Enhancing Legal Document Processing through Natural Language Understanding and Generation techniques

Candidate: Irene Benedetto
Supervisor: Luca Cagliero

Politecnico di Torino

April 4, 2025

The rapid advancement of Artificial Intelligence (AI), particularly in Natural Language Processing (NLP), offers significant opportunities for innovation in the legal domain. However, the characteristics of legal language and the intricate nature of legal tasks present substantial challenges for AI adoption in real-use cases. Current legal systems often rely on manual annotation and predefined taxonomies that are time-consuming to maintain and struggle to capture the evolving nature of legal concepts. Existing AI approaches face limitations such as dependence on outdated information and data scarcity. In the realm of decision-making, legal professionals hesitate to fully trust AI-generated outcomes due to concerns about accuracy, transparency, and accountability, especially since AI systems often lack true legal reasoning. This dissertation explores how NLP and in particular Language Models can address these challenges and enhance legal workflows across three areas: automatic document exploration, content accessibility, and reasoning applications for more complex tasks. To address these challenges, this research develops a classification system, based on Language Models, to automatically infers and annotates taxonomy relationships between legal documents, effectively overcoming the limitations of traditional taxonomy-based approaches. This dissertation benchmarks state-of-the-art abstractive summarization models tailored to the Italian legal domain, enhancing content accessibility by generating high-quality summaries of lengthy and complex legal texts. The thesis also introduces AI models for court judgment prediction and explanation, incorporating legal entities to improve accuracy and explainability. Additionally, it presents novel pipelines that use Large Language Models and Retrieval Augmented Generation to align AI-generated legal solutions more closely with specific case details, thereby improving legal reasoning and decision-making accuracy. Focusing primarily on the Italian legal domain, this work employs both quantitative and qualitative evaluation, to compare different approaches, including open- and closed- source Large Language models.

Results of this work reveal that integrating domain-specific techniques—such as named-entities and specialized pre-training—significantly enhances the performance, robustness, and explainability of AI models in legal document analysis tasks like classification, summarization, and court judgment prediction. In addition, incorporating collaborative multi-model approaches and advanced retrieval techniques significantly enhances AI systems' ability to perform complex reasoning tasks, which is crucial for improving the reliability of decision-making systems. However, for these tasks, the generalization capability of larger models prevails on smaller fine-tuned models. From an applicative perspective classification tasks language models are effective due to their robustness and scalability, though they are limited in addressing complex challenges. While LLMs show promising performance in text summarization and legal reasoning with fine-tuning, they face challenges in explainability and complex reasoning.