Legal Fairness Analysis via Treatment Effect Estimation

Anonymous ACL submission

Abstract

Legal fairness is one of the most important principles pursued by modern legal systems. Unfortunately, unfairness may be inevitably introduced in real-world cases due to both objective and subjective uncertainty, such as ambiguity in the law or practical bias in judgments. Existing works for fairness analysis mainly rely on labor-intensive element annotation for cases, which suffer from limited generalization ability. To address this issue, we propose to utilize large-scale textual data to perform quantitative legal fairness analysis via our **Ca**usal-based **L**egal **F**airness Measuring Framework (CaLF). To verify its effectiveness, we construct a legal-fairness dataset, and experimental results show that CaLF can accurately characterize the unfairness. Further, we adopt CaLF on a large-scale real-world dataset and come to several interesting experimental observations from the perspective of gender, age, and region.

1 Introduction

Legal fairness is the principle that each individual is supposed to be treated equally before the law without discrimination, and it is regarded as an essential element of advanced law systems (Browne et al., 2001). Nevertheless, it is hard to achieve absolute fairness in the real world, and unfair judgments are sometimes inevitable in reality (Arvey, 1979; Hammond, 1996). Table 1 shows an example of different judgments between two similar cases in the real world. If no reasonable justifications present, such judgment undoubtedly undermines the principle of fairness, whether it results from subjective or objective uncertainty. Therefore, it is crucial to measure the judgment differences caused by specific factors (e.g., region, gender), which may uncover legal unfairness and can help regulate the judicial practice and prevent unfair judgments.

Case A: Alice stole a diamond ring worth $35,000$ RMB from her friend. After arrested, Alice returned the stolen goods. Other circumstances of the defendant include confession and obtaining forgiveness. **Prison Term:** 1 year 2 months (suspended for 2 years). **Alice:** Female, age 20, from Region A.
Case B: Bob secretly stole an car (valued at $35,000$ RMB) from his ex-girlfriend and sold it. The stolen property was recovered and returned to the victim. The defendant confessed to the crime. **Prison Term:** 4 years 5 months. **Bob:** Male, age 39, from Region B.

Table 1: An example of different judgments between two similar theft cases in the real world. Crucial legal elements are denoted with underlines. Case details can be found in Appendix F.

Legal fairness analysis has been studied for decades (Douglas, 1949; Sheppard, 1985; Hoff, 1994; Reamer, 2005; Valvoda et al., 2021; Wang et al., 2021). Some quantitative studies attempt to use statistical methods to perform correlation analysis (Grogger and Ridgeway, 2006; Fryer Jr, 2019; Johnson et al., 2019), which cannot capture comprehensive information from complex factors and thus suffer from spurious correlation problem. To tackle this problem, causal inference is introduced to conduct causal effect analysis (Pierson et al., 2020; Gaebler et al., 2020; Knox et al., 2020). However, these works require to represent the cases with a few structured elements. Compared with original legal documents, the elemental representations need time-consuming and labor-intensive annotations. Thus, these methods are extremely restricted with generalization in the large-scale real-world analysis. Therefore, in this work, we aim to utilize large-scale legal documents to measure the unfairness, i.e., the causal effect on the judgment result.

However, the task is non-trivial, and there exist two crucial challenges: (1) Case Representation: Regarding the given textual legal cases, how to effectively generate expressive case representations for downstream analysis is a challenge. (2) Causal Effect Estimation: Legal judgments are usually influenced by various factors. How to estimate the causal effect between the factors and the judgment result is another challenge.

To address these issues, we propose a simple and effective **Ca**usal-based **L**egal **F**airness Measuring Framework (CaLF), which uses neural models to extract expressive text representation, and then adopt a re-weighting causal model, inverse propensity weighting (IPW) (Rosenbaum, 1987; Rosenbaum and Rubin, 1983), to estimate the causal effect. Specifically, we first normalize the case distribution across groups by assigning each case a weight calculated by neural models. Then we measure the unfairness as the difference between the weighted average judgment result of different groups. Taking gender as an example, if males often commit more serious crimes than females and thus receive heavier sentences, we adopt reweighting to balance the proportion of serious cases for the two genders, and the average judgment results can be compared for analysis.

Notably, CaLF can be applied to analyze the outcome of various judicial processes, including arrest, conviction, and sentencing, etc. In this paper, we choose the term of penalty (i.e., the outcome of the sentencing process) as our target for analysis, since it is the main punishment for offenders.

To verify the effectiveness of CaLF, we construct the first legal treatment effect estimation dataset, LegalTrEE. We annotate each case with factual legal elements and use a matching algorithm based on elemental trial (Cohen, 1982; Tadros and Tierney, 2004; Quintard-Morénas, 2010; Zhang, 2010) to get the counterfactual outcomes. Experimental results on LegalTrEE prove that CaLF can more accurately estimate causal effect than other models.

Furthermore, we adopt CaLF on the large-scale legal dataset from China, CAIL2018 (Xiao et al., 2018), to conduct the real-world legal fairness analysis. The experiment covers the perspective of age, gender, and region, while we also focus on 5 typical charges. From the result, we find CaLF can detect some noteworthy biases. The young tend to be sentenced to 0.8 months shorter than others on average, perhaps because of leniency towards students. Criminals in regions with high crime rates tend to be sentenced to 4.4 months longer than ones in regions with low crime rates, reflecting the traditional Chinese concept of "governing the country with severe law during trouble times".

To summarize, we make several noteworthy contributions in this paper:[1]

[1]We will release our code and dataset once accepted.

(1) We design a framework, CaLF, which utilizes large-scale legal documents for fairness analysis. Compared with previous works, CaLF has better applicability and performance.

(2) We build the first legal-domain treatment effect estimation dataset, LegalTrEE, on which we conduct comprehensive experiments to prove the reliability of CaLF.

(3) We perform fairness analysis on large-scale real-world court decision data from the perspective of age, gender, and regional equality.

We hope our approach and analysis can provide legal researchers or judicial practitioners a macro perspective on fairness, and thus promote related works and judicial equality around the world.

2 Related Work

2.1 Legal Fairness Analysis

Most of the current works on legal fairness are from a case-by-case or microcosmic perspective (Douglas, 1949; Sheppard, 1985; Tyler, 1988; Browne et al., 2001; Reamer, 2005; Hoff, 1994). Recently, many researchers attempt to analyze legal fairness quantitatively with statistical methods, such as correlation and regression analysis (Grogger and Ridgeway, 2006; Fryer Jr, 2019; Johnson et al., 2019), which cannot capture information from complex factors and suffer from spurious correlation problem. To tackle this issue, some researchers utilize the causal inference theory (Pierson et al., 2020; Gaebler et al., 2020; Knox et al., 2020). However, these works simplify the cases' facts to a few structured elements, which need high-cost annotation. Besides, Wang et al. (2021) attempt to analyze legal fairness from large-scale textual data, but the method is limited by the unsatisfactory performance of sentencing prediction models (Zhong et al., 2020b). These existing methods are restricted with generalization in practice.

2.2 Treatment Effect Estimation

Treatment effect estimation aims to evaluate the causal effect of a given treatment on the outcome (Yao et al., 2020). Previous works mainly use elementary vectors as covariates, so they cannot be applied to our textual study (Rosenbaum and Rubin, 1983; Rosenbaum, 1987; Rosenbaum and Rubin, 1985; Nie and Wager, 2017). In recent years, many researchers start to employ neural networks for text-oriented treatment effect estimation (Keith et al., 2020; Pham and Shen, 2017; Veitch et al., 2019).

