ClearCase : Legal Text Classification with Explainable AI

Arun Karthik Sengottuvel sengottu@usc.edu
Ashwinkumar Venkatnarayanan venkatna@usc.edu
Indira Swaminathan iswamina@usc.edu
Bhavya Avuthu avuthu@usc.edu
Mona Teja Kurakula mkurakul@usc.edu
Mohan Sai Ganesh Kanna mkanna@usc.edu

Abstract

Classifying legal documents is essential for organizing case law and helping professionals manage legal information. While legal experts can navigate complex terminology, it's often challenging for non-experts. This study introduces a Transformer-based legal text classification system that not only supports professionals in categorizing cases, but also makes the reasoning behind these classifications clear and helpful to non-experts. Using pre-trained models like LegalBERT, Longformer, or RoBERTa, the system analyzes complex legal texts, accurately extracting citation classes and key phrases. By integrating Explainable AI (XAI) techniques such as SHAP, LIME, or attention visualization, it ensures both precise classifications and transparent, human-readable explanations of critical legal phrases influencing its decisions.

1 Introduction

With the exponential increase in legal texts, automating the classification process offers a scalable solution that saves time and reduces manual effort, allowing legal professionals to focus on case analysis and decision-making. Given the high stakes and complex language of legal documents, AI-driven decisions must be both reliable and understandable.

1.1 Why is it useful?

Legal professionals often spend considerable time manually sorting and analyzing vast amounts of legal documents. Automating this process with transformer-based models boosts efficiency, reduces human error, and speeds up access to relevant legal texts. What sets this approach apart is its emphasis on explainability, allowing legal practitioners to trust and validate AI-generated classifications with confidence. This not only improves the technology's effectiveness but also makes it ideal for real-world legal settings, where transparency, trust, and accountability are of utmost importance.

1.2 How is it interesting?

The intersection of deep learning, legal technology, and explainability presents a unique challenge due to the complexity of legal language. Legal documents often contain specialized jargon, citations, and intricate structures that require domain-specific expertise. This project explores how transformers fine-tuned for the legal domain can enhance classification accuracy compared to general-purpose models. By integrating XAI techniques such as SHAP, LIME, or attention visualization, we gain deeper insights into the decision-making process of deep learning models. This dual focus on boosting performance and ensuring transparency not only pushes the boundaries of AI research but also makes it more applicable and trustworthy for real-world legal applications.

2 Potential Solution

The proposed system utilizes state-of-the-art transformer models like LegalBERT, Longformer, or RoBERTa to classify legal documents based on citation classes and key catchphrases, enhancing legal research and document retrieval efficiency. To ensure interpretability, the system integrates XAI techniques such as attention visualization, SHAP, or LIME. These methods reveal which legal phrases, citations, or contextual elements influence classification decisions, providing transparency into the model's reasoning process.

The system's performance will be evaluated using classification accuracy, F1-score, and interpretability metrics, comparing different transformer architectures. Additionally, XAI techniques will be assessed by determining how well the identified key phrases align with legal citations and expert reasoning, ensuring that the system offers both high accuracy and trustworthiness for legal professionals.

2.1 Dataset

The dataset consists of 24,985 legal case records, each described by four columns: `case_id`, `case_outcome`, `case_title`, and `case_text`. The primary input for the model is the `case_text`, which contains detailed excerpts from legal case proceedings, often including citations and legal reasoning. The goal is to classify these texts into one of ten possible `case_outcome` categories, such as cited, applied, followed, considered, and many more, indicating how the case was referenced or treated in a legal context.

3 Expected Outcomes

This project aims to develop an accurate and transparent legal document classification system that supports legal professionals in efficiently categorizing and retrieving key legal texts. Using transformer models like LegalBERT, Longformer, or RoBERTa, the system will handle complex legal language and extract relevant citation classes and catchphrases with high precision.

A significant outcome will be the integration of XAI techniques such as attention visualization, SHAP or LIME to provide interpretable justifications for AI-driven decisions. This will increase trust in the system by allowing legal practitioners to understand the reasoning behind the model's classification. Comparative evaluations will highlight the most effective transformer architectures and explainability methods, contributing to the advancement of legal AI systems that balance performance and transparency.

4 Proposed Timeline

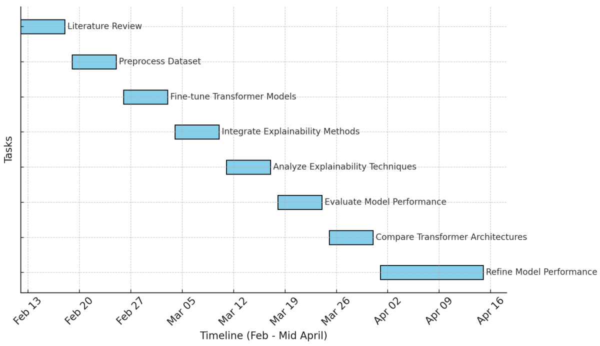

Figure 1: Proposed Project Timeline.

4.1 Weeks 1-3

- Conduct a literature review on legal document classification and XAI.
- Preprocess the Legal Citation Text Classification dataset.
- Fine-tune pre-trained transformer models (LegalBERT, Longformer, RoBERTa) on the dataset.

4.2 Weeks 4-6

- Integrate explainability techniques, such as attention visualization, SHAP, or LIME.
- Analyze how well these techniques align with legal reasoning and citation relevance.

4.3 Weeks 7-12

- Evaluate the model's performance using classification accuracy, F1-score, precision, and interpretability metrics.
- Compare different transformer architectures and explainability methods.
- Refine model performance and improve explanation clarity based on evaluation results.

References

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Ion Androutsopoulos, and Nikolaos Aletras. 2020. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Preprint*, arXiv:1705.07874.

A. Mohankumar. 2023. Legal text classification dataset. Kaggle.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938.

Carlos A. C. Sáenz and Katharina Becker. 2024. Understanding stance classification of BERT models: an attention-based framework. *Knowledge and Information Systems*, 66:419–451.