Extração de Palavras-chave em Chats de Dispositivos Móveis em Análises Periciais

Horácio Aaron Christhian Galdezanni Pedroso[a], Jeovane Honório Alves[a], Jean Paul Barddal[a,b]

[a] *Programa de Pós-Graduação em Informática (PPGIa), Pontifícia Universidade Católica do Paraná (PUCPR), R. Imaculada Conceição, 1155, Curitiba, Paraná, Brasil*
[b] *Autor para correspondência: jean.barddal@ppgia.pucpr.br*

Palavras-chaves: Extração de palavras-chave, Aprendizagem profunda, Análises periciais

O projeto em questão concentra-se na aplicação de técnicas de extração de palavras-chave para suporte da atividade de perícia digital forense no âmbito da luta contra o tráfico de drogas. Em parceria com a Polícia Científica do Estado do Paraná, o estudo visa explorar, aplicar e comparar o desempenho de modelos extratores na análise de conversas de texto extraídas de celulares apreendidos a fim de verificar se essa pode ser classificada como relacionada a alguma atividade criminosa. Devido sua natureza sensível e confidencial, visto que se trata de dados associados diretamente a processos de investigação policial e judicial, seu acesso foi restrito às dependências físicas e lógicas da polícia. A pesquisa empregou inicialmente os modelos TF-IDF, YAKE, RAKE, TextRank e KeyBERT, este utilizando bases pré-treinadas como MPnet, MiniLM, TensorFlow e Bertimbau. Durante a análise experimental, ao comparar o resultado dos algoritmos com a análise de dez laudos produzidos pelos peritos, foi possível perceber que palavras-chave relacionadas a movimentação de drogas ilícitas assim como ao tráfico foram extraídas por múltiplos modelos, em particular os modelos YAKE, TextRank e KeyBERT (Mpnet), onde taxas de F1 score de 67% foram obtidas. Em fases futuras, o projeto visa ainda a experimentação de modernos modelos de processamento de linguagem natural (NLP), como os Grandes Modelos de Linguagem (Large Language Models - LLM), bem como a viabilidade de aplicação de técnicas de finetuning com o objetivo de refinar a precisão dos modelos com base no contexto estudado (i.e., tráfico de drogas). Além disso, a pesquisa objetiva avaliar e comparar a eficácia na extração de palavras-chave – bem como expandindo sua aplicação para o conceito de frases-chave – desses variados modelos extratores, os quais englobam categorias como modelos estatísticos e de modelo pré-treinados (incluindo os LLMs).

ISSN:2526-1371 DOI 10.5281/zenodo.15014707