However, these works rely on the counterfactual outcome prediction, which is greatly challenging, especially in the legal domain. Due to the unsatisfactory performance of existing prison term prediction models, introducing outcome prediction in our task will bring bias to the results.

2.3 Legal AI

Legal AI focuses on applying artificial intelligence technology to help legal tasks (Zhong et al., 2020b). In recent years, with the development of deep learning, many researchers introduce natural language processing (NLP) technology to Legal AI and achieve remarkable progress on many tasks, such as legal judgment prediction (Chen et al., 2019; Zhong et al., 2020a; He et al., 2019), similar case matching (Tran et al., 2019; Xiao et al., 2019), legal information extraction (Chen et al., 2020; Shen et al., 2020), and jurisprudential perspectives verification (Valvoda et al., 2021). However, few works attempt to employ advanced NLP technologies to analyze legal fairness.

3 Methodology

In this section, we first describe notations and the problem formulation of legal fairness, and then introduce the proposed **Ca**usal-based **L**egal **F**airness Measuring Framework (CaLF). Notably, since prison term is the main punishment for criminals, we select prison term as the analysis target. Our approach can be transferred to the analysis of other judicial processes, which is left for future work due to the limitation of accessible data.

3.1 Notations

We formalize the problem as a treatment effect estimation task. We use the triplet (X, Y, T) to represent a case:

Covariate (background) X. In causal inference theory, covariate X is the background information of each sample. In our problem, the covariate $X = (w_1, w_2, \ldots, w_l) \in \mathbb{R}^l$ represents the case's factual information in plain text, where l denotes the text length and w_i denotes the i-th token.

Outcome Y. We let the outcome $Y \in \mathbb{R}$ to denote the judgment result. To better quantificationally measure the unfairness, we take the prison term (unit: month) as the judgment result in this paper, so we have $Y \geq 0$. In practice, the outcome can also indicate other judgment results, such as fine, charged rate, etc.

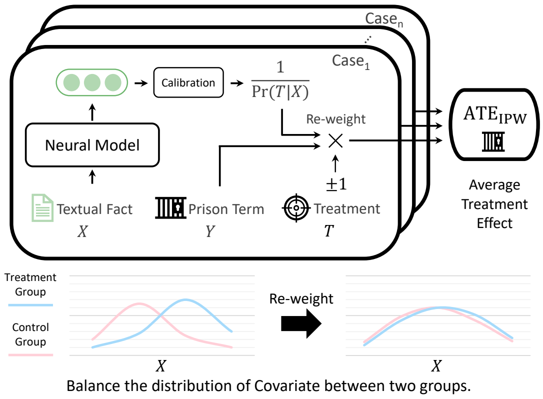

Figure 1: A schematic diagram of CaLF. We employ IPW to estimate ATE and use a neural model to estimate the propensity score.

Treatment T. The treatment $T \in \{0, 1\}$ is the potential unfair factor we study. In this paper, we take age (≤ 28 or > 28), gender (male or female), and region (south or north, GDP high or low, etc.) as T to detect the unfairness. In this way, samples are divided into two groups, the treatment group ($T = 1$) and the control group ($T = 0$).

Our goal is to estimate the average treatment effect (ATE):

$$\text{ATE} = \mathbb{E}\left[Y\left(T=1\right) - Y\left(T=0\right)\right], \qquad (1)$$

representing the measured value of unfairness. Intuitively, it indicates how many months samples in the treatment group are expected to be sentenced more than samples in the control group, on average.

For example, if we take gender as the treatment and set $T = 1$ for male defendants' cases while $T = 0$ for females', the ATE can be interpreted as the average term that men are sentenced to more than women if the criminal acts are the same.

3.2 CaLF

Figure 1 shows the schematic diagram of CaLF. The main idea of CaLF is to balance the data distribution between the treatment group and the control group via sample re-weighting. On this basis, ATE is calculated as the difference between the weighted mean prison term of the two groups.

Specifically, we utilize neural models to estimate each case's propensity score, i.e., the inverse case weight, from textual data and employ the inverse propensity weighting (IPW) method to estimate ATE. In the following sections, we will introduce the IPW and how to estimate the propensity score.

Inverse Propensity Weighting

The critical challenge in treatment effect estimation is that data distribution differs from groups, so we

Group	# Cases	Average				Treatment Effect
		# Chars	# Words	Y^{factual}	$Y^{\text{c-factual}}$	
All	3,086	492.43	277.58	4.981	4.935	0.956
Female ($T = 0$) Cases	1,580	476.80	269.32	3.892	4.780	0.889
Male ($T = 1$) Cases	1,506	508.83	286.23	6.124	5.097	1.027

Table 2: Statistics of LegalTrEE. Here $Y^{\text{factual}} = Y(T = t)$ represents the factual outcome, that is, the factual judgment of the case in the real world. $Y^{\text{c-factual}} = Y(T = 1 - t)$ represents the counterfactual outcome that we matched following elemental trial-based matching algorithm. The unit for Y^{factual}, $Y^{\text{c-factual}}$, and ATE is month.

cannot simply compare the two groups' mean values. Inverse propensity weighting (IPW) (Rosenbaum, 1987; Rosenbaum and Rubin, 1983) use reweighting to balance the data distribution between groups and thus to get accurate measured value.

As the gender example in the introduction, fairness does not mean having identical average sentences for men and women. If a case is likely to be in the male group but actually in the female group, IPW will adjust its weighting upwards to balance the bulk of similar male cases. More generally, the more abnormal the factual treatment is, the more the case weights. Specifically, the re-weight for each sample (x, y, t) is the inverse of the conditional probability $\Pr(T = t|X = x)$. Finally, the ATE is estimated as the weighted mean prison term.

Formally, we are to estimate the ATE given an estimating dataset $C^{\text{e}} = \left\{\left(x^{(i)}, y^{(i)}, t^{(i)}\right)\right\}_{i=1}^{|C^{\text{e}}|}$. Following previous works, treatment effect estimation relies on the two assumptions. One is unconfoundedness, which means legal documents contain sufficient information:

$$T \perp\!\!\!\perp Y(T = 0), Y(T = 1)\,|X. \tag{2}$$

The other is overlap, which means no case definitely belongs to a specific group:

$$0 < \Pr(T = 1|X) < 1. \tag{3}$$

In practice, both two assumptions can be satisfied for fairness analysis, when we employ the factual description as the covariate and the prison term as the outcome. Based on the two assumptions, we can employ inverse propensity weighting to estimate the ATE as:

$$\text{ATE}_{\text{IPW}} = \frac{1}{|C^{\text{e}}|} \sum_{i=1}^{|C^{\text{e}}|} y^{(i)} \left(\frac{t^{(i)}}{e\left(x^{(i)}\right)} - \frac{1 - t^{(i)}}{1 - e\left(x^{(i)}\right)} \right). \tag{4}$$

Here $e(x)$ represents the propensity score (Rosenbaum and Rubin, 1983), defined as the conditional probability of treatment given covariates:

$$e(x) = \Pr(T = 1|X = x). \tag{5}$$

Estimating Propensity Score

Following Equation 4, we can estimate the ATE with propensity score. In this paper, since we are to encode plain-text legal documents, we employ neural models to estimate propensity scores. For previous works in the field of causal inference, topic models and word counts are widely adopted to deal with texts. However, these methods will lose much of the complex semantic information in legal documents and thus are not suitable for our work.

Specifically, we formalize the task as a binary classification problem. We train the model predicting treatment T with the covariate X as input, and the propensity score $e(X)$ represents the output probability of $T = 1$.

