# Explorando BERT-MLCNN para a classificação de discursos de ódio em português

Relatório de Atividade 3 - CI1174 - Tópicos em Aprendizado de Máquina

Raul Jose Silverio da Silva
raul.silverio@ufpr.br
Departamento de Informática
Universidade Federal do Paraná
Curitiba, Paraná, Brasil

## Abstract

O aumento da prevalência de discurso de ódio na internet apresenta desafios significativos que exigem métodos eficazes de detecção. Este artigo explora a aplicação de técnicas de aprendizado profundo, especificamente a combinação de um modelo BERT específico para o português (BERTimbau [9] e BERTabaporu [2]) com uma Rede Neural Convolucional de Múltiplas Camadas (MLCNN), para detectar discurso de ódio em textos em português. Utilizando os datasets: Portuguese Hate Speech Expanded Dataset (TuPyE [7]), que abrange várias categorias de discurso de ódio (por exemplo, etarismo, lgbtfobia, racismo), e o UlyssesSD-Br [6], para detecção de posicionamentos. Implementamos e avaliamos um modelo BERT-MLCNN [1] em todos esses conjuntos de dados para comparação de desempenho com os modelos originais.

## Keywords

Hate Detection, Stance Detection, BERT, Convolutional Neural Networks, Deep Learning, NLP, Political Texts, Portuguese Language

## 1 Introdução

O reconhecimento de discurso de ódio em textos é uma tarefa desafiadora e crucial no campo de Processamento de Linguagem Natural (PLN). Com o aumento da interação digital e o crescimento das plataformas online, o discurso de ódio tornou-se uma preocupação significativa, afetando indivíduos e grupos sociais em diversas formas, como racismo, misoginia, homofobia, e xenofobia. O discurso de ódio pode se manifestar de maneira sutil ou explícita, tornando difícil sua identificação automática, especialmente em contextos complexos e dinâmicos da linguagem natural [8].

Pesquisas recentes têm utilizado abordagens baseadas em redes neurais profundas para abordar essa tarefa, em especial modelos pré-treinados como BERT [3] e suas variações adaptadas para línguas específicas, como o BERTimbau [9], voltado para o português. Esses modelos, ao capturarem representações contextuais da linguagem, têm demonstrado grande sucesso em várias tarefas de PLN, incluindo a detecção de discurso de ódio. No entanto, apesar do sucesso de modelos como BERT, ainda há desafios em melhorar a precisão e a generalização para o contexto do português, especialmente em contextos que envolvem discurso de ódio específico de diferentes categorias (agressividade, racismo, homofobia, entre outros).

Neste trabalho, propomos a combinação do BERT com uma Rede Neural Convolucional de Múltiplas Camadas (CNN-Multilayer), uma abordagem que tem se mostrado eficaz em tarefas de análise de sentimentos e classificação de texto [7]. A ideia central é explorar as forças do BERT para gerar embeddings contextuais ricos e das CNNs para capturar padrões locais importantes que podem ser fundamentais na detecção de discurso de ódio em português. Essa combinação permite que o modelo aproveite o melhor de ambos os mundos: a capacidade do BERT de entender o contexto linguístico e a habilidade das CNNs de capturar características espaciais e hierárquicas nos textos.

A principal contribuição deste estudo é a aplicação do modelo BERT-MCNN proposto por Atandoh et al. no Portuguese Hate Speech Expanded Dataset (TuPyE) [7], um dataset específico para a detecção de discurso de ódio no português. Este conjunto de dados contém uma ampla gama de categorias de discurso de ódio, permitindo uma análise detalhada das diferentes formas de discriminação e violência verbal. Além disso, buscamos avaliar o desempenho do modelo em diversas métricas, como precisão, recall, F1-Score e acurácia, para validar sua eficácia em cenários do mundo real.

## 2 Trabalhos Relacionados

Pesquisas recentes no português têm explorado variações de modelos baseados em BERT. [5] aplicaram técnicas como BERT e AutoML para análise de linguagem tóxica em português, enquanto [10] desenvolveram o HateBR, um corpus anotado voltado para comentários em redes sociais. Esses estudos evidenciam o impacto do uso de PLN para melhorar a identificação de discursos de ódio e destacar a importância de abordagens culturalmente sensíveis.

