Detecção de Discurso de Ódio em Comentários de Redes Sociais

Lucas Zito	**Marcelo Pirro**	**Nicolas Benitiz**	**Rafael Campos**	**Rafael Penido**
UFSCar	UFSCar	UFSCar	UFSCar	UFSCar
Sorocaba - SP	Sorocaba - SP	Sorocaba - SP	Sorocaba - SP	Sorocaba - SP
802626	800510	813037	801968	802726

***Abstract*—Com o crescimento exponencial das redes sociais, tornou-se cada vez mais comum a disseminação de discursos de ódio em ambientes digitais. Esta realidade representa um desafio significativo tanto para plataformas de mídia quanto para a sociedade como um todo, exigindo soluções automatizadas eficazes para a identificação e mitigação de conteúdos ofensivos. Este trabalho propõe um sistema de detecção de discurso de ódio em língua portuguesa, utilizando uma abordagem multimodal capaz de analisar tanto texto quanto imagem. Diferentemente de abordagens tradicionais binárias, a detecção aqui é formulada como um problema multilabel, permitindo identificar diferentes categorias de discurso de ódio simultaneamente. Além disso, incorporam-se técnicas de explicabilidade para interpretar as decisões dos modelos, promovendo maior transparência e confiança. Os modelos foram avaliados sobre um conjunto de dados anotado e mostraram resultados promissores na identificação precisa e interpretável de conteúdos nocivos. Esta proposta visa contribuir para a construção de ambientes online mais seguros e para o avanço da moderação automatizada baseada em inteligência artificial.**

***Index Terms*—detecção de discurso de ódio; análise multimodal; aprendizado de máquina; classificação multilabel; explicabilidade; redes sociais; processamento de linguagem natural.**

I. Introdução

Com o crescimento das redes sociais, tornou-se cada vez mais comum a disseminação de discursos de ódio em ambientes digitais. Essa realidade impõe desafios importantes tanto para as plataformas quanto para a sociedade, exigindo soluções automatizadas eficazes para identificar e mitigar conteúdos ofensivos. A moderação manual, embora necessária, é limitada em escala e sujeita a interpretações subjetivas, o que reforça a importância de sistemas baseados em inteligência artificial.

O discurso de ódio assume diversas formas e pode ser dirigido a múltiplos grupos sociais, com linguagem que varia entre ofensas explícitas e manifestações sutis e ambíguas. Esse cenário se torna ainda mais complexo pelo uso de linguagem informal, abreviações, emojis e termos codificados característicos das redes sociais.

Nesse contexto, a detecção automática de discurso de ódio tornou-se um campo relevante no Processamento de Linguagem Natural (PLN), motivando o uso de modelos supervisionados, incluindo abordagens baseadas em regras e aprendizado profundo. No entanto, muitos estudos tratam o problema de forma binária (ofensivo ou não), o que limita a captura de nuances e sobreposições entre categorias.

Este trabalho propõe uma abordagem mais abrangente para a detecção de discurso de ódio em português, utilizando modelos de aprendizado profundo sobre um conjunto de dados multilabel, o *TuPyE-Dataset*, anotado com múltiplas categorias de discurso ofensivo.

Além de explorar abordagens tradicionais e contextuais (TF-IDF, Word2Vec e BERT), o estudo também investiga representações multimodais — combinando texto e imagem — para enriquecer a análise, e incorpora técnicas de explicabilidade para tornar as decisões do modelo mais interpretáveis. Assim, o trabalho se destaca por integrar três frentes complementares na moderação automatizada: **multimodalidade**, **classificação multilabel** e **explicabilidade**.

II. Trabalhos Relacionados

Diversas pesquisas recentes vêm explorando a detecção automática de discurso de ódio em língua portuguesa, com abordagens que vão desde modelos tradicionais até arquiteturas modernas de aprendizado profundo.

