FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing

Anonymous ACL submission

Abstract

We present a benchmark suite of four datasets for evaluating the fairness of pre-trained legal language models and the techniques used to fine-tune them for downstream tasks. Our benchmarks cover four jurisdictions (European Council, USA, Swiss, and Chinese), five languages (English, German, French, Italian and Chinese) and fairness across five attributes (gender, age, nationality/region, language, and legal area). In our experiments, we evaluate pre-trained language models using several group-robust fine-tuning techniques and show that none of these combinations guarantee fairness, nor consistently mitigate group disparities. Furthermore, we analyze what causes performance differences across groups, and how group-robust fine-tuning techniques fail to mitigate group disparities under both representation inequality and temporal distribution swift.

1 Introduction

The sector of law produces massive volumes of textual data (Katz et al., 2020), and as a result, legal research for settling personal injury claims, for example, can take several years, potentially discouraging clients. Legal systems around the world, e.g., in India,[1] Brazil,[2] or the US[3], experience year-long backlogs of pending cases. Natural Language Processing (NLP) for law (Chalkidis and Kampas, 2019; Aletras et al., 2019; Zhong et al., 2020) receives increasing attention. Assistive technologies can speed up legal research or discovery significantly assisting lawyers, judges and clerks. They can also help legal scholars to study case law (Katz, 2012), improve access of law to laypersons, help sociologists and research ethicists to expose biases in the justice system (Angwin et al., 2016; Dressel and Farid, 2018), and even scrutinize decision-making itself (Bell et al., 2021).

[1] https://tinyurl.com/mjy2uf9a
[2] https://tinyurl.com/2uttucmn
[3] https://tinyurl.com/4ybhhff8

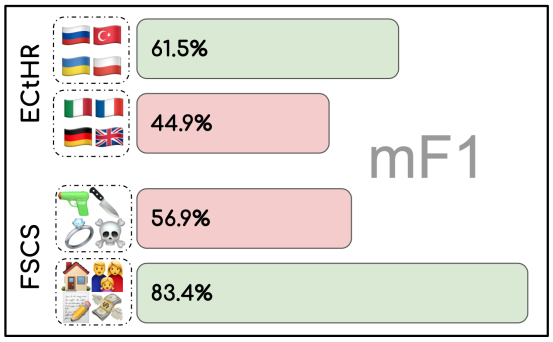

Figure 1: *Group disparity* for *defendant state* (C.E. Europe vs. The Rest) in ECtHR and *legal area* (Penal law vs. Civil law) in FSCS.

In the context of law, *non-discrimination* (i.e. *equality*) is of paramount importance, e.g., EU non-discrimination law (Council of European Union, 2000, 2006) prohibits both direct and indirect discrimination. Discrimination occurs when one person is treated *less favourably than others would be treated in comparable situations* on grounds of sex, racial or ethnic origin, disability, sexual orientation, religion or belief and age.[4] Given the gravity that legal outcomes have for individuals, assistive technologies cannot be adopted to speed up legal research at the expense of fairness (Wachter et al., 2021), potentially also decreasing the trust in our legal systems (Barfield, 2020).

In recent years, the NLP and machine learning literature has introduced fairness objectives, typically derived from the Rawlsian notion of *equal opportunities* (Rawls, 1971), to evaluate the extent to which models discriminate across protected attributes. Some of these rely on notions of resource allocation, i.e., reflecting the idea that groups are treated fairly if they are equally represented in the training data used to induce our models, or if the same number of training iterations is performed per group. This is sometimes referred to as the *resource allocation* perspective on fair-

[4] An in-depth analysis of the notion of discrimination and fairness in law is presented in Appendix A.

ness (Lundgard, 2020). Contrary, there is also a *capability*-centered approach to fairness (Anderson, 1999; Robeyns, 2009), in which the goal is reserve enough resources per group to achieve similar performance levels, which is ultimately what is important for how individuals are treated in legal processes. We adopt a capability-centered approach to fairness and define fairness in terms of performance parity (Hashimoto et al., 2018) or equal risk (Donini et al., 2018).[5]

Performance disparity (Hashimoto et al., 2018) refers to the phenomenon of high overall performance, but low performance on minority groups, as a result of minimizing risk across samples (not groups), Since some groups benefit more than others from models and technologies that exhibit performance disparity, this likely widens gaps between those groups. Performance disparity works against the ideal of fair and equal opportunities for all groups in our societies. We therefore define a *fair* classifier as one that has similar performance (equal risk) across all groups (Donini et al., 2018).

