# Uma Abordagem Integrada para Detecção de Discurso de Ódio em Mídias Sociais Utilizando Vetorização de Textos e Emojis

**Arthur Lima de Araújo Miranda**[1], **Cleyton Mário de Oliveira Rodrigues**[1]

[1]Programa de Pós-Graduação em Engenharia de Computação (PPGEC) – Universidade de Pernambuco (UPE) – Recife – PE – Brazil – CEP: 50720-001

***Abstract.*** *This paper proposes an integrated approach for hate speech detection in social media, combining three main dimensions: (1) fusion of Brazilian Portuguese datasets (HateBR and TuPy-E), (2) joint processing of texts and emojis, and (3) a two-stage classification architecture (binary and multiclass). Using the BERTimbau model adapted to capture semantic relations and emoji representations, the system first performs binary classification (hate vs non-hate) followed by specific categorization (xenophobia, gender/sexuality, etc.). Results achieved 85% accuracy in the binary stage and up to 86% in specific categories. We discuss the relationship between data volume and performance, as well as future strategies for model improvement, including the use of LLMs (Large Language Models) and metadata integration.*

***Resumo.*** *Este artigo propõe uma abordagem integrada para detecção de discurso de ódio em mídias sociais, combinando três dimensões principais: (1) fusão de datasets em português brasileiro (HateBR e TuPy-E), (2) processamento conjunto de textos e emojis, e (3) arquitetura classificatória em duas etapas (binária e multiclasse). Utilizando o modelo BERTimbau adaptado para capturar relações semânticas e representações de emojis, o sistema realiza primeiro uma classificação binária (ódio vs não-ódio) seguida de categorização específica (Xenofobia, gênero/sexualidade, etc). Os resultados alcançaram 85% de acurácia na etapa binária e até 86% em categorias específicas. Discute-se a relação entre volume de dados e desempenho, bem como estratégias futuras para aprimoramento do modelo, incluindo o uso de LLMs (Large Language Models) e integração de metadados.*

## 1. Introdução

Dentro da Computação Jurídica, um problema de bastante destaque é o uso de técnicas de mineração de textos para classificação de discursos de ódio. Combater o discurso de ódio não significa limitar ou proibir a liberdade de expressão. Significa evitar que o discurso de ódio se transforme em algo mais perigoso, particularmente o incitamento à discriminação, hostilidade e violência, que é proibido pelo direito internacional (UN BR, 2023). Com o acesso às redes sociais cada vez mais amplo, escondido por trás do falso senso de anonimato, de apitos de cachorro (do inglês "*dog whistles*") e, em conjunto com moderações cada vez menos presentes e mais automatizadas, cresce também a presença do discurso de ódio difundido nessas mídias. Crimes como Xenofobia, Racismo, Homofobia, Sexismo, entre outros estão ficando cada vez mais comuns no mundo digital. Assim, o discurso de ódio se tornou um dos métodos mais frequentes para difundir mentiras e desinformação,

online e offline, ameaçando a paz, o entendimento e o diálogo entre as pessoas e nações, e o progresso rumo ao desenvolvimento sustentável (UN BR, 2024).

Embora não seja um problema apenas brasileiro, o país tem sofrido uma alta cada vez mais notável desses ataques. Dados divulgados por empresas como a Deep Digital LLYC e por ONGs como a Safernet demonstram como, não só crimes envolvendo discurso de ódio nas mídias sociais triplicaram nos últimos anos no Brasil, como ele também lidera o volume de ataques à comunidade LGBTQIAP+, somando cerca de 37.67% (JORNAL NACIONAL, 2023).

Legalmente falando, o Brasil não possui uma legislação específica que criminalize o discurso de ódio. No entanto, existem leis que podem ser aplicadas em casos de crimes motivados por preconceito ou discriminação, como a Lei nº 7.716/1989, que trata dos crimes resultantes de preconceito de raça ou cor, ou a Lei 9.459/1997, que trata dos crimes de Xenofobia. As mídias sociais são, sem dúvidas, os principais canais utilizados pelos grupos propagadores desse discurso. Sugere-se, portanto, que algoritmos computacionais inteligentes podem ajudar na identificação automática dos discursos de ódio.

