Research, Society and Development, v. 14, n. 5, e10814548633, 2025
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v14i5.48633

Reconhecimento de padrões na detecção do discurso de ódio: uma abordagem de redes neurais e ensembles

Pattern recognition in hate speech detection: a neural network and ensemble approach

Reconocimiento de patrones en la detección del discurso de odio: un enfoque con redes neuronales y ensambles

Recebido: 03/04/2025 | Revisado: 25/05/2025 | Aceitado: 26/05/2025 | Publicado: 29/05/2025

Filipe Cordeiro de Medeiros Azevedo
ORCID: https://orcid.org/0000-0003-0821-8779
Centro de Estudos e Sistemas Avançados do Recife, Brasil
E-mail: fcma@cesar.org.br

Resumo
O discurso de ódio em plataformas online é um problema crescente, com impactos sociais significativos. Este trabalho propõe uma abordagem para a classificação binária de discurso de ódio em textos em português, utilizando algoritmos de aprendizado de máquina e aprendizado profundo. Os experimentos foram conduzidos em um conjunto de dados anotado, com representações textuais geradas por word embeddings pré-treinados do GloVe. O modelo baseado em votação, que combina os resultados dos classificadores base, apresentou o melhor desempenho geral, alcançando um F1-score de 0.76. Os resultados demonstram a eficácia das redes neurais, especialmente na captura de padrões textuais complexos, e destacam o potencial de abordagens combinadas para a tarefa de classificação de discurso de ódio. Este estudo reforça a importância de explorar arquiteturas diversificadas e técnicas de pré-processamento alinhadas às peculiaridades do idioma português.
Palavras-chave: Discurso de ódio; Aprendizado de máquina; Processamento de Linguagem Natural.

Abstract
Hate speech on online platforms is a growing problem with significant social impacts. This work proposes an approach for binary classification of hate speech in Portuguese texts using machine learning and deep learning algorithms. The experiments were conducted on an annotated dataset, with textual representations generated by pre-trained GloVe word embeddings. The voting-based model, which combines the outputs of the base classifiers, achieved the best overall performance, reaching an F1-score of 0.76. The results demonstrate the effectiveness of neural networks, especially in capturing complex textual patterns, and highlight the potential of combined approaches for the hate speech classification task. This study reinforces the importance of exploring diverse architectures and preprocessing techniques tailored to the specific characteristics of the Portuguese language.
Keywords: Hate speech; Machine learning; Natural Language Processing.

Resumen
El discurso de odio en las plataformas en línea es un problema creciente, con impactos sociales significativos. Este trabajo propone un enfoque para la clasificación binaria del discurso de odio en textos en portugués, utilizando algoritmos de aprendizaje automático y aprendizaje profundo. Los experimentos se realizaron sobre un conjunto de datos anotado, con representaciones textuales generadas mediante word embeddings preentrenados de GloVe. El modelo basado en votación, que combina los resultados de los clasificadores base, obtuvo el mejor rendimiento general, alcanzando un F1-score de 0.76. Los resultados demuestran la eficacia de las redes neuronales, especialmente en la captura de patrones textuales complejos, y destacan el potencial de los enfoques combinados para la tarea de clasificación de discurso de odio. Este estudio refuerza la importancia de explorar arquitecturas diversas y técnicas de preprocesamiento adaptadas a las particularidades del idioma portugués.
Palabras clave: Discurso de odio; Aprendizaje automático; Procesamiento de Lenguaje Natural.

1. Introdução

Com o crescimento das interações em redes sociais, o discurso de ódio online tornou-se um problema significativo, com impactos sociais profundos, incluindo a promoção de violência e exclusão de indivíduos ou grupos minoritários. Além disso, a curadoria informacional e a formação de bolhas digitais têm ampliado a polarização política e contribuído para a disseminação

de discursos extremistas, afetando diretamente o ambiente democrático (Ferreira, 2024). Detectar automaticamente o discurso de ódio é um desafio no campo do Processamento de Linguagem Natural (PLN), especialmente em idiomas com poucos conjuntos de dados, como o português.