In practice, we can employ BERT (Devlin et al., 2019) or any other NLP models to get text encoding, and then we use a linear layer and a softmax layer to predict the propensity score $e(x) = \hat{t}$. Besides, for training, we employ the cross-entropy loss function to optimize the model.

Neural networks suffer from the overconfidence issue, which means the model-predicted propensity scores are too close to 0 or 1 and are not the probability of maximum likelihood (Guo et al., 2017). Therefore, we employ calibration methods to adjusted the predicted propensity score $e^{\text{adj}}(x) = \text{Calib}(e(x))$. In this paper, we utilize temperature scaling (Hinton et al., 2015; Guo et al., 2017) to adjust the propensity score predicted by neural models. The main idea of temperature scaling is to train a single parameter to scale the hidden layer value of the neural model, thus adjusting the scale of the predicted probability. Please refer to Appendix D.1 for its detailed description.

There is another challenge for the neural network. As the sample numbers of $T = 0$ and $T = 1$ are usually unbalanced, neural models will overfit to the label with more samples, and the estimation of propensity score and ATE will be seriously affected. To resolve this problem, we balance the number of positive ($T = 1$) and negative ($T = 0$) samples by undersampling to make the model esti-

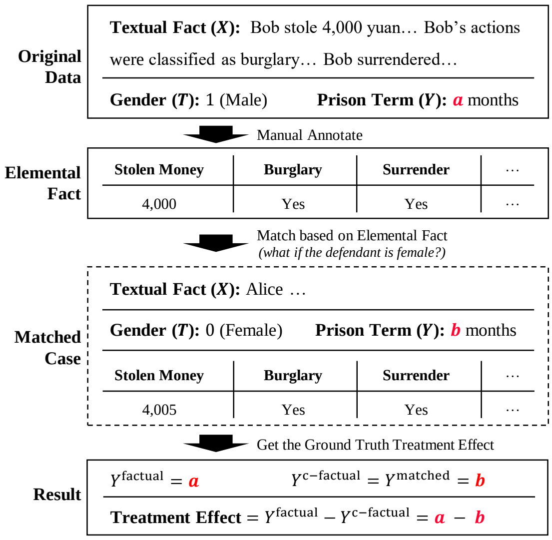

Figure 2: The schematic diagram of the construction of LegalTrEE. This is an example for one specific case, and all cases are processed following these steps.

mate the propensity score more accurately.

4 Dataset: LegalTrEE

To evaluate the effectiveness of CaLF and baselines, we construct the first legal causal dataset, **Legal Tr**eatment **E**ffect **E**stimation Dataset (LegalTrEE). Based on the elemental trial theory, we manually annotate the legal elements of each case and match cases with similar elements in two groups. Then average judgment differences between matched cases can be regarded as the ground-truth ATE (i.e., the unfairness), and thus we can compare the model-estimated ATE with the ground-truth ATE to simply get the model error. Figure 2 shows a schematic diagram of the construction of LegalTrEE.

We focus on China for analysis in this paper. The Supreme People's Court of China has published a large-scale legal document dataset, CAIL2018 (Xiao et al., 2018), which is currently one of the largest legal datasets and consists of millions of cases. It provides great data support for our work. Moreover, China has a large population and a vast territory, so there exist many complex factors (e.g., race, region) that may cause unfairness. Therefore, judicial unfairness in the Chinese legal system is worthy to explore. Therefore, we construct LegalTrEE based on CAIL2018 to verify the effectiveness of CaLF. Notably, the CAIL2018 dataset is also used for our large-scale analysis.

Theft is the charge we focus on in LegalTrEE because it has the most cases in the CAIL2018 dataset. We select only one charge because involving multiple charges in the dataset require too many legal elements to be considered and annotated. Gender is the treatment we focus on in LegalTrEE because it is one of the most talked-about forms of discrimination. We define $T = 1$ to represent the defender in the case is male, and $T = 0$ for female.

To build a treatment effect estimation dataset, how to get the counterfactual outcome (i.e., $Y(T = 1 - t)$) is a challenge because it is often non-observable. Existing works mainly use domain-specific knowledge to build semi-synthetic datasets (Yao et al., 2020). In this paper, the counterfactual outcome denotes what the judgment will be if the treatment is reversed and the covariate remains. According to elemental trial theory in the legal domain (Tadros and Tierney, 2004; Cohen, 1982; Quintard-Morénas, 2010; Zhang, 2010), judgments must be solely based on crucial legal elements from the case fact. Therefore, the case in which the treatment is reversed and the elements are close enough can represent the counterfactual outcome. On this basis, we use matching to find such cases to build the complete LegalTrEE dataset.

Referring to the relevant articles and legal professionals, we enumerate 16 essential legal elements for theft cases' sentencing. Then we pick thousands of theft cases from the CAIL2018 dataset and annotate them with these elements. We use a matching algorithm based on elemental trial to obtain these cases' counterfactual outcomes, i.e., $Y^{\text{c-factual}} = Y(T = 1 - t)$. Briefly, we match cases where the value of the stolen property is close and other elements are identical. Please refer to Appendix A for more details of the legal elements and the matching algorithm.

Removing atypical cases that cannot be matched, we finally obtain our LegalTrEE dataset with $3,086$ cases, of which $1,580$ are female cases ($T = 0$), and $1,506$ are male cases ($T = 1$). The statistics are shown in Table 2. From the table, we find males' average prison term (6.124 months) is longer than females' (3.892 months). However, there is an ATE of 0.956 months, which means that males expect to be sentenced to 0.956 months longer than females in the same criminal background.

Moreover, to check the accuracy of the matching scheme, we randomly sample 100 pairs of matched cases from LegalTrEE and invite legal professionals to help evaluate their similarity. Specifically, we define 4 levels of similarity and ask legal professionals to grade these case pairs. The result shows that 100% of the pairs achieve level 3 similarity (similar), and 81% of the pairs achieve level 4 (almost identical). This result proves the effectiveness and the reliability of the element design, the annotation, and the matching algorithm. Please refer to Appendix A.4 for more details.

5 Experiments on LegalTrEE

In this part, we test the performance of CaLF and the baseline methods on LegalTrEE.

5.1 Experimental Settings

In this section, we first take the textual case description as X, the prison term as Y, and the gender as treatment T to estimate ATE (the sentencing unfairness defined in this paper). Then, we compare the model-estimated ATE with the ground-truth ATE (0.956) to get the estimation error.

Models. We employ CNN (Kim, 2014) and BERT (Devlin et al., 2019) as the encoder for CaLF's propensity score estimation.

Dataset. We use LegalTrEE in this part of the experiment. We employ 3-fold cross-validation and randomly divide the train and test set by $2:1$.

Baselines. We compare our CaLF with several representative baselines. We test traditional element-oriented methods as baselines, where linear regression is used for the prison term estimation, and logistic regression is used for the propensity score estimation (Yao et al., 2020). We also introduce two neural causal methods as baselines: (1) Regression only (Regr.) (Keith et al., 2020), the method that only uses regression to predict factual and counterfactual prison terms and simply subtracts them to obtain ATE. (2) Targeted maximum likelihood estimator (TMLE) (Van Der Laan and Rubin, 2006), a doubly robust method that models both propensity score and outcome prediction to get better and more robust estimation performance. More specifically, TMLE subtracts the estimated prison term to get ATE like regression only but further uses propensity score and well-designed methods to adjust the regression-predicted prison term.

Please refer to Appendix C for more settings.

5.2 Experimental Results

The results are shown in Table 3. From the results, we can observe that CaLF with CNN and calibration can outperform other text-oriented methods, and the average analysis error is less than 10 days.