Silva & Roman (2020) analisaram algoritmos como Naïve Bayes, Regressão Logística e SVM em tweets, destacando o bom desempenho do SVM e do MLP frente a modelos recorrentes simples [1]. Azevedo (2025) utilizou embeddings GloVe com redes neurais profundas e votação em ensemble, obtendo F1 de 0.76 em textos ofensivos [2].

Com o avanço dos modelos baseados em transformadores, Leite et al. (2020) demonstraram que o BERTimbau, pré-treinado para o português, alcança macro-F1 de 0.76 em detecção de toxicidade [3].

Esses estudos fornecem a base sobre a qual este trabalho se apoia. Nossa proposta avança ao integrar modelos contextuais, classificação multilabel e análise multimodal, ampliando o alcance da detecção de discursos ofensivos em ambientes digitais.

III. Dados e Pré-processamento

A. Descrição dos Dados

Neste trabalho, utilizamos o **TuPyE-Dataset** [4], um corpus anotado voltado para tarefas de Processamento de Linguagem Natural (PLN) em língua portuguesa, com ênfase na detecção de discurso de ódio. O conjunto de dados é composto por **43.668 textos**, extraídos de diferentes plataformas de redes

sociais, principalmente Twitter e Instagram, com linguagem informal e características típicas da comunicação digital.

O dataset integra exemplos provenientes de quatro fontes principais:

- **Leite et al.**: 21.000 tweets.
- **TuPy**: 10.000 tweets adicionais com curadoria própria.
- **Vargas et al.**: 7.000 posts do Instagram.
- **Fortuna et al.**: 5.668 tweets.

Os dados são rotulados em três categorias:

- **Não-agressivos**: 31.121 exemplos.
- **Agressivos sem ódio**: 3.180 exemplos.
- **Agressivos com discurso de ódio**: 9.367 exemplos.

Além da diversidade de origem, o TuPyE-Dataset destaca-se por estar estruturado em formato binário compatível com bibliotecas de aprendizado de máquina modernas, como as da plataforma Hugging Face. O corpus já se encontra tokenizado e rotulado, o que facilita seu uso direto em tarefas de classificação textual.

A natureza multilabel do problema está refletida nos exemplos que podem conter sobreposição entre categorias de discurso ofensivo, viabilizando a investigação de abordagens mais sofisticadas para detecção simultânea de múltiplos tipos de discurso de ódio. O idioma predominante do corpus é o português brasileiro, em registros informais e espontâneos.

B. Análise exploratória

Para uma melhor compreensão do conteúdo do *TuPyE-Dataset*, foram realizadas análises sobre a distribuição lexical e a co-ocorrência de rótulos.

Nuvens de Palavras por Categoria

A Figura 1 mostra as palavras mais frequentes em quatro categorias: **aporophobia**, **body_shame**, **lgbtphobia** e **political**. Termos como "pobre", "gorda", "viado" e "comunista" ilustram o tipo de vocabulário ofensivo predominante em cada rótulo, evidenciando ataques direcionados à condição social, aparência física, identidade de gênero e posicionamento político.

Esses padrões mostram que cada tipo de discurso de ódio possui características léxicas distintas, o que pode ser explorado pelos modelos durante a classificação.

Fig. 1. Nuvens de palavras para as categorias *aporophobia*, *body_shame*, *lgbtphobia* e *political*.

Co-ocorrência de Rótulos

A Figura 2 apresenta a frequência de co-ocorrência entre os rótulos. Há forte associação entre **aggressive** e **hate**, e entre **misogyny**, **lgbtphobia** e **body_shame**, indicando que discursos ofensivos muitas vezes atacam múltiplos grupos simultaneamente.

Essa sobreposição reforça a natureza multilabel do problema, exigindo modelos que identifiquem relações entre categorias.