In sum, we adopt the view that (approximate) equality under the law in a modern world requires that our NLP technologies exhibit (approximately) equal risk across sensitive attributes. For everyone to be treated equally under the law, regardless of race, gender, nationality, or other characteristics, NLP technologies need to be (approximately) insensitive to these attributes. In a supervised learning setting, models are trained on historical data that not always represent all groups in our societies equally. Moreover, historical legal data tends to reflect social biases in our societies and legal institutions. For example, criminal justice is already often strongly influenced by racial bias, with people of colour being more likely to be arrested and receive higher punishments than others, both in the US[6] and in the UK.[7] When models are deployed in production, they may reinforce these biases. We consider three types of attributes in this work:

- *Demographics*: The first category includes demographic information of the involved parties, e.g., the gender, sexual orientation, nationality, age, or race of the plaintiff/defendant in a case. In this case, we aim to mitigate biases against specific groups, e.g., a model performs worse for female defendants or is biased against black defendants.
- *Regional*: The second category includes regional information of the courts in charge of a case. In this case, we aim to mitigate disparity in-between different regions in a given jurisdiction, e.g., a model performs better in specific cases originated or ruled in courts of specific regions.
- *Legal Topic*: The third category includes legal topic information on the subject matter of the controversy. In this case, we aim to mitigate disparity in-between different topics (areas) of law, e.g., a model performs better in a specific field of law, for example civil cases.

Contributions We introduce FairLex, a multilingual fairness benchmark of four legal datasets covering four jurisdictions (Council of Europe, United States of America, Swiss Confederation and People's Republic of China), five languages (English, German, French, Italian and Chinese) and various sensitive attributes (gender, age, region, etc.). We release four pre-trained transformer-based language models, each tailored for a specific FairLex dataset (task) within our benchmark, which can be used as baseline models (text encoders). We conduct experiments with several group-robust algorithms and provide a quantitative and qualitative analysis of our results, highlighting open challenges in the development of robustness methods in legal NLP.

2 Related Work

Fair machine learning The literature on inducing approximately fair models from biased data is rapidly growing. See Mehrabi et al. (2021) for a recent survey. We rely on this literature in how we define fairness, and for the algorithms that we compare in our experiments below. As already discussed, we adopt a capability-centered approach to fairness and define fairness in terms of performance parity (Hashimoto et al., 2018) or equal risk (Donini et al., 2018). The fairness-promoting learning algorithms we evaluate are discussed in detail in §4. Some of these – Group Distributionally Robust Optimization (Sagawa et al., 2020) and Invariant Risk Minimization (Arjovsky et al., 2020) – have previously been evaluated for fairness in the context of hate speech (Koh et al., 2021).

Fairness in law Studying fair machine learning in the context of legal (computational) applications

[5] The dominant alternative to equal risk is to define fairness in terms of equal odds. Equal odds fairness does not guarantee Rawlsian fairness, and often conflicts with the rule of law.

[6] https://tinyurl.com/4cse552t

[7] https://tinyurl.com/hkff3zcb

Dataset	Original Publication	Classification Task	No of Classes	Attributes	
				Attribute Type	#N
ECtHR	(Chalkidis et al., 2021)	Legal Judgment Prediction: *ECHR Violation Prediction*	10+1	Defendant State	2
				Applicant Gender	2
				Applicant Age	3
SCOTUS	(Spaeth et al., 2020)	Legal Topic Classification: *Issue Area Classification*	14	Respondent Type	4
				Decision Direction	2
FSCS	(Niklaus et al., 2021)	Legal Judgment Prediction: *Case Approval Prediction*	2	Language	3
				Region of Origin	6
				Legal Area	6
SPC	(Wang et al., 2021b)	Legal Judgment Prediction: *Crime Severity Prediction*	6	Defendant Gender	2
				Region of Origin	7

Table 1: Main characteristics of FairLex datasets (ECtHR, SCOTUS, FSCS, SPC). We report the examined tasks, the number of classes, the examined attributes and the number (#N) of groups per attribute.

has a limited history. In a classic study, Angwin et al. (2016) analyzed the performance of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, which was used for parole risk assessment (recidivism prediction) in the US. The system relied on 137 features from questionnaires and criminal records. Angwin et al. (2016) found that blacks were almost twice as likely as whites to be mislabeled as high risk (of re-offending), revealing a severe racial bias in the system. The system was later compared to crowdworkers in Dressel and Farid (2018).

These studies relied on tabular data and did not involve text processing. More recently, Wang et al. (2021b) studied legal judgment consistency using a dataset of Chinese criminal cases. They evaluated the consistency of LSTM-based models across region and gender and reported severe fairness gaps across gender. They also found that the fairness gap was particular severe for more serious crimes.