Method		avg ATE	avg$\lvert\delta\rvert$ (δ = ATE − GT)	std (δ)
Ground Truth (GT)		0.956	0	N/A
Element -Oriented Baseline	Regr.	0.897	0.059	**0.124**
	TMLE	0.904	0.052	0.241
	IPW	**0.991**	**0.035**	0.140
Neural Baseline	Regr. + CNN	-0.026 ±0.091	0.982	**0.208**
	Regr. + BERT	3.836±0.188	2.880	0.277
	TMLE + CNN	1.724 ± 0.050	0.768	0.243
	TMLE + BERT	2.574±0.299	1.618	0.670
CaLF	IPW + BERT	1.499 ± 0.347	0.543	0.973
	IPW + BERT w/ Calib.	1.398 ± 0.365	0.442	1.029
	IPW + CNN	1.448 ± 0.095	0.492	0.283
	IPW + CNN w/ Calib.	**1.170 ± 0.163**	**0.214**	0.508

Table 3: Experimental results for CaLF and baseline methods on LegalTrEE (unit: month). We employ 3-fold cross-validation and report the average ATE, average error, and errors' standard deviation. We repeat each experiment 10 times and report the 95% confidence interval of the results as $\mu \pm 1.96\frac{\sigma}{\sqrt{10}}$, where (μ, σ^2) are the mean and variance of the results.

Besides, we have the following observations about the experimental results.

(1) The calibration method improves the performance. Whether for CNN or BERT, the calibration (temperature scaling) can improve the ATE prediction performance by adjusting the propensity scores. We also compare the performance of several calibration approaches and conduct error analysis in Appendix D.

(2) Both two neural baselines have worse performance than CaLF. This problem is likely to be caused by the unsatisfied performance of the prison term prediction model (Zhong et al., 2020b; Chen et al., 2019), which brings bias to the results. Therefore, these baselines that need prison term prediction is not suitable for our work.

(3) The performance of BERT-based methods is worse than CNN-based methods. From the observation, we find that BERT suffers from the overfitting problem and usually captures subtle features that are irrelevant to the judgment. Thus, BERT can predict treatment labels accurately but fails to accurately estimate the propensity scores.

(4) Element-oriented regression achieve the best performance among all methods, even better than CaLF with CNN and calibration. Since element-oriented approaches introduce legal knowledge to the problem and simplify cases to a few elements, they can perform well on the regression task of prison term prediction. However, these methods are not comparable to text-oriented methods, which can be easily applied to the analysis of large-scale textual legal documents. In contrast, if we want to use element-oriented methods for such analy-

Factor	Specifically According to	$T = 1$	$T = 0$
Gender	Gender assignment	Male	Female
Age	Age at court session	≤ 28	> 28
Region (South or North)	Qinling-Huaihe Line	South	North
Region (GDP)	Ranking of GDP per capita	Top 10	Bottom 10
Region (Crime Rate)	Ranking of crime rate	Top 5	Bottom 5

Table 4: The descriptions of the factors.

Charge	Average Treatment Effect (ATE)				
	Age	Gender	Region Split by		
			S or N	GDP	CR
Overall	-0.8±0.4	0.9±0.7	1.8±0.9	0.6±0.6	4.4±0.8
Drug Trafficking	-9.6±2.9	0.4±1.1	9.6±3.2	-7.0±4.2	3.7±2.5
Theft	-0.8±0.5	0.8±0.8	3.2±0.6	2.3±0.6	6.5±1.0
Intentional Injury	-0.6±0.6	-0.4±0.4	-2.8±2.1	3.6±1.5	4.4±2.5
Traffic Offence	0.2±0.6	0.9±0.7	-1.7±1.6	0.4±0.8	-1.3±1.2
Providing Venues for Drug Users	0.1±0.1	0.6±0.2	-1.4±0.2	2.3±0.5	1.2±0.5

Table 5: The experimental results of the average treatment effect (ATE) (unit: month). For example, in theft cases, the youth (age ≤ 28) expect to be sentenced to 0.8 ± 0.5 months shorter than others (age > 28) given the same criminal background. The results show that there is little judicial unfairness in China generally. The results which are considered unfair (with absolute values over 3 months) are denoted with underlines.

sis, the high-cost manual annotation is necessary, which makes the task highly unacceptable.

6 Analyses on Massive Real-World Data

In this part, we conduct experiments on CAIL2018 and attempt to measure the sentencing unfairness of the criminals in China. We take gender, age, and region as treatments to measure the sentencing unfairness. Besides, we select 5 typical charges to further evaluate the unfairness of specific crimes. Notably, we conducted the analyses strictly following the guidance of legal experts to ensure the reasonableness and reliability of the results.

6.1 Experimental Settings

In this section, we first take the textual case description as X, the prison term as Y, and different factors as T to evaluate the ATE. Then, we conduct analyses based on our experimental results.

Treatments. As Table 4 shows, we select gender, age, and region as treatments (factors) for experiments and analyses. Gender is defined biologically as male or female. Age is divided as ≤ 28 or > 28 because the age of 28 is considered as the standard of whether a citizen is mature enough to take the responsibilities (Zhou, 2018). Region is used to test if the human geographical environment, regional economic status, and crime rate will affect the judgment results. In this paper, regions are divided by the provincial administrative units of China.

Model. We use CaLF with CNN and calibration (temperature scaling) here for analysis, because it outperforms other models in Section 5.

Dataset. CAIL2018 is used in this part, and we totally introduce about 3×10^5 cases for the large-scale analysis. For each experiment, we randomly divide the train and test set by $2 : 1$. The dataset statistics can be found in Appendix B.

6.2 Experimental Results and Analyses

Table 5 shows the experimental results of the unfairness (ATE) measured by CaLF based on our settings and within our dataset. Intuitively, the measured results represent that in our experimental dataset, how much more the treatment group will be sentenced than the control group on average. For example, gender only causes 0.9 months' sentencing bias (favoring women) for all criminals, and for drug trafficking cases, the number decreases to 0.4.

We can find that the measured ATE value varies from different scenarios. According to legal experts, 3 months is generally the minimum unit of sentencing in the Chinese legal system, so we take it as our threshold of unfairness. In this way, 70% scenarios in our experiment can be identified as fair, while there are also 30% of bias results. For example, age plays a significant role in drug trafficking cases according to our model. Besides, the bias we detected is concentrated in cases of specific charges, and the overall fairness is acceptable, except the perspective of regional crime rates.

From the results, we have the following observations with jurisprudential supports.

(1) The youth are favored. For either overall or specific charges, young are often sentenced shorter according to the experimental results. As mentioned above, the age of 28 is thought as the standard of whether a citizen is mature (Zhou, 2018). Further, those not older than 28 include a large group of students. Therefore, the observation that youth are often favored can be explained that judges tend to give more forgiveness and leniency to immature young people and students.

(2) "Governing the country with severe law during trouble times". Overall, criminals in areas with high crime rates tend to be sentenced 4.4 months longer than ones in areas with low crime rates, as is the situation for most charges. This traditional Chinese concept is recorded in the Rites of Zhou. In the modern Chinese legal system, it is also well documented. The thought of retribution sets the upper limit of a crime, while the aim of prevention might reduce the sentence (Zhang, 2011). In other words, it is necessary for judges to have discretion power for the purpose of prevention. In western criminal policy theory, the deterrence theory is a similar concept (Paternoster, 2010). The core idea of deterrence is that offenders may weigh the costs and benefits of crime, so when people feel that security is deteriorating, it is easy to think that "crime can be reduced by increasing penalties."