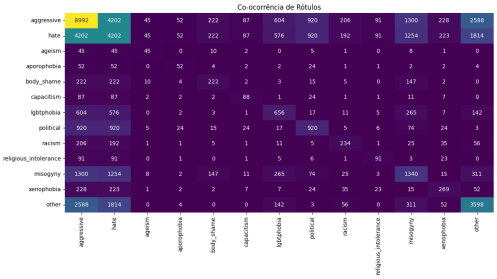

Fig. 2. Matriz de co-ocorrência entre rótulos do *TuPyE-Dataset*.

Essas análises reforçam a importância de estratégias robustas de representação textual e aprendizado, capazes de capturar tanto padrões lexicais quanto correlações entre rótulos.

C. Pré-processamento

No desenvolvimento deste trabalho, diferentes estratégias de pré-processamento textual foram avaliadas para adequar os dados aos modelos TF-IDF, Word2Vec e BERT. O pré-processamento incluiu etapas como conversão para minúsculas, remoção de URLs, menções, hashtags, emojis e números, além da substituição de entidades nomeadas por categorias genéricas. Essas transformações visaram uniformizar o corpus e eliminar ruídos, facilitando a generalização dos modelos.

Destaca-se também a unificação de expressões multipalavras (MWEs), como insultos compostos, e a expansão de abreviações e contrações, que ajudaram a preservar significados importantes e reduzir a dispersão do vocabulário. Para TF-IDF e Word2Vec, a remoção de stopwords, lematização e filtragem por classe gramatical mostraram-se eficazes, pois reduzem o vocabulário a termos mais relevantes. Já para o BERT, optou-se por manter o texto mais próximo do original, preservando stopwords e pontuação, pois o modelo depende do contexto completo para gerar boas representações.

A escolha das transformações foi guiada tanto por recomendações da literatura quanto por testes empíricos. Observou-se que, enquanto a simplificação do texto favorece modelos baseados em contagem, a preservação do contexto é essencial para modelos contextuais como o BERT. Assim, o pré-processamento foi adaptado conforme o modelo, impactando diretamente a performance dos classificadores e a robustez na identificação de discurso de ódio.

IV. Estratégia com Regras

A solução baseada em regras foi concebida como uma abordagem determinística para a detecção multilabel de discurso

de ódio, utilizando o TuPyE-Dataset. O dataset foi dividido em dois subconjuntos: um para construir dicionários de palavras-chave por categoria (cerca de 32 mil exemplos) e outro para avaliação (aproximadamente 8 mil exemplos). O objetivo era investigar até que ponto listas de termos discriminativos poderiam aproximar os resultados reais do dataset, servindo como baseline e fonte de recursos para modelos posteriores.

O sistema começa com o pré-processamento dos textos, removendo menções, convertendo escrita hacker, normalizando para minúsculas e sem acentos, e mantendo apenas letras, números, espaços e emojis. Essa padronização facilita a correspondência exata com os termos dos dicionários de categorias. O núcleo da abordagem está na construção de dicionários expandidos, associando cada categoria (como racismo, misoginia, lgbtqfobia, etc.) a listas de palavras e expressões representativas, extraídas do próprio dataset. Um conjunto de emojis agressivos também foi criado para reforçar a detecção de agressividade.

A classificação é feita pela função detectar_categorias, que verifica, para cada texto, a presença de termos dos dicionários em suas formas originais e variantes em escrita hacker. A presença de emojis agressivos pode indicar ou reforçar a categoria ”aggressive”. O resultado é um vetor multilabel para cada texto, acompanhado de uma pontuação de agressividade e uma classificação qualitativa baseada em limiares simples.

A avaliação utiliza métricas padrão multilabel, como F1-score, Hamming Loss e Subset Accuracy. Os resultados mostram que, embora a abordagem identifique muitos casos explícitos, sua precisão depende fortemente da abrangência dos dicionários e não interpreta contexto, ironia ou polissemia. Por exemplo, a palavra ”macaco” sempre resulta em ”racismo”, mesmo fora de contexto, gerando falsos positivos.

