Incorporating Precedents for Legal Judgement Prediction on European Court of Human Rights Cases

Santosh T.Y.S.S, Mohamed Hesham Elganayni,
Stanisław Sójka, Matthias Grabmair
School of Computation, Information, and Technology;
Technical University of Munich, Germany

Abstract

Inspired by the legal doctrine of stare decisis, which leverages precedents (prior cases) for informed decision-making, we explore methods to integrate them into LJP models. To facilitate precedent retrieval, we train a retriever with a fine-grained relevance signal based on the overlap ratio of alleged articles between cases. We investigate two strategies to integrate precedents: direct incorporation at inference via label interpolation based on case proximity and during training via a precedent fusion module using a stacked-cross attention model. We employ joint training of the retriever and LJP models to address latent space divergence between them. Our experiments on LJP tasks from the ECHR jurisdiction reveal that integrating precedents during training coupled with joint training of the retriever and LJP model, outperforms models without precedents or with precedents incorporated only at inference, particularly benefiting sparser articles.

1 Introduction

The task of case outcome classification deals with identifying the outcome from a textual description of case facts and is generally referred to as Legal Judgement Prediction (LJP) (e.g., Aletras et al. 2016; Chalkidis et al. 2019). It has been studied using corpora from different jurisdictions, such as the European Court of Human Rights (ECHR) (Chalkidis et al., 2019, 2021, 2022a; Aletras et al., 2016; Medvedeva et al., 2020; Santosh et al., 2022, 2023a), Chinese Criminal Courts (Luo et al., 2017; Zhong et al., 2018; Yue et al., 2021; Zhong et al., 2020; Yang et al., 2019), US Supreme Court (Katz et al., 2017; Kaufman et al., 2019), Indian Supreme Court (Malik et al., 2021; Shaikh et al., 2020), French court of Cassation (Şulea et al., 2017b), Federal Supreme Court of Switzerland (Niklaus et al., 2021), Turkish Constitutional court (Sert et al., 2021) UK courts (Strickson and De La Iglesia, 2020) and German courts (Waltl et al., 2017).

In this study, we focus on classifying case outcomes in the ECHR A and B benchmark tasks introduced by LexGLUE (Chalkidis et al., 2022a). Task B involves identifying the set of articles of the European Convention of Human Rights alleged to have been violated by the claimant, while Task A aims to classify which of the convention's articles has been deemed violated by the court. Both tasks utilize the fact description of the case extracted from the published judgement document as input. Early approaches to LJP relied on rule-based methods (Segal, 1984; Kort, 1957; Nagel, 1963), followed by classification techniques using bag-of-words features (Aletras et al., 2016; Şulea et al., 2017a). Recent advancements involve deep learning (Zhong et al., 2018, 2020; Yang et al., 2019) with adoption of pre-trained transformer models (Chalkidis et al., 2019; Niklaus et al., 2021), including legal-domain-specific variants (Zheng et al., 2021; Chalkidis et al., 2023). Furthermore, various strategies are explored, such as leveraging dependencies between auxiliary tasks (Tyss et al., 2023b; Yue et al., 2021; Valvoda et al., 2023) or incorporating constraints like contrastive learning (Tyss et al., 2023b; Zhang et al., 2023) and injecting legal knowledge (Liu et al., 2023; Tyss et al., 2023a).

Drawing inspiration from the legal doctrine of *stare decisis*, where precedents (prior cases decided in courts of law) are pivotal in common law jurisdictions to support arguments to arrive at the final outcome. Even in civil law systems, though the prior cases are not directly involved in the final judgment, they are still crucial references during the decision-making process to ensure consistency and proper application of law. We explore how to leverage precedents in LJP models to predict the outcome of a query case. While previous research has focused on using precdents for enhancing case representations through contrastive learning (Tyss et al., 2023b; Zhang et al., 2023; Gan et al., 2022), our work investigates two strategies for integrat-

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 3743–3750

ing precedents: (i) direct incorporation at inference through interpolation and (ii) integration during the training process via a fusion layer, enabling the model to reason with similar cases to derive the outcome of the query case, unlike prior works which use similar cases in the form of exemplars in in-context learning to facilitate LLMs in a zero-shot setting (Wu et al., 2023; Shui et al., 2023).

To enhance relevant precedent retrieval in both strategies, we train a retriever with a fine-grained relevance signal based on the overlap ratio of alleged articles between two cases. For direct integration at inference, we perform label distribution interpolation across each article with the retrieved cases' outcomes, determined by their proximity to the query case. This interpolation has been widely explored in retrieval-augmented KNN-based language models (Khandelwal et al., 2019; Xu et al., 2023; He et al., 2021), providing memorization capabilities for rare patterns that are otherwise challenging to capture within a parametric model.

Additionally, LJP models struggle to effectively leverage retrieved precedents as they are not explicitly trained to relate the query case to retrieved documents and conduct reasoning based on them. Moreover, they focus on memorization during training rather than offloading this process to the retrieval component. To address this, we introduce a precedent fusion module that incorporates precedents during training via stacked cross-attention. Further, we propose joint training that optimizes the retriever model alongside the LJP model, utilizing the relevance signal derived from the fusion module, helping to overcome the latent space divergence issue with static retrievers (Izacard and Grave, 2021; Izacard et al., 2023). Our experiments demonstrate that integrating precedents at training time along with joint training outperform models with out precedents and with precedents at inference only and without joint training, with larger improvements for sparser articles.

2 Method: Incorporating Precedents

We describe our baseline model which takes the case fact description x and outputs the set of articles as multi-hot vectors (alleged ones in case of Task B and violated ones in Task A). Then we outline the retriever to obtain precedent cases based on the current case. We then introduce two strategies for incorporating information from precedent cases: one during inference and the other during training.

2.1 Baseline Model

We adopt a hierarchical model as as outlined in (Chalkidis et al., 2022a) to account for longer case fact inputs. Each paragraph in the case facts is independently encoded using LegalBERT (Chalkidis et al., 2020), based on the [CLS] representation. These paragraph representations are passed through a two-layer transformer to obtain contextualized representations for each paragraph which are then max-pooled to derive the final case representation. This is inputted into a classification layer to produce the multi-hot outcome vector.

2.2 Precedent Retrieval

We aim to retrieve legal precedents sharing semantically similar facts with the query case to provide additional supervision for outcome prediction. Due to the lengthy legal documents involved, using a standard retriever isn't feasible. Instead, we adopt a hierarchical architecture akin to the baseline model without the classification head, as our retriever h. We employ a pair-wise similarity loss to train the retriever wherein we obtain each case representation through retriever and similarity is computed as the dot product between them. Golden similarity is determined at a fine-grained level using the label overlap ratio (LOR), computed via Jaccard similarity, between the allegation labels of the cases. We compute similarity over allegation labels because they aid in retrieving precedents to tackle challenging negative instances of violation task where the article was alleged but found not to be violated. The loss function is expressed as:

$$L(\theta) = \text{MSE}\left((h(x_i) \cdot h(x_j)), \text{LOR}(y_i, y_j)\right)$$

$$\text{LOR}(y_i, y_j) = \frac{|y_i \cap y_j|}{|y_i \cup y_j|}$$

where MSE indicates mean-squared error, $y_{i/j}$ indicate allegation multi-hot vector of cases $x_{i/j}$ respectively. This approach resembles contrastive learning, making cases with similar allegations lie closer in embedding space (Khosla et al., 2020; Santosh et al., 2023b) but at a fine-grained level.

2.3 Precedents at Inference

We construct a datastore $\{K, V\}$ comprising all precedent case representations as keys paired with their multi-hot outcome vectors as values. We retrieve the k-most similar precedent cases N to the query case and incorporate them via label interpolation in a non-parametric way. Given the multi-label

classification nature of the task, we interpolate each article separately as a binary classification problem by deriving a probability estimate for each outcome under each article, considering both the probability assigned to that label and its complement (1 - probability). Then the distribution of each label under p_{kNN} is derived using softmax of their negative distances indicating the closer a precedent case is to the query case, the larger its influence is.

$$p_{\text{kNN}}(l_{ij}|x, x_i) \propto \sum_{(k,v)\in N} \mathbb{1}_{l_{ij}=v_j} \exp\left(\frac{-d(h(x_i),k)}{\tau}\right)$$

τ denotes the temperature hyperparameter and d(.) indicates euclidean distance, j denotes the specific article. Finally, we interpolate the p_{baseline} with p_{kNN} to obtain final as:

$$p_{\text{final}}(l_{ij}|x, x_i) = \lambda p_{\text{baseline}}(l_{ij}|x, x_i) + (1-\lambda) p_{\text{kNN}}(l_{ij}|x, x_i)$$

where λ serves to balance p_{kNN} and p_{baseline}.

2.4 Precedents during training

We introduce a precedent fusion module to effectively incorporate precedent information during training, separating knowledge memorization from reasoning. This allows the model to focus on understanding the query case and conducting reasoning based on retrieved precedents, rather than solely on memorization. Further, we employ a joint training to optimize both the retriever and LJP models, preventing divergence in the latent space between these two modules if optimized independently.

Upon retrieving precedent cases, we fuse this information into the query using a stacked cross-attention module. This module computes cross-attention between the query case and the retrieved case representations (i.e keys) from the datastore, determining importance weights for their outcome vectors (i.e values). The weighted outcome vectors are transformed into the input space through a feed-forward layer and added to the query representation via a residual connection. This is represented as:

$$h_i^L = h_i^{L-1} + \text{g}(\text{softmax}(\frac{h_i^{L-1} W_q \cdot K W_k}{\sqrt{d_k}}) V W_v)$$

where h_i^L represents the fused query representation at layer L, $\text{g}(\cdot)$ denotes the feed-forward layer, W_q, W_k, W_v are learnable parameters and d_k represents the dimensionality of the keys representation.

We jointly train the retriever and LJP model by minimizing the KL-divergence between the retriever similarity scores (s) and the cross-attention scores in the fusion module, aggregating them across layers to obtain a single score for each retrieved precedent (a). This approach, inspired by Izacard and Grave 2021, leverages cross-attention scores as a proxy for similarity to improve retriever tailored in conjunction for the task of LJP.

$$L(\theta) = \sum_{k=1}^{K} a_k \log \frac{a_k}{s_k}$$

where K is the number of retrieved precedents. We only optimize retriever parameters with above loss and not the LJP encoder. To account for computational overhead for updating the datastore after every update to the retriever, We allow it to be stale and asynchronously update at certain frequency.

3 Experiments & Results

3.1 Dataset & Metrics

We experiment on the ECHR task A and B of LexGLUE (Chalkidis et al., 2022a), which consist of 11k case fact descriptions chronologically split into training (2001–2016, 9k cases), validation (2016–2017, 1k cases), and test sets (2017-2019, 1k cases). Both tasks include 10 prominent articles as labels. Following Chalkidis et al. 2022a, we report micro-F1 and macro-F1 (Mic-/Mac-F1) for both the tasks. We also report hard-macro-F1 (H.Ma-F1) for Task A following Santosh et al. 2022, which is the mean F1-score computed for each article where cases with that article having been violated are considered as positive instances, and cases with that article being alleged but not found to have been violated as negative instances. We also report label overlap ratio (LOR) based on allegation labels for the models with retriever component. Implementation details are described in App. A.

3.2 Results

Precedents at Inference: We investigate the impact of various retriever models by incorporating the precedents retrieved by them directly at inference using label interpolation. We create self-retrieval method that uses the trained LJP model without its classification head as the retriever to obtain precedents. We create a retriever trained with binary relevance loss indicated by at least one shared alleged article, as an alternative to the fine-grained label overlap loss described earlier. Overall, as shown in Table 1, we observe that adding precedents directly at inference brings improvements in macro-F1 scores for both tasks compared to the baseline, regardless of the retriever used. This suggests that the retrieved precedents, when

		Task B			Task A			
Model	**Retriever**	**Mac-F1**	**Mic-F1**	**LOR**	**H.Ma-F1**	**Mac-F1**	**Mic-F1**	**LOR**
Baseline	-	73.56	79.45	-	62.02	64.18	70.42	-
Inference	Self-Retrieval	74.12	79.88	0.58	62.83	64.78	69.17	0.54
	Binary Rel.	74.94	79.77	0.61	63.77	65.02	*70.16*	0.58
	Fine-grained Rel.	*75.63*	*80.28*	*0.65*	*65.14*	*66.84*	69.81	*0.63*
Freeze LJP and Retreiver	Mean	75.48	79.92	0.65	65.81	67.25	71.32	0.63
	Cross Attention	75.81	*80.98*	0.65	66.23	68.19	71.26	0.63
	Stack Cross Att.	*76.61*	79.62	0.65	67.12	*68.66*	***72.17***	0.63
Train LJP only	Stack Cross Att.	74.73	78.15	0.62	64.88	65.69	70.88	0.59
Train both		76.45	80.47	0.66	66.77	68.29	71.65	0.62
Train both with KLD.		***77.82***	***81.29***	***0.70***	***67.79***	***69.12***	*71.74*	***0.66***

Table 1: Results on Task A and Task B. Rel., Att., KLD. indicate relevance, attention and KL-Divergence respectively. Best results overall and in each group are bolded and italicized respectively.

interpolated at inference, help the model perform better on sparse classes, which may have been difficult to capture with the implicit parameters of the models due to label imbalance in the training data. Among the retrievers, we find that training with outcome-based relevance signal improves performance compared to using self-retrieval, indicating sub-optimal representations learned by the model through the LJP task alone, which are enhanced by operating directly on the embedding space. Furthermore, the addition of fine-grained relevance loss helps the model learn representations more effectively, as evidenced by the increased performance in downstream tasks and label overlap ratio (LOR). Consequently, we utilize the fine-grained relevance-based retriever for our subsequent experiments.

Fusing Precedents at Training: We explore different methods for integrating precedent information into the query representation using pre-trained LJP encoder and retriever models, keeping them frozen to isolate the fusion's effect. We devise three fusion variants: (i) computing the mean of retrieved precedent outcomes (values) and adding them to the original representation via a projection layer, (ii) employing cross-attention between the current case representation as the query and the retrieved representation, with labels as key and value vectors and (iii) utilizing stacked cross-attention with multiple layers. Our findings indicate that cross-attention outperforms simple mean, suggesting that the retrieved memory values are not equally important and require learning weights using a cross-attention. Moreover, learning the contribution of each retrieved precedent also acts as a filter for noisy, non-related precedents provided by the retriever. Furthermore, stacking these cross-attention layers enhances macro-F1 scores across both tasks, underscoring the necessity of complex interactions to learn similarity between the query and retrieved cases. Hence we adopt stacked cross-attention layers for our subsequent experiments.

Training Retriever and LJP: We unfreeze each of the components to create three variants: training only the LJP model, training both modules jointly and our joint training with additional retriever-specific KL-Divergence loss. Surprisingly, training the LJP model alone leads to a drop in performance for both tasks, primarily reflected in lower LOR values compared to freezing both modules. This suggests that the latent space of the LJP diverges from the frozen retriever, making it challenging to capture relevance. Training the retriever along with the LJP brings performance back to a comparable level to the frozen models but still slightly lags behind, indicating that training the retriever with LJP-specific loss alone does not provide a strong enough signal to learn relevance. Overall, the attention-score-based KL-divergence serves as a better proxy to guide the retriever to follow the latent space of the LJP model to provide better precedents (as seen in higher LOR), resulting in improved downstream performance across both tasks.

4 Conclusion

We enhance ECHR outcome classification by integrating precedents into LJP models during training. This improvement is driven by three components: (i) an effective retriever trained with fine-grained relevance signals of allegation labels, (ii) precedent fusion models enabling the offloading of memorization to the retrieval step and reasoning with cues from precedents and (iii) joint training of the retriever alongside the LJP model, enhancing its representations. These precedents also equips models to provide explanations through analogous case-based reasoning, warranting further investigation.

Limitations

It's important to acknowledge that despite being labeled as "legal judgment prediction" tasks, the fact statements are typically not finalized until the decision outcome is known. This characteristic transforms the task into one of retrospective classification rather than prediction (Medvedeva et al., 2021). Although this introduces distracting and confounding phenomena, as highlighted in Santosh et al. 2022, the dataset remains valuable for developing NLP models that analyze fact statements for text patterns corresponding to specific convention articles drafted by the court.

In our study, we demonstrated enhancements in outcome classification performance for ECtHR cases through the incorporation of precedents. While our techniques for precedent integration and retriever training objectives are generalizable and applicable to any jurisdiction, it's important to note that the experimental findings are specific to the context of the ECHR court. The degree of improvement achieved may depend on patterns of drafting texts that may form the basis for computing similarities to retrieve relevant precedents, which in turn can affect downstream performance.

It's worth mentioning that one can directly employ cases cited in the reasoning section of the documents to train retrievers or to validate the effectiveness of retrieved precedents. However, we did not use them directly due to the disguised positive problem (Santosh et al., 2024). This problem arises because a case can be cited for various reasons, such as being authoritative or due to familiarity bias of the drafter. Additionally, it's not feasible to cite all relevant prior cases, leaving the possibility of non-cited cases being related in disguise. Hence, we used LOR based on allegation labels as a signal of relevance, which may be weaker and coarse-grained in nature, as two cases with the same allegation labels might have different involving factors, making them weakly relevant. This approach only provides a lower bound of precise relevance. Future works can explore stronger relevance signals to learn and evaluate those precedents.

Furthermore, our proposed precedent incorporation strategies focus solely on the facts and respective outcomes of prior cases. This approach may be sub-optimal compared to the actual scenario where humans utilize the entire case document, including the reasoning section, which involves argumentation to arrive at the outcome by deducing the application of relevant law in the given context. In future, it would be beneficial to design models that can integrate these argument sections of precedent cases. By incorporating argumentation, models can deduce outcomes in a more explainable manner by learning applicable arguments within the query context, enhancing the transparency and interpretability of the decision-making process.

Ethics Statement

Our experiments were conducted on a dataset of ECHR decisions, which is publicly available as part of the LexGLUE benchmark (Chalkidis et al., 2022a) and sourced from the public court database HUDOC[1]. While these decisions contain real names and are not anonymized, we do not anticipate any harm beyond the disclosure of this information. However, it's important to acknowledge that utilizing historical data to train models may lead to classifiers that exhibit biased behavior. For instance, Chalkidis et al. 2022b investigated disparities in classification performance concerning an applicant's gender, age, and the respondent state. Additionally, by leveraging pre-trained encoders, our models inherit any biases encoded within them. However, legal NLP systems leveraging case outcome information and intended for practical deployment should naturally be scrutinized against applicable equal treatment imperatives regarding their performance, behavior, and intended use (Baumgartner et al., 2024).

The task of LJP raises significant ethical and legal concerns, both in general and specifically within the context of the European Court of Human Rights (ECtHR) (Fikfak, 2021). We do not advocate for the practical implementation of LJP/COC systems by courts. As demonstrated by Santosh et al. 2022, these systems rely on superficial, statistically predictive signals that lack legal relevance, highlighting the risks associated with deploying predictive systems in high-stakes domains such as law. They argue that models utilizing case outcome signals must be developed cautiously, aiming to align their inferences with legal expert reasoning. This aligns with the broader legal NLP community's increasing focus on the ethical aspects of developed systems in technical research (Medvedeva et al., 2021, 2023; Tsarapatsanis and Aletras, 2021; Leins et al., 2020).

In this study, we utilize LJP as a technical bench-

[1] https://hudoc.echr.coe.int

marking task for the development and analysis of neural NLP models on legal text. Our primary objective is to make incremental technical advancements towards enabling systems to work with precedents in a manner that mirrors human experts' analysis of case facts through interactions with past cases. We do not advocate for the practical application of LJP systems by courts but rather aim to explore how their core functionality of processing legal text can align with expert practices as closely as possible. Consequently, our results should be interpreted as technical contributions aimed at advancing models capable of deriving insights from legal data in a legally, ethically, and methodically sound manner. Our research group is dedicated to furthering research on such models to promote transparency, accountability, and explainability of data-driven systems in the legal domain.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Nina Baumgartner, Matthias Stürmer, Matthias Grabmair, Joel Niklaus, et al. 2024. Towards explainability and fairness in swiss judgement prediction: Benchmarking on a multilingual dataset. *arXiv preprint arXiv:2402.17013*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. *arXiv preprint arXiv:2305.07507*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022a. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022b. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406.

Veronika Fikfak. 2021. What future for human rights? decision-making by algorithm. *Decision-making by algorithm (September 3, 2021). Strasbourg Observers*, 19.

Leilei Gan, Baokui Li, Kun Kuang, Yi Yang, and Fei Wu. 2022. Exploiting contrastive learning and numerical evidence for improving confusing legal judgment prediction. *arXiv preprint arXiv:2211.08238*.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5703–5714.

Gautier Izacard and Edouard Grave. 2021. Distilling knowledge from reader to retriever for question answering. In *ICLR 2021-9th International Conference on Learning Representations*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, 12(4):e0174698.

Aaron Russell Kaufman, Peter Kraft, and Maya Sen. 2019. Improving supreme court forecasting using boosted decision trees. *Political Analysis*, 27(3):381–387.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Diederik P Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization 3rd int. In *International Conference on Learning Representations*.

Fred Kort. 1957. Predicting supreme court decisions mathematically: A quantitative analysis of the "right to counsel" cases. *American Political Science Review*, 51(1):1–12.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of nlp are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913.

Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. Mlljp: Multi-law aware legal judgment prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1023–1034.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062.

Masha Medvedeva, Ahmet Üstün, Xiao Xu, Michel Vols, and Martijn Wieling. 2021. Automatic judgement forecasting for pending applications of the european court of human rights. In *Proceedings of the Fifth Workshop on Automatec Semantic Analysis of Information in Legal Text (ASAIL 2021)*, pages 12–23. CEUR Workshop Proceedings.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28(2):237–266.

Masha Medvedeva, Martijn Wieling, and Michel Vols. 2023. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212.

Stuart S Nagel. 1963. Applying correlation analysis to case prediction. *Tex. L. Rev.*, 42:1006.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35.

T. Y. S. S Santosh, Oana Ichim, and Matthias Grabmair. 2023a. Zero shot transfer of article-aware legal outcome classification for european court of human rights cases. *arXiv preprint arXiv:2302.00609*.

TYS Santosh, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023b. Leveraging task dependency and contrastive learning for case outcome classification on european court of human rights cases. *arXiv preprint arXiv:2302.00768*.

TYS Santosh, Rashid Gustav Haddad, and Matthias Grabmair. 2024. Ecthr-pcr: A dataset for precedent understanding and prior case retrieval in the european court of human rights. *arXiv preprint arXiv:2404.00596*.

T.y.s.s Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022. Deconfounding legal judgment prediction for European court of human rights cases towards better alignment with experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1138, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeffrey A Segal. 1984. Predicting supreme court cases probabilistically: The search and seizure cases, 1962-1981. *American Political Science Review*, 78(4):891–900.

Mehmet Fatih Sert, Engin Yıldırım, and İrfan Haşlak. 2021. Using artificial intelligence to predict decisions of the turkish constitutional court. *Social Science Computer Review*, page 08944393211010398.

Rafe Athar Shaikh, Tirath Prasad Sahu, and Veena Anand. 2020. Predicting outcomes of legal cases based on legal factors using classifiers. *Procedia Computer Science*, 167:2393–2402.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348.

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal judgement prediction for uk courts. In *Proceedings of the 2020 the 3rd international conference on information science and system*, pages 204–209.

Octavia-Maria Şulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef van Genabith. 2017a. Exploring the use of text classification in the legal domain.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017b. Predicting the law area and decisions of french supreme court cases. In

Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 716–722.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599.

Santosh Tyss, Oana Ichim, and Matthias Grabmair. 2023a. Zero-shot transfer of article-aware legal outcome classification for european court of human rights cases. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 593–605.

Santosh Tyss, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023b. Leveraging task dependency and contrastive learning for case outcome classification on european court of human rights cases. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1103–1103.

Josef Valvoda, Ryan Cotterell, and Simone Teufel. 2023. On the role of negative precedent in legal outcome prediction. *Transactions of the Association for Computational Linguistics*, 11:34–48.

Bernhard Waltl, Georg Bonczek, Elena Scepankova, Jörg Landthaler, and Florian Matthes. 2017. Predicting the outcome of appeal decisions in germany's tax law. In *International conference on electronic participation*, pages 89–99. Springer.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with llm and domain-model collaboration. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Frank F Xu, Uri Alon, and Graham Neubig. 2023. Why do nearest neighbor language models work? In *International Conference on Machine Learning*, pages 38325–38341. PMLR.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4085–4091.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. Neurjudge: a circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 973–982.

Han Zhang, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023. Contrastive learning for legal judgment prediction. *ACM Transactions on Information Systems*, 41(4):1–25.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1250–1257.

A Implementation Details

We set the maximum token sequence length and maximum number of segments in hierarchical models to 128 and 64, respectively. We train the LJP models using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 3e-5, employing linear decay and 1000 warmup steps for up to 30 epochs. To optimize training efficiency, we utilize mixed precision and gradient accumulation techniques. The retriever module is initialized using the LegalBERT (Chalkidis et al., 2020) model and further pre-trained for 50,000 pairs of cases, which are sampled uniformly across the entire range of relevance scores from 0 to 1. We employ the Faiss library (Johnson et al., 2019) to construct a datastore of precedents, facilitating efficient similarity computations during retrieval. For incorporating precedents at inference time, we perform a grid search over the interpolation factor (λ) in increments of 0.1 within the range of [0, 1] to select the best model based on the validation set. Additionally, we vary the value of k over powers of 2 from 8 to 256. In training incorporation via the fusion layer, we set the number of stacked cross-attention layers to 4. The index is refreshed in joint training every epoch, and we set the number of retrieved precedents during training to 7.