Besides, we can find that although there is no significant south-north or regional economical bias overall, some partial differences for specific charges seem to exist. Regional difference is a complex topic in China (Wu, 2001; Talhelm et al., 2014; Liu et al., 2018), and the regional judgment bias may be caused by complex factors, e.g., the cultures and customs, the development. Therefore, we think it is a topic worthy of in-depth study and analysis. There are also some other interesting experimental observations that we cannot explain right now. For example, the measured unfairness of drug trafficking cases is significant from most perspectives. Since we do find sufficient jurisprudential support for them, we cannot arbitrarily come to any conclusions based on these observations, so we will leave these as our future work, and also to the legal community.

6.3 Discussion

Since legal fairness is a principled and serious topic, it is necessary to further discuss the potential risks of our approach. Here we list several important issues which may lead to biased results.

(1) Data collection. We collect our dataset from the cases published by the Chinese government. Due to confidentiality, there are still some non-public cases, which means that there may be a distribution difference between the collected data and the real-world data. If such differences exist, the results will be unreal. **(2) The subjectivity in legal documents.** Legal practitioners strive to follow the guidelines of objectivity and comprehensiveness in the process of writing legal documents. However, there are no golden rules for writing legal documents and it is difficult to achieve absolutely objective. The inevitable subjectivity in the legal documents may introduce bias to the result. **(3) The two assumptions.** As mentioned in Section 3, IPW is based on the unconfoundedness and overlap assumption. If relevant criminal information is missing (unconfoundedness violated) or the case distribution of the two groups does not overlap (overlap violated), then the IPW-measured ATE will be influenced. **(4) Limitations of models.** Regardless of the model employed, there are inevitably prediction errors, leading to biased propensity scores and thus affect measured ATE. In this paper, we employ calibration to ensure the model accuracy to the utmost extent (detailed analysis can be found in Appendix D.3). We also encourage the community to improve the model performance in future works.

7 Conclusion and Future Work

In this paper, we formalize legal fairness analysis as the treatment effect estimation task and propose CaLF, a **Ca**usal-based **L**egal **F**airness Measuring Framework. We build the first legal treatment effect estimation dataset LegalTrEE to verify the effectiveness of CaLF. Then we conduct large-scale experiments on CAIL2018 to analyze the sentencing unfairness of the criminals in China.

We will explore the following directions in the future: (1) We will combine legal knowledge to carry out more in-depth analysis and give comprehensive explanations of more experimental observations. (2) Since the legal systems of different countries are very different, we will attempt to conduct legal fairness analysis for other countries. Given sufficient open data, such analysis and comparison will be interesting, as well as of great importance.

We hope with the development of legal fairness analysis, legal judgments around the world can become more transparent and fair, and equality before the law can be truly achieved.

Ethical Considerations

In this paper, we aim to leverage AI technology for legal fairness analysis. The goal of this work is to give a macro perspective for the legal domain and legal experts, thus promote equality and non-discrimination around the world. We do NOT aim to praise or criticize any country's legal system or for any political purpose.

Since this work is concerned with an NLP application in the legal domain, it is necessary to discuss several potential ethical issues here.

Intended Use

Usage. CaLF mainly focuses on utilizing large-scale legal documents to analyze legal fairness. We hope that modern NLP and Legal AI techniques can help quantify and promote legal fairness in the real world. Notably, this does not mean that we are challenging the authority or the standing of traditional jurisprudence. The goal of legal intelligence is to use AI technology to help legal tasks and provide various supports to judicial practitioners, instead of replacing them or competing with them. As such, we argue that CaLF can and can only assist judicial practitioners or legal experts in their works.

Failure Mode. There are inevitably prediction errors in CaLF. Therefore, as mentioned above, CaLF's result can only be used as a reference or a corroboration instead of the main evidence for any conclusions. In this way, the results of CaLF can also be validated by jurisprudence, and the potential impact of errors can be well limited.

Misuse Potential. We demand that anyone cannot make conclusions about any country's legal system only based on CaLF. Without jurisprudential evidence or professional research, such conclusions are undoubtedly arbitrary, and this kind of misuse seriously violates our motivation as well as the principle of legal intelligence.

Scope of Our Analysis

In this paper, we focus on the sentencing process of the trial stage for fairness analysis, only for those who are convicted. Besides the sentencing process and the trial stage, there can be unfairness in many other parts of the legal system, such as the filing stage and the prosecution stage. Due to the data limitation, we leave these for our future work, and we greatly hope to construct a more comprehensive dataset to improve related works and further promote the transparency of the legal system.

This is also a special reminder of the limitations of our analysis and experimental results. Our results are not representative of the global legal system. Everyone should notice the serious risks (especially political risk) of misinterpreting our results or misusing our analysis.

Manual Annotation

In this paper, we construct a dataset LegalTrEE via manual annotation. During the annotation stage, we first annotate some cases on our own to approximate the workload, and then we determine annotators' wages based on local standards.

Data Privacy and Anonymization

All the legal documents we used in our work are published by the Supreme People's Court of China, and the participant names are anonymized.

References

Richard D Arvey. 1979. Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin*, 86(4):736.

Judith A Browne, Daniel J Losen, and Johanna Wald. 2001. Zero tolerance: Unfair, with little recourse. *New directions for youth development*, 2001(92):73–99.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. In *Proceedings of EMNLP*.

Yanguang Chen, Yuanyuan Sun, Zhihao Yang, and Hongfei Lin. 2020. Joint entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of COLING*, pages 1561–1571.

Jerome Alan Cohen. 1982. The criminal procedure law of the people's republic of china. *The Journal of Criminal Law and Criminology*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

William O Douglas. 1949. Stare decisis. *Columbia Law Review*, 49(6):735–758.

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. 2001. *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York.

Roland G Fryer Jr. 2019. An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127(3):1210–1261.

Johann Gaebler, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel, and Jennifer Hill. 2020. A causal framework for observational studies of discrimination. *arXiv preprint arXiv:2006.12460*.

Jeffrey Grogger and Greg Ridgeway. 2006. Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, 101(475):878–887.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of ICML*, pages 1321–1330. PMLR.

Kenneth R Hammond. 1996. *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. Oxford University Press on Demand.

Congqing He, Li Peng, Yuquan Le, Jiawei He, and Xiangyu Zhu. 2019. Secaps: A sequence enhanced capsule model for charge prediction. In *Proceedings of ICANN*, pages 227–239. Springer.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Joan Hoff. 1994. *Law, gender, and injustice: A legal history of US women*, volume 1. NYU Press.

David J Johnson, Trevor Tress, Nicole Burkel, Carley Taylor, and Joseph Cesario. 2019. Officer characteristics and racial disparities in fatal officer-involved shootings. *PNAS*, 116(32):15877–15882.

Katherine A. Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using rext to remove confounding from causal estimates. In *Proceedings of ACL*, pages 5332–5344.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Dean Knox, Will Lowe, and Jonathan Mummolo. 2020. Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of ACL*, pages 138–143.

Siyang Liu, Shujia Huang, Fang Chen, Lijian Zhao, Yuying Yuan, Stephen Starko Francis, Lin Fang, Zilong Li, Long Lin, Rong Liu, et al. 2018. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and chinese population history. *Cell*, 175(2):347–359.

Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Wilson Leão, Steeven Janny, and Christian Gagné. 2018. Attended temperature scaling: a practical approach for calibrating deep neural networks. *arXiv preprint arXiv:1810.11586*.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of AAAI*.

Xinkun Nie and Stefan Wager. 2017. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*.

Raymond Paternoster. 2010. How much do we really know about criminal deterrence. *The journal of criminal law and criminology*, pages 765–824.