Previous work has focused on the analysis of specific cases, languages or algorithms, but FairLex aims at easing the development and testing of bias-mitigation models or algorithms within the legal domain. FairLex allows researchers to explore fairness across four datasets covering four jurisdictions (Council of Europe, United States of America, Swiss Confederation and People's Republic of China), five languages (English, German, French, Italian and Chinese) and various sensitive attributes (gender, age, region, etc.). Furthermore, we provide competitive baselines including state-of-the-art transformer-based models, adapted to the examined datasets, and an in-dept examination of performance of four group robust algorithms described in detail in Section 4.

3 Benchmark Datasets

ECtHR The European Court of Human Rights (ECtHR) hears allegations that a state has breached human rights provisions of the European Convention of Human Rights (ECHR). We use the dataset of Chalkidis et al. (2021), which contains 11K cases from ECtHR's public database. Each case is mapped to *articles* of the ECHR that were violated (if any). This is a multi-label text classification task. Given the facts of a case, the goal is to predict the ECHR articles that were violated, if any, as decided (ruled) by the court. The cases are chronologically split into training (9k, 2001–16), development (1k, 2016–17), and test (1k, 2017–19) sets.

To facilitate the study of fairness of text classifiers, we record for each case the following attributes: (a) The *defendant states*, which are the European states that allegedly violated the ECHR. The defendant states for each case is a subset of the 47 Member States of the Council of Europe;[8] To have statistical support, we group defendant states in two: Central-Eastern European states, on one hand, and all other states, as classified by the EuroVoc thesaurus.[9] (b) The *applicant's age* at the time of the decision. We extract the birth year of the applicant from the case facts, if possible, and classify its case in an age group (≤35, ≤64, or older) ; and (c) the *applicant's gender*, extracted from the facts, if possible based on pronouns, classified in two categories (male, female).

SCOTUS The US Supreme Court (SCOTUS) is the highest federal court in the United States of America and generally hears only the most controversial or otherwise complex cases which have not been sufficiently well solved by lower courts. We combine information from SCOTUS opinions with

[8] https://www.coe.int/

[9] https://op.europa.eu/en/web/eu-vocabularies

the Supreme Court DataBase (SCDB)[10] (Spaeth et al., 2020). SCDB provides metadata (e.g., date of publication, decisions, issues, decision directions and many more) for all cases. We consider the available 14 thematic issue areas (e.g, Criminal Procedure, Civil Rights, Economic Activity, etc.). This is a single-label multi-class document classification task. Given the court opinion, the goal is to predict the issue area whose focus is on the subject matter of the controversy (dispute). SCOTUS contains a total of 9,262 cases that we split chronologically into 80% for training (7.4k, 1946–1982), 10% for development (914, 1982–1991) and 10% for testing (931, 1991–2016).

From SCDB, we also use the following attributes to study fairness: (a) the *type of respondent*, which is a manual categorization of respondents (defendants) in five categories (person, public entity, organization, facility and other); and (c) the *direction of the decision*, i.e., whether the decision is liberal, or conservative, provided by SCDB.

FSCS The Federal Supreme Court of Switzerland (FSCS) is the last level of appeal in Switzerland and similarly to SCOTUS, the court generally hears only the most controversial or otherwise complex cases which have not been sufficiently well solved by lower courts. The court often focus only on small parts of previous decision, where they discuss possible wrong reasoning by the lower court. The Swiss-Judgment-Predict dataset (Niklaus et al., 2021) contains more than 85K decisions from the FSCS written in one of three languages (50K German, 31K French, 4K Italian) from the years 2000 to 2020. The dataset provides labels for a simplified binary (*approval*, *dismissal*) classification task. Given the facts of the case, the goal is to predict if the plaintiff's request is valid or partially valid. The cases are also chronologically split into training (59.7k, 2000-2014), development (8.2k, 2015-2016), and test (17.4k, 2017-2020) sets.

The dataset provides three additional attributes: (a) the *language* of the FSCS written decision, in either German, French, or Italian; (b) the *legal area* of the case (public, penal, social, civil, or insurance law) derived from the chambers where the decisions were heard; and (c) the *region* that denotes in which federal region was the case originated.

SPC The Supreme People's Court of China (SPC) is the last level of appeal in China and considers cases that originated from the high people's courts concerning matters of national importance. The Chinese AI and Law challenge (CAIL) dataset (Xiao et al., 2018) is a Chinese legal NLP dataset for judgment prediction and contains more 1m criminal cases. The dataset provides labels for *relevant article of criminal code* prediction, *charge* (type of crime) prediction, imprisonment *term* (period) prediction, and monetary *penalty* prediction.

