Universidade de Pernambuco
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Tese de Doutorado

Área: Inteligência Computacional

Título: Arquitetura de Adaptação Paramétrica Eficiente para Detecção Multirrótulo de Xenofobia Implícita em Textos em Português Brasileiro

Orientador: Profº Dr. Cleyton Mário de Oliveira Rodrigues

A xenofobia em ambientes digitais configura problema social e jurídico de crescente gravidade no Brasil, onde é tipificada como crime pela Lei nº 9.459/1997. Dados da SaferNet Brasil registraram aumento de 874% nas denúncias de xenofobia online entre 2021 e 2022, evidenciando a urgência de ferramentas automatizadas de detecção. A natureza implícita dessas manifestações, que recorrem a ironia, metáforas, negações e recursos linguísticos sofisticados, torna a tarefa particularmente desafiadora para os sistemas existentes.

Esta tese de doutorado propõe desenvolver e validar uma arquitetura de aprendizado profundo para detecção multirrótulo de xenofobia implícita em textos em português, baseada em técnicas de Adaptação Paramétrica Eficiente (PEFT), especificamente a técnica LoRA (Low-Rank Adaptation). A proposta endereça três limitações centrais dos sistemas atuais: o custo computacional proibitivo do ajuste fino completo de LLMs, a simplificação excessiva das abordagens binárias de classificação e a inadequação das métricas tradicionais para avaliar coerência contextual.

A arquitetura proposta aplica LoRA sobre o modelo BERTimbau Large, reduzindo os parâmetros treináveis de 110 milhões para aproximadamente 1–2% do modelo original, com impacto direto no consumo de memória e na latência de inferência. Quatro cabeças de classificação especializadas serão acopladas ao modelo base para predição simultânea de estereótipo, desumanização, exclusão e score contínuo de toxicidade. Para aumentar a transparência das decisões, serão integrados métodos de IA Explicável via Integrated Gradients, permitindo identificar quais tokens mais influenciaram cada classificação.

Como contribuição metodológica original, propõe-se a métrica de Integridade Contextual, que avalia a precisão da classificação e sua coerência com o contexto global do texto, detectando negações e estruturas contraditórias que podem inverter o sentido das ofensas. A pesquisa também prevê a construção e anotação de um corpus de 5.000 a 10.000 textos em português brasileiro com anotações multirrótulo, a ser disponibilizado à comunidade científica. Os resultados esperados incluem um sistema funcional entregue como API RESTful e dashboard, com potencial de aplicação por plataformas digitais, ONGs e órgãos governamentais na moderação de conteúdo.

Referências Bibliográficas:

- HU, E. J. et al. LoRA: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.
- JAHAN, M. S.; OUSSALAH, M. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, v. 546, p. 126232, 2023.
- SANTOS, A. R. S.; RODRIGUES, C. M. O. Um ensemble para a avaliação de emojis na identificação de postagens xenofóbicas. Dissertação de Mestrado, Universidade de Pernambuco, 2024.

- SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. A. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: *Brazilian Conference on Intelligent Systems (BRACIS)*, 2020.
- SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.