### 2.1 Datasets para Detecção de Discurso de Ódio

A detecção de discurso de ódio em português enfrenta desafios específicos devido à complexidade da língua, incluindo sua gramática rica, vocabulário extenso e variações regionais [4]. Nesse contexto, o TuPyE se destaca como o maior corpus público anotado para a tarefa, com 43.668 documentos organizados em categorias como ageismo, misoginia, racismo, xenofobia, entre outras. O dataset foi desenvolvido a partir da integração de dados de estudos prévios [4] [5] [10] e da inclusão de 10.000 documentos originais do TuPy-Dataset. O TuPyE oferece uma base sólida para treinar e avaliar modelos robustos de detecção de discurso de ódio [7].

### 2.2 Modelos Pré-treinados e Abordagens Híbridas

Os avanços em PLN têm sido impulsionados por modelos baseados em transformadores, como o BERT (Devlin et al., 2019), que

capturam representações contextuais profundas. No contexto do português, modelos como o BERTimbau (Souza et al., 2020) e o BERTaBaporu (Costa et al., 2023) foram desenvolvidos para lidar com as particularidades da língua, incluindo variações sintáticas e semânticas. Ambos têm mostrado excelente desempenho em tarefas de classificação de texto, incluindo a identificação de discurso de ódio.

Modelos híbridos, como o BERT-MCNN [1], combinam o poder contextual do BERT com a capacidade das Redes Convolucionais (CNNs) de identificar padrões locais e hierárquicos nos textos. Essa integração é particularmente eficaz em textos curtos e em cenários onde há sobreposição de categorias de discurso de ódio, como misoginia e racismo [1] [5].

## 3 Metodologia

O objetivo deste estudo é aplicar a arquitetura BERT-MultiLayer Convolutional Neural Network (BERT-MCNN), proposta por [1], ao problema de detecção de discurso de ódio no contexto da língua portuguesa, utilizando o TuPyE[7], o maior corpus público para essa tarefa. A seguir, detalhamos o protocolo metodológico adotado.

### 3.1 Arquitetuda do modelo

A arquitetura BERT-MCNN combina a capacidade contextual do BERT com o poder das Redes Convolucionais de Múltiplas Camadas (CNNs) de capturar padrões locais e hierárquicos em textos. O pipeline consiste em três etapas principai

(1) Extração de Embeddings com BERT:
   - Utilizamos o BERTimbau[9] como modelo pré-treinado, otimizado para o português, para gerar embeddings contextuais. O modelo é fine-tuned no dataset TuPyE para capturar nuances linguísticas e contextuais específicas da tarefa. Cada texto no dataset é tokenizado usando o tokenizador do BERTimbau e processado para produzir embeddings de alta dimensão.
   - Cada texto no dataset é tokenizado usando o tokenizador do BERTimbau e processado para produzir embeddings de alta dimensão.

(2) Extração de Características com CNNs:
   - Os embeddings extraídos são alimentados em uma Rede Convolucional de Múltiplas Camadas (CNN). Essa rede utiliza diferentes tamanhos de kernel para capturar padrões textuais locais, como expressões ofensivas ou combinações de palavras associadas ao discurso de ódio.
   - Após a convolução, camadas de pooling são aplicadas para reduzir a dimensionalidade, preservando as características mais relevantes.

(3) Classificação Multilabel com Camadas Densas:
   - As características extraídas pelas CNNs são passadas por camadas densas (fully connected) para realizar a classificação final. A última camada utiliza uma função de ativação sigmoide para permitir a classificação multilabel, onde cada texto pode pertencer a múltiplas categorias de discurso de ódio.
   - A saída do modelo é uma probabilidade para cada categoria de discurso de ódio, permitindo que o modelo identifique e classifique textos com base em múltiplas categorias simultaneamente.

(4) Classificação Binária com Camadas Densas:
   - Para a classificação binária (ódio ou não ódio), uma camada densa adicional é adicionada após as camadas convolucionais, utilizando uma função de ativação sigmoide para produzir uma probabilidade binária.
   - A saída do modelo é uma probabilidade de que o texto pertença à classe de discurso de ódio, permitindo uma análise mais direta e eficiente.