Recently, Wang et al. (2021b) re-annotated a subset of approx. 100k cases with demographic attributes. Specifically the new dataset has been annotated with: (a) the *applicant's gender*, classified in two categories (male, female); and (b) the *region* of the court that denotes in which out of the 7 provincial-level administrative regions was the case judged. We re-split the dataset chronologically into training (80k, 2013-2017), development (12k, 2017-2018), and test (12k, 2018) sets. In our study, we examine a *crime severity* prediction task, a single-label multi-class classification task, where given the facts of a case, the goal is to predict how severe was the committed crime with respect to the imprisonment term. We approximate crime severity by the length of imprisonment term, split in 6 clusters (0, ≤12, ≤36, ≤60, ≤120, >120 months).

4 Fine-tuning Algorithms

Across experiments, our main goal is to find a hypothesis for which the risk $R(h)$ is minimal:

$$h^* = \arg\min_{h \in \mathcal{H}} R(h) \quad (1)$$

$$R(h) = \mathbb{E}(L(h(x), y)) \quad (2)$$

where y are the targets (*ground truth*) and $h(x) = \hat{y}$ is the system hypothesis (model's predictions).

Similar to previous studies, $R(h)$ is an expectation of the selected loss function ($\mathcal{L}$). In this work, we study multi-label text classification (Section 3), thus we aim to minimize the binary cross-entropy loss across L classes:

$$\mathcal{L} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \quad (3)$$

ERM (Vapnik, 1992), which stands for Empirical Risk Minimization, is the most standard and widely used optimization technique to train neural methods. The loss is calculated as follows:

$$\mathcal{L}_{ERM} = \sum_{i=1}^{N} \frac{\mathcal{L}_i}{N} \quad (4)$$

where N is the number of instances (training examples) in a batch, and $\mathcal{L}_i$ is the loss per instance.

[10] http://scdb.wustl.edu

Besides ERM, we also consider a representative selection of group-robust fine-tuning algorithms which aims at mitigating performance disparities with respect to a given attribute (A), e.g., the gender of the applicant or the region of the court. Each attribute is split into G groups, i.e., male/female for gender. All algorithms rely on a balanced group sampler, i.e., an equal number of instances (samples) per group (N_G) are included in each batch. Most of the algorithms are built upon group-wise losses ($\mathcal{L}_g$), computed as follows:

$$\mathcal{L}(g_i) = \frac{1}{N_{g_i}} \sum_{j=1}^{N_{g_i}} \mathcal{L}(x_j) \quad (5)$$

Group DRO (Sagawa et al., 2020), stands for Group Distributionally Robust Optimization (DRO). Group DRO is an extension of the Group Uniform algorithm, where the group-wise losses are weighted inversely proportional to the group training performance. The total loss is:

$$\mathcal{L}_{DRO} = \sum_{i=1}^{G} w_{g_i} * \mathcal{L}(g_i), \text{ where} \quad (6)$$

$$w_{g_i} = \frac{1}{W}(\hat{w}_{g_i} * e^{L(g_i)}) \quad \text{and} \quad W = \sum_{i=1}^{G} w_{g_i} \quad (7)$$

where G is the number of groups (labels), $\mathcal{L}_g$ are the averaged group-wise (label-wise) losses, w_g are the group (label) weights, $\hat{w}_g$ are the group (label) weights as computed in the previous update step. Initially the weight mass in equally distributed across groups.

REx (Krueger et al., 2020), which stands for Risk Extrapolation, is yet another proposed group-robust optimization algorithm. Krueger et al. (2020) hypothesize that variation across training groups is representative of the variation later encountered at test time, so they also consider the variance across the group-wise losses. In V-REx the total loss is calculated as follows:

$$\mathcal{L}_{REX} = \mathcal{L}_{ERM} + \lambda * \text{Var}([\mathcal{L}_{g_1}, \ldots, \mathcal{L}_{g_G}]) \quad (8)$$

where Var is the variance among the group-wise losses and λ, a weighting hyper-parameter scalar.

IRM (Arjovsky et al., 2020), which stands for Invariant Risk Minimization, mainly aims to penalize variance across multiple training dummy estimators across groups, i.e., performance cannot vary in samples that correspond to the same group. The total loss is computed as follows:

$$\mathcal{L}_{IRM} = \frac{1}{G}\left(\sum_{i=1}^{G} \mathcal{L}(g_i) + \lambda * P(g_i)\right) \quad (9)$$

Please refer to Arjovsky et al. (2020) for the definition of the group penalty terms (P_g).

