ISSN: 2358-8829

DETECÇÃO AUTOMÁTICA DE COMENTÁRIOS TÓXICOS EM LÍNGUA PORTUGUESA: UMA ABORDAGEM BASEADA EM PLN

Jorge Miguel Viana Torres [1]
Murilo Gabriel da Silva Mourão [2]
Gabriel de Assis Buhatem [3]
Wendel Brito Silva [4]
Erick Macgregor Santos Lima [5]

INTRODUÇÃO

A intensificação de interações nas redes sociais ampliou a ocorrência de linguagem abusiva, assédio e incivilidade on-line, com impactos sobre bem-estar e participação cívica. No campo técnico, a detecção automática de toxicidade é formulada como tarefa de classificação de texto em PLN, historicamente apoiada em representações esparsas como bag-of-words e TF-IDF combinadas a classificadores lineares ou probabilísticos (MANNING; RAGHAVAN; SCHÜTZE, 2008; JURAFSKY; MARTIN, 2023). Mais recentemente, modelos contextuais do tipo transformer (por exemplo, BERTimbau) passaram a capturar nuances semânticas, gírias e polissemia com maior sensibilidade (SOUZA; NOGUEIRA; LOTUFO, 2020). Do ponto de vista ético-jurídico, a literatura adverte para transparência, mitigação de vieses e prestação de contas em moderação automatizada, a fim de compatibilizar o enfrentamento de abusos com a liberdade de expressão (FIGUEIREDO; SILVA, 2022).

Este trabalho apresenta um pipeline reprodutível para detecção binária de comentários tóxicos vs. não tóxicos em português do Brasil, concebido como baseline didático para iniciação científica e cursos técnicos. Justifica-se pela necessidade de soluções simples, transparentes e comparáveis, úteis tanto para formação quanto para evolução incremental com modelos mais avançados. Tem como objetivo geral construir e avaliar um classificador supervisionado de toxicidade em português. E como objetivos específicos: (i) realizar pré-processamento (limpeza, normalização, tokenização, remoção de stopwords e stemming RSLP); (ii) vetorização por TF-IDF; (iii) treinamento com Multinomial Naive Bayes; (iv) avaliação por acurácia, precisão, revocação e F1-score;

[1] Cursando Técnico em Informática no IEMA - MA, jorgeditor.social@gmail.com;
[2] Cursando Técnico em Informática no IEMA - MA, mouraomurilo7@gmail.com;
[3] Cursando Técnico em Informática no IEMA - MA, gabrielbuhatem@gmail.com;
[4] Cursando Técnico em Informática no IEMA - MA, wendelbrito589987@gmail.com;
[5] **Orientador:** Mestrando em Computação UFPI e Professor do IEMA, erickmacgregor2_@hotmail.com

ISSN: 2358-8829

(v) discutir limitações e caminhos de melhoria com embeddings contextuais (MANNING; RAGHAVAN; SCHÜTZE, 2008; JURAFSKY; MARTIN, 2023; SOUZA; NOGUEIRA; LOTUFO, 2020). Em síntese metodológica, adotou-se particionamento hold-out 80/20 e interpretação por matriz de confusão.

O modelo alcançou acurácia de 72% no conjunto de teste, com precisão satisfatória para a classe "Tóxico" e revocação inferior, indicando perda de casos sutis (ironia, eufemismos). A discussão relaciona esse padrão às limitações de representações esparsas, sugerindo como próximos passos BERTimbau e técnicas de balanceamento para elevar o recall da classe minoritária, além de tuning e validação cruzada. Em termos conclusivos, o pipeline proposto cumpre o papel de baseline transparente e pedagógico, servindo de plataforma para aprimoramentos técnicos e para uso responsável em contextos educacionais e institucionais, com atenção a riscos, vieses e salvaguardas normativas (FIGUEIREDO; SILVA, 2022; JURAFSKY; MARTIN, 2023)..

METODOLOGIA

A pesquisa caracteriza-se como aplicada, quantitativa e experimental, com foco na implementação de um pipeline de Processamento de Linguagem Natural (PLN) para classificação binária de comentários tóxicos e não tóxicos em língua portuguesa. As etapas seguiram uma abordagem sequencial, reprodutível e ética, dividida em seis fases principais: preparação do corpus, pré-processamento, vetorização, divisão de dados, treinamento do modelo e avaliação de desempenho.