Thai T Pham and Yuanyuan Shen. 2017. A deep causal inference approach to measuring the effects of forming group loans in online non-profit microfinance platform. *arXiv preprint arXiv:1706.02795*.

Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. 2020. A large-scale analysis of racial disparities in police stops across the united states. *Nature human behaviour*, pages 1–10.

François Quintard-Morénas. 2010. The presumption of innocence in the french and anglo-american legal traditions. *The American Journal of Comparative Law*, 58(1):107–149.

Frederic G Reamer. 2005. Ethical and legal standards in social work: Consistency and conflict. *Families in society-The journal of contemporary social services*, 86(2):163–169.

Paul R Rosenbaum. 1987. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Paul R Rosenbaum and Donald B Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Shirong Shen, Guilin Qi, Zhen Li, Sheng Bi, and Lusheng Wang. 2020. Hierarchical chinese legal event extraction via pedal attention mechanism. In *Proceedings of COLING*, pages 100–113.

Blair H Sheppard. 1985. Justice is no simple matter: Case for elaborating our model of procedural fairness. *Journal of Personality and Social Psychology*, 49(4):953.

Victor Tadros and Stephen Tierney. 2004. The presumption of innocence and the human rights act. *The Modern Law Review*, 67(3):402–434.

Thomas Talhelm, Xiao Zhang, Shige Oishi, Chen Shimin, Dechao Duan, Xiaoli Lan, and Shinobu Kitayama. 2014. Large-scale psychological differences within china explained by rice versus wheat agriculture. *Science*, 344(6184):603–608.

Vu Tran, Minh Le Nguyen, and Ken Satoh. 2019. Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. In *Proceedings of ICAIL*.

Tom R Tyler. 1988. What is procedural justice-criteria used by citizens to assess the fairness of legal procedures. *Law & Soc'y Rev.*, 22:103.

Josef Valvoda, Tiago Pimentel, Niklas Stoehr, Ryan Cotterell, and Simone Teufel. 2021. What about the precedent: An information-theoretic analysis of common law. In *Proceedings of NAACL*, pages 2275–2288.

Mark J Van Der Laan and Daniel Rubin. 2006. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1).

Victor Veitch, Dhanya Sridhar, and David M Blei. 2019. Using text embeddings for causal iinference. *arXiv preprint arXiv:1905.12741*.

Yuzhong Wang, Chaojun Xiao, Shirong Ma, Haoxi Zhong, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2021. Equality before the law: Legal judgment consistency analysis for fairness. *arXiv preprint arXiv:2103.13868*.

Dianting Wu. 2001. A study on north-south differences in economic growth (in chinese). *Geographical Research*, 2.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, et al. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962*.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2020. A survey on causal inference. *arXiv preprint arXiv:2002.02770*.

Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, volume 1, pages 609–616.

Mingkai Zhang. 2010. Responsibilism and sentencing principles - centered on the point theory (in chinese). *Chinese Journal of Law*, pages 128–145.

Mingkai Zhang. 2011. *Criminal Law (in Chinese)*. Law Press. China.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of AAAI*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. How does nlp benefit legal system: A summary of legal artificial intelligence. In *Proceedings of ACL*.

Qiang Zhou. 2018. Notes on the "law of the people's republic of china on people's jurors (draft)" (in chinese). *Communiqué of the Standing Committee of the National People's Congress of the People's Republic of China*, 000(003):381–384.

A LegalTrEE: Legal Treatment Effect Estimation Dataset

In this part, we describe the details of our new dataset, the **Legal Tre**atment **E**ffect **E**stimation Dataset, LegalTrEE.

A.1 Elements Description

For each case, we represent it as an element vector, $X \in \mathbb{R}^{16}$. In other words, there are 16 elements related to the theft cases' sentencing, according to the Chinese Criminal Law, and "The interpretation of several issues on the application of the law in handling criminal cases of theft"[2] published by the Chinese Supreme People's Court and Chinese Supreme People's Procuratorate. Here we describe these elements in turn.

Amount of theft: $x_1 \in \mathbb{R}$. The amount of stolen money or properties. The core element for theft cases' sentencing. For the stolen objects, the value is based on the valuation in the legal instrument.

Ratio of refund: $x_2 \in [0, 1]$. If a person commits theft, he/she may be mitigated if he/she returns the stolen goods or makes restitution. We define x_2 to represent the restitution as a percentage of the amount of theft.

Level of theft amount: $x_3 \in \{0, 1, 2, 3\}$. The level of theft amount is divided into relatively large, huge, and especially huge. The standards of these levels vary slightly from region to region. For this dataset, we select cases from several specific regions, so that the standards for relatively large, huge, and especially huge are $2,000$, $60,000$, and $400,000$ yuan, respectively. Here we use $x_3 = 1$ to represent the amount is relatively large, use $x_3 = 2$ to represent the amount is huge, use $x_3 = 3$ to represent the amount is especially huge, and use

[2] http://www.court.gov.cn/shenpan-xiangqing-6622.html

$x_3 = 0$ to represent the amount does not reach a relatively large amount.

Burglary: $x_4 \in \{0, 1\}$. The element is used to represent whether the criminal intrudes into another person's residence to steal.

Multiple thefts: $x_5 \in \{0, 1\}$. The element is used to represent committed thefts more than three times within two years.

With a murder weapon: $x_6 \in \{0, 1\}$. The element is used to represent whether theft with firearms, explosives, control knives, other instruments prohibited by the government, or other instruments sufficient to endanger others' safety.

Pickpocketing: $x_7 \in \{0, 1\}$. The element is used to represent theft of property carried by others in public places or public transport.

Minors: $x_8 \in \{0, 1\}$. Minors under the age of 18 are persons of limited criminal responsibility and shall be punished less severely.

75 years old or older: $x_9 \in \{0, 1\}$. A person over the age of 75 is of limited criminal responsibility and shall be punished less severely.

Psychosis: $x_{10} \in \{0, 1\}$. Psychosis who have not yet completely lost the ability to recognize or control can be punished less severely.

Voluntary surrender: $x_{11} \in \{0, 1\}$. (Article 67 of the Chinese Criminal Law) Voluntary surrender refers to the act of voluntarily delivering oneself up to justice and truthfully confessing one's crime after one has committed the crime. Any criminal who voluntarily surrenders may be given a mitigated punishment. The ones whose crimes are relatively minor may be exempted from punishment.

Recidivism: $x_{12} \in \{0, 1\}$. (Article 65 of the Chinese Criminal Law) If a criminal commits another crime punishable by fixed-term imprisonment or heavier penalty within five years after serving his sentence of not less than fixed-term imprisonment or receiving a pardon, he is a recidivist and shall be given a heavier punishment.

Criminal attempt: $x_{13} \in \{0, 1\}$. (Article 23 of the Chinese Criminal Law) A criminal attempt refers to a case where an offender has already started to commit a crime but is prevented from completing it for reasons independent of his will. An offender who attempts to commit a crime may, in comparison with one who completes the crime, be given a lighter or mitigated punishment.

Forgiven: $x_{14} \in \{0, 1\}$. For those obtain forgiveness from the victim, the punishment can be reduced.

Other aggravating circumstances: $x_{15} \in \{0, 1\}$. The element is used to represent whether there are other aggravating circumstances.

Other mitigating circumstances: $x_{16} \in \{0, 1\}$. The element is used to represent whether there are other mitigating circumstances.

For those binary elements that take a value in $\{0, 1\}$, we have 1 means yes, and 0 means no. For x_{15} and x_{16}, the description in the judgment document prevails.

A.2 Inter-annotator Agreement

The Krippendorff's alpha of the annotation is over 0.94.

