# Modelos Lexicais Calibrados para Detecção de Ofensividade em Português Brasileiro: um Estudo no ToLD-Br

**Iago de Sousa Aragão, Gabrielly Alves Gomes, Patricia Medyna Lauritzen de Drumond**

Universidade Federal do Piauí (UFPI), Teresina, PI, Brasil

{iago.aragao, gabrielly.gomes, patriciamedyna}@ufpi.edu.br

**Abstract.** This paper presents an experimental study on binary offensive language detection in Brazilian Portuguese using the ToLD-Br dataset. We evaluate lightweight lexical models based on word and character TF-IDF, threshold calibration, and a score reweighting strategy based on differential evidence from interpretable text perturbations. Results show that calibrated lexical models are competitive baselines for this task: the TF-IDF baseline reaches 0.7525 Macro-F1, while the reweighting strategy reaches 0.7531 Macro-F1 and 0.8348 ROC-AUC. Because the gain over the calibrated baseline is marginal and was not subjected to formal significance testing, the main contribution is to characterize simple, inexpensive, reproducible, and interpretable lexical baselines for this benchmark.

**Resumo.** Este artigo apresenta um estudo experimental sobre detecção binária de ofensividade em português brasileiro no conjunto ToLD-Br. Avaliamos modelos lexicais leves baseados em TF-IDF de palavras e caracteres, calibração de limiar e uma estratégia de reponderação de escore por evidência diferencial obtida por perturbações interpretáveis do texto. Os resultados mostram que modelos lexicais calibrados são baselines competitivos para a tarefa: a baseline TF-IDF alcança Macro-F1 de 0,7525, enquanto a reponderação atinge Macro-F1 de 0,7531 e ROC-AUC de 0,8348. Como o ganho sobre a baseline calibrada é marginal e não foi submetido a teste formal de significância, a principal contribuição é caracterizar baselines lexicais simples, baratos, reprodutíveis e interpretáveis para esse benchmark.

## 1. Introdução

A disseminação de linguagem ofensiva, tóxica e discriminatória em mídias sociais tem motivado o desenvolvimento de métodos automáticos para moderação e análise de conteúdo. No português brasileiro, a tarefa é dificultada por variação linguística, gírias, abreviações, ironia, erros ortográficos e formas indiretas de agressão. Embora modelos baseados em Transformers apresentem resultados expressivos em várias tarefas de Processamento de Linguagem Natural, eles nem sempre são a alternativa mais adequada em cenários de baixo custo computacional, infraestrutura limitada ou necessidade de maior interpretabilidade.

O ToLD-Br foi proposto para detecção de toxicidade em tweets em português brasileiro, contemplando formulações binária e multilabel [Leite et al. 2020]. Como as categorias multilabel são fortemente desbalanceadas, com baixa frequência de classes como racismo, xenofobia, misoginia e homofobia, este artigo concentra-se na formulação binária, isto é, na distinção entre textos ofensivos e não ofensivos.

Resultados neurais reportados anteriormente são usados apenas como contexto de literatura, não como comparação experimental direta.

Este trabalho investiga a seguinte questão de pesquisa: até que ponto modelos lexicais leves, com calibração de limiar e reponderação por evidência textual, conseguem detectar ofensividade em português brasileiro de forma competitiva, sem recorrer a embeddings pré-treinados ou Transformers? As contribuições são: (i) uma avaliação experimental de modelos lexicais para o ToLD-Br binário; (ii) uma comparação entre TF-IDF de palavras e caracteres, atributos interpretáveis e reponderação por evidência diferencial; e (iii) uma discussão sobre o papel de modelos simples, calibrados e reprodutíveis em tarefas de toxicidade em português brasileiro.