Adversarial Removal (Elazar and Goldberg, 2018) algorithm mitigates group disparities by means of an additional adversarial classifier (Goodfellow et al., 2014). The adversarial classifier share the encoder with the main network and is trained to predict the protected attribute (A) of an instance. The total loss factors in the adversarial one, thus penalizing the model when it is able to discriminate groups. Formally, the total loss is calculated as:

$$\mathcal{L}_{AR} = \mathcal{L}_{ERM} - \lambda * \mathcal{L}_{ADV} \quad (10)$$

$$\mathcal{L}_{ADV} = \mathcal{L}(\hat{g}_i, g_i) \quad (11)$$

where $\hat{g}_i$ is the adversarial classifier's prediction for the examined attribute A (in which group (g_i) of A, does the example belong to) given the input (x).

5 Experimental SetUp

Models Since we are interested in classifying long documents (up to 6000 tokens per document, see Figure 2 in Appendix C.1), we developed a hierarchical BERT-based model similar to that of Chalkidis et al. (2021), so as to avoid using only the first 512 tokens of a text. Our hierarchical model, first, encodes the text through a pre-trained Transformer-based architecture, thus representing each paragraph independently with the [CLS] token. Then, the paragraph representations are fed into a two-layers transformer encoder with the exact same specifications of the first one (e.g., hidden units, number of attention heads), so as to contextualize them, i.e., it makes paragraphs representations aware of the surrounding paragraphs. Finally, the model max-pools the context-aware paragraph representations computing the document-level representation and feed it to a classification layer.

For the purpose of this work, we release four domain-specific BERT models with continued pre-training on the corpora of the examined datasets (ECtHR, SCOTUS, FSCS, SPC).[11] We train mini-sized BERT models with 6 Transformer blocks, 384 hidden units, and 12 attention heads. We warm-start all models from the public MiniLMv2 models checkpoints (Wang et al., 2021a) using the distilled

[11] All models will be released on Hugging Face upon acceptance.

Algorithm	ECtHR (ECHR Violation Prediction)									SCOTUS (Issue Area Classification)					
	Defendant State			*Applicant Gender*			*Applicant Age*			*Respondent Type*			*Direction*		
	↑ $\overline{\text{mF1}}$	↓ GD	↑ mF1_W	↑ $\overline{\text{mF1}}$	↓ GD	↑ mF1_W	↑ $\overline{\text{mF1}}$	↓ GD	↑ mF1_W	↑ $\overline{\text{mF1}}$	↓ GD	↑ mF1_W	↑ $\overline{\text{mF1}}$	↓ GD	↑ mF1_W
Bag-of-Words Linear Classifier															
ERM	46.8	3.0	43.8	44.1	4.9	40.6	46.9	6.3	40.9	73.8	6.6	61.8	77.5	2.6	74.9
Transformer-based Classifier															
ERM	53.2	8.3	44.9	57.5	3.1	54.4	54.1	5.9	46.2	75.1	4.0	70.8	78.1	1.6	76.6
ERM+GS	54.4	5.5	48.9	**57.8**	3.3	54.5	**56.0**	5.6	48.7	**75.2**	3.9	70.9	77.1	1.3	76.0
ADV-R	53.8	5.8	47.9	54.6	3.2	51.5	48.9	6.1	40.6	56.9	4.7	53.1	41.0	**0.8**	40.3
G-DRO	**55.0**	**5.2**	**49.8**	56.3	**1.9**	**55.0**	52.6	6.2	44.3	74.5	3.3	**71.6**	77.1	1.7	75.4
IRM	53.8	5.7	48.1	53.8	2.3	52.5	54.8	**4.4**	49.5	73.4	4.8	68.2	78.1	2.7	75.4
REx	54.6	6.3	48.3	54.6	2.0	53.2	55.0	4.5	**49.8**	73.8	**3.8**	68.2	**78.2**	1.1	**77.1**