Apesar de transparente e útil como baseline, a solução baseada em regras é limitada: não reconhece termos fora dos dicionários, não lida com ambiguidades ou ironias, e sua generalização é restrita. Ainda assim, representa uma estratégia inicial robusta, fornecendo recursos lexicais valiosos e ressaltando a necessidade de métodos mais sofisticados para lidar com a complexidade da linguagem natural.

V. Estratégia com Aprendizado

Para a tarefa de detecção de discurso de ódio multilabel, foram exploradas três abordagens principais de representação textual: *TF-IDF*, *Word2Vec* e *BERT*. Cada uma dessas abordagens foi combinada com diferentes algoritmos de aprendizado supervisionado, com o objetivo de avaliar seus desempenhos e limitações em um cenário de dados desbalanceados e linguagem informal.

A. TF-IDF com Classificadores Tradicionais

A primeira abordagem consistiu na vetorização dos textos utilizando a técnica de *Term Frequency-Inverse Document Frequency (TF-IDF)*, que representa cada comentário como um vetor esparso com base na frequência relativa de palavras. A seguir, foram testados três modelos de classificação:

- **Regressão Logística**
- **Naïve Bayes**
- **Random Forest**

Os resultados demonstraram que o modelo **Naïve Bayes** teve desempenho inferior aos demais, apresentando baixa capacidade de generalização. A **Regressão Logística** e a **Random Forest** apresentaram resultados semelhantes, com leve vantagem para esta última. No entanto, em termos gerais, os modelos baseados em TF-IDF alcançaram apenas resultados modestos, com F1-Score em torno de 0.4.

B. Word2Vec

A segunda abordagem utilizou representações densas a partir de embeddings pré-treinados do *Word2Vec*, considerando a média dos vetores de cada palavra em um comentário. Como o *Word2Vec* não considera o contexto em que as palavras aparecem, era esperado que seu desempenho fosse limitado, especialmente em uma tarefa sutil como a detecção de discurso de ódio.

De fato, os resultados obtidos foram comparáveis aos da abordagem com TF-IDF, sem ganhos significativos. Apesar de sua natureza distribuída, a falta de sensibilidade ao contexto reduziu sua eficácia na tarefa multilabel.

C. BERT

A terceira e mais promissora abordagem empregou o modelo *BERTimbau Base*, uma versão do BERT pré-treinada para a língua portuguesa. Esse modelo é capaz de capturar o contexto completo das palavras em uma sentença, oferecendo representações semânticas muito mais robustas.

Nos testes iniciais, o modelo BERT apresentou desempenho consideravelmente superior às abordagens anteriores, com F1-Score chegando a 0.63, mesmo sem ajustes específicos para o desbalanceamento do dataset. Isso evidencia sua capacidade de lidar com nuances linguísticas e múltiplas categorias simultâneas.

D. Tratamento do Desbalanceamento

Dado o forte desbalanceamento entre classes presentes no *TuPyE-Dataset*, principalmente com predominância da classe *neutral*, foram testadas diferentes estratégias para mitigar esse efeito durante o treinamento com BERT:

- **Pesos na Função de Perda**: atribuição de pesos maiores às classes minoritárias, de forma proporcional à frequência inversa de cada rótulo.
- **Focal Loss**: substituição da função de perda tradicional (BCE) por *Focal Loss*, que penaliza mais fortemente os erros em exemplos difíceis, sendo especialmente útil em cenários desbalanceados.
- **Undersampling da Classe Majoritária**: redução do número de exemplos da classe predominante (*neutral*), visando equilibrar a distribuição dos rótulos durante o treinamento.

Essas estratégias foram avaliadas experimentalmente, e seus impactos nos resultados finais serão discutidos na seção de métricas. A expectativa é que essas técnicas contribuam para a melhoria do desempenho nas classes menos representadas, sem comprometer a precisão geral do modelo.