A.3 Matching Algorithm

We first define the matching score of cases and then find matched cases between the treated and control group according to the matching score.

Definition of Matching Score

For case $c_{\mathrm{A}} = \left(x^{\mathrm{A}}, y^{\mathrm{A}}, t^{\mathrm{A}}\right)$, and case $c_{\mathrm{B}} = \left(x^{\mathrm{B}}, y^{\mathrm{B}}, t^{\mathrm{B}} = 1 - t^{\mathrm{A}}\right)$, we specify that they are matchable if and only if $x_i^{\mathrm{A}} = x_i^{\mathrm{B}}, \forall i = 4, 5, \ldots, 16$. Then, for the two matchable cases, we define the matching score of c_{B} to c_{A} as:

$$\begin{aligned} \mathrm{match}_{c_{\mathrm{A}}}\left(c_{\mathrm{B}}\right) =& \theta - \alpha \frac{|x_1^{\mathrm{A}} - x_1^{\mathrm{B}}|}{\max\left\{x_1^{\mathrm{A}}, \delta\right\}} \\ & - \beta |x_2^{\mathrm{A}} - x_2^{\mathrm{B}}| \\ & - \gamma |x_3^{\mathrm{A}} - x_3^{\mathrm{B}}|. \end{aligned} \quad (6)$$

Here $\alpha, \beta, \gamma, \delta, \theta$ are parameters.

The main idea of the matching score is based on the elemental trial, i.e., judgment should be only correlated with the legal element. Therefore, for those $x_i\,(i = 4, 5, \ldots, 16)$ do not match, we consider that they are unmatchable; otherwise, we hope the other three elements, the amount of theft (x_1), the ratio of refund (x_2), and the level of theft amount (x_3) to be as close as possible.

Matching for Counterfactual Outcomes

For each case c in the treated group, we will try to find a matched case c_{M} in the control group. Vice versa, for each case c in the control group, we will try to find c_{M} in the treated group. If we find such a well-matched case, we set the counterfactual outcome of case c as the factual outcome of case c_{M}, i.e., we set $Y^{\text{c-factual}}(c) = Y(c_{\mathrm{M}})$.

Specifically, for case c, we first get the maximum matching score that other cases can achieve with it:

$$m_c = \max_{c_{\mathrm{m}} \in C_{\text{matchable}}} \mathrm{match}_c\left(c_{\mathrm{m}}\right). \tag{7}$$

As a particular case, for those where no matchable cases can be found, and for those where $m_c < 0$, we simply remove them from the dataset. Then, we construct the candidate set containing cases with a similar matching score:

$$C_{\text{candidate}} = \left\{c_{\mathrm{m}} \middle| \mathrm{match}_c\left(c_{\mathrm{m}}\right) + \epsilon \geq m_c\right\}. \tag{8}$$

Here ϵ is another parameter for the algorithm. Finally, we randomly select a case in the candidate set with equal probability as the matching case for c. Since the cases in the candidate set are all similar enough to the cases to be matched, the random selection simulates the randomness of judges.

Parameter Values

For the matching algorithm, we have 6 parameters, α, β, γ, δ, θ, and ϵ. The value of these parameters for building LegalTrEE is shown in Table 6. All the parameters are determined in strict compliance with the legal experts.

α	β	γ	δ	θ	ϵ
5.0	2.5	0.5	100	1.0	0.01

Table 6: Parameter values of matching in practice.

A.4 Evaluation of Matching

To verify the effectiveness of the matching scheme, we randomly pick 32 matched case pairs (about 1% of all) and invite legal professionals to check their similarity. Specifically, we define four similarity levels, and the legal professionals are required to grade each pair of cases after careful discussion. The description of the four levels and the feedback from legal professionals are shown in Table 7.

From the result, we can find that the picked case pairs are all at least similar, and most (81%) of them are almost identical. This result demonstrates that the element designing, the annotation, and the matching scheme are all reliable to a great extent.

B Dataset Statistics for the Large-Scale Analysis

Table 8 show the dataset size of the experiments in Section 6. The data is randomly picked from the CAIL2018 (Xiao et al., 2018) dataset. The number in the table represents the amount of data for each experiment, and the size of the treated ($T = 1$) group and the control ($T = 0$) group is balanced by undersampling.

Level	Description	#Case
1	**Not similar at all.** It is hard to find any similarity between the two cases.	0
2	**Not Similar.** There are key differences between the two cases. The sentences should not be the same (or should be discussed independently).	0
3	**Similar.** There are only differences in the details between the two cases, and these differences will have little impact on sentencing.	19
4	**Almost identical.** It is hard to find any differences between the two cases, even in the details.	81
	Total	100

Table 7: Legal professionals' evaluation to the 100 matched case pairs.

C Experimental Settings

In this section, we introduce the experimental settings that are omitted in the main text.

C.1 Baselines and Models

CNN (Kim, 2014): This work proposes Convolutional Neural Networks with multiple filter widths specifically for text classification. In this paper, we follow the architecture of Kim (2014) for our implementation.

BERT (Devlin et al., 2019): BERT is the model formed by multiple bidirectional Transformer layers. The parameters of BERT has been fully pretrained on large-scale text corpora. In this paper, we employ the BERT-base pre-trained on Chinese corpora for experiments.

C.2 Training Settings

We use Adam (Kingma and Ba, 2015) to train CNN and use BertAdam (Devlin et al., 2019) to train BERT. We employ character-level embedding to train BERT and use external Chinese word vectors[3] (Li et al., 2018) to train CNN.

We train models with NVIDIA GTX 2080 Ti.

For models with calibration module, we divide the train set by $1:1$ for training the encoder and the calibration model, respectively.

We repeat each experiment 10 times and ensure the results pass the normality test (Shapiro-Wilk test). Suppose $\left(\mu, \sigma^2\right)$ are the mean and variance of the results, we report the 95% confidence interval of the results as $\mu \pm 1.96\frac{\sigma}{\sqrt{10}}$.

[3]https://github.com/Embedding/Chinese-Word-Vectors

Charge	Age	Gender	Region Split by S or N	GDP	CR
Overall	10,000	10,000	10,000	10,000	10,000
Drug Trafficking	9,532	10,000	10,000	10,000	10,000
Theft	10,000	10,000	10,000	10,000	10,000
Intentional Injury	10,000	10,000	10,000	10,000	10,000
Traffic Offence	10,000	10,000	10,000	10,000	10,000
Providing Venues for Drug Users	6,220	10,000	10,000	10,000	10,000

Table 8: Dataset size of the experiments in Section 6

C.3 Hyper-Parameters of Neural models

The hyper-parameters of neural models are shown in Table 9.

	CNN	BERT	Calibration
Learning Rate	3×10^{-4}	10^{-6}	10^{-3}
Weight Decay	0	0	0
Max Sequence Length	512	512	N/A
Dropout	0.8	0.8	N/A
Hidden Layer Size	300	768	N/A
Epoch	12	8	8

Table 9: The hyper-parameters of neural models.

C.4 Result Selection

To further prevent models from overfitting, we take the result of the epoch with the lowest loss on valid set as the experimental result for each model.

D Calibration

D.1 Methodology of Temperature Scaling

The main idea of temperature scaling is to train a single parameter $\tau > 0$ to scale hidden layer scores of the neural model. In this way, the scale of estimated propensity scores can be calibrated.

