CustodiAI: A System for Predicting Child Custody Outcomes

Yining Juan,[1] **Chung-Chi Chen,**[2] **Hsin-Hsi Chen,**[1] **Daw-Wei Wang**[3,4]
[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[2] AIST, Japan
[3] Department of Physics, National Tsing Hua University, Taiwan
[4] Center for the Applications and Developments of AI in Humanity and Social Sciences,
National Tsing Hua University, Taiwan
ynjuan@nlg.csie.ntu.edu.tw, c.c.chen@acm.org
hhchen@csie.ntu.edu.tw , dwwang@phys.nthu.edu.tw

Abstract

Predicting child custody decisions post-divorce is crucial but challenging due to numerous non-numerical, text-based factors, particularly in joint custody scenarios. This study presents the Intermediate Self-Supervised Training (ISST) method, a two-stage approach that classifies document paragraphs using original rationale labels before leveraging this to predict custody at the document level. Achieving up to 90.57% accuracy and notably, a 78.95% F1-score for joint custody cases, it surpasses previous models by 13%. We further refine the model to mitigate gender bias in the training data and provide error estimations, enhancing fairness and reliability. Our user-friendly online system exemplifies our model's applicability in out-of-court dispute resolution, potentially reducing time and financial strains for families in crisis.

1 Introduction

The principle of "best interests of the minor child" serves as the keystone of judicial decision-making in post-divorce custody matters. Nonetheless, the definition and interpretation of this standard are fraught with complexity. For example, Article 1055-1 of Taiwan's Civil Code,[1] which requires judges to factor in the parent's commitment and aptitude for the child's protection and education when determining custody. This, and similar statements found in the Civil Code, are laden with unstructured, descriptive human language that makes precise prediction of judicial outcomes a challenging endeavor.

The predictive challenge escalates when it comes to joint custody awards, a vital yet relatively uncommon outcome in legal proceedings. Deciding on sole custody—whether to grant it to the mother or father—translates into a binary classification problem, where it's feasible to determine which party is best suited for the child's welfare, both theoretically and practically. Conversely, real-world scenarios often entertain the possibility of joint custody, which introduces a more complex ternary classification task in machine learning terms. Neglecting this dimension could result in overlooking crucial factors when evaluating the child's best interest, consequently rendering the model ineffective for practical applications.

This work aims to devise a predictive machine learning model that leverages the nuance and depth of human language rationale, thus enhancing the efficiency of public dispute resolution. We propose the Intermediate Self-Supervised Training (ISST) method, built on expert-labeled training data, which incorporates a two-stage training process for custody's ternary classification. In our model, we've integrated 13 custody factors to bolster its explainability for the public. The model's overall accuracy peaks at 90.57%, and in cases involving joint custody, it attains an F_1-score as high as 78.95%, marking a 13% increase from previous models. Keeping in view the ethical implications of system deployment, we've also incorporated data augmentation and model error estimation to mitigate potential gender bias and avoid misinterpretation of predicted results. A user-friendly online demo landing page is available[2], offering users the option to select inputs in either text format, categorical items, or both. By significantly reducing the time and financial costs tied to legal processes, our work offers a promising tool for out-of-court dispute resolution.

2 Related Work

Over recent decades, the rise of Artificial Intelligence (AI) applications has instigated transformative changes in all facets of human existence. One such significantly impacted area is the judi-

[1] https://is.gd/VDimZk

[2] CustodiAI: https://demo-cjdp-aifr.herokuapp.com/demo-home

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations, pages 10–16
November 1–4, 2023.

cial sector, where researchers can leverage the vast corpus of high-quality data from court judgments to train ML models. Numerous studies have investigated applications utilizing data from legal decisions across different countries (Sulea et al., 2017; Luo et al., 2017; Kowsrihawat et al., 2018; Malik et al., 2021). In most of these applications, experts have labeled the data sources using either categorical or numerical variables.