E. Abordagem Multimodal

Além das estratégias puramente textuais, foi conduzido um experimento exploratório utilizando uma abordagem multimodal, com o objetivo de incorporar informações visuais às representações linguísticas. Para isso, utilizaram-se dois codificadores: o **BERTimbau** para o texto e o **ViT-B/32 (Vision Transformer)** para imagens.

Os embeddings gerados por ambos os encoders foram concatenados, formando uma única representação multimodal para cada comentário. Essa representação final foi então passada por uma **rede neural totalmente conectada**, composta por duas camadas densas, seguidas por uma *função de ativação ReLU*, uma camada de *Dropout* e, por fim, uma camada sigmoidal para gerar as saídas multilabel.

O **Dropout** foi inserido como uma técnica de *regularização*, com o objetivo de reduzir o risco de overfitting — uma preocupação relevante dado o tamanho relativamente reduzido do dataset, especialmente na dimensão visual. Essa arquitetura buscou integrar de forma eficiente os dois tipos de entrada, permitindo ao modelo capturar correlações entre o conteúdo textual e elementos visuais que podem reforçar ou complementar o discurso ofensivo.

VI. Análise e Resultados

Para avaliar o desempenho das diferentes abordagens de detecção de discurso de ódio, foram conduzidos experimentos com múltiplas estratégias de representação e classificação. Os resultados são apresentados em termos de F1-Score médio, considerando a tarefa multilabel. A seguir, descrevemos e comparamos os principais cenários testados.

A. TF-IDF com Classificadores Tradicionais

Na primeira abordagem, os textos foram vetorizados com TF-IDF, técnica que representa as palavras com base em sua frequência relativa no corpus. Essa representação foi combinada com três algoritmos de aprendizado supervisionado: Regressão Logística, Naive Bayes e Random Forest. Além disso foi feito um modelo base, sem nenhum pré-processamento, nele foi utilizada Regressão Logística, apenas para ter como base para comparar com os outros.

O modelo Naive Bayes teve desempenho significativamente inferior, indicando baixa capacidade de capturar a complexidade da tarefa. A Regressão Logística apresentou resultados razoáveis, e a Random Forest superou levemente os demais, destacando-se como a melhor dentre os algoritmos tradicionais com TF-IDF. Ainda assim, os valores de F1-Score permaneceram modestos, abaixo de 0.41, o que reforça as limitações dessas técnicas para tarefas multilabel com linguagem informal e ruído textual.

Modelo	**F1-Score**
Modelo base (Regressão Logística)	0.3912
Regressão Logística	0.3523
Naive Bayes	0.0855
Random Forest	0.4046

TABLE I
F1-Score utilizando TF-IDF com diferentes classificadores.

B. Word2Vec

Na segunda abordagem, os textos foram representados por meio de embeddings pré-treinados do Word2Vec, onde cada comentário foi vetorizado como a média dos vetores de suas palavras. Por não considerar o contexto de uso das palavras, esperava-se um desempenho inferior.

Essa expectativa foi confirmada nos experimentos: o modelo de Regressão Logística treinado com Word2Vec apresentou F1-Score de apenas 0.0412, o que revela a limitação dessa técnica para a tarefa proposta, especialmente em comparação com modelos baseados em contexto.

Modelo	**F1-Score**
Regressão Logística	0.0412

TABLE II
F1-Score utilizando Word2Vec.

C. BERT

A terceira abordagem utilizou o modelo BERTimbau Base, um transformador pré-treinado em português. Essa arquitetura permite capturar relações contextuais profundas entre as palavras, o que é especialmente útil para detectar discursos de ódio mais sutis e com sobreposição de categorias.

Inicialmente, o modelo foi treinado com a função *BCEWithLogitsLoss*. Para mitigar o forte desbalanceamento entre os rótulos, foram testadas estratégias complementares, como a utilização de *class weights* e o *undersampling* da classe majoritária. Esses ajustes contribuíram significativamente para a melhora do desempenho, elevando o F1-Score de 0.5895 para 0.6308.