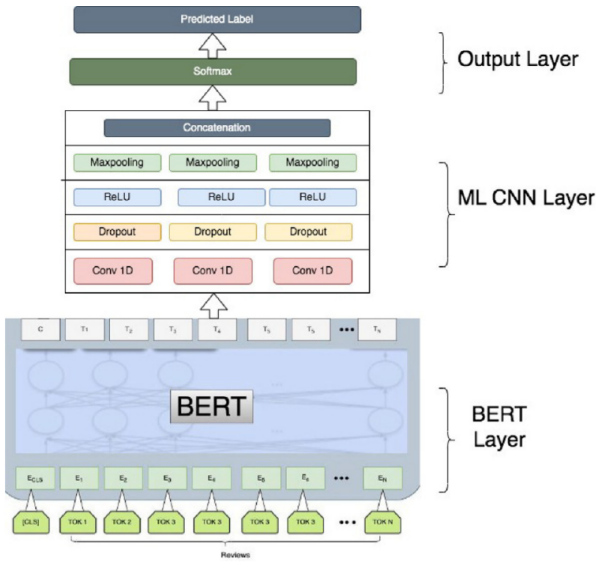


**Figure 1: Framework do BERT Multi-Camadas de Rede Neural Convolucional (MCNN) [1].**

A figura 1 ilustra o framework do modelo BERT-MCNN originalmente proposto por [1], destacando as etapas de extração de embeddings, convolução e classificação. Em nosso trabalho, adaptamos essa arquitetura para a classificação multilabel e binária, utilizando como função de ativação a sigmoide na camada de saída para permitir a classificação de múltiplas categorias de discurso de ódio ou posicionamento.

### 3.2 Parâmetros do modelo

Os parâmetros do modelo BERT-MCNN foram ajustados para otimizar o desempenho na tarefa de detecção de discurso de ódio. A seguir, detalhamos os principais parâmetros utilizados:

- **Tamanho do Embedding**: Utilizamos embeddings de 768 dimensões, correspondentes ao modelo BERTimbau Base.
- **Tamanho do Kernel**: Empregamos tamanhos de kernel variados (4, 6 e 8) para capturar diferentes n-gramas e padrões textuais.
- **Número de Filtros**: Cada camada convolucional possui 128 filtros, permitindo a extração de características ricas dos embeddings.

- **Função de Ativação**: Utilizamos a função ReLU nas camadas convolucionais e sigmoide na camada de saída para classificação multilabel.
- **Dropout**: Aplicamos uma taxa de dropout de 0.5 para evitar overfitting durante o treinamento.
- **Otimização**: O modelo é otimizado usando o algoritmo AdamW, com uma taxa de aprendizado inicial de 5e-6 para as camadas do BERT e 5e-5 para as camadas convolucionais, além de uma taxa de decaimento de 0.01.
- **Batch Size**: Utilizamos um tamanho de batch de 8, adequado para o treinamento eficiente do modelo.
- **Número de Épocas**: O modelo é treinado por um número máximo de 50 épocas, com Early Stopping aplicado após 10 épocas sem melhoria na validação.

A configuração dos parâmetros foi escolhida com base em experimentos preliminares e na literatura existente, utilizando como base a arquitetura BERT-MCNN proposta por [1], visando maximizar a performance do modelo na detecção de discurso de ódio em português.