While initial AI explorations in the judicial realm focused mainly on criminal cases, recent trends indicate a broader scope. This includes the adoption of Natural Language Processing (NLP) techniques to analyze verdicts, exemplified by studies on judgments from the European Court of Human Rights (Aletras et al., 2016), and later refined in subsequent works (Medvedeva et al., 2018). Furthermore, efforts by researchers like Chen & Eagel have contributed specialized datasets for unique contexts such as asylum adjudication (Chen and Eagel, 2017).

Despite the progress made in the field, Legal Judgment Prediction (LJP) faces several challenges, including the issue of AI model predictions' asymmetry in distinguishing positive and negative outcomes (Valvoda et al., 2023). This discrepancy is prevalent even with the emergence of models that simulate real-world court processes and human judge decision-making, indicating the complexity of the problem. Notably, models utilizing supervised contrastive learning frameworks have demonstrated their ability to differentiate between similar law articles and charges, mirroring the analytical skills of real-world judges (Zhang et al., 2023). However, there are still concerns regarding the occasional shortcomings of LJP models, particularly in accurately identifying nuanced critical events (Feng et al., 2022). To address this issue, the Event-based Prediction Model (EPM) focuses on capturing detailed information about specific events, providing a potential solution to enhance the performance of LJP models (Feng et al., 2022).

Our research focuses on analyzing child custody outcomes within the field of family law. While models developed by Long et al. show promising results when applied to carefully curated divorce datasets (Long et al., 2019), they may struggle to capture the complexities of real-world situations. Predicting court judgment outcomes, particularly in cases involving child custody, is a complex endeavor. The diverse range of potential outcomes, including shared custody arrangements and sole custody allocations, adds an additional layer of intricacy to the prediction process. This multifaceted landscape presents challenges when it comes to deploying machine learning models in public domains.

The significance of predicting legal outcomes extends beyond academic realms. As our research demonstrates, there is a critical demand for accurate prediction models, particularly in intricate situations such as post-divorce child custody. To address this need, we propose the implementation of the Intermediate Self-Supervised Training (ISST) method. This approach aims to enhance the reliability, fairness, and robustness of custody outcome predictions.

3 Data Preparation

We sourced court verdicts for this study from the Open Data of Taiwan Judicial Yuan(Taiwan Judicial Yuan, 2018), spanning January 1, 2015, to December 31, 2017. Using three keywords: "divorce", "discretion", and "the best interests of the child", we filtered unrelated verdicts, yielding a collection of 1,343 cases related to post-divorce child custody. After expert manual examination and labeling, we identified 529 cases where both parents sought custody. The court judgments fell into three categories: sole custody to the plaintiff (252 cases), sole custody to the defendant (184 cases), and joint custody (93 cases). We used 'plaintiff' and 'defendant' instead of 'father' and 'mother' to avoid potential gender bias in the data.

3.1 Rationale Paragraphs

Court judgment rationales, derived solely from the judge-authored text in verdicts, were labeled and classified into four categories related to custody decisions: "favorable to plaintiff (FP)", "unfavorable to the plaintiff (UP)", "favorable to defendant (FD)", and "unfavorable to the defendant (UD)". Each rationale ranged from 20 to 200 Chinese characters. Importantly, to avoid potential gender biases, plaintiffs' and defendants' genders were not labeled in the dataset. Below are examples of the four types of paragraphs labeled in our dataset:

(1) An FP rationale:

The plaintiff has stable parenting capacities, parenting time, a caring environment, educational planning, and a high motivation to obtain custody.

(2) A UP rationale:

The plaintiff, during this marital relationship, was incarcerated due to drug charges during pregnancy, which indicates an improper parental attitude and inability to provide a stable and positive child-raising environment.

(3) An FD rationale:

The child has resided with the defendant for an extended period without any negative circumstances in the living environment and parental support system.

(4) A UD rationale:

The defendant is employed as a university lecturer and also teaches a baking class, resulting in a weaker parental bond due to the distant family relationship.