Configuração do Modelo BERT	**F1-Score**
BCEWithLogitsLoss	0.5895
BCEWithLogitsLoss + Class Weights	0.6205
BCEWithLogitsLoss + Class Weights + Undersampling	0.6308

TABLE III
F1-Score utilizando BERT com diferentes estratégias de balanceamento.

D. Avaliação da Explicabilidade

Com o objetivo de aumentar a transparência e a auditabilidade do modelo, aplicamos técnicas de explicabilidade para interpretar suas decisões. Neste trabalho, utilizamos o **LIME** (Local Interpretable Model-Agnostic Explanations), uma abordagem *pós-hoc*, ou seja, aplicada após o treinamento do modelo. Essa técnica busca explicar predições individuais ao gerar aproximações lineares locais do comportamento do classificador, permitindo a identificação dos termos mais influentes em cada decisão.

Para avaliar a qualidade das explicações geradas pelo LIME, foi utilizada a métrica de **overlap**. Essa métrica quantifica a interseção entre as palavras destacadas automaticamente

pelo LIME e aquelas identificadas manualmente pelos autores como as mais relevantes para justificar a classificação de determinado comentário. Em outras palavras, o overlap mede o grau de concordância entre a explicação gerada pelo modelo e a percepção humana sobre quais palavras realmente carregam valor discriminativo em cada exemplo.

Nos experimentos realizados, o valor médio de overlap obtido foi de **63%**, indicando uma boa correspondência entre as decisões do modelo e as interpretações humanas em uma proporção significativa dos casos.

Esse processo envolveu uma análise manual detalhada, na qual um conjunto de exemplos foi revisado para identificar, por inspeção dos autores do relatório, os termos centrais responsáveis por expressar o discurso de ódio ou sua categoria específica. Em seguida, essas anotações foram comparadas com as palavras mais importantes atribuídas pelo LIME para o mesmo exemplo.

Essa análise mista, combinando explicabilidade automática e validação humana, demonstrou que o modelo não apenas alcançou bons resultados quantitativos, mas também operou com decisões interpretáveis e alinhadas com o senso comum, o que é essencial para aplicações sensíveis como a moderação de conteúdo em redes sociais.

E. Análise Qualitativa de Exemplo

Para complementar os resultados quantitativos apresentados anteriormente, foi realizada uma análise qualitativa de predições individuais com o objetivo de avaliar a coerência do modelo em situações reais.

Um exemplo analisado foi o comentário *"viadinho de merda"*. O modelo classificou esse comentário com probabilidade superior a 50% para as categorias **aggressive**, **hate** e **lgbtphobia**, o que gerou sua rotulação multilabel final nessas três classes.

Essa predição é coerente com a natureza do conteúdo: trata-se de um xingamento com forte conotação ofensiva, direcionado à identidade sexual, utilizando um termo pejorativo comumente associado à comunidade LGBTQIA+. A inclusão simultânea dos rótulos *hate* e *lgbtphobia* reforça que o modelo foi capaz de reconhecer tanto o caráter discriminatório do termo quanto sua motivação homofóbica. O rótulo *aggressive* também é justificado pela linguagem explícita e pelo uso de palavrões, que caracterizam agressividade verbal.

Esse exemplo ilustra a capacidade do modelo em lidar com sobreposições entre categorias de discurso de ódio, uma característica essencial em tarefas multilabel, além de demonstrar um entendimento adequado das nuances da linguagem ofensiva no contexto informal e coloquial das redes sociais.

F. Abordagem Multimodal

A Tabela IV apresenta os resultados finais da abordagem multimodal, ela foi realizada em cima do dataset HateBR (que está contido no TuPyE). Essa abordagem combinou representações textuais geradas pelo BERT com representações visuais extraídas por meio do ViT-B/32. O modelo obteve uma acurácia geral de 88,1%, com F1-Score equilibrado de 0.881 para ambas as classes (discurso de ódio e não-discurso).

