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

Proposta de Dissertação de Mestrado

Área: Computação Inteligente

Título: Análise e aplicação de técnicas de NLP, Aprendizado de Máquina e *explainable AI (*XAI*)* para classificação de documentos jurídicos.

Orientador(a): Eraylson Galdino da Silva (eraylson.galdino@upe.br)

Descrição:

A classificação de documentos é uma tarefa importante no contexto de mineração de dados. Podemos encontrar aplicações de classificação de documentos em diferentes áreas, tais como: medicina, educação, economia e no direito.

O processo para o desenvolvimento de um modelo de classificação de documentos é composto por quatro etapas principais: (I) Limpeza e Extração de Características, (II) Redução da Dimensionalidade, (III) Modelagem do classificador e (IV) Avaliação do modelo. Em cada etapa temos um conjunto de técnicas cuja eficácia depende do contexto da aplicação, de forma que se faz necessário um estudo comparativo entre as técnicas para a construção de um sistema de classificação automática de documentos. Além dessas etapas, atualmente vem sendo aplicadas algumas técnicas de *explainable AI* (XAI) ou *Interpretable Machine Learning* (IML) vêm sendo aplicadas para proporcionar interpretações ou explicações das classificações feitas pelos modelos.

Apesar do avanço das técnicas para classificação de documentos, algumas áreas ainda realiza a classificação de documentos através de um especialista ou através de técnicas baseadas em regras. Em ambos os casos, o processo pode ser custoso e sujeito a falha humana no momento de interpretar o texto e de criar regras para classificação do documento. Como exemplo podemos utilizar o contexto de procuradores que precisam classificar se um determinado processo é relevante ou irrelevante para o governo. Cada procurador pode ficar responsável por avaliar dezenas de processos semanalmente, o que torna uma atividade exaustiva e sujeita à erros por conta do cansaço humano. Caso seja criado um conjunto de regras para classificar tais processos, é possível que a base de regras seja incompleta e que não generalize para novos casos.

Nesse contexto, a presente proposta consiste em um estudo de quais técnicas de NLP e Aprendizado de Máquina apresentam resultados acurados no contexto de classificação de documentos jurídicos, assim como quais técnicas XAI e IML podem ser aplicadas para possibilitar uma explicação à classificação. Através dessa pesquisa é esperado o desenvolvimento de aplicações de classificação automática de documentos jurídicos de forma que possa contribuir com órgãos públicos, aumentando a eficiência e eficácia na análise de processos e tomada de decisão.

Referências Bibliográficas:

1. CSÁNYI, Gergely Márk et al. Building a Production-Ready Multi-Label Classifier for Legal Documents with Digital-Twin-Distiller. Applied Sciences, v. 12, n. 3, p. 1470, 2022.
2. JOSI, Frieda; WARTENA, Christian; HEID, Ulrich. Preparing legal documents for NLP analysis: Improving the classification of text elements by using page features. In: Computer Science & Information Technology (CS & IT). AIRCC Publishing Corporation, 2022. p. 17-29.

3. KOWSARI, Kamran et al. Text classification algorithms: A survey. Information, v. 10, n. 4, p. 150, 2019.
4. WANG, Shirui; ZHOU, Wenan; JIANG, Chao. A survey of word embeddings based on deep learning. Computing, v. 102, p. 717-740, 2020.
5. GHOSH, Mahuya; DAS, Amit Kumar; CHAKRABARTI, Amlan. A Short Review on XAI Techniques in Text Data. In: International Conference on Advances in Data Science and Computing Technologies. Singapore: Springer Nature Singapore, 2022. p. 353-364.
6. MORADI, Milad; SAMWALD, Matthias. Explaining black-box models for biomedical text classification. IEEE journal of biomedical and health informatics, v. 25, n. 8, p. 3112-3120, 2021.