O discurso de ódio é definido como uma linguagem que ataca ou diminui, incita violência ou ódio contra grupos, com base em características específicas, como aparência física, religião, etnia, orientação sexual ou identidade de gênero (Fortuna et al., 2018). Identificar e moderar essas mensagens é crucial para garantir um ambiente digital mais seguro e inclusivo.

Nos últimos anos, várias técnicas de aprendizado de máquina e aprendizado profundo têm sido aplicadas para lidar com essa tarefa, incluindo abordagens baseadas em Redes Neurais Convolucionais (CNNs), Redes Neurais Recorrentes (RNNs) e Máquinas de Vetores de Suporte (SVMs). No entanto, a maioria dos estudos foca no idioma inglês, com poucos trabalhos desenvolvidos para o português (Fortuna et al., 2019). (Gamback et al., 2017) também propuseram uma abordagem baseada em redes neurais convolucionais para a detecção de discurso de ódio em tweets, comparando diferentes combinações de embeddings e character n-grams. Seus resultados demonstraram que modelos baseados em CNN superaram classificadores tradicionais como regressão logística, alcançando um F1-score de 78.3%, o que reforça a eficácia dessa arquitetura em tarefas de classificação textual.

Neste trabalho, é proposto uma abordagem para a classificação do discurso de ódio de textos em português, utilizando diferentes modelos de aprendizado de máquina e aprendizado profundo, incluindo Redes Neurais e SVMs.

A seguir, o artigo está organizado da seguinte forma: na seção 2, é apresentado os trabalhos relacionados; na seção 3, uma descrição do método e os dados utilizados; na seção 4, discutimos os resultados experimentais; Por fim, na seção 5, as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

O problema da detecção de discurso de ódio tem sido amplamente investigado em diferentes contextos, com abordagens que variam desde métodos baseados em aprendizado de máquina tradicionais até técnicas de aprendizado profundo. Esta seção apresenta uma revisão de estudos significativos, com foco na classificação de discurso de ódio e em métodos aplicados ao idioma português.

Um dos desafios fundamentais na detecção de discurso de ódio é a disponibilidade de conjuntos de dados anotados, especialmente para línguas com poucos recursos. (Fortuna et al., 2018) discutem um conjunto de definições para discurso de ódio e destacaram a importância de taxonomias que diferenciem nuances em tipos de discurso ofensivo. Além disso, o trabalho introduz um conjunto de dados hierárquico em português, permitindo uma classificação mais granular e possibilitando análises detalhadas de interseções entre categorias de discurso de ódio (Fortuna et al., 2019).

No contexto de aprendizado de máquina, abordagens tradicionais, como Máquinas de Vetores de Suporte (SVMs) e classificadores lineares, têm sido aplicadas com sucesso. (Waseem et al., 2016) exploraram o impacto de características linguísticas e demográficas na detecção de discurso de ódio, utilizando modelos supervisionados em dados do Twitter (Waseem et al., 2016). Esses métodos, entretanto, dependem fortemente de representações explícitas, como Bag of Words e TF-IDF, que podem limitar a captura de relações semânticas mais profundas.

Abordagens clássicas continuam sendo exploradas na literatura devido à sua simplicidade e interpretabilidade. Em (Asogwa et al., 2022), por exemplo, compararam SVM e Naive Bayes na tarefa de classificação de discurso de ódio, demonstrando que o SVM alcançou desempenho significativamente superior, com precisão próxima a 99%, enquanto o Naive Bayes teve desempenho consideravelmente inferior. O estudo reforça a robustez do SVM mesmo frente a alternativas mais recentes, sendo frequentemente utilizado como linha de base em estudos comparativos.

