Kickstarting Legal Multi-label Classification Experimentation

Claudia Schulz and **Martina Forster** and **Prudhvi Nokku** and **Stavroula Skylaki**
clauschulz1812@gmail.com

Abstract

Multi-Label Classification (MLC) is a common task in the legal domain, where more than one label may be assigned to a legal document. A wide range of methods can be applied, ranging from traditional ML approaches over fine-tuned Transformer-based architectures to zero-shot LLM prompting. Depending on the data characteristics, such as available training data, text length and number of labels, different approaches may yield the best results. In addition to prediction performance, another important consideration in practice is prediction speed and cost of the different suitable approaches.

Experimenting with different baseline approaches to find the most promising one for a given legal MLC task is usually time-intensive. To cut down the baseline testing time on new projects, we designed a MLC baseline suite that allows to seamlessly train and evaluate a variety of different MLC models in one go. This includes traditional similarity methods like TF-IDF as well as BERT-style models such as RoBERTa, Bi-/Cross-Encoders, and T5, and allows to compare domain-specific with general-domain models.

We tested our baseline suite on two public legal datasets, POSTURE50K and EURLEX57K, and compare the results with state-of-the-art LLM prompting approaches. To explore the comparative advantage offered by different approaches in relation to the dataset properties, we varied the amount of training data and the number of labels in these datasets, simulating different types of datasets. Our results highlight performance-speed-cost trade-offs.