No contexto de algoritmos inteligentes, a Mineração de Textos é uma área interdisciplinar (envolve a Inteligência Artificial, o PLN, a Estatística, entre outras áreas) para que se possa garimpar informações textuais em busca de informações úteis para tomadas de decisões. Geralmente, envolve buscas quantitativas e qualitativas em grandes volumes de textos (muitas vezes, não-estruturados) na tentativa de identificar padrões ou tendências no texto escrito. Para a análise textual, Ebecken et al. (2003) discorre sobre algumas estratégias, como a análise estatística, que se baseia na frequência/importância dos termos presentes no texto, ou seja, os documentos são vistos através de um bag of words. Contudo, técnicas mais avançadas têm surgido para superar as limitações dessa abordagem, como o TF-IDF (Term Frequency-Inverse Document Frequency), que pondera a ocorrência dos termos não apenas pela frequência, mas também pela relevância no corpus, reduzindo o peso de palavras comuns e destacando termos discriminativos (AIZAWA, 2003). Além disso, métodos baseados em *embeddings*, como o Word2Vec, permitem capturar relações semânticas e contextuais entre palavras, representando-as como vetores densos em um espaço multidimensional, o que facilita a identificação de nuances e expressões codificadas (MIKOLOV et al., 2013).

Apesar dos avanços recentes em técnicas de PLN, a detecção de discurso de ódio ainda enfrenta desafios significativos. Um dos principais obstáculos é a natureza contextual e subjetiva do discurso de ódio, que pode variar de acordo com o contexto cultural, social e histórico. Além disso, a evolução constante das linguagens e o uso de eufemismos, gírias e expressões codificadas dificultam a identificação precisa desses discursos. Portanto, é essencial que os modelos de detecção sejam continuamente atualizados e treinados com dados diversificados e representativos, a fim de garantir uma classificação mais precisa e justa.

Este trabalho apresenta uma abordagem integrada para detecção de discurso de ódio em mídias sociais, utilizando técnicas de PLN com o modelo BERTimbau, pré-treinado especificamente para o português brasileiro. A solução proposta combina *datasets* em português brasileiro (HateBR e TuPy-E) para treinar classificadores binários (ódio vs não-ódio) e multiclasse (categorias específicas, como racismo, homofobia e xe-

nofobia). Além disso, discute-se a relação entre volume de dados e desempenho, bem como estratégias futuras para aprimoramento do modelo, incluindo o uso de LLMs (*Large Language Models*) e integração de metadados.

Este documento está estruturado da seguinte maneira: a Seção 2 descreve trabalhos relacionados, destacando suas contribuições e limitações. A Seção 3 detalha a metodologia proposta, incluindo a preparação dos dados, o treinamento do modelo e as técnicas de avaliação. Em seguida, a Seção 4 apresenta e analisa os resultados obtidos, comparando-os com abordagens existentes. Por fim, a Seção 5 conclui o trabalho, discutindo implicações práticas e sugerindo direções futuras para pesquisa.

## 2. Trabalhos Relacionados

A detecção automática de discurso de ódio em mídias sociais tem sido amplamente explorada na literatura, com abordagens variando desde técnicas clássicas de aprendizado de máquina até modelos profundos baseados em redes neurais. Esta seção destaca estudos fundamentais citados no artigo de Nascimento et al. (2023), contextualizando suas contribuições e limitações em relação à proposta deste trabalho, além de incorporar *insights* do corpus TuPy-E, do *dataset* HateBR e do modelo ABMM para árabe (Almaliki et al., 2023).

