Introduction

The PILOT model was created to help predict the outcomes of legal cases in case law systems. In these systems, judges make decisions based on past cases (called precedents), not just written laws. The main goals of the PILOT model are to find similar past cases that can help predict the result of a new case, and to understand how legal decisions change over time, so the model can stay accurate even when laws or interpretations evolve. This helps legal professionals make better decisions faster and with more confidence.

The data used to train and test the PILOT model comes from the European Court of Human Rights (ECHR). This court handles cases related to human rights violations in Europe. The researchers created a special dataset called ECHR2023, which includes thousands of real legal cases from the ECHR database.

The original legal documents were very long (often over 2,000 words), written in different European languages, and hard to read and understand. To make the data easier to use, the researchers used a smart AI tool called GPT-3.5-turbo. This tool helped by reading the long documents, picking out the most important facts, and summarizing those facts into short bullet points. This process focused only on the FACT section of each case, which describes what happened. It did not include the final decision or any extra information that might give away the result.

Each case in the dataset includes a short summary of the facts, the date of the decision, the legal articles involved, and the final outcome (which articles were violated). The dataset was split into training set with 8,138 cases used to teach the model, validation set with 3,000 cases used to fine-tune the model, and test set with 3,000 cases used to check how well the model works. The cases were split by time, so the model was trained on older cases and tested on newer ones. This makes the testing more realistic, like how judges work in real life.

Methodologies

The methodology of the PILOT addresses temporal pattern shifts in legal case outcome prediction within case law systems. The framework utilizes Legal-BERT for semantic embeddings and integrates three modules: Relevant Case Retrieval, Case Encoder with Evidence Fusion, and Temporal Shift Mining, utilizing the ECHRN2023 dataset from the European Court of Human Rights (ECtHR).

Legal-BERT Overview and Suitability: Legal-BERT, a BERT variant fine-tuned on legal texts like case law, statutes, and contracts, captures domain-specific terminology, syntactic structures, and contextual nuances. Its suitability for PILOT lies in its ability to generate

accurate semantic embeddings for complex legal documents, enabling effective case retrieval and evidence fusion by understanding legal context better than general-purpose models like BERT.

The Relevant Case Retrieval module employs Legal-BERT to generate semantic embeddings of the current case's factual description, querying a precedent database to retrieve cases with high cosine similarity. This provides contextual legal knowledge, though limited by semantic similarity constraints.

The Case Encoder with Evidence Fusion module encodes the current case and precedents using Legal-BERT, fusing their representations to integrate judicial evidence. Robustness is enhanced by applying BERT's dropout layers, generating varied representations within a batch.

The Temporal Shift Mining module introduces a temporal decay term to prioritize recent legal patterns, mitigating performance (Micro-F1: 0.77 training vs. 0.07 testing) due to temporal shifts.

Finally, the future work to enhance the algorithm should include (1) enhancing retrieval by incorporating legal principles and jurisdictional factors, to ensure more relevant precedent selection and improve prediction accuracy. (2) including additional case details, like legal arguments or procedural history, to enrich the model's understanding and produce more comprehensive predictions. These address limitations like oversimplified legal complexities, advancing PILOT's practical utility.

Experiments

The authors conducted a series of experiments to evaluate the performance of PILOT, with the results demonstrating strong and consistent performance across all evaluation metrics. The assessment was based on the average of five runs using different random seeds. Four metrics were deployed to measure the model performance: micro-F1, micro-Jaccard, micro- PR-AUC, and micro-ROC-AUC. A strict experimental condition was implemented to ensure that no future cases were used as references during the training period. In the training phase, all prior cases from the training set were available as precedents. For contrastive learning, only data from the training split of the dataset were utilised.

During the evaluation, the dataset was partitioned in chronological order into 8138 training instances, 3000 validation instances, and 3000 test instances. The chronological split to ensure that no future data will not be utilised for training and evaluation. Thereby,

the models were enhanced for better adaptability to concept drift and reinforcing temporal coherence.

