo.

Por fim, o modelo treinado no dataset ToLD-BR e avaliado no dataset HateBR apresentou, como esperado, resultados relativamente piores, porém não tão distantes, sugerindo que é possível obter alguma generalização entre comentários de redes sociais diferentes. A piora foi justamente mais focada nos falsos negativos, o que intuitivamente já prova uma característica inerente do teste: o modelo, ao ser testado, entrará em contato com termos e construções sintáticas inéditas oriundas da diferença ressaltada anteriormente na cultura das diferentes plataformas virtuais.

Dataset	Avaliação	F1	Acurácia
ToLD-BR	5-Fold CV	70.83	71.95
HateBR	5-Fold CV	80.82	80.97
ToLD-BR+HateBR	5-Fold CV	71.38	72.10
Told-BR	HateBR	64.62	67.51

Tabela 7. Resultados da avaliação do modelo treinado em cada dataset. O símbolo + significa que os datasets foram concatenados

6. Conclusão

6.1. Recapitulação

Neste trabalho abordamos a questão geral do discurso de ódio em redes sociais e medidas preventivas à sua disseminação usando automatização por meio de modelos classificadores. Nesse escopo, levantamos a questão de que, embora existam conjuntos de dados textuais que classifiquem postagens como discurso de ódio ou não, os dados para treino de modelos são ainda relativamente escassos, em especial em português. Nossos testes em dois datasets distintos, envolvendo avaliação entre datasets e com uma mistura dos datasets levam a um entendimento que há um potencial em obter datasets ainda maiores ao concatená-los, dado que apesar de haver uma diferença na estrutura de textos em redes sociais distintas, discursos de ódio tendem a ser muito semelhantes.

6.2. Limitações

Apesar dos resultados surpreendentes, tudo foi fundamentado nos vieses dos anotadores de cada conjunto de dados, cujo critério para a rotulação carece de um embasamento concreto para definição de Discurso de Ódio. Dessa forma, ao compilar vários dados diferentes, estamos unindo possíveis inconsistências de interpretações que podem piorar o desempenho do modelo treinado.

Além disso, utilizamos um modelo baseado em BERT que é leve, porém útil o suficiente para fornecer resultados de pesquisa interessantes que provam os potenciais de LLMs nessas tarefas. Isso permitiu que os pesquisadores deste trabalho, mesmo sem acesso a máquinas com alto poder computacional, pudessem testar hipóteses e chegar nas conclusões descritas anteriormente.

Por fim, o curto tempo de desenvolvimento deste projeto não permitiu uma possível otimização dos hiperparâmetros de treinamento que poderiam trazer mais informações e interpretações sobre a qualidade dos dados perante o modelo usado.

6.3. Trabalhos Futuros

Como possíveis trabalhos futuros, é valido verificar se modelos de linguagens mais poderosos ainda - como GPT - fariam uma grande diferença nessa tarefa de criar um classificador genérico o suficiente para moderar diferentes plataformas virtuais, fazendo com que os custos muito mais altos façam sentido nesse cenário. Além disso, seria útil investigar se a união com mais datasets de outras redes sociais agregam tanto valor quanto os usados no trabalho presente, no intuito de generalizar ainda mais o conhecimento obtido no treinamento.

Referências

da Silva Oliveira, A. (2024). Detecção de discurso de ódio em língua portuguesa: uma análise do impacto da qualidade e diversidade de dados no desempenho de grandes modelos de linguagem. Master's thesis, Universidade Federal de Ouro Preto (UFOP), Ouro Preto, Brasil. Dissertação de Mestrado.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Firmino, A. A. (2022). *Uma abordagem para detecção de discurso de ódio utilizando aprendizado de máquina baseado em cruzamento de idiomas*. PhD thesis, Universidade Federal de Campina Grande, Campina Grande, PB, Brasil. Tese de Doutorado em Ciência da Computação.

Galinari, M. M. (2020). Identificando os "discursos de ódio": um olhar retórico-discursivo. *Revista de Estudos da Linguagem*.

Guide, B. F. (2022). Detecção automática de discurso de ódio punitivista em redes sociais. *Biblioteca Digital - USP*.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Wong, K.-F., Knight, K., and Wu, H., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2023). Exploring the limits of transfer learning with a unified text-to-text transformer.

Review, L. J. (2021). Journalists in brazil, mexico, argentina, and colombia combat misogynistic online discourse with the help of artificial intelligence. Accessed: 2025-01-06.

Silva, R. L. (2011). *Discursos de ódio em redes sociais: jurisprudência brasileira*. Revista Direito GV.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 7174–7183, Marseille, France. European Language Resources Association.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.