In this study, we have labeled 4,153 paragraphs, encompassing 1,637 paragraphs favorable to the plaintiff, 423 unfavorable to the plaintiff, 1,475 favorable to the defendant, and 618 unfavorable to the defendant.

3.2 Custody Factors

In addition to the rationale paragraphs, we labeled 13 factors relevant to custody decisions. These factors, derived from Article 1055-1 of the Civil Code in Taiwan and previous studies, include: (1) Child's Age, (2) Child's Willingness, (3) Child's Development, (4) Parent's Occupation, (5) Parent's Financial Status, (6) Parent's Health Status, (7) Parent's Character, (8) Parent's Living Condition, (9) Primary Caregiver, (10) Parent's Willingness and Capability, (11) Parent's Friendliness to the Other Parent, (12) Parent-Child Affection, (13) Supporting System.

Upon examining all the verdicts in our dataset, we found that almost all the factors mentioned by judges regarding child custody decisions could be classified into the aforementioned 13 factors. For instance, the sample FP paragraph provided in the previous subsection could also be labeled by Parent's Willingness and Capability. Notably, there were cases where a paragraph was labeled with more than one factor due to the complex descriptions in verdicts.

4 System Description

The backbone of our system is the ISST model, a two-stage BERT-based model used for predicting child custody. The initial stage is a self-supervised training stage using rationale labels for paragraph classification, followed by the second stage, which is supervised fine-tuning on our target task. This stage utilizes the encoder weights produced by the initial stage. Figure 1 compares our approach with previous methods.

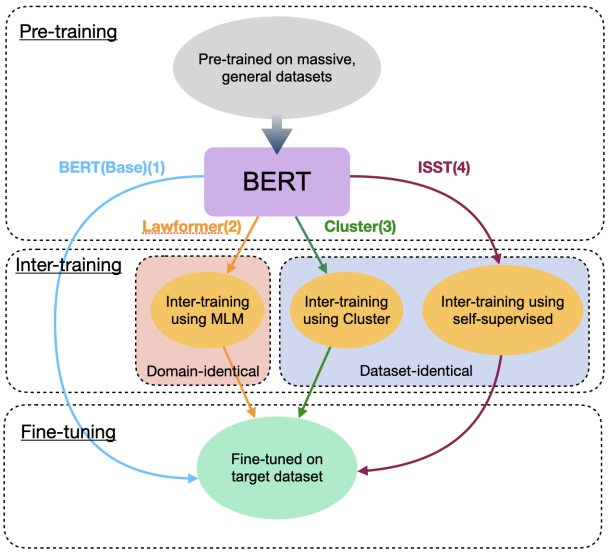

Figure 1: The diagram shows the different stages involved in developing a pre-trained model. During the pre-training phase, the model is trained on general corpora. In the inter-training phase, the Lawformer(Xiao et al., 2021) (illustrated in Path 2 of the figure) is introduced to data specific to the target domain and is trained to employ the Masked Language Model (MLM) approach. On the other hand, the method advanced by Shnarch, Eyal, et al. (Shnarch et al., 2022) (represented in Path 3 of the figure) as well as our proposed ISST(path 4 of the figure) exclusively utilize the dataset associated with the target task, devoid of data from other related domains, focusing on classification tasks.

4.1 First Stage: Intermediate Self-Supervised Training

Fine-tuning pre-trained BERT (Kenton and Toutanova, 2019) word vectors is a standard practice for contemporary NLP tasks. However, performance can vary greatly for task-specific datasets. To improve prediction accuracy for judicial decisions regarding child custody, we propose a self-supervised, intermediate task training method that aligns learned features more closely with our final goal.

While BERT's pre-trained word vectors are generalized, they may not be sufficiently precise for specialized tasks. We convert the original input features into an intermediate training target, enabling the model to learn representations more relevant to the task at hand. Although the accuracy of the final task can be enhanced through a super-

vised learning method with different data (Gururangan et al., 2020) or an unsupervised learning method (Shnarch et al., 2022), these approaches do not correlate strongly with the final task.