## 2. Trabalhos relacionados

A detecção automática de linguagem abusiva e ofensiva tem sido estudada sob diferentes formulações, incluindo discurso de ódio, toxicidade, abuso verbal e ofensa. Nobata et al. (2016) investigaram métodos para detecção de linguagem abusiva em conteúdo online, explorando recursos linguísticos e estatísticos. Davidson et al. (2017) discutiram a dificuldade de separar discurso de ódio, linguagem ofensiva e textos neutros, mostrando que a definição da tarefa influencia diretamente o comportamento dos classificadores. Revisões da área também destacam que n-gramas e classificadores lineares permanecem baselines relevantes e devem ser considerados em comparações com modelos neurais [Fortuna and Nunes 2018].

Para o português brasileiro, Leite et al. (2020) propuseram o ToLD-Br, um dataset de tweets anotados para toxicidade em diferentes categorias. O trabalho destaca a dificuldade da classificação multilabel em função do desbalanceamento e da menor frequência de algumas categorias. Outros recursos, como o HateBR, também evidenciam a relevância da detecção de linguagem ofensiva em redes sociais brasileiras, incluindo classificação binária e níveis de ofensividade [Vargas et al. 2022].

Apesar do avanço de modelos neurais, abordagens lexicais continuam relevantes em tarefas de texto, especialmente quando combinam n-gramas de palavras e caracteres. N-gramas de caracteres são úteis em mídias sociais por capturarem variações ortográficas, repetições de letras e formas mascaradas de ofensa. Neste trabalho, avaliamos se uma configuração lexical leve, calibrada e interpretável pode alcançar desempenho competitivo no ToLD-Br binário.

Nesse contexto, este artigo se posiciona como uma avaliação experimental de baselines lexicais calibradas, e não como proposta de uma arquitetura neural ou de um novo classificador de ampla generalização. A ênfase está em custo computacional, reprodutibilidade e interpretação dos efeitos de calibração e reponderação em um benchmark específico.

## 3. Metodologia

### 3.1. Dataset e tarefa

Foi utilizado o dataset ToLD-Br, composto por 21.000 tweets em português brasileiro anotados quanto à presença de categorias de toxicidade. As categorias consideradas são homofobia, obscenidade, insulto, racismo, misoginia e xenofobia. A tarefa binária foi construída a partir das categorias multilabel: um texto foi considerado ofensivo quando

ao menos uma das categorias recebeu valor positivo, e não ofensivo quando todas as categorias estavam ausentes.

Após a transformação, a distribuição observada foi de 11.745 textos não ofensivos e 9.255 textos ofensivos. O pré-processamento foi leve: URLs, menções e espaços foram normalizados; hashtags tiveram apenas o símbolo removido; palavrões, pontuação expressiva, risos e variações ortográficas foram preservados por serem sinais potencialmente relevantes para a tarefa.

### 3.2. Modelos avaliados

Foram avaliadas três configurações, implementadas com scikit-learn [Pedregosa et al. 2011]. B1 combina TF-IDF de palavras e TF-IDF de caracteres com Regressão Logística. Para palavras, usamos n-gramas de 1 a 3; para caracteres, n-gramas de 3 a 5. A Regressão Logística foi treinada com class_weight="balanced".

B2 acrescenta atributos linguístico-discursivos simples ao TF-IDF, incluindo presença de termos agressivos genéricos, termos de violência, pronomes de alvo, marcadores de generalização, menções a grupos sociais, risos, exclamações e proporção de letras maiúsculas. Esses atributos foram concebidos como recursos interpretáveis auxiliares, não como substitutos da representação lexical.

B3 parte da baseline B1 e adiciona uma reponderação leve do escore de ofensividade. Para isso, são geradas perturbações interpretáveis do texto, substituindo evidências como termos agressivos, termos de violência, alvos, generalizações e menções a grupos sociais. O modelo base é aplicado ao texto original e às versões perturbadas; a diferença entre as probabilidades é usada como sinal de evidência diferencial. Também foi incorporado um sinal de memória lexical contrastiva, calculado pela similaridade TF-IDF entre a amostra de validação e exemplos ofensivos e não ofensivos difíceis do conjunto de treino. O ajuste final é aplicado de forma restrita a casos próximos ao limiar; na melhor configuração, escore_final = escore_base + ajuste_por_evidência, com alpha=0,20, janela=0,02 e limiar=0,48.