### 2.1. Definições e Bases Teóricas

Nascimento et al. (2023) revisam definições de discurso de ódio adotadas por plataformas como Facebook, Twitter e YouTube, além de propostas acadêmicas. Cohen-Almagor (2013) define discurso de ódio como ”linguagem hostil e maliciosa motivada por preconceito contra características inatas de indivíduos ou grupos”. Já Davidson et al. (2017) focam em linguagem que ”humilha, insulta ou expressa ódio contra grupos específicos”. Fortuna e Nunes (2019) ampliam o escopo ao incluir formas sutis, como uso de humor para reforçar estereótipos. Essas definições orientaram a criação de conjuntos de dados e a seleção de características em estudos subsequentes, mas a subjetividade inerente ao tema permanece um desafio crítico.

### 2.2. Conjuntos de Dados e Metodologias

Diversos conjuntos de dados públicos foram desenvolvidos para treinar modelos de detecção. Waseem e Hovy (2016) criaram um corpus em inglês com 16.914 tweets categorizados como sexistas, racistas ou neutros, enquanto Davidson et al. (2017) propuseram um *datasets* de 24.802 tweets em inglês rotulados como ”ódio”, ”ofensivo”ou ”neutro”. Para línguas além do inglês, Almaliki et al. (2023) introduziram o ABMM, um modelo BERT-mini para detecção de discurso de ódio em árabe, alcançando 98.6% de acurácia em tweets categorizados como ”normal”, ”abusivo”e ”ódio”. No contexto do português brasileiro, destacam-se o HateBR (Vargas et al., 2022), com 7.000 comentários do Instagram, e o TuPy-E (Oliveira et al., 2023), que combina múltiplas fontes, incluindo tweets e dados do ToLD-Br. Entretanto, como destacado por Poletto et al. (2021), a escassez de dados publicamente disponíveis em idiomas como o português brasileiro limita avanços nesses contextos, problema também observado no ABMM para dialetos árabes minoritários.

### 2.3. Técnicas de Extração de Características

Estudos pioneiros utilizaram abordagens baseadas em dicionários (Gitari et al., 2015) e métricas léxicas, como a distância de Levenshtein para identificar termos ofensivos

obfuscados (Nandhini e Sheeba, 2015). Modelos clássicos como *Bag-of-Words* (BoW) e *n-grams* foram amplamente adotados (Burnap e Williams, 2016; Watanabe et al., 2018), porém enfrentam limitações em capturar contexto semântico. Para superar isso, Nobata et al. (2016) combinaram *word2vec* com características sintáticas (POS tags), alcançando F1-score de 0.79-0.81 em comentários do Yahoo!.

O advento de modelos de *deep learning* trouxe avanços significativos. Zhang et al. (2018) propuseram uma arquitetura híbrida CNN-LSTM, enquanto enquanto Pitsilis et al. (2018) utilizaram LSTMs com metadados de usuários para detectar mensagens sexistas (F1=0.99). Além disso, Almaliki et al. (2023) aplicaram uma versão compacta do BERT (BERT-mini) para árabe, demonstrando eficácia em cenários de dados limitados. Recentemente, modelos baseados em BERT têm sido explorados: del Arco et al. (2021) compararam *embeddings* pré-treinados em espanhol, e Cao et al. (2020) desenvolveram o *DeepHate*, um *framework* multimodal que combina LDA e LSTM. Contudo, como observado por MacAvaney et al. (2019), a dependência excessiva de palavras-chave explícitas e o viés em conjuntos de dados ainda prejudicam a generalização.

### 2.4. Desafios e Lacunas

Nascimento et al. (2023) identificam desafios críticos: (1) **Viés em *datasets***, como no corpus de Waseem e Hovy (2016), onde 1.972 tweets racistas foram gerados por apenas 9 usuários; (2) **Polissemia**, onde termos ambíguos dificultam a interpretação (Senarath e Purohit, 2020); e (3) **Generalização intercultural**, já que modelos treinados em um idioma (ex: inglês) performam mal em outros (Kumar et al., 2019). Além disso, poucos estudos exploraram cenários multilíngues ou integraram metadados contextuais de forma eficaz (Ribeiro et al., 2018). O trabalho de Almaliki et al. (2023) reforça a importância de adaptações linguísticas, como a normalização de caracteres árabes, desafio análogo à variação de dialético no português brasileiro.