In contrast, our method involves training rationale paragraphs to be classified by their inherent attributes (Favorable or Unfavorable), which are highly correlated with the target task sentences, albeit not identical. In the first stage of training, it's crucial to ensure a balanced proportion of 'Favorable' and 'Unfavorable' rationale paragraphs. An imbalance in the dataset can introduce biases during the intermediate fine-tuning, which might adversely affect the subsequent target task fine-tuning. Then, the encoder weights thus obtained are used for the target task, which is the ternary classification of child custody using "effective verdicts". This intermediate task training is essentially a self-supervised learning approach that maps the input data (rationale paragraphs) onto the intrinsic attributes of the "effective verdicts". However, these attributes, despite their high correlation, might not yield optimal results when directly applied to predict child custody. As such, we leverage the resulting embedding weights of the rationale paragraphs instead of the original BERT word vectors for the second training stage.

4.2 Second Stage: Target Task Fine-Tuning

In the second stage, we add an initialized classifier layer on top of the BERT model, well-trained from the previous stage. We structure the input data as: $D = \{(p^{(FP)}, p^{(UP)}, p^{(FD)}, p^{(UD)})_i, y_i\}_{i=1}^{N}$, corresponding to rationale paragraphs Favorable to the Plaintiff (FP), Unfavorable to the Plaintiff (UP), Favorable to the Defendant (FD), and Unfavorable to the Defendant (UD). This input tuple is then mapped onto one of three judicial decision outcomes: $y \in \{Plaintiff, Defendant, BothParties\}$.

To utilize the cross-attention module in BERT, facilitating the fusion of fine-grained information within the paragraphs and their rationale and factor labels, we design the following prompt for each input item p_i^j with $j \in \{FP, UP, FD, UD\}$ to the BERT model:

$$p_i^j = \text{[CLS]}\ f_i^j\ \text{[SEP]}\ r_i^j\ \text{[SEP]}.$$

In this formulation, r_i^j represents the rationale paragraph favorable/unfavorable to the plaintiff/defendant, and f_i^j follows the form: "The party has [Sentiment] factors for [Factors]". The "Factors" slot is populated with Chinese words describing previously mentioned factors such as Child's Age or Primary Caregiver, and the "Sentiment" slot is populated with Chinese words that represent "Favorable" or "Unfavorable".

This prompt significantly enhances the user interface's flexibility: users can decide their preferred input style— rationale text only,r_i^j, itemized factors only,f_i^j, or keeps both r_i^j and f_i^j as input. We also augment only factors or rationale text data during the training phase, so that the model can support these three different inputs at the same time.

However, it's worth noting that the longest input sequence length accepted by $BERT_{BASE}$ is only 512 tokens. The total length of the inputs $p^{(FP)}, p^{(UP)}, p^{(FD)}, p^{(UD)}$ averages around 2549 tokens, meaning direct concatenation would result in a significant loss of valuable information. To circumvent this, we propose an iterative solution for this long text problem, as detailed in Algorithm 1: we process the four input items ($p^{(FP)}$, etc.) in sequence through BERT, obtaining their respective [CLS] embedding vectors of 768 dimensions. These embeddings are concatenated into a 3072-dimension ($= 768 \times 4$) vector representing the "effective verdict", which is then fed into the neural network classifier to produce the final prediction for child custody.

Algorithm 1 ISST algorithm

Input: $(\{p^{(FP)}, p^{(UP)}, p^{(FD)}, p^{(UD)}\}, y)$
Output: Θ

$\Theta \leftarrow$ Pre-trained from inter-training method.
$V = \{\}$
for *each* $(\{p^{(FP)}, p^{(UP)}, p^{(FD)}, p^{(UD)}\}, y)$ **do**
  for $j = \{FP, UP, FD, UD\}$ **do**
    V = V $\oplus$ BatchNorm(BERT(p^j))
  end for
  Obtain the predicted result
  $\hat{y} = argmin_{y^*}\ \ell(y^*|V)$
  Update the network via the gradient
  $\nabla_G \mathcal{L}((y^*|V))$
end for