1. Preparação e Análise do Corpus

O conjunto de dados utilizado neste estudo é o "Comentários Tóxicos PT-BR", um *dataset* público em língua portuguesa formado por comentários coletados e compilados a partir de diversas bases anteriores de toxicidade textual. A base contém 16.412 comentários não tóxicos e 13.510 comentários tóxicos, totalizando 29.922 registros rotulados. Cada instância é composta por duas colunas: text (comentário em formato de texto) e toxic (rótulo binário, sendo 1 para tóxico e 0 para não tóxico). Segundo a documentação da base, os dados foram extraídos de *datasets* multilíngues e adaptados para o português, preservando apenas conteúdos públicos e anonimizados.

Antes da modelagem, realizou-se uma verificação de integridade para detectar valores nulos, duplicados e inconsistências de codificação de caracteres, assegurando a limpeza inicial do corpus. Em seguida, foi conduzida uma análise exploratória da distribuição de classes, que revelou leve desbalanceamento entre as categorias

ISSN: 2358-8829

(aproximadamente 55% não tóxicos e 45% tóxicos). Esse perfil foi mantido nas etapas seguintes, visto que o objetivo principal era construir um modelo baseline educacional, permitindo reprodutibilidade e comparabilidade com futuras versões aprimoradas do pipeline.

2. Pré-processamento Textual

O pré-processamento visa padronizar e limpar o texto para otimizar o desempenho dos algoritmos de aprendizado. Foram aplicadas as seguintes operações, conforme a literatura de PLN (MANNING; RAGHAVAN; SCHÜTZE, 2008):

1. **Limpeza de caracteres e números:** remoção de dígitos, URLs e símbolos não alfabéticos.
2. **Normalização:** conversão integral dos textos para letras minúsculas.
3. **Tokenização:** segmentação dos textos em unidades linguísticas (*tokens*).
4. **Remoção de *stopwords*:** exclusão de palavras muito frequentes e semanticamente neutras (por exemplo, “de”, “para”, “com”).
5. **Stemming:** redução das palavras aos seus radicais por meio do algoritmo **RSLP (Removedor de Sufixos da Língua Portuguesa)**, agrupando variações morfológicas.
6. **Reconstrução:** recomposição dos tokens processados em frases normalizadas, preparando-as para vetorização.

Essas etapas garantiram um vocabulário mais limpo, eliminando ruídos linguísticos e reduzindo a dimensionalidade do texto.

3. Representação Vetorial dos Textos

Para converter as frases em dados numéricos compreensíveis por algoritmos, aplicou-se a técnica TF-IDF (Term Frequency–Inverse Document Frequency), amplamente utilizada em classificação de textos (MANNING; RAGHAVAN; SCHÜTZE, 2008). O modelo gerou uma matriz de 29.922 amostras × 5.000 atributos, na qual cada valor representa a importância relativa de um termo no corpus. Palavras comuns receberam pesos baixos, enquanto termos discriminativos, como ofensas, receberam pesos altos.

4. Divisão do Conjunto de Dados

Os dados foram divididos segundo a estratégia hold-out, separando 80% para treinamento e 20% para teste, garantindo uma avaliação independente da etapa de aprendizagem. Essa divisão busca verificar se o modelo consegue generalizar para novos exemplos sem sobreajuste (*overfitting*).

ISSN: 2358-8829

5. Treinamento do Modelo

Para o aprendizado supervisionado, foi escolhido o algoritmo Naive Bayes Multinomial, devido à sua eficiência, simplicidade e bons resultados em tarefas de PLN (JURAFSKY; MARTIN, 2023). O modelo estima a probabilidade de um comentário pertencer a cada classe com base na frequência das palavras, assumindo independência condicional entre os termos. O treinamento foi realizado com as bibliotecas scikit-learn, pandas e nltk, amplamente reconhecidas pela comunidade científica. Após o ajuste, o modelo foi testado no conjunto de validação para medir o desempenho geral.

6. Avaliação de Desempenho

O desempenho foi mensurado pelas métricas clássicas de classificação:

- **Acurácia (Accuracy):** proporção total de previsões corretas.
- **Precisão (Precision):** proporção de comentários realmente tóxicos entre os classificados como tóxicos.
- **Revocação (Recall):** proporção de comentários tóxicos corretamente identificados.
- **F1-score:** média harmônica entre precisão e revocação, útil em casos de classes desbalanceadas.

O modelo apresentou acurácia de 72%, precisão de 73% e revocação de 59% na classe “tóxico”, conforme detalhado no relatório de classificação e matriz de confusão (dispostos no banner). Tais resultados indicam um bom desempenho geral, embora revelem sensibilidade limitada a ofensas sutis, aspecto que pode ser melhorado com modelos contextuais, como o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020).

7. Aspectos Éticos e Reprodutibilidade

A pesquisa utilizou dados abertos e anonimizados, sem identificação pessoal dos autores dos comentários. O estudo foi conduzido estritamente para fins educacionais e científicos, observando princípios éticos e de transparência algorítmica. O pipeline foi documentado passo a passo, permitindo reprodutibilidade total por outros estudantes e pesquisadores interessados.

REFERENCIAL TEÓRICO

O Processamento de Linguagem Natural (PLN) busca permitir que computadores compreendam e processem a linguagem humana. Técnicas clássicas como TF-IDF e algoritmos probabilísticos, a exemplo do Naive Bayes, têm sido amplamente

ISSN: 2358-8829

empregadas em tarefas de classificação de textos por sua eficiência e simplicidade (MANNING; RAGHAVAN; SCHÜTZE, 2008).

A detecção de comentários tóxicos é um problema de classificação supervisionada, cujo objetivo é identificar conteúdos ofensivos ou discriminatórios. No português, Silva et al. (2021) demonstraram que modelos baseados em TF-IDF alcançam bons resultados, mesmo diante das particularidades linguísticas da língua. Mais recentemente, modelos contextuais como o BERTimbau, adaptado ao português brasileiro, passaram a oferecer melhor desempenho em casos de ironia e ambiguidades (SOUZA; NOGUEIRA; LOTUFO, 2020).

Além dos aspectos técnicos, há também questões éticas associadas à moderação automatizada. Segundo Figueiredo e Silva (2022), é essencial equilibrar a detecção de discursos nocivos com a preservação da liberdade de expressão, garantindo transparência e responsabilidade no uso de Inteligência Artificial (IA).

Assim, este estudo apoia-se em fundamentos técnicos consolidados e em reflexões éticas recentes, propondo uma solução simples, reprodutível e socialmente responsável para o combate à toxicidade digital.

RESULTADOS E DISCUSSÃO

O pipeline proposto alcançou acurácia global de 72%, com precisão elevada para a classe “Tóxico” e revocação inferior, sugerindo perda de casos de linguagem ofensiva sutil. As métricas por classe e a matriz de confusão evidenciam maior sensibilidade a palavrões explícitos e menor a eufemismos/ironia, padrão típico de representações esparsas (TF-IDF). Esses resultados compõem um baseline reprodutível e didático para cursos técnicos e IC, útil como ponto de partida para modelos contextuais e ensembles.

CONSIDERAÇÕES FINAIS

O estudo demonstrou a viabilidade de um baseline simples e transparente para detecção de toxicidade em português, útil para formação técnica e para iteração futura com embeddings contextuais. As próximas etapas priorizam aumento do recall da classe “Tóxico”, mitigação de vieses e avaliação humana cega, conciliando efetividade técnica e direitos fundamentais (FIGUEIREDO; SILVA, 2022).

Palavras-chave: Processamento de Linguagem Natural; Classificação de Texto; Toxicidade; TF-IDF; Naive Bayes; BERTimbau.

ISSN: 2358-8829

AGRADECIMENTOS

Agradecemos ao professor Erick MacGregor Santos Lima, nosso orientador, por incentivar e possibilitar nossa entrada no mundo da pesquisa, e ao IEMA – IP Coelho Neto pela estrutura física e apoio institucional.

REFERÊNCIAS

FIGUEIREDO, M. A.; SILVA, J. C. Liberdade de expressão e moderação de conteúdo online: desafios éticos e jurídicos. **Revista de Direito Digital**, v. 7, n. 2, p. 89–104, 2022.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing. 3. ed., draft. 2023. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 20 out. 2025.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

SILVA, A. R. et al. Toxic comments detection in Brazilian Portuguese using NLP techniques. **Journal of Information and Data Management**, v. 12, n. 3, p . 45–58, 2021.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: **Portuguese Conference on Artificial Intelligence (EPIA)**, 2020. arXiv:2009.01961.

GEDORNETO. Comentários tóxicos PT-BR: Dataset com comentários tóxicos coletados de outros datasets. Kaggle. Disponível em: https://www.kaggle.com/datasets/gedorneto/comentrios-toxicos-ptbr. Acesso em: 20 out. 2025.