### 2.5. Abordagens em Português Brasileiro

No contexto do português brasileiro, destaca-se o corpus HateBR (Vargas et al., 2022), composto por 7.000 comentários do Instagram anotados manualmente por especialistas em três camadas: classificação binária (ofensivo/não ofensivo), nível de ofensividade (alta, moderada, baixa) e nove categorias de ódio (xenofobia, racismo, homofobia, entre outras). Os experimentos com n-grams e TF-IDF alcançaram 85% de F1-score, superando trabalhos anteriores em português. Entretanto, como apontado pelos autores, a concentração de dados em poucos usuários (ex: 9 usuários geraram 1.972 comentários racistas) introduz viés, limitando a generalização.

Recentemente, o TuPy-E (Oliveira et al., 2023), emergiu como o maior corpus aberto para português, combinando dados diversas bases disponíveis e um novo conjunto de tweets. Utilizando o BERTimbau (pré-treinado para português), o modelo obteve 90.3% de F1-score na classificação binária e 86% na multiclasse, com desempenho superior em categorias como misoginia (F1=0.65) e LGBTfobia (F1=0.76). Esses resultados alinham-se com as estratégias do ABMM para árabe, que emprega normalização léxica e tokenização adaptada para dialetos regionais.

### 2.6. Contribuição deste Trabalho

Este trabalho avança ao propor uma abordagem integrada para o português brasileiro, utilizando o *dataset* TuPy-E juntamente com o HateBR e inspirando-se em técnicas do

ABMM para lidar com variações linguísticas. Foram criados modelos especialistas: um para detecção binária de discurso de ódio e outros para categorias específicas (xenofobia e questões de gênero/sexualidade). A detecção é realizada sequencialmente, onde cada texto é processado por classificadores dedicados, garantindo maior precisão em subtarefas. Essa arquitetura, semelhante à abordagem multicamada do ABMM, enfrenta desafios em categorias com poucos dados, resultando em viés de classificação. Para mitigar esse problema. Diferentemente de estudos anteriores focados em inglês ou espanhol, esta proposta adapta o BERTimbau às nuances do português brasileiro, combinando técnicas de pré-processamento específicas (ex: tratamento de emojis) com uma arquitetura modular inovadora.

## 3. Materiais e Métodos

Esta seção descreve a metodologia utilizada para detecção automática de discurso de ódio em mídias sociais, alinhando-se aos avanços em Mineração de Textos e PLN. A abordagem proposta combina o modelo BERTimbau, pré-treinado em português brasileiro, com conjuntos de dados especializados (HateBR e TuPy-E). A Figura 1 ilustra o fluxo do processo, desde a classificação binária das mensagens até a classificação multiclasse.

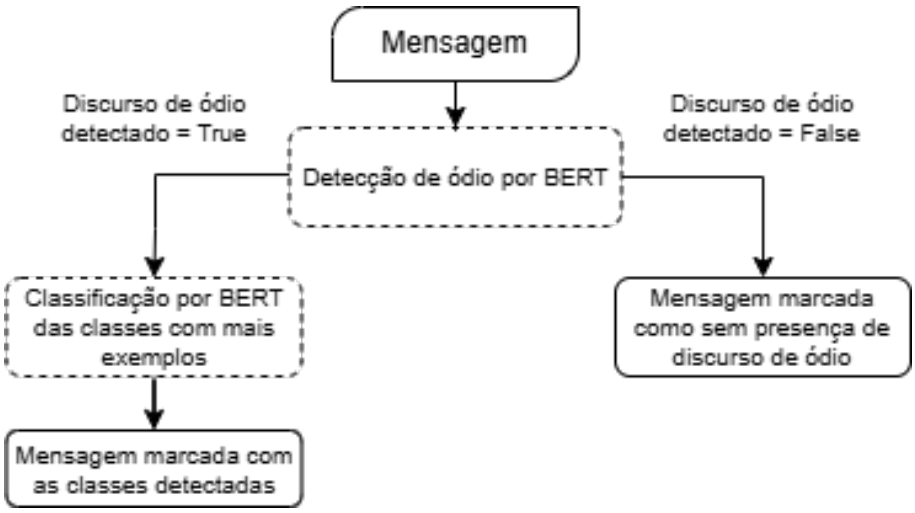


**Figura 1. Visão geral da abordagem sugerida para detecção de discurso de ódio**

A metodologia adotada neste trabalho segue o método científico e está organizada nas seguintes etapas:

### 3.1. Revisão bibliográfica

Nesta etapa, foi realizada uma revisão bibliográfica abrangente sobre PLN, modelos de geração e classificação de texto, e o estado da arte em detecção automatizada de discurso de ódio. A revisão incluiu a análise de artigos científicos e projetos relacionados.

### 3.2. Produção de um *Dataset* de Discurso de Ódio em português Brasileiro

Devido às limitações impostas pelas redes sociais para extração de dados (como altos custos e restrições de acesso) e à complexidade de classificar grandes volumes de textos de forma não enviesada, optou-se por utilizar *datasets* já existentes. Foram selecionadas as bases HateBR e TuPy-E, anteriormente mencionadas.

Para integrar esses *datasets* e gerar uma base de dados robusta e diversificada, que futuramente será publicizada, foram unidas as partições de treino, validação e teste de cada base usando a biblioteca *datasets* e criadas quatro macro-categorias através de operações lógicas OU sobre as colunas originais, garantindo compatibilidade entre os esquemas de anotação distintos. **GenSex**(no HateBR as colunas homofobia e sexismo; no TuPy-E, misoginia e lgbtfobia); **Xeno** (no HateBR, antisemitismo, racismo e xenofobia; no TuPy-E, racismo e xenofobia); **Political** (no HateBR, partidarismo e apologia à ditadura; no TuPy-E, continuou com o mesmo nome); **Physical** (no HateBR é equivalente a gordofobia; no TuPy-E, capacitismo, etarismo e *body shame*)

Foi aplicado então um filtro de qualidade em ambos os DataFrame para remover todas as linhas em que alguma macro-categoria era verdadeira, mas o rótulo *hate* permanecia zero. Por fim, foram normalizados os esquemas de colunas entre os dois DataFrames (adicionando colunas ausentes no HateBR preenchidas com zeros), foram concatenados verticalmente, convertendo todas as colunas de macro-categorias e o rótulo *hate* para tipo inteiro e embaralhando o conjunto completo com random_state=42, resultando no DataFrame *FinalSet*.

### 3.3. Estudo de Soluções para Classificação Textual

Foram pesquisadas soluções para classificação automática de discurso de ódio, com base no *dataset* produzido. Dois modelos se destacaram:

• BERTimbau: Uma versão do BERT (*Bidirectional Encoder Representations from Transformers*) pré-treinadoa especificamente para o português brasileiro. O BERTimbau é capaz de capturar relações contextuais bidirecionais em textos, sendo altamente eficaz para tarefas de classificação textual.

• DeepSeek: Um modelo de linguagem generativo pré-treinado conhecido por sua capacidade de gerar texto coerente e contextualizado. Apesar de ser originalmente projetado para geração de texto, sua arquitetura avançada e capacidade de entender contextos complexos o tornam uma alternativa interessante para análise de sentimentos e detecção de discursos de ódio.

### 3.4. Desenvolvimento de um Modelo para Classificação Textual

O processo de desenvolvimento dos modelos seguiu uma abordagem sistemática dividida em cinco etapas principais:

#### 3.4.1. Preparação dos Dados

Utilizou-se o *Random Under-Sampling* para equalizar a distribuição das classes, garantindo uma proporção de 1:1 entre exemplos de ódio e não-ódio. Em seguida, os dados foram separados em conjuntos de treino (80%), validação (10%) e teste (10%) preservando a distribuição original de classes e foi empregado um tokenizador customizado (BERTimbau + Emojis presentes no *dataset*) com truncamento em 512 tokens e estratégia de *dynamic padding*