### 3.3. Protocolo experimental

Os experimentos foram conduzidos com validação cruzada estratificada em 5 folds. Em cada fold, os vetorizadores TF-IDF foram ajustados apenas no conjunto de treino e aplicados ao conjunto de validação. As predições fora da amostra foram acumuladas para o cálculo das métricas finais. Avaliamos acurácia, Macro-F1, F1 da classe ofensiva, F1 da classe não ofensiva, precisão, revocação e ROC-AUC. Macro-F1 corresponde à média não ponderada do F1 das classes, sendo adequada para cenários desbalanceados por atribuir o mesmo peso às classes ofensiva e não ofensiva. O limiar de decisão foi ajustado por busca em grade priorizando Macro-F1 nas predições fora da amostra acumuladas; por isso, esse ajuste é tratado como exploração de calibração e discutido nas ameaças à validade.

## 4. Resultados

A Tabela 1 apresenta os principais resultados obtidos na tarefa binária. Pelo critério principal de Macro-F1, o maior valor observado foi obtido por B3, com Macro-F1 de 0,7531, acurácia de 0,7548, F1 da classe ofensiva de 0,7325 e ROC-AUC de 0,8348. A matriz de confusão desse modelo apresentou 7.050 verdadeiros positivos, 8.801

verdadeiros negativos, 2.944 falsos positivos e 2.205 falsos negativos. O ganho em relação a B1 calibrado foi pequeno, de 0,0006 ponto de Macro-F1, e não foi submetido a teste formal de significância; portanto, B3 deve ser interpretado como refinamento exploratório, não como superioridade robusta.

**Tabela 1. Resultados na detecção binária de ofensividade no ToLD-Br**

| Modelo | Limiar | Acurácia | Macro-F1 | F1 ofensivo | Prec. of. | Rev. of. |
|---|---|---|---|---|---|---|
| B1 - TF-IDF | 0,50 | 0,7543 | 0,7513 | 0,7236 | 0,7177 | 0,7296 |
| B1 - TF-IDF + limiar | 0,48 | 0,7545 | 0,7525 | 0,7304 | 0,7076 | 0,7547 |
| B2 - TF-IDF + atributos | 0,50 | 0,7537 | 0,7504 | 0,7219 | 0,7183 | 0,7256 |
| B2 - atributos + limiar | 0,47 | 0,7534 | 0,7518 | 0,7314 | 0,7034 | 0,7616 |
| B3 - ToxDx | 0,48 | 0,7548 | 0,7531 | 0,7325 | 0,7054 | 0,7618 |

## 5. Discussão

Os resultados indicam que modelos lexicais calibrados constituem baselines fortes para a detecção de ofensividade em português brasileiro. O modelo B1, baseado apenas em TF-IDF de palavras e caracteres, alcançou Macro-F1 de 0,7525 após ajuste de limiar. Esse resultado reforça a importância de baselines simples e bem calibradas antes da adoção de modelos mais custosos; comparações com resultados neurais reportados na literatura devem ser interpretadas apenas como contexto, pois os protocolos de divisão, ajuste e avaliação não são idênticos.

A inclusão de atributos interpretáveis no modelo B2 não trouxe ganho claro em relação ao TF-IDF puro. Isso sugere que, no cenário binário, os n-gramas de palavras e caracteres já capturam boa parte dos sinais explícitos de ofensividade presentes no dataset. Ainda assim, esses atributos permanecem úteis para análise qualitativa e interpretação dos fatores linguísticos associados às predições.

A estratégia de reponderação por evidência diferencial, avaliada no B3, obteve o maior Macro-F1 observado, mas o ganho sobre B1 calibrado foi de apenas 0,0006 e não permite afirmar superioridade estatística. O principal efeito observado foi o aumento da revocação da classe ofensiva, de 0,7547 para 0,7618, acompanhado de pequena redução na precisão ofensiva. Assim, a reponderação tornou o classificador ligeiramente mais sensível a textos ofensivos sem alterar substancialmente o desempenho global.