Algorithm	FSCS (Case Approval Prediction)									SPC (Criminal Offense Prediction)					
	Language			*Legal Area*			*Region*			*Defendant Gender*			*Region*		
	↑ $\overline{\text{mF1}}$	↓ GD	↑ mF1_W	↑ $\overline{\text{mF1}}$	↓ GD	↑ mF1_W	↑ $\overline{\text{mF1}}$	↓ GD	↑ mF1_W	↑ $\overline{\text{mF1}}$	↓ GD	↑ mF1_W	↑ $\overline{\text{mF1}}$	↓ GD	↑ mF1_W
Bag-of-Words Linear Classifier															
ERM	55.5	6.2	46.8	54.4	9.7	40.9	56.8	5.0	46.6	33.5	0.7	32.8	31.7	5.0	25.5
Transformer-based Classifier															
ERM	67.8	2.1	65.0	**69.4**	9.6	**56.9**	**69.7**	**2.9**	**63.9**	**60.2**	**0.6**	**60.1**	**59.3**	3.5	**56.4**
ERM+GS	66.4	3.5	61.7	67.1	9.3	55.5	67.9	3.0	62.3	59.4	0.7	59.1	58.2	3.1	55.9
ADV-R	62.6	5.1	59.0	65.6	12.4	50.0	67.4	3.2	61.5	53.3	1.3	52.1	53.5	**2.5**	50.8
G-DRO	**70.5**	**0.6**	**69.9**	57.5	**5.6**	52.6	67.7	4.2	60.2	59.2	1.3	57.9	58.9	3.7	55.7
IRM	68.3	1.9	66.7	67.8	9.5	55.8	68.7	3.0	63.2	56.4	1.5	55.7	58.0	3.1	54.9
REx	67.2	3.5	62.4	66.6	8.9	56.0	68.4	3.1	62.4	58.5	0.7	58.3	58.6	3.3	54.4

Table 2: Test results for all examined group-robust algorithms per dataset attribute. We report the average performance across groups ($\overline{\text{mF1}}$), the *group disparity* among groups (GD), and the worst-group performance (mF1_W). ↑ denotes that higher scores are better, while ↓ denotes that lower scores are better.

version of RoBERTa (Liu et al., 2019) for the English datasets (ECtHR, SCOTUS) and the one distilled from XLM-R (Conneau et al., 2020) for the rest (trilingual FSCS, and Chinese SPC). Given the limited size of these models, we can effectively use up to 4096 tokens in ECtHR and SCOTUS and up to 2048 tokens in FSCS and SPC for up to 16 samples per batch in a 24GB NVIDIA GPU CARD.[12]

For completeness, we also consider linear Bag-of Words (BoW) classifiers using TF-IDF scores of the most frequent n-grams (where $n = 1, 2, 3$) in the training corpus of each dataset.

Data Repository and Code We release a unified version of the benchmark on Hugging Face Datasets (Lhoest et al., 2021).[13] In our experiments, we use and extend the WILDs (Koh et al., 2021) library. For reproducibility and further exploration with new group-robust methods, we release our code on Github.[12]

Evaluation Details Across experiments we compute the macro-F1 score per group (mF1_i), excluding the group of *unidentified* instances, if any.[14] We report macro-F1 to avoid bias toward majority classes because of class imbalance and skewed label distributions across train, development, and test subsets. We report the average macro-F1 across groups ($\overline{\text{mF1}}$) and the *group disparity* (GD) among groups measured as the group-wise std dev.:

$$GD = \sqrt{\frac{1}{G}\sum_{i=1}^{G}(\text{mF1}_i - \overline{\text{mF1}})^2} \qquad (12)$$

We also report the *worst-group performance* ($\text{mF1}_W = \min([\text{mF1}_1, \text{mF1}_2, \ldots \text{mF1}_G])$).

6 Baseline Results

Main Results In Table 2, we report the results of all our baselines on the four datasets introduced in this paper. We first observe that the results of linear classifiers trained with the ERM algorithm (top row per dataset) are consistently worse (lower average and worst-case performance, higher group disparity) compared to transformed-based models in the same setting. In other words linear classifier have lower overall performance, while being less *fair* with respect to the applied definition of fairness (i.e. equal performance across groups).

As one can see, transformer-based models trained with the ERM algorithm, i.e., without taking into account information about groups and their distribution, perform either better on in the same ballpark than models trained with methods specialized to mitigate biases (Section 4), with an average loss of 0.17 only in terms of $mF1$ and of 0.78 in terms of $mF1_W$. While, these algorithms

[12] This is particularly important for group-robust algorithms that consider group-wise losses.

[13] Both links will be revealed upon acceptance.

[14] The group of *unidentified* instances includes the instances, where the value of the examined attribute is unidentifiable (unknown). See details in Appendix C.2.

improve worst case performance in the literature, when applied in a controlled experimental environment, they fail in a real-world setting, where both groups across attributes and labels are imbalanced, while also both group and label distribution change over time. Furthermore, we cannot identify one algorithm that performs better across datasets and group with respect to the others, indeed results are quite mixed without any recognizable pattern.