### 3.4.2. Arquitetura do Modelo

Os modelos para detecção binária de ódio foram construídos sobre o BERTimbau Base (*neuralmind/bert-base-portuguese-cased*) e o BERTimbau Large (*neuralmind/bert-large-portuguese-cased*), havendo: O redimensionamento da camada de *embeddings* para acomodar tokens especiais e emojis; O congelamento parcial das camadas, onde apenas as 2 últimas camadas do *encoder* (11 e 10) e a camada de *embeddings* foram treináveis; A adição de camada densa com *dropout* (30%) sobre a representação do *token* [CLS].

### 3.4.3. Configuração de Treinamento

Os hiperparâmetros foram otimizados através de experimentação iterativa:

- **Otimizador**: AdamW com taxa de aprendizado de $2 \times 10^{-5}$
- **Regularização**: *Weight decay* de 0.01 e *gradient clipping* em 1.0
- ***Batch Size***: 16 amostras por dispositivo
- **Duração**: 30 *epochs*

### 3.4.4. Especialização para Categorias Específicas

Após o treinamento do modelo de detecção binária de discurso de ódio, realizou-se a especialização para categorias específicas (Xeno e GenSex). Foi utilizado o modelo binário pré-treinado como ponto de partida, aproveitando o conhecimento já adquirido sobre discurso de ódio. Conservou-se a mesma configuração de treinamento anterior, reduzindo apenas a duração para 15 *epochs*.

Buscando garantir a reprodutibilidade, todo o processo foi implementado utilizando uma *seed* de valor 42 e paralelismo desabilitado.

### 3.5. Avaliação da Acurácia do Modelo Treinado

Foram conduzidos testes automatizados para calcular a acurácia dos modelos treinados, utilizando métricas como precisão, *recall* e *F1-score*. Os resultados foram comparados com modelos já existentes na literatura, identificando possíveis melhorias. As correções e otimizações necessárias estão sendo implementadas para aumentar a eficácia do modelo.

## 4. Resultados e Discussão

Após seis meses de desenvolvimento, foram obtidos resultados significativos para modelos de detecção de discurso de ódio baseados em BERTimbau. A Tabela 1 sintetiza o desempenho dos modelos, evidenciando diferenças críticas entre as abordagens.

O modelo hate_base alcançou 84,0% de acurácia (AUC = 0,911), enquanto o hate_large atingiu 85,2% (AUC = 0,897), demonstrando que o aumento do tamanho do modelo não garantiu ganhos proporcionais em todas as métricas. Já os modelos especializados apresentaram comportamentos distintos: o GenSex_base, treinado com maior volume de dados, destacou-se com 86,9% de acurácia e AUC elevada (0,965), embora com precisão moderada (0,740) e *recall* alto (0,902). Em contraste, o modelo Xeno_base,

com menor base de treinamento, registrou 80,1% de acurácia e F1 significativamente inferior (0,544), refletindo desafios na harmonização entre precisão (0,554) e recall (0,801).

A análise revela que a quantidade de dados impacta diretamente o desempenho: o modelo GenSex_large, mesmo com arquitetura ampliada, não superou a versão base em acurácia (84,7% vs. 86,9%), enquanto o Xeno_large apresentou queda expressiva para 74,0% de acurácia, apesar de uma AUC elevada (0,929). Isso sugere que, em cenários com dados limitados, modelos menores podem ser mais robustos a *overfitting*.

Comparando com trabalhos da literatura, que frequentemente atingem acurácias próximas a 90%, os resultados obtidos são inferiores. Entretanto, é crucial destacar que esses modelos foram treinados com *fine-tuning* simples e custo computacional reduzido, uma escolha estratégica para balancear eficiência e viabilidade prática. A discrepância reflete *trade-offs* intencionais: enquanto modelos da literatura empregam técnicas complexas (como aumento de dados ou treinamento prolongado), esta abordagem priorizou acessibilidade computacional e a viabilidade prática.