Esse comportamento sugere que a evidência diferencial é mais adequada como mecanismo auxiliar e exploratório do que como substituição do classificador lexical. O classificador TF-IDF permanece como núcleo do sistema, enquanto os sinais baseados em perturbações e memória lexical atuam como refinamento de casos de fronteira. Uma análise estatística pareada e avaliações em outros datasets são necessárias para verificar se esse refinamento se mantém fora do ToLD-Br.

## 6. Ameaças à validade

Este estudo possui limitações. Primeiro, os resultados foram obtidos em validação cruzada, e não em uma partição oficial de teste idêntica à utilizada por trabalhos de referência; por isso, comparações com resultados neurais da literatura devem ser

interpretadas com cautela. Segundo, a busca de limiar e de hiperparâmetros da reponderação foi realizada sobre predições fora da amostra acumuladas, o que pode produzir uma estimativa ligeiramente otimista. Terceiro, não foram conduzidos testes estatísticos pareados; assim, diferenças pequenas, como o ganho de B3 sobre B1 calibrado, não devem ser interpretadas como evidência de superioridade robusta. Uma versão futura deve separar treino, validação e teste ou realizar calibração em uma validação interna.

Além disso, os atributos interpretáveis e perturbações textuais foram definidos manualmente, podendo não cobrir toda a diversidade linguística do português brasileiro. Por fim, os experimentos foram conduzidos apenas no ToLD-Br; portanto, ainda não é possível afirmar que os achados generalizam para outros domínios ou conjuntos, como HateBR, OLID-BR, comentários de Instagram, fóruns online ou outras redes sociais.

## 7. Conclusão

Este trabalho apresentou um estudo experimental de modelos lexicais leves para detecção binária de ofensividade em português brasileiro no ToLD-Br. Em validação cruzada estratificada, a combinação de TF-IDF de palavras e caracteres com Regressão Logística e calibração de limiar alcançou desempenho forte, com Macro-F1 de 0,7525. A estratégia adicional de reponderação por evidência diferencial obteve o maior Macro-F1 observado, com 0,7531, F1 ofensivo de 0,7325 e ROC-AUC de 0,8348, mas seu ganho sobre a baseline calibrada é marginal e deve ser tratado como exploratório.

Embora o ganho sobre a baseline lexical seja marginal, o resultado sugere que sinais interpretáveis derivados de perturbações textuais e memória lexical podem auxiliar a decisão em casos de fronteira. O principal achado, contudo, é que modelos lexicais calibrados continuam competitivos para a detecção de ofensividade em português brasileiro, especialmente quando se busca baixo custo computacional, reprodutibilidade e interpretabilidade. Trabalhos futuros devem incluir testes estatísticos pareados, análise de erros mais ampla e avaliação em múltiplos datasets.

## Declaração de uso de IA generativa

Ferramentas de IA generativa foram utilizadas como apoio à organização textual, estruturação da metodologia e revisão linguística deste manuscrito. A responsabilidade integral pelos experimentos, interpretação dos resultados, verificação bibliográfica e conteúdo final permanece com os autores humanos.

## Referências

Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of the International AAAI Conference on Web and Social Media.

Fortuna, P. and Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys, 51(4), pages 1-30.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT.

Leite, J. A., Silva, D. F., Bontcheva, K. and Scarton, C. (2020). Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In Proceedings of AACL-IJCNLP, pages 914-924.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y. and Chang, Y. (2016). Abusive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pages 2825-2830.

Vargas, F. A., Carvalho, I., Rodrigues de Góes, F., Pardo, T. A. S. and Benevenuto, F. (2022). HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. In Proceedings of LREC, pages 7174-7183.

Wang, S. and Manning, C. D. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In Proceedings of ACL.