5 Evaluation

In this section, we employ a 5-fold cross-validation approach to compare the performance of our ISST method with three other methods. Figure 1 highlights the differences among these techniques. The first method is the standard $BERT$(Kenton and Toutanova, 2019) model, which relies only on pre-trained encoder weights and excludes intermediate training. The second method, Lawformer(Xiao

	Sole Custody (Binary)			Sole Custody & Joint Custody (Three Categories)			
	Accuracy	$F_1(P)$	$F_1(D)$	Accuracy	$F_1(P)$	$F_1(D)$	$F_1(B)$
BERT	95.91±1.54	96.53±1.41	94.93±1.68	83.77±4.48	92.73±1.75	85.24±5.15	53.12±10.55
Lawformer	94.55±0.85	95.35±0.92	93.29±0.84	85.38±4.36	92.06±2.37	89.25±4.95	51.94±10.64
Cluster	97.73±1.44	98.02±1.28	97.3±1.72	85.85±3.3	95.74±1.62	84.62±6.01	65.00±3.77
ISST (Proposed)	**98.85±2.43**	**99.18±2.1**	**98.63±2.89**	**90.57±3.07**	**95.83±1.71**	**89.74±3.69**	**78.95±7.7**

Table 1: Experimental results under different task settings.

et al., 2021), involves intermediate training with Masked Language Model (MLM) tasks before fine-tuning. The third method, Cluster (Shnarch et al., 2022), employs an intermediate unsupervised learning approach for training.

Our analysis begins with the prediction results for sole custody cases, which involve binary classification, with custody awarded either to the Plaintiff or the Defendant. For these predictions, we utilize both rationale paragraphs and custody factors as input features. As Table 1 illustrates, our ISST approach demonstrates superior performance in terms of accuracy (Acc.) and F_1-scores for both Plaintiff ($F_1(P)$) and Defendant ($F_1(D)$).

Table 1 extends the comparison to include cases involving both sole custody and joint custody. This extends the problem to a ternary classification task, with custody awarded to the plaintiff, to the defendant, or to both parties. Our ISST model's overall accuracy exceeds 90%, outperforming all other models by at least 5%. When comparing the F_1-scores of the three classes, our model again secures the highest scores. The most significant difference, exceeding 27%, is observed in the cases of joint custody. This demonstrates that our ISST model significantly improves the prediction of child custody when joint custody is a potential outcome.

It is important to note that we also modified the task of Intermediate Self-Supervised Training (ISST) from determining the favorability or unfavorability of a rationale paragraph to identifying the specific custody factor to which the paragraph relates among the 13 available factors. Following this modification, we observed a decline in the performance of the ISST (specifically, the accuracy for predictions related to Sole Custody & Joint Custody was 83.46%, which is similar to the baseline performance achieved by simply fine-tuning BERT). This outcome aligns with the fact that when judges make custody rulings, they address the favorable/unfavorable conditions much more than which custody factors presented by each party in a holistic manner for their final decision.

This instructive result indicates that the purpose of the intermediate task should be to serve as a bridge between the pre-trained BERT and the distinct objectives of the target task. By integrating a task with a high correlation to the target during intermediate training, the expressiveness of BERT for the target task can be enhanced. Therefore, the careful selection of a task with a higher correlation to the target task for ISST is pivotal to optimizing the performance of the target task.

6 System

6.1 Ethical Considerations for System Deployment

Before deploying our ISST model, trained on real-world cases for public use, it's crucial to address potential ethical issues, particularly gender bias. Our labels "plaintiff" and "defendant", rather than "father" and "mother", might imply gender neutrality, but our dataset of 529 cases shows imbalance: fathers filed 42.34% and mothers 57.66% of the cases. Since only 19.64% of fathers and 68.20% of mothers were granted sole custody, gender bias may unintentionally manifest in our model.

To counter this, we've augmented the data by interchanging plaintiffs and defendants' features and outcomes in sole custody cases, creating a larger, gender-balanced dataset. Consequently, our *balanced* ISST model retains its accuracy while reducing the potential gender bias, enhancing its public application reliability.

In addition, We've also integrated an error estimation method using token-level perturbations on rationale statements, repeated 100 times on 5 differently seeded models. This yields 500 predictions, the standard deviation of which indicates prediction reliability. It is interesting to find that dominant probabilities (P>80%) typically have minimal errors (<5%), whereas tight contests (P<50%) show larger errors (>24%), suggesting users either provide additional information or ignore results. To offer users a more intuitive understanding of the trustworthiness of a given prediction, we trans-

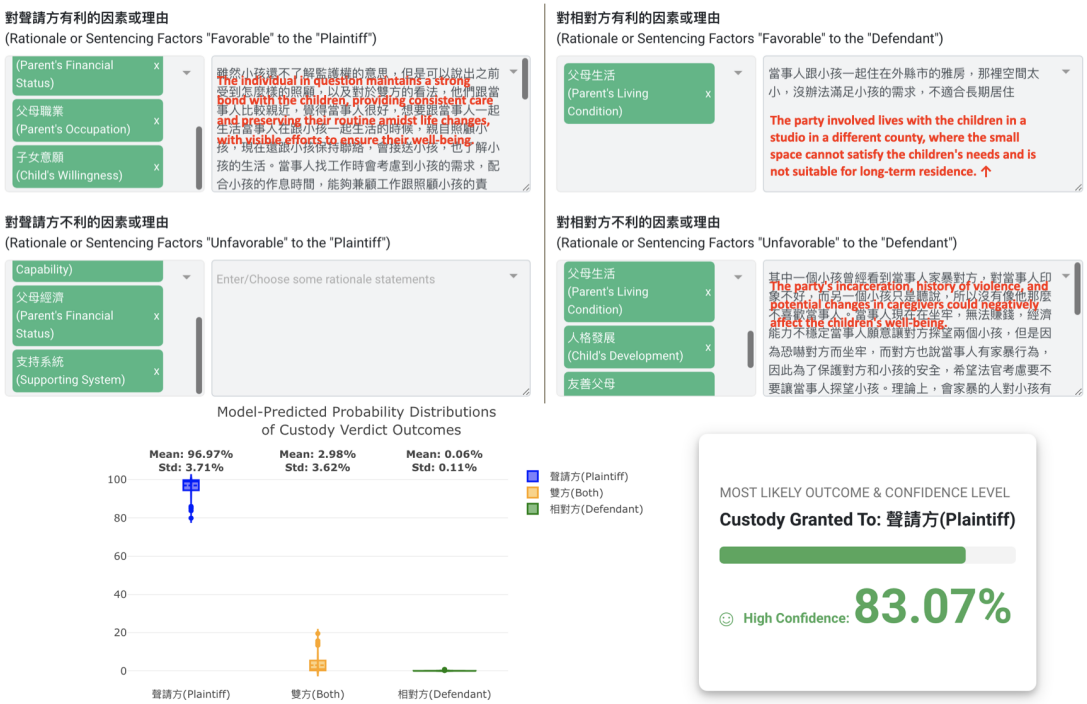

Figure 2: Screenshot of the proposed CustodiAI.

late this error estimation into a confidence value (0-1) using the formula $confidence = e^{-k*std}$. Based on our tests, we classify confidence values above 77.88% as 'high', below 30.11% as 'low', and those in between as 'moderate'.[3]

This adaptive adjustment helps avoid user misinterpretation of our system's outputs, boosting its trustworthiness and paving its way as a reliable AI solution for public use.

6.2 Demonstration

In this section, we showcase the functionality of the Intermediate Self-Supervised Training (ISST) system. Figure 2 shows the screenshot of the proposed CustodiAI. Additional details can be found in the accompanying demo video[4].

To simplify the user experience, we allow inputs of either rational statements or relevant custody factors. Sample statements based on court rulings are provided to help users articulate their cases effectively. Users can then select and adjust these templates to best fit their circumstances.

Contrary to most other comparable systems that only offer a single most probable result, our interface displays a violin plot of 500 predictions, indicating possible outcomes and their associated error estimations. A summary of the most probable judgment and its confidence level is displayed at the bottom right.

By pairing visually engaging prediction distributions with confidence scores, we can empower users to make informed decisions. The interactive interface's simplicity and flexibility deepen understanding of AI predictions, fostering trust and engagement. This approach helps offset possible model bias, reinforcing our system as a reliable public tool.

7 Conclusion

Our study introduces the novel Intermediate Self-Supervised Training (ISST) method for predicting complex child custody decisions, significantly improving upon previous models with an accuracy up to 90.57% and an F1-score of 78.95% in joint custody cases. We've ensured fairness and reliability by mitigating gender bias in the training data and providing error estimations for the predicted results. This work offers a tangible, user-friendly solution for out-of-court dispute resolution, marking a significant stride towards more efficient, accurate, and fair decision-making in child custody disputes.

[3]We empirically chose k=5 in the formula $confidence = e^{-k*std}$ to accentuate differences within the most common standard deviation range (0-0.24) in the testing dataset. Confidence thresholds of 30.11% and 77.88% were defined based on the confidence scores calculated from the first (q1) and third (q3) standard deviation quartiles, respectively, to delineate low, moderate, and high confidence levels.

[4]Demo video: https://reurl.cc/1eo9AV

Acknowledgments

We thank Prof. Yun-Hsien Diana Lin, Yun-Diao Lin, Ya-Lun Li and Hong-Hsiang Liu at National Tsing Hua University for the annotation of court verdicts and important discussion. This research is partially supported by the National Science and Technology Council, Taiwan, under grants MOST 109-2420-H-007-012-, MOST 109-2634-F-007-011-, MOST 110-2634-F-007-020-, MOST 110-2221-E-002-128-MY3 and NSTC 111-2634-F-002-023-. Yining Juan and Daw-Wei Wang were also supported by National Tsing Hua University through the Higher Education Sprout Project by the Ministry of Education(MOE) in Taiwan. The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956 and a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Limitations

The limitation of our method stays in the form of the original data: it has to contain rationale paragraphs with proper inner attribute labeling so that the intermediate self-supervised training can be applied. This type of input feature may not be available for other types of text and hence limits the application of our model. Furthermore, since these rationale paragraphs are originally written by judges, layman users of our system may not be able to generate the same type of writing for their input information, reducing the accuracy of the custody prediction when applied to the public.

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.

Daniel L Chen and Jess Eagel. 2017. Can machine learning help predict the outcome of asylum adjudications? In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 237–240.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Kankawin Kowsrihawat, Peerapon Vateekul, and Prachya Boonkwan. 2018. Predicting judicial decisions of criminal cases from thai supreme court using bi-directional gru with attention mechanism. In *2018 5th Asian Conference on Defense Technology (ACDT)*, pages 50–55. IEEE.

Shangbang Long, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Automatic judgment prediction via legal reading comprehension. In *Chinese Computational Linguistics: 18th China National Conference*, pages 558–572. Springer.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. *arXiv preprint arXiv:1707.09168*.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. *arXiv preprint arXiv:2105.13562*.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the european court of human rights: Looking into the crystal ball. In *Proceedings of the conference on empirical legal studies*, page 24.

Eyal Shnarch, Ariel Gera, Alon Halfon, Lena Dankin, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2022. Cluster & tune: Boost cold start performance in text classification. *arXiv preprint arXiv:2203.10581*.

Octavia-Maria Sulea, Marcos Zampieri, Mihaela Vela, and Josef Van Genabith. 2017. Predicting the law area and decisions of french supreme court cases. *arXiv preprint arXiv:1708.01681*.

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.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

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.