Com avanço de técnicas de aprendizado profundo possibilitada também pelo aperfeiçoamento de hardwares, arquiteturas como Redes Neurais Convolucionais (CNNs) e Redes Neurais Recorrentes (RNNs) mostraram desempenho superior em tarefas de classificação de texto. (Badjatiya et al., 2017) investigaram o uso de embeddings de palavras treinadas com GloVe e modelos como LSTMs e CNNs para a detecção de discurso de ódio em tweets, obtendo melhorias significativas em relação a abordagens baseadas em características manuais.

Para o idioma português, estudos como os de (Hartmann et al., 2017) avaliaram diferentes representações de palavras, como Word2Vec, FastText e GloVe, em tarefas semelhantes, destacando a importância de modelos mais ajustados às particularidades linguísticas. Além disso, (Fortuna et al., 2019) apresentaram além de um conjunto de dados hierarquicamente anotado para discurso de ódio em português, um modelo de classificação de mensagens de acordo com múltiplos critérios utilizando uma combinação de redes LSTMs e xGBoost para tarefas de classificação binária e multiclasse.

Estudos recentes também demonstram o potencial das redes neurais profundas na tarefa de classificação de discurso de ódio. (d'Sa et al., 2020) compararam diferentes arquiteturas baseadas em DNN, como CNN, Bi-LSTM e CRNN, em conjunto com representações vetoriais como FastText e BERT, obtendo resultados superiores aos métodos baseados em características manuais e classificadores tradicionais, como o SVM. Esses achados reforçam a relevância de explorar arquiteturas neurais especializadas e embasaram a escolha dos modelos avaliados neste trabalho.

Apesar dos avanços, muitos desafios permanecem, incluindo a adaptação de modelos generalistas para contextos linguísticos específicos, o manejo de mensagens ambíguas ou contextualmente ofensivas e a integração de dados multimodais em sistemas de classificação. Este trabalho busca contribuir para a literatura ao explorar abordagens baseadas em SVMs e redes neurais aplicadas a textos em português, avaliando diferentes representações e técnicas de pré-processamento.

3. Metodologia

O presente estudo é um relato de experiência de uma pesquisa laboratorial com uso de software e técnicas de redes neurais num estudo de natureza qualitativa e quantitativa (Pereira et al., 2018; Barros, 2024; Mussi, Flores & Almeida, 2021). A metodologia deste trabalho foi desenvolvida com o objetivo de avaliar a eficácia de diferentes modelos de aprendizado de máquina e aprendizado profundo na tarefa de classificação binária de discurso de ódio em textos em português. Nesta seção, são apresentados o conjunto de dados utilizado, as técnicas de pré-processamento, os modelos aplicados e os métodos de avaliação.

3.1 Conjunto de dados

Neste estudo, foi utilizado o conjunto de dados Hate Speech Dataset (HSD), fornecido por (Fortuna et al., 2019). A base de dados é composta por tweets em português, originalmente classificados de forma hierárquica em diferentes categorias de discurso de ódio. Para os objetivos deste trabalho, os rótulos binários foram definidos pela combinação das hierarquias originais, resultando em duas classes: discurso de ódio e não discurso de ódio.

A base contém aproximadamente 5.668 tweets, com uma distribuição desbalanceada entre as classes: cerca de 68% das mensagens são classificadas como não contendo discurso de ódio, enquanto 32% são classificadas como contendo discurso de ódio. Essa proporção reflete a distribuição natural dos dados, sem a aplicação de técnicas de balanceamento.

Os dados foram divididos inicialmente em dois subconjuntos: treino (80%) e teste (20%). Dentro do conjunto de treino, foi realizada uma validação cruzada estratificada, utilizando partições em que 80% dos dados eram destinados ao treino e 20% à validação. Essa estratégia garantiu a preservação da proporção entre as classes em todas as etapas do treinamento e validação, permitindo uma avaliação consistente do desempenho dos modelos.

A base HSD é amplamente reconhecida por capturar nuances linguísticas específicas do idioma português e foi selecionada neste trabalho por sua qualidade e relevância, proporcionando um cenário desafiador para a tarefa de classificação de discurso de ódio.