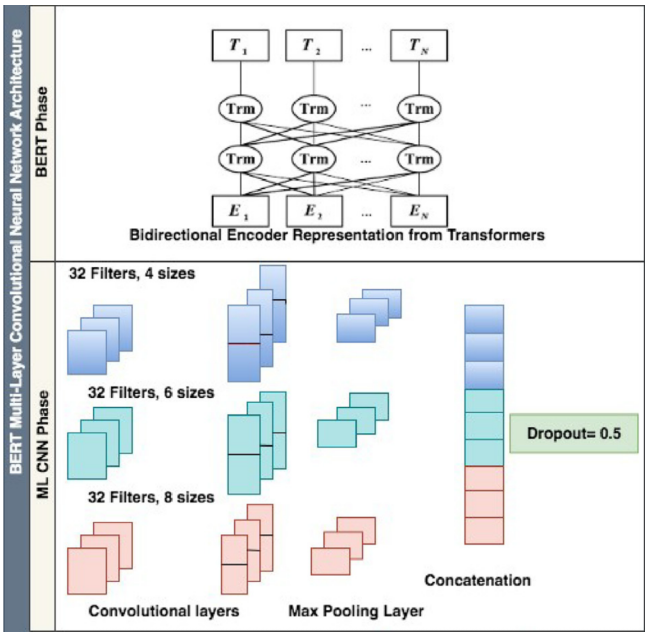


**Figure 2: Arquitetura do modelo BERT-MCNN.**

## 3.3 Datasets

O TuPyE contém 43.668 textos anotados com múltiplas categorias de discurso de ódio, permitindo análises binárias (ódio ou não) e multilabel (múltiplas categorias simultâneas). As etapas de preparação do dataset incluem:

- Divisão dos Dados: Os dados foram divididos em 80% para treinamento, 20% para teste.
- Pré-processamento:
  - Tokenização com o tokenizador do BERTimbau.
  - Conversão para minúsculas e remoção de caracteres especiais desnecessários.
  - Remoção de links, menções, e emojis, mantendo apenas o texto relevante.

## 3.4 Configuração do Treinamento

- Função de Perda: Utilizamos entropia cruzada binária para classificação multilabel.
- Otimização: O algoritmo AdamW foi empregado, com taxa de aprendizado inicial de 5e-6 para as camadas do BERT e 5e-5 para as camadas convolucionais, bem como uma taxa de decaimento de 0.01.
- Hiperparâmetros:
  - Batch Size: 8
  - Número de Épocas: Aplicado Early Stopping com paciência de 10 épocas
  - Taxa de Dropout: 0.5

## 3.5 Métricas de avaliação do modelo

Para avaliar o desempenho do modelo BERT-MCNN, utilizamos as seguintes métricas:

- **Acurácia**: $\frac{VP+VN}{VP+VN+FP+FN}$ - Proporção geral de classificações corretas
- **Precisão**: $\frac{VP}{VP+FP}$ - Medida de exatidão das classificações positivas
- **Recall**: $\frac{VP}{VP+FN}$ - Capacidade de identificação de casos positivos
- **F1-Score**: $\frac{2 \cdot Precisão \cdot Recall}{Precisão+Recall}$ - Média harmônica entre precisão e recall

# 4 Resultados

Os resultados do modelo BERTimbau-MLCNN foram avaliados em dois datasets: o HATE-BR e o TuPyE. A seguir, apresentamos os resultados obtidos para cada um desses conjuntos de dados.

## 4.1 Resultados do HATE-BR

## 4.2 TuPyE

Os resultados do modelo BERTimbau para a classificação multilabel no TuPyE são apresentados na tabela 1. Os resultados mostram que o modelo BERTimba alcançou uma precisão de samples de 0.86, recall de 0.85 e F1-Score de 0.85 para o modelo Base, enquanto o modelo BERTimbau Large alcançou uma precisão de samples de 0.87, recall de 0.86 e F1-Score de 0.86.