Group Disparity Analysis We identify three general (attribute agnostic) factors that could potentially lead to performance disparity across groups:

- *Representation Inequality*: Not all groups are equally represented in the training set. To examine this aspect, we report the number of training cases per group.
- *Temporal Concept Drift*: The label distribution for a given group changes over time, i.e., in-between training and test subsets. To examine this aspect, we report per group, the KL divergence in-between the training and test label distribution.
- *Worst Class Influence*: The performance is not equal across labels (classes), which may disproportionally affect the macro-averaged performance across groups. To examine this aspect, we report the *Worst Class Influence (WCI)* score per group, which is computed as follows:

$$\text{WCI}(i) = \frac{\text{\#test-cases (worst-class)}}{\text{\#test-cases}} \qquad (13)$$

In Table 3, we present the results across all attributes. We observe that only in 4 out of 10 cases (attributes), the less represented groups are those with the worst performance compared to the rest. It is generally not the case that high KL divergence (drift) correlates with low performance. In other words, group disparities does not seem to be driven by temporal concept drift. Finally, the influence of the worst class is relatively uniform across groups in most cases, but in the cases where groups differ in this regard, worst class influence correlates with error in 2 out of 3 cases.[15]

In ECtHR, considering performance across defendant state, we see that all the three factors correlate internally, i.e., the worst performing group is

[15]For ECtHR performance across defendant states and SCOTUS across directions, but not for ECtHR performance across applicant age.

ECtHR (ECHR Violation Prediction)				
Group	mF1	#train-cases (%) (↑)	LD_{KL} (↓)	WCI (↓)
Defendant State				
E.C. European	**70.2**	**7,224 (80%)**	**0.17**	**0.07**
The Rest	48.7	1,776 (20%)	0.28	0.57
Applicant Gender				
Male	54.4	**4,187 (77%)**	**0.17**	**0.18**
Female	**60.6**	1,507 (23%)	0.26	0.19
Applicant Age				
≤ 65 years	**59.7**	**4279 (68%)**	**0.18**	0.15
> 65 years	56.5	1130 (18%)	0.32	0.26
≤ 35 years	46.2	868 (14%)	0.19	**0.12**

SCOTUS (Issue Area Classification)				
Group	mF1	#train-cases (%) (↑)	LD_{KL} (↓)	WCI (↓)
Respondent Type				
Public Entity	77.4	**2796 (51%)**	**0.07**	0.04
Person	74.9	1847 (34%)	0.05	**0.03**
Organization	**81.1**	741 (13%)	0.11	**0.03**
Facility	80.7	140 (3%)	0.26	0.06
Direction				
Liberal	76.2	**3335 (52%)**	**0.04**	**0.08**
Conservative	**80.8**	3146 (48%)	0.05	0.17

FSCS (Case Approval Prediction)				
Group	mF1	#train-cases (%) (↑)	LD_{KL} (↓)	WCI (↓)
Language				
German	68.2	**35458 (60%)**	**0.03**	0.20
French	**70.6**	21179 (35%)	**0.03**	**0.19**
Italian	65.2	3072 (5%)	0.04	**0.19**
Legal Area				
Penal law	56.9	**15173 (31%)**	**0.00**	0.20
Civil law	**83.4**	11795 (25%)	**0.00**	0.20
Social law	66.4	11477 (24%)	0.02	**0.16**
Insurance Law	70.8	9727 (20%)	0.06	0.20
Region				
R. Lémanique	71.3	**13436 (27%)**	0.04	0.20
Zürich	68.5	8788 (18%)	0.04	0.18
E. Mittelland	69.8	8257 (17%)	**0.08**	**0.16**
E. Switzerland	**73.6**	5707 (12%)	0.02	0.24
N.W. Switzerland	72.8	5655 (11%)	0.03	0.19
C. Switzerland	69.5	4779 (10%)	0.03	0.19
Ticino	68.3	2255 (6%)	**0.00**	0.17
Federation	63.9	1308 (3%)	**0.00**	0.27

SPC (Criminal Offense Prediction)				
Group	mF1	#train-cases (%) (↑)	LD_{KL} (↓)	WCI (↓)
Defendant Gender				
Male	60.3	**73952 (92%)**	**0.03**	0.01
Female	60.1	6048 (8%)	0.08	**0.03**
Region				
Beijing	**66.8**	**16588 (21%)**	**0.05**	**0.02**
Liaoning	56.7	13934 (17%)	**0.05**	**0.02**
Hunan	59.5	12760 (16%)	**0.05**	**0.02**
Guangdong	58.0	12278 (15%)	**0.05**	0.01
Sichuan	56.4	11606 (14%)	0.06	**0.02**
Guangxi	58.9	8674 (11%)	0.07	**0.02**
Zhejiang	58.8	4160 (5%)	0.07	**0.02**

Table 3: Statistics for the three general (attribute agnostic) cross-examined factors (*representation inequality*, *temporal concept drift*, and *worst-class influence*), as introduced in Section 6. We highlight the ***worst*** and **best** performing group per attribute. In **boldface**, we highlight the best (less harmful) value per factor across groups. Performance (mF1) reported for ERM.

less represented, has higher temporal drift and has more cases in the worst performing class. This is not the case considering performance across other attributes. It is also not the case for SCOTUS. In FSCS, considering the attributes of language and region, representation inequality seems to be an important factor that leads to group disparity. This is not the case for legal area, where the best represented group is the worst performing group. In other words, there are other reasons that lead to performance disparity in this case; for example, inconsistencies in rules and gathering of evidence in criminal cases potentially affects the predictability of rulings (Macula, 2019). In sum, we do not see any of these factors fully explain the performance disparities across groups.

Cross-Attribute Influence Analysis We have evaluated fairness across attributes that are not necessarily independent of each other. We therefore evaluate the extent to which performance disparities along different attributes correlate, i.e., how attributes interact, and whether performance differences for attribute A_1 can potentially explain performance differences for another attribute A_2. We examine this for the two attributes with the highest group disparity: the *defendant state* in ECtHR, and the *legal area* in FSCS. For the bins induced by these two attributes (A_1), we compute mF1 scores across other attributes (A_2).

In ECtHR, approx. 83% and 81% of *male* and *women* applicants are involved in cases against *E.C. European* states (best-performing group). Similarly, in case of age groups, we observe that ratio of cases against E.C. European states is: 87% and 86% for ≤65 and ≤35, the best- and worst-performing groups respectively.

In FSCS, the ratio of cases relevant to *penal law* is: approx. 29%, and 41% written in written in *French* (best-performing group) and *Italian* (worst-performing group). Similarly, approx. 27% originated in *E. Switzerland* (best-performing group) and 42% in *Federation* (worst performing group) are relevant to penal law. In both attributes, there is a 15% increase of cases relevant to penal law for the worst performing groups. In other words, the group disparity in one attribute A2 (language, region) could be also explained by the influence of another attribute A1 (legal area).

In Table 4, we report the performance in the aforementioned cross-attribute (A1, A2) pairings. With the exception of the (age, defendant state) cross-examination in ECtHR, we observe that group disparities in attribute A2 (Table 3) are consistent across groups of the plausible influencer (i.e. attribute A1). Hence, cross-attribute influence does not explain the observed group disparities.

ECtHR (A1: Defendant State)			
Group (A2)	E.C.E.	Rest	Avg.
Male	55.8	35.1	54.4
Female	**61.3**	**47.1**	**60.6**
≤35	48.1	**44.2**	46.2
≤65	**61.0**	34.7	**59.7**
FSCS (A1: Legal Area)			
Group (A2)	Penal Law	Civil Law	Avg.
French	**57.4**	**82.4**	**70.6**
Italian	56.2	69.4	65.2
E. Switzerland	**55.9**	**87.0**	**73.6**
Federation	54.5	72.8	63.9

Table 4: Results in cross-attribute influence. Scores for pairings of groups for attributes (A1, A2).

We believe that such an in-depth analysis of the results is fundamental to understand the influence of different factors in the outcomes. This analysis wouldn't be possible, if we had "counterfeited" an ideal scenario, where all groups and labels where equally represented. While a controlled experimental environment is frequently used to examine specific factors, it could hide, or partially alleviate such phenomena, hence producing misleading results on fairness of the examined models.

7 Conclusions

We introduced FairLex, a multi-lingual benchmark for the development and testing of bias-mitigation models or algorithms within the legal domain, based on four datasets covering four jurisdictions, five languages and various sensitive attributes. Furthermore, we provided competitive baselines including state-of-the-art transformer-based models adapted to the examined datasets, and an in-dept examination of performance of four group robust algorithms (Adversarial Removal, IRM, Group DRO, and REx). While, these algorithms improve worst case performance in the literature, when applied in a controlled experimental environment, they fail in a real-world setting, where both groups across attributes, and labels are imbalanced, while also both group and label distributions change over time. Furthermore, we cannot identify a single algorithm that performs better across datasets and groups compared to the rest.

Ethics Statement

The scope of this work is to provide an evaluation framework along with extensive experiments to further study fairness within the legal domain. Following the work of Angwin et al. (2016), Dressel and Farid (2018), and (Wang et al., 2021b), we provide a diverse benchmark covering multiple tasks, jurisdictions, and protected (examined) attributes. We conduct experiments based on state-of-the-art pre-trained transformer-based language models and compare model performance across four