3.2 Pré-processamento

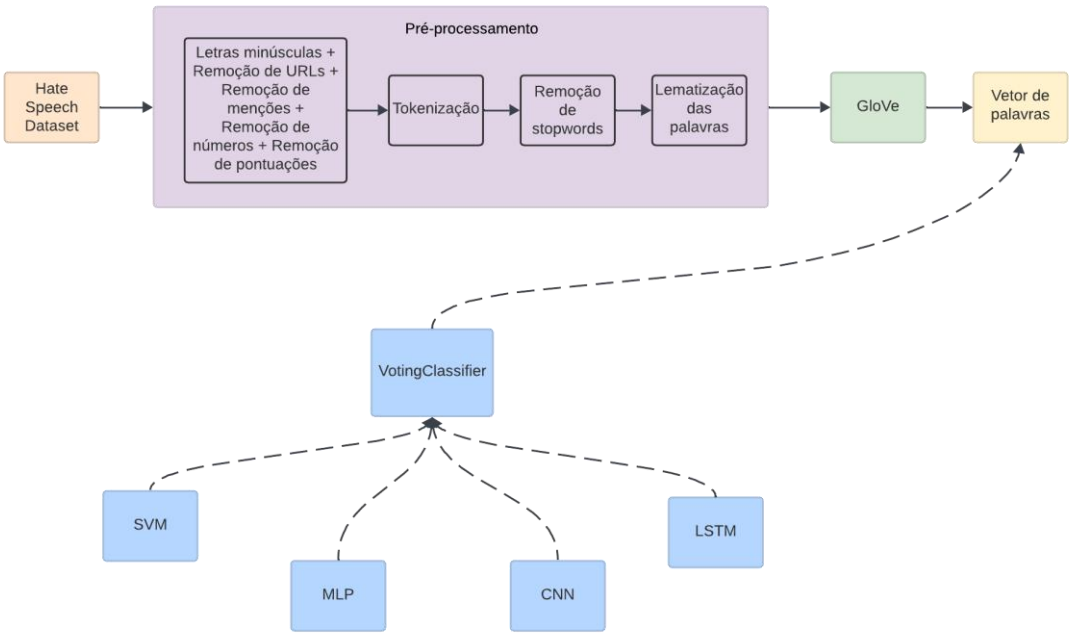
O pré-processamento dos textos foi realizado utilizando a biblioteca NLTK em Python, com o objetivo de reduzir o ruído e normalizar os dados, otimizando o desempenho dos modelos de classificação. Inicialmente, os textos foram convertidos para letras minúsculas, garantindo consistência na análise. Em seguida, elementos irrelevantes, como números, URLs e menções no Twitter, foram removidos utilizando expressões regulares. Também foi eliminada a pontuação, reduzindo a complexidade textual sem prejudicar a semântica.

Para a preparação das mensagens, foi aplicada a tokenização com a função word_tokenize do NLTK, dividindo os textos em palavras individuais. Posteriormente, as stopwords em português, como preposições e artigos, foram excluídas com base na lista disponibilizada pela biblioteca. A lematização foi utilizada para reduzir as palavras às suas formas base, empregando o lematizador WordNetLemmatizer. Após o processamento, os tokens foram unidos novamente em uma string, formando os textos finais utilizados pelos modelos.

Para representar os textos como vetores numéricos, foram utilizados os embeddings pré-treinados do GloVe (Global Vectors for Word Representation), amplamente reconhecidos por capturarem relações semânticas e contextuais em grande escala. O GloVe é treinado em grandes corpus e produz vetores densos, onde palavras com significados similares possuem representações próximas no espaço vetorial. Neste estudo, foi empregada uma versão do GloVe treinada em corpus relevantes para o português, garantindo uma adaptação adequada ao idioma e uma representação eficiente dos textos para os algoritmos.

Todo o processo pode ser identificado pelo esquema da Figura 1. Essas etapas asseguraram que os dados estivessem não apenas limpos e uniformes, mas também representados de forma eficiente para exploração pelos algoritmos.

Figura 1 - Pré-processamento e algoritmos.