Specifically, let $h \in \mathbb{R}^2$ represent the hidden layer output:

$$h(x) = \text{Linear}\big(\text{Encoder}(x)\big). \tag{9}$$

Then, the unadjusted propensity score can be obtained by a softmax layer:

$$e^{\text{unadj}}(x) = \frac{\exp\left(h(x)_1\right)}{\exp\left(h(x)_0\right) + \exp\left(h(x)_1\right)}. \tag{10}$$

In contrast, temperature scaling adjusts the propensity score as:

$$e^{\text{adj}}(x) = \frac{\exp\left(\frac{h(x)_1}{\tau}\right)}{\exp\left(\frac{h(x)_0}{\tau}\right) + \exp\left(\frac{h(x)_1}{\tau}\right)}. \tag{11}$$

We use the cross-entropy loss to optimize the single parameter τ. It can be proved that, in expectation, the loss is minimized if and only if the predicted propensity score infinitely approximates the true conditional probability (Friedman et al., 2001).

D.2 Experiments on LegalTrEE

Besides temperature scaling, two calibration approaches are selected for comparative experiments on LegalTrEE: (1) Histogram binning (Zadrozny and Elkan, 2001), a non-parametric calibration method. The main idea of histogram binning is to divide all uncalibrated predictions into different bins, and assign each bin a calibrated score to minimize the bin-wise squared loss. (2) Attended temperature scaling (Mozafari et al., 2018), a variant of temperature scaling. Attended Temperature scaling uses ATS loss to improve the performance with fewer training samples.

The same as in Section 5, we use different calibration approaches to measure the average treatment effect (ATE) and check the difference between the measured value and the ground truth. In addition, we evaluate the expected calibration error (ECE) (Naeini et al., 2015) of the propensity score. The definition of ECE is:

$$\text{ECE} = \mathop{\mathbb{E}}_{x}\left[\Big|\Pr\left(T = 1 | X = x\right) - e\left(x\right)\Big|\right]. \tag{12}$$

ECE is negatively correlated with the predicting precision of propensity scores. In other words, a low ECE can reflect high predicting precision of the propensity score, and vice versa. In this paper, we follow previous works (Guo et al., 2017) and approximate ECE by the binning approach. Specifically, M equally-spaced bins $B_1, \ldots, B_M$ are used. And the ECE is calculated as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n}\Big|\text{pos}\left(B_m\right) - \text{conf}\left(B_m\right)\Big|. \tag{13}$$

Here n denotes the number of samples, $\text{pos}\left(B_m\right)$ denotes the rate of positive (treated) samples in B_m, and $\text{conf}\left(B_m\right)$ denotes the average propensity score of samples in B_m. In this paper, we take

$M = 10$ for approximating ECE following (Guo et al., 2017).

Method		avg ATE	avg$\lvert\delta\rvert$ (δ = ATE − GT)	std (δ)	avg ECE
Ground Truth (GT)		0.956	0	N/A	N/A
CNN	w/o Calib.	1.448 ± 0.095	0.492	**0.283**	0.069
	HB	3.167 ± 0.624	2.211	2.462	0.098
	TS	$\mathbf{1.170 \pm 0.163}$	**0.214**	0.508	**0.044**
	ATS	1.195 ± 0.171	0.239	0.472	0.045
BERT	w/o Calib.	1.499 ± 0.347	0.543	0.973	0.069
	HB	1.822 ± 0.253	0.866	0.480	0.051
	TS	1.398 ± 0.365	0.442	1.029	0.059
	ATS	1.398 ± 0.365	0.442	1.029	0.059

Table 10: Comparison between three calibration methods, histogram binning (HB), temperature scaling (TS), and attended temperature scaling (ATS). Lower is better for both avg$\lvert\delta\rvert$ and avg ECE. Temperature scaling (TS) is the calibration approach we use in the main text.

The results are shown in Table 10. In addition to the fact that CNN with temperature scaling outperforms other models, we also have several noteworthy observations: (1) The prediction error of ATE is positively correlated with ECE in general. This observation uncovers that well-performed calibration can lead to less prediction error of ATE as expected, and also reflects sideways that LegalTrEE and our evaluation are reliable to a great extent. (2) Attended temperature scaling can achieve almost identical performance as temperature scaling. We argue that the main reason is these two methods are essentially the same except that they use different loss functions. (3) The performance of histogram binning is bad, even worse than models without calibration. We argue that the main reason is this non-parametric approach is too trivial for our complicated task.

D.3 Experiments on CAIL2018

To further enhance the persuasiveness and reliability of the large-scale analysis (Section 6), we show the ECE of the model predicted propensity scores in Table 11.

As Table 11 showed, the ECEs of the models in our large-scale analysis are low. Specifically, the ECEs of these models are smaller than the best model in the baseline experiment (0.044, shown in Table 10). Since propensity score is the only variable to be predicted in the IPW scheme, the prediction accuracy of propensity score (ECE) can greatly reflect the accuracy of IPW. Therefore, with a very high probability, the accuracy of our large-scale analysis is better than that of the baseline experiment.

E Charge Description

In this section, we describe the criminal charges mentioned in this paper. All the descriptions refer to Chinese Criminal Law.

Drug Trafficking: The act of knowingly smuggling, trafficking, transporting, or manufacturing drugs.

Theft: The act that, for the purpose of illegal possession, steals a relatively large amount of public or private property or commits theft repeatedly.

Intentional Injury: The act that intentionally inflicts injury upon another person.

Traffic Offence: The act violates regulations governing traffic and transportation and thereby causes a serious accident, resulting in serious injuries or deaths or heavy losses of public or private property.

Providing Venues for Drug Users: The act that provides shelter for another person to ingest or inject narcotic drugs.

F Details of Example Case in Introduction

Here we show the detailed description of the two real-world example cases in the introduction. Due to anonymity, some details (such as names, locations, dates, etc.) have been manually omitted.

Case A. In a karaoke room in Region A, when the defendant Alice was helping the victim Carol find a lost ring, she took it away in her purse when Carol was not paying attention. A few months later, the defendant Alice was found by Carol wearing her lost ring while she was playing with her cell phone in Carol's store. The stolen ring was tested by a mineral testing center and found to be a diamond ring. The price certification center determined that the stolen diamond ring's retail market price was RMB $35,000$ on the day of the crime. After the crime, the relatives of the defendant Alice compensated Carol on her behalf, and Carol expressed her understanding to the defendant Alice. The defendant Alice confessed to the crime in a good manner.

Case B. In Region B, defendant Bob stole his ex-girlfriend Daisy's car parked on the north side of the neighborhood gate and sold the vehicle for RMB $20,000$, squandering the proceeds. The stolen vehicle was appraised to be worth RMB

Charge	Expected Calibration Error (ECE)				
	Age	Gender	Region Split by		
			S or N	GDP	CR
Overall	0.014	0.033	0.017	0.012	0.020
Drug Trafficking	0.031	0.019	0.016	0.014	0.012
Theft	0.026	0.032	0.019	0.030	0.028
Intentional Injury	0.016	0.024	0.028	0.022	0.017
Traffic Offence	0.020	0.017	0.010	0.018	0.011
Providing Venues for Drug Users	0.025	0.024	0.018	0.023	0.018

Table 11: The model ECE of experiments in Section 6. Lower is better.

$35,000$. The stolen vehicle was extracted and returned to the owner after the crime. The defendant Bob confessed to the crime in a good manner.

G Explanation of Consistency Prerequisite

In this paper, we only focus on Chinese criminal cases for experiments and analyses. In China, the criminal judgments should only be based on the "Criminal Law of the People's Republic of China", which is nationally consistent. In other words, if two criminals behaved the same, they should be sentenced the same, even if they are from different regions. Therefore, the principle of "each individual should be equal" holds, and the model of treatment effect estimation is applicable.