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

Proposta de Tese de Doutorado

Área: Inteligência Computacional

Título: Desenvolvimento de Técnicas de Inteligência Artificial Explicável (XAI) Personalizáveis para Modelos de Correção de Textos Discursivos em Contextos Educacionais Diversificados

Orientador: Profº Dr. Cleyton Mário de Oliveira Rodrigues

A explicação das decisões tomadas por modelos de IA em contextos educacionais é fundamental para garantir transparência e confiança, especialmente em tarefas de correção de textos discursivos. Embora técnicas como LIME e SHAP sejam amplamente utilizadas para gerar explicações em modelos de aprendizado de máquina, sua aplicação a dados textuais apresenta limitações. Originalmente projetadas para dados estruturados, essas técnicas necessitam de adaptações para lidar com a complexidade semântica dos textos não estruturados, que dependem de representações numéricas como embeddings gerados por modelos avançados como BERT e T5. Além disso, as explicações geradas frequentemente são genéricas e pouco alinhadas com critérios pedagógicos e regionais, dificultando sua aplicabilidade em contextos educacionais diversificados, como no Brasil, Angola ou outros países.

Diante disso, esta pesquisa propõe o desenvolvimento de uma nova técnica de Inteligência Artificial Explicável (XAI), projetada especificamente para dados textuais em avaliações educacionais. O modelo proposto buscará superar as limitações das técnicas existentes ao introduzir explicações personalizáveis e alinhadas aos critérios pedagógicos e culturais. Diferentemente de abordagens tradicionais, a nova técnica integrará recursos visuais interativos, hierárquicos e ajustáveis, permitindo que educadores compreendam como os modelos tomaram suas decisões e como esses critérios se relacionam com elementos específicos do texto.

A metodologia será estruturada em três etapas principais: a construção de um corpus diversificado de textos discursivos, representando variações regionais e culturais; a utilização de modelos pré-existentes para realizar a correção, permitindo o foco na explicação; e o desenvolvimento da nova técnica de XAI, incluindo ferramentas para visualizações mais intuitivas e explicações ajustáveis a diferentes contextos educacionais.

Espera-se que o modelo de explicação proposto promova maior confiança, transparência e aceitação em sistemas de IA aplicados à educação, ao mesmo tempo em que contribui para o avanço do uso de XAI em dados textuais. Além disso, o sistema poderá ser utilizado em múltiplos países, adaptando-se a critérios específicos. Espera-se também que esta possar servir como referência para futuras pesquisas na área de educação e IA explicável.

Referências Bibliográficas:

- **CESARINI, Mirko; MALANDRI, Lorenzo; PALLUCCHINI, Filippo; SEVESO, Andrea; XING, Frank.** Explainable AI for Text Classification: Lessons from a Comprehensive Evaluation of Post Hoc Methods. *Cognitive Computation*, v. 16, p. 3077–3095, 2024. Disponível em: https://link.springer.com/article/10.1007/s12559-024-10325-w. Acesso em: 17 jan. 2025.

- **MARDATOUI, Oumaima; BOUJEMAA, Nozha; VAKILI, Vahid.** An Analysis of LIME for Text Data. In: *Proceedings of Machine Learning Research*, 2021. Disponível em: https://proceedings.mlr.press/v130/mardaoui21a/mardaoui21a.pdf. Acesso em: 17 jan. 2025.

- **LI, Jiazheng; GUI, Lin; ZHOU, Yuxiang; WEST, David; ALOISI, Cesare; HE, Yulan.** Distilling ChatGPT for Explainable Automated Student Answer Assessment. *arXiv preprint arXiv:2305.12962*, 2023. Disponível em: https://arxiv.org/abs/2305.12962. Acesso em: 17 jan. 2025.

- **MANABE, Hitoshi; HAGIWARA, Masato.** EXPATS: A Toolkit for Explainable Automated Text Scoring. *arXiv preprint arXiv:2104.03364*, 2021. Disponível em: https://arxiv.org/abs/2104.03364. Acesso em: 17 jan. 2025.

- **RAMON, Yanou; MARTENS, David; PROVOST, Foster; EVGENIOU, Theodoros.** Counterfactual Explanation Algorithms for Behavioral and Textual Data. *arXiv preprint arXiv:1912.01819*, 2019. Disponível em: https://arxiv.org/abs/1912.01819. Acesso em: 17 jan. 2025.