Esses resultados indicam que a integração de informações visuais e textuais contribuiu positivamente para a capacidade discriminativa do sistema, possivelmente capturando sinais adicionais presentes nas imagens que reforçam ou complementam o conteúdo ofensivo dos comentários.

Classe	Precisão	Revocação	F1-Score	Suporte
0.0 (não-discurso de ódio)	0.871	0.890	0.881	693
1.0 (discurso de ódio)	0.890	0.871	0.881	707
Acurácia			0.881	1400
Média macro	0.881	0.881	0.881	1400
Média ponderada	0.881	0.881	0.881	1400

TABLE IV
RESULTADOS DA ABORDAGEM MULTIMODAL (TEXTO + IMAGEM) SOBRE O CONJUNTO DE TESTE.

G. Resumo Comparativo

Os resultados indicam que, embora abordagens tradicionais com TF-IDF tenham fornecido um ponto de partida razoável, o modelo BERT se destacou amplamente em termos de desempenho. As estratégias de balanceamento também se mostraram eficazes, elevando a capacidade do modelo de lidar com classes minoritárias, o que é essencial em contextos reais de moderação de conteúdo.

VII. DIFICULDADES E OBSTÁCULOS ENCONTRADOS

Durante o desenvolvimento do trabalho, foram identificadas diversas dificuldades técnicas e conceituais que impactaram diretamente na modelagem, pré-processamento e avaliação do sistema proposto. A seguir, listamos os principais obstáculos enfrentados:

1) **Ambiguidade e Subjetividade Linguística:** Muitos comentários apresentam ambiguidades, sarcasmo ou ironia, o que dificulta a interpretação automática. Uma mesma frase pode ser interpretada como ofensiva ou neutra dependendo do contexto, do tom ou da intenção do autor.
2) **Desbalanceamento das Classes:** Algumas categorias, como *racism* e *misogyny*, estão sobrerrepresentadas no dataset, enquanto outras, como *aporophobia* e *capacitism*, possuem poucos exemplos. Esse desbalanceamento prejudica o desempenho dos modelos em classes minoritárias, exigindo técnicas específicas de balanceamento ou amostragem.
3) **Sobreposição de Rótulos (Overlapping):** Por se tratar de uma tarefa multilabel, é comum que uma mesma amostra contenha múltiplos tipos de discurso de ódio simultaneamente. Isso demanda que os modelos sejam capazes de capturar correlações entre rótulos, o que torna a modelagem mais complexa.
4) **Pré-processamento de Linguagem Informal:** Muitos textos utilizam linguagem informal com variações criativas, como escrita "hacker" (ex: *M4t4*), abreviações, emojis e menções a usuários (@user). Isso exige estratégias específicas de normalização que preservem

o significado ofensivo original, sem comprometer a semântica do texto.

5) **Limitações da Detecção Baseada em Regras:** Embora a abordagem baseada em regras tenha se mostrado eficaz para detectar expressões explícitas, sua cobertura e precisão foram limitadas em casos mais sutis ou implícitos. Isso reforça a importância da complementação com modelos de aprendizado supervisionado.
6) **Interpretação dos Resultados:** Apesar do uso de técnicas de explicabilidade, como LIME, nem sempre as palavras-chave destacadas explicam de forma satisfatória o motivo de determinada predição. Essa limitação compromete a auditabilidade dos modelos e a confiança necessária em aplicações sensíveis.
7) **Dificuldades com Dados Multimodais:** Ao explorar técnicas multimodais, especialmente em comentários provenientes do Instagram, foi identificado que diversas postagens associadas a figuras públicas haviam sido removidas da plataforma. Isso resultou na perda de acesso às imagens correspondentes, dificultando a análise multimodal e reduzindo o tamanho efetivo da base de dados visual.

VIII. Estratégia Final

Ao longo do desenvolvimento do projeto, foram conduzidos diversos experimentos com diferentes combinações de pré-processamento textual, vetorização e algoritmos de aprendizado. A escolha da configuração final considerou não apenas o desempenho dos modelos, mas também a robustez da abordagem frente às características específicas do *TuPyE-Dataset*, como a linguagem informal e o desbalanceamento entre as classes.

Após a fase exploratória, definiu-se uma configuração de pré-processamento mais eficiente, que incluiu as seguintes etapas:

- Conversão de todo o texto para letras minúsculas.
- Remoção de acentos, pontuações, números, URLs, emojis e menções/hashtags.
- Expansão de abreviações comuns.
- Normalização de risadas.
- Remoção de palavras irrelevantes (stopwords).
- Aplicação de lematização.
- Filtro de tokens com comprimento mínimo de dois caracteres.
- Combinação de expressões multi-palavra (multiword expressions).

Essa configuração se mostrou eficaz para limpar o texto sem perder o conteúdo semântico necessário para a classificação de discursos de ódio.

Como estratégia de modelagem final, optou-se pela utilização do modelo **BERTimbau Base**, implementado por meio da biblioteca *Transformers* da Hugging Face. Esse modelo foi treinado com a função de perda **Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss)**, ajustada com *class weights* para compensar o forte desbalanceamento entre os rótulos do dataset.

A adoção dessa configuração buscou maximizar o desempenho preditivo do modelo, especialmente nas classes menos representadas, garantindo maior sensibilidade para os diversos tipos de discurso de ódio presentes nos dados.

IX. Conclusão

Neste trabalho, foi proposta e avaliada uma abordagem para a detecção de discurso de ódio em comentários de redes sociais em língua portuguesa, considerando o problema como uma tarefa multilabel. Foram exploradas diversas estratégias de representação textual, desde métodos baseados em frequência (TF-IDF), até embeddings semânticos (Word2Vec) e modelos de linguagem baseados em contexto (BERT).

Os resultados demonstraram que os modelos tradicionais apresentaram desempenho limitado, especialmente frente à complexidade da linguagem presente em redes sociais. Por outro lado, o uso do BERTimbau mostrou ganhos expressivos em F1-Score, especialmente quando combinado a estratégias de balanceamento de classes, como o uso de *class weights* e *undersampling*.

Adicionalmente, foi investigada uma abordagem multimodal em cima de um dataset simplificado, combinando representações visuais (ViT-B/32) com textuais (BERT), integradas em uma arquitetura neural com camadas densas e dropout.

Foram ainda incorporadas técnicas de explicabilidade, como o uso de LIME, com o objetivo de interpretar e validar as decisões do modelo. Apesar de algumas limitações observadas na qualidade das explicações geradas, essa etapa contribuiu para maior transparência do sistema proposto.

Em suma, o trabalho apresentou uma solução robusta e atualizada para o desafio da detecção de discurso de ódio online, integrando elementos fundamentais como **classificação multilabel**, **abordagem multimodal** e **explicabilidade**. Os resultados obtidos reforçam o potencial de modelos contextuais e multimodais na construção de ferramentas automatizadas para moderação de conteúdo e promoção de ambientes digitais mais seguros.

References

[1] A. Silva and N. Roman, "Hate speech detection in Portuguese with Naïve Bayes, SVM, MLP and Logistic Regression," in *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, Porto Alegre, Brasil, 2020, pp. 1–12. doi:10.5753/eniac.2020.12112.

[2] F. C. M. Azevedo, "Reconhecimento de padrões na detecção do discurso de ódio: uma abordagem de redes neurais e ensembles," *Research, Society and Development*, vol. 14, no. 5, p. e10814548633, May 2025.

[3] J. A. Leite, D. F. Silva, K. Bontcheva, and C. Scarton, "Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis," CoRR, abs/2010.04543, 2020.

[4] Silly-Machine, "TuPyE-Dataset: A Binary-Formatted Corpus of Brazilian Online Comments for NLP Tasks," Hugging Face, 2023. Disponível em: https://huggingface.co/datasets/Silly-Machine/TuPyE-Dataset