Nota-se ainda que os modelos especializados em xenofobia (Xeno) apresentaram *recall* elevado ($\geq$80%), indicando sensibilidade para capturar casos positivos, mas precisão limitada ($\approx$55%), o que implica alto número de falsos positivos. Essa característica pode ser aceitável em contextos de moderação preliminar, onde subdetecção é crítica, mas demanda revisão humana complementar.

**Tabela 1. Resultados dos modelos BERTimbau com métricas principais. Todas as métricas apresentam três casas decimais.**

| Modelo | Acurácia | AUC | Precisão | *Recall* | F1 |
|---|---|---|---|---|---|
| Hate_base | 0.840 | 0.911 | 0.841 | 0.840 | 0.840 |
| Hate_large | 0.852 | 0.897 | 0.855 | 0.852 | 0.852 |
| GenSex_base | 0.869 | 0.965 | 0.740 | 0.902 | 0.782 |
| GenSex_large | 0.847 | 0.960 | 0.718 | 0.883 | 0.755 |
| Xeno_base | 0.801 | 0.900 | 0.554 | 0.801 | 0.544 |
| Xeno_large | 0.740 | 0.929 | 0.551 | 0.846 | 0.517 |

## 5. Considerações Finais

Como desdobramento da pesquisa pretende-se desenvolver classificações mais precisas, com foco nas categorias com maior incidência. Para isso, serão implementadas estratégias de otimização de hiperparâmetros e técnicas de aumento de dados (*data augmentation*), visando mitigar o desbalanceamento de classes e melhorar a generalização do modelo. Além disso, será investigada a implementação de arquiteturas como o DeepSeek especificamente para categorias minoritárias, onde a escassez de dados limita a eficácia de abordagens convencionais. Adicionalmente, será explorada a integração sistemática de metadados (como histórico de interações do usuário e contexto conversacional) para enriquecer a análise e a classificação. Por fim, os resultados serão divulgados em revistas e periódicos da área de PLN e Inteligência Artificial (IA). Além disso, resumos expandidos serão submetidos a congressos científicos, contribuindo para a disseminação do conhecimento e o debate sobre o combate ao discurso de ódio no ambiente digital.

## Referências


Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing Management*, 39(1):45–65.

Almaliki, M., Almars, A. M., Gad, I., and Atlam, E.-S. (2023). Abmm: Arabic bert-mini model for hate-speech detection on social media. *Electronics*, 12(1048).

ANDES (2023). Brasil lidera discurso de ódio nas redes sociais contra população lgbtqiap+. Acesso em: 20 de março de 2025.

BR, U. (2023). O discurso de ódio 'é um dos sinais de alerta de genocídio e de outros crimes atrozes,' alerta guterres. Acesso em: 20 de março de 2025.

BR, U. (2024). Como podemos ajudar a combater o discurso de ódio nas redes sociais. Acesso em: 20 de março de 2025.

Caseli, H. M. and (orgs.), M. G. V. N. (2023). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN.

Ebecken, N., Lopes, M., and Costa, M. (2003). Mineração de textos. In *Capítulo 13, p. 337–370*. Manole.

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 1st edition.

Lorena, A. C. and de Carvalho, A. C. (2007). Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, 14(2):43–67.

Nacional, J. (2023). Denúncias de crimes envolvendo discurso de ódio nas redes sociais triplicaram nos últimos 6 anos, aponta levantamento. Acesso em: 20 de março de 2025.

Nascimento, F. R. S., Cavalcanti, G. D. C., and Costa-Abreu, M. D. (2023). Exploring automatic hate speech detection on social media: A focus on content-based analysis. *SAGE Open*, April-June 2023:1–19.

Oliveira, F., Reis, V., and Ebecken, N. (2023). Tupy-e: Detecting hate speech in brazilian portuguese social media with a novel dataset and comprehensive analysis of models. *arXiv preprint*.

Vargas, F., Carvalho, I., Goés, F., Pardo, T. A., and Benevenuto, F. (2022). Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. *arXiv preprint*.