| Categoria | Precisão | Recall | F1-Score | Suporte |
|---|---|---|---|---|
| **BERTimbau Base** | | | | |
| Ageism | 1.00 | 0.00 | 0.00 | 15 |
| Aporophobia | 1.00 | 0.00 | 0.00 | 16 |
| Body Shame | 0.58 | 0.54 | 0.56 | 54 |
| Capacitism | 1.00 | 0.00 | 0.00 | 20 |
| Lgbtphobia | 0.85 | 0.67 | 0.75 | 171 |
| Political | 0.59 | 0.56 | 0.58 | 220 |
| Racism | 0.29 | 0.27 | 0.28 | 62 |
| Religious Intolerance | 0.25 | 0.11 | 0.15 | 19 |
| Misogyny | 0.65 | 0.60 | 0.62 | 324 |
| Xenophobia | 0.41 | 0.31 | 0.35 | 78 |
| Other | 0.56 | 0.49 | 0.52 | 909 |
| Not Hate | 0.92 | 0.93 | 0.92 | 7177 |
| Micro Avg | 0.86 | 0.84 | 0.85 | 9065 |
| Macro Avg | 0.67 | 0.37 | 0.39 | 9065 |
| Weighted Avg | 0.85 | 0.84 | 0.84 | 9065 |
| Samples Avg | 0.86 | 0.85 | 0.85 | 9065 |
| **BERTimbau Large** | | | | |
| Ageism | 0.40 | 0.13 | 0.20 | 15 |
| Aporophobia | 0.75 | 0.19 | 0.30 | 16 |
| Body Shame | 0.78 | 0.65 | 0.71 | 54 |
| Capacitism | 0.50 | 0.15 | 0.23 | 20 |
| Lgbtphobia | 0.78 | 0.75 | 0.76 | 171 |
| Political | 0.61 | 0.53 | 0.57 | 220 |
| Racism | 0.39 | 0.42 | 0.40 | 62 |
| Religious Intolerance | 0.27 | 0.16 | 0.20 | 19 |
| Misogyny | 0.67 | 0.63 | 0.65 | 324 |
| Xenophobia | 0.39 | 0.22 | 0.28 | 78 |
| Other | 0.62 | 0.46 | 0.53 | 909 |
| Not Hate | 0.91 | 0.94 | 0.93 | 7177 |
| Micro Avg | 0.87 | 0.85 | 0.86 | 9065 |
| Macro Avg | 0.59 | 0.44 | 0.48 | 9065 |
| Weighted Avg | 0.85 | 0.85 | 0.85 | 9065 |
| Samples Avg | 0.87 | 0.86 | 0.86 | 9065 |

**Table 1: Resultados Originais do Modelo BERTimbau para Classificação Multilabel no TuPyE [7]**

| Categoria | Precisão | Recall | F1-Score | Suporte |
|---|---|---|---|---|
| **BERTimbau-MLCNN Base** | | | | |
| Ageism | 0.00 | 0.00 | 0.00 | 6 |
| Aporophobia | 0.00 | 0.00 | 0.00 | 9 |
| Body Shame | 0.63 | 0.61 | 0.62 | 28 |
| Capacitism | 0.00 | 0.00 | 0.00 | 4 |
| Lgbtphobia | 0.75 | 0.69 | 0.72 | 77 |
| Political | 0.55 | 0.44 | 0.49 | 117 |
| Racism | 0.50 | 0.32 | 0.39 | 28 |
| Religious Intolerance | 0.50 | 0.13 | 0.20 | 8 |
| Misogyny | 0.71 | 0.52 | 0.60 | 177 |
| Xenophobia | 0.67 | 0.32 | 0.43 | 44 |
| Other | 0.52 | 0.50 | 0.51 | 449 |
| Not Hate | 0.94 | 0.95 | 0.95 | 3815 |
| Micro Avg | 0.88 | 0.86 | 0.87 | 4762 |
| Macro Avg | 0.48 | 0.37 | 0.41 | 4762 |
| Weighted Avg | 0.87 | 0.86 | 0.86 | 4762 |
| Samples Avg | 0.89 | 0.89 | 0.88 | 4762 |
| **BERTimbau-MLCNN Large** | | | | |
| Ageism | 0.00 | 0.00 | 0.00 | 6 |
| Aporophobia | 0.00 | 0.00 | 0.00 | 9 |
| Body Shame | 0.82 | 0.82 | 0.82 | 28 |
| Capacitism | 0.00 | 0.00 | 0.00 | 4 |
| Lgbtphobia | 0.87 | 0.88 | 0.88 | 77 |
| Political | 0.81 | 0.71 | 0.75 | 117 |
| Racism | 0.75 | 0.75 | 0.75 | 28 |
| Religious Intolerance | 1.00 | 0.50 | 0.67 | 8 |
| Misogyny | 0.87 | 0.79 | 0.83 | 177 |
| Xenophobia | 0.96 | 0.57 | 0.71 | 44 |
| Other | 0.84 | 0.77 | 0.80 | 449 |
| Not Hate | 0.96 | 0.96 | 0.96 | 3815 |
| Micro Avg | 0.94 | 0.92 | 0.93 | 4762 |
| Macro Avg | 0.66 | 0.56 | 0.60 | 4762 |
| Weighted Avg | 0.94 | 0.92 | 0.93 | 4762 |
| Samples Avg | 0.95 | 0.94 | 0.94 | 4762 |

**Table 2: Resultados do Modelo BERTimbau-MLCNN para Classificação Multilabel no TuPyE**

Os resultados do modelo BERTimbau-MLCNN para a classificação multilabel no TuPyE são apresentados na tabela 2. Os resultados mostram que o modelo BERTimbau-MLCNN Base alcançou uma precisão de samples de 0.88, recall de 0.86 e F1-Score de 0.87, enquanto o modelo BERTimbau-MLCNN Large alcançou uma precisão de samples de 0.95, recall de 0.94 e F1-Score de 0.94, indicando uma melhora significativa em relação ao modelo BERTimbau original.

## 5 Conclusão

Neste trabalho, apresentamos uma abordagem integrada para a análise de sentimentos e detecção de discurso de ódio utilizando o BERT com uma Rede Neural Convolucional Multi-Camadas (MLCNN). Nossa metodologia combina a capacidade do BERT de capturar relações contextuais de longo alcance com a eficiência das CNNs na extração de características locais. Os resultados experimentais demonstraram que essa integração aumenta significativamente

a performance do modelo em termos de acurácia, precisão, recall e F1-score.

Os resultados obtidos validam a proposta de integrar modelos baseados em BERT com CNNs, ressaltando que essa estratégia captura tanto contextos de longo alcance quanto padrões locais importantes no texto. Isso se traduziu em uma melhora notável nas métricas de avaliação quando comparado a architectures individuais de BERT ou CNN.

## References


[1] Peter Atandoh, Fengli Zhang, Daniel Adu-Gyamfi, Paul H. Atandoh, and Raphael Elimeli Nuhoho. 2023. Integrated deep learning paradigm for document-based sentiment analysis. *Journal of King Saud University - Computer and Information Sciences* 35, 7 (2023), 101578. https://doi.org/10.1016/j.jksuci.2023.101578

[2] Pablo Botton Costa, Matheus Camasmie Pavan, Wesley Ramos Santos, Samuel Caetano Silva, and Ivandr'e Paraboni. 2023. BERTabaporu: Assessing a Genre-Specific Language Model for Portuguese NLP. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, Ruslan Mitkov and Galia Angelova (Eds.). INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 217–223. https://aclanthology.org/2023.ranlp-1.24

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[4] Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A Hierarchically-Labeled Portuguese Hate Speech Dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem (Eds.). Association for Computational Linguistics, Florence, Italy, 94–104. https://doi.org/10.18653/v1/W19-3510

[5] João A. Leite, Diego F. Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. arXiv:2010.04543 [cs.CL] https://arxiv.org/abs/2010.04543

[6] Dyonnatan F Maia, Nádia FF Silva, Ellen PR Souza, Augusto S Nunes, Lucas C Procópio, Guthemberg da S Sampaio, Márcio de S Dias, Adrio O Alves, Dyéssica F Maia, Ingrid A Ribeiro, et al. 2022. UlyssesSD-Br: Stance Detection in Brazilian Political Polls. In *EPIA Conference on Artificial Intelligence*. Springer, 85–95.

[7] Felipe Oliveira, Victoria Reis, and Nelson Ebecken. 2023. TuPy-E: detecting hate speech in Brazilian Portuguese social media with a novel dataset and comprehensive analysis of models. *arXiv preprint arXiv:2312.17704* (2023).

[8] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Lun-Wei Ku and Cheng-Te Li (Eds.). Association for Computational Linguistics, Valencia, Spain, 1–10. https://doi.org/10.18653/v1/W17-1101

[9] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

[10] Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 7174–7183. https://aclanthology.org/2022.lrec-1.777