Fonte: Dados da Pesquisa (2025).

3.3 Algoritmos utilizados

Nesta seção, detalhamos os algoritmos empregados para a classificação de discurso de ódio e os procedimentos de otimização realizados durante o treinamento.

Foram utilizados cinco abordagens principais: Máquina de Vetores de Suporte (Support Vector Machine - SVM), Multi-Layer Perceptron (MLP), Redes Neurais Convolucionais (Convolutional Neural Networks - CNN), Redes Neurais Recorrentes com Long Short-Term Memory (LSTM) e um modelo de classificação baseado em votos. Cada um desses modelos foi configurado e otimizado para maximizar o desempenho na tarefa de classificação binária.

A Máquina de Vetores de Suporte (SVM) foi empregada como um modelo linear de base, utilizando o núcleo (kernel) radial base (RBF) para lidar com possíveis relações não lineares entre os atributos. O SVM é amplamente reconhecido por sua eficácia em conjuntos de dados com distribuição desbalanceada devido à sua capacidade de maximizar margens entre classes e seu desempenho robusto em espaços de alta dimensionalidade (Cortes, 1995).

O Multi-Layer Perceptron (MLP), uma rede neural feedforward totalmente conectada, foi configurado com múltiplas camadas ocultas e funções de ativação ReLU. Essa arquitetura é amplamente utilizada em tarefas de classificação por sua flexibilidade e capacidade de modelar relações não lineares nos dados.

A Rede Neural Convolucional (CNN) foi configurada para explorar padrões locais no texto representado por embeddings. Camadas convolucionais com diferentes tamanhos de filtro foram combinadas com camadas de pooling e dropout para reduzir a dimensionalidade e evitar sob reajuste. Redes CNN são eficazes para tarefas de processamento de texto, pois podem capturar padrões semânticos em n-gramas de maneira eficiente.

A Rede Neural Recorrente LSTM foi utilizada para explorar dependências de longo prazo no texto. Configurada com uma ou mais camadas LSTM empilhadas, a arquitetura inclui camadas de regularização (dropout) e funções de ativação sigmoidal para melhorar a estabilidade durante o treinamento. As LSTMs são amplamente utilizadas em PLN devido à sua habilidade de lidar com sequências textuais de comprimento variável e de preservar informações ao longo de longos contextos (Hochreiter, 1997).

Por fim, foi implementado um modelo de classificação baseado em votos. Nesse modelo, a predição final é definida pela combinação das predições individuais dos quatro classificadores anteriores. A decisão final considera uma entrada como contendo discurso de ódio caso pelo menos dois classificadores a classifiquem positivamente. Essa abordagem visa combinar os pontos fortes de diferentes algoritmos, promovendo maior robustez e precisão na classificação.

Durante a fase de treinamento, cada modelo foi otimizado por meio do Grid Search, permitindo testar combinações de hiperparâmetros, como taxa de aprendizado, regularização e número de camadas. Após a otimização, os modelos foram avaliados no conjunto de teste para medir seu desempenho em dados não vistos.

A utilização de múltiplos algoritmos, combinados com o modelo baseado em votos, permitiu a comparação de diferentes abordagens e a exploração de suas respectivas vantagens na tarefa de classificação de discurso de ódio em textos em português.

3.4 Experimentos

Os experimentos realizados tiveram como objetivo principal avaliar o desempenho de diferentes algoritmos na tarefa de classificação binária do discurso de ódio em textos em português. O SVM foi utilizado para capturar padrões lineares e não lineares, enquanto o MLP explorou padrões globais por meio de camadas densas. Já a CNN destacou-se na identificação de padrões locais e contextuais, e a LSTM analisou dependências sequenciais de longo prazo, capturando nuances textuais mais complexas.

Nas linhas seguintes a Quadro 1 apresenta os hiperparâmetros para os respectivos algoritmos. Esses valores foram definidos após a etapa de otimização baseada em Grid Search, considerando o equilíbrio entre desempenho e complexidade computacional.

Quadro 1 – Hiperparâmetros.

Algoritmo	**Hiperparâmetros**
SVM	kernel = rbf gamma = scale c = 10
MLP	epochs = 10 batch size = 128 learning rate = 0.001 camadas densas = [1000 e 500 unidades + ReLU] dropout = 0.2 optimizer = Adam
CNN	epochs = 10 batch size = 128 learning rate = 0.001 camada convolucional = 100 k = 1 camadas densas = [1000 unidades + ReLU] dropout = [0.5, 0.2] optimizer = Adam
LSTM	epochs = 10 batch size = 128 learning rate = 0.001 lstm = 200 camadas densas = [1000 unidades + ReLU] dropout = [0.5, 0.2] optimizer = Adam

Fonte: Dados da Pesquisa (2025).

A principal métrica utilizada para avaliar o desempenho dos algoritmos foi o F1-score, que equilibra precisão e recall, sendo especialmente útil em cenários com classes desbalanceadas. Os modelos foram treinados e validados utilizando validação cruzada estratificada no conjunto de treino. Após a seleção dos melhores hiperparâmetros, os modelos foram avaliados no conjunto de teste para mensurar seu desempenho em dados ainda não vistos.

Para garantir a reprodutibilidade dos experimentos, foi utilizada uma semente fixa com o valor 42 em todas as etapas, incluindo o treinamento, divisão de dados. Adicionalmente, foi realizada uma análise comparativa entre as duas abordagens de word embeddings (GloVe e Gemini) para avaliar o impacto das diferentes representações textuais no desempenho dos modelos. Todo o código utilizado pode ser encontrado no repositório do Github[1].

4. Resultados e Discussão

Os resultados obtidos com os modelos testados são apresentados na Quadro 2, onde são comparadas as métricas de precisão (Precision), revocação (Recall) e F1-score para os quatro algoritmos base (SVM, MLP, CNN e LSTM) e o modelo baseado em votação (VotingClassifier).

[1] https://github.com/filipecmedeiros/Portuguese-Hate-Speech-Classification

Quadro 2 – Resultados.

Modelo	**Métrica**		
	Precision	Recall	F1-score
SVM	0.71	0.73	0.72
MLP	0.70	0.71	0.70
CNN	0.75	0.73	0.73
LSTM	0.75	0.75	0.75
VotingClassifier	0.76	0.76	**0.76**

Fonte: Dados da Pesquisa (2025).

Os modelos base apresentaram desempenhos variados, com os algoritmos CNN e LSTM alcançando F1-scores de 0.73 e 0.75, respectivamente. Ambos se destacaram por sua capacidade de capturar padrões complexos nos textos, atribuída ao uso de embeddings representativos e arquiteturas projetadas para lidar com a estrutura sequencial dos dados. O SVM, embora simples, apresentou um F1-score de 0.72, demonstrando robustez em tarefas de classificação binária com dados desbalanceados. Por outro lado, o MLP obteve o menor desempenho, com um F1-score de 0.70, possivelmente devido à sua limitação em capturar dependências locais e contextuais.

O modelo de votação, que combina as predições dos classificadores base, obteve o melhor desempenho geral, com um F1-score de 0.76. Esse resultado indica que a combinação dos pontos fortes dos diferentes algoritmos melhora a robustez e a precisão do sistema, especialmente na detecção do discurso de ódio.

Comparando com os resultados apresentados por (Fortuna et al., 2019), que utilizaram redes neurais recorrentes para a classificação hierárquica de discurso de ódio em português, o desempenho dos modelos neste trabalho, focado em classificação binária, é competitivo. (Fortuna et al., 2019) relataram um F1-score de aproximadamente 0.74 em tarefas similares. Os resultados obtidos pelo modelo de votação neste estudo (F1-score de 0.76) sugerem que a combinação de arquiteturas diversificadas pode superar abordagens individuais, destacando a relevância de técnicas como votação para tarefas de PLN em português.

Os resultados evidenciam também a eficácia das abordagens baseadas em redes neurais, especialmente quando combinadas com técnicas de votação, na tarefa de detecção do discurso de ódio. Adicionalmente, a análise reforça a importância de explorar diferentes arquiteturas para maximizar o desempenho em problemas de PLN, adaptando as soluções às características específicas do idioma.

5. Conclusão

Este trabalho apresentou uma abordagem para a classificação de discurso de ódio em textos em português, utilizando quatro algoritmos base — SVM, MLP, CNN e LSTM — e um modelo de classificação baseado em votação. Os experimentos demonstraram que o modelo de votação alcançou o melhor desempenho geral, com um F1-score de 0.76, destacando-se como uma solução robusta ao combinar os pontos fortes dos diferentes algoritmos.

Os resultados também evidenciam a eficácia das redes neurais, especialmente CNNs e LSTMs, que se mostraram bem-sucedidas na captura de padrões semânticos e dependências sequenciais dos textos. A utilização de word embeddings pré-treinados, como o GloVe, contribuiu de forma significativa para os resultados, fornecendo representações vetoriais que capturam relações semânticas relevantes para a tarefa de classificação.

Comparado com estudos anteriores, como o trabalho de (Fortuna et al., 2019), os resultados sugerem que a classificação binária, aliada à combinação de modelos por votação, pode superar abordagens mais tradicionais em tarefas específicas de PLN. Este estudo reforça a importância de explorar arquiteturas diversificadas e técnicas de pré-processamento alinhadas às peculiaridades do idioma português.

Para trabalhos futuros, planeja-se expandir esta pesquisa em várias direções. Primeiramente, investigar abordagens para lidar com o desbalanceamento natural do conjunto de dados, como técnicas de oversampling ou undersampling. Além disso, será interessante explorar embeddings contextuais mais avançados, como os gerados por modelos como o BERT e seus derivados ajustados para o português.

A classificação de discurso de ódio permanece como um desafio significativo, especialmente em idiomas com menos recursos, como o português. Este trabalho contribui para a evolução da área, mas reforça a necessidade de investigações contínuas e a busca por soluções que possam lidar com as complexidades do problema em diferentes contextos e linguagens.

Referências

Asogwa, D. C., Chukwuneke, C. I., Ngene, C. C., & Anigbogu, G. N. (2022). Hate speech classification using SVM and Naive Bayes. IOSR Journal of Mobile Computing & Application (IOSR-JMCA), 9(1), 27–34. https://doi.org/10.9790/0050-09012734

Cortes, C. (1995). Support-vector networks. Machine Learning.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 759–760).

d'Sa, A. G., Illina, I., & Fohr, D. (2020). Classification of hate speech using deep neural networks. Revue d'Information Scientifique & Technique, 25(1). HAL Id: hal-03101938. https://hal.science/hal-03101938v1

Ferreira, M. C. dos S., & Teixeira, T. (2024). Social media and political polarization as threats to democracy. Research, Society and Development, 13(7), e7713746214–e7713746214.

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled Portuguese hate speech dataset. In Proceedings of the Third Workshop on Abusive Language Online (pp. 94–104).

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), 1–30.

Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In Proceedings of the First Workshop on Abusive Language Online (pp. 85–90). Vancouver, Canada: Association for Computational Linguistics. https://aclanthology.org/W17-3013

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., & Aluísio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. arXiv preprint, arXiv:1708.06025.

Hochreiter, S. (1997). Long short-term memory. Neural Computation MIT-Press.

Pereira, A. S., Shitsuka, D. M., Parreira, F. J., & Shitsuka, R. (2018). Metodologia da pesquisa científica. Brasil.

Rakhlin, A. (2016). Convolutional neural networks for sentence classification. GitHub, 6, 25.

Silva, S. C., & Serapião, A. B. S. (2018). Detecção de discurso de ódio em português usando CNN combinada a vetores de palavras. In Anais do VI Symposium on Knowledge Discovery, Mining and Learning (pp. 1–8). SBC.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop (pp. 88–93).