In the Legal Case Outcome Prediction experiments, PILOT significantly outperformed all eight other methods across the four evaluations metrics even when accounting for standard deviation from five different seeds. The outstanding performance is due to PILOT explicitly accounting for temporal pattern shifts, which the other model's neglect. According to the results, PILOT improved micro-F1 score by 2.74% compared to LWDROV2. Interestingly, during the experiment, ChatGPT refused to provide predictions, which limited its ability to generate accurate predictions. As a result, the ChatGPT 5-shots ranked among the lowest performance in three of the four evaluation metrics.

An ablation study was conducted to evaluate the impact of relevant case retrieval module and the temporal pattern handling module on PILOT overall performance. The result demonstrated that both models contributed positively to performance improvement. However, the performance declines when integrated with law article semantics. The author suggested this might be due to the evolving content and interpretation of law articles over time which the model struggles to capture without explicitly incorporating time factor.

A qualitative case study further confirmed that the retrieved precedent cases were semantically relevant to the target case. The author observed that the violated articles mentioned in retrieved cases were closely aligned with and encompassed those in the target case, suggesting that PILOT captures comprehensive coverage of relevant legal provision. These findings support PILOTS performances and its effectiveness.

Lastly, the authors conducted hyperparameters analysis focusing on case retrieval module and training objectives. For the case retrieval module, two key hyperparameters were examined: k, which determine the number of top relevant precedent cases to be retrieved, and α, the coefficient associated with the temporal decay function. The analysis suggested that setting $\alpha \in [1, 10]$ yields optimal performance. For the training objectives, the drift loss weight was evaluated, with $\lambda = 0.10$ identified as the best balance. The higher values were found to negatively affect the model performance.

Conclusion

While the PILOT model displays a promising step toward outcome prediction in legal cases, the authors openly acknowledge that it remains a **preliminary work**, a baseline for future investigation. The PILOT model described in the text **cannot be deployed** as it oversimplifies some factors surrounding legal cases that are undeniably complex.

1. Over-Reliance on Semantic Similarity for Precedent Selection
The selection of precedent cases which the paper largely attributes to its success over previous models are determined on semantic similarity alone. This is problematic as cases that are semantically similar may differ critically in factual details. This risks false equivalence, where similar sounding cases in fact lead to opposite legal outcomes.

2. Limited Use of Case Information
Currently, PILOT only uses a legal case's factual section to make predictions. This fails to consider testimonies, witness credibility, legal arguments, and procedural context, the subtleties of which can significantly alter the judicial outcome. The paper doesn't shed light on how to integrate such data, establishing an area for future development.

Beyond these points, the paper mentions the need to **eliminate bias**. To critique the text, it's striking that there exists no further elaboration on this. As such, this makes for excellent discussion on whether bias is a problem a model can solve, or inherent to the broader topic of whether legal outcome prediction is an ethical practice to begin with.

The paper goes on to suggest an improvement to its **interpretability via an LLM** module that explains its judgement. Lastly, to bolster predictive power, a mixture-of-experts approach was proposed utilizing multiple instances of PILOT with varying parameters.

While legal outcomes shouldn't be determined by a single model alone, I posit it also shouldn't be determined by AI alone. Such discussion will be saved for the seminar.

Individual Key Takeaways

The complexity and sparsity of legal cases make AI unsuitable, as it struggles to capture the intricate nuances which is require for judgment. AI is unsuitable for law due to its non-human-like nature, lacking the ability to sense human emotions and ethical.
— Fahad

Judges or lawyers may rely on PILOT as a crutch, reducing genuine human legal deliberation. PILOT should not replace legal professionals. Every law case should be treated with a level of equity rather than a pattern to be mimicked. — Noel

The PILOT model shows that AI in law must be used carefully. It should be fair, protect people's privacy, and help judges, not replace them. The model also needs to be clear about how it makes decisions, so people can trust it. — Sekar

Legal judgement goes beyond technical analysis. It is shaped by power, place and space. The law should be tempered with mercy. The AI models lack the capacity to replicate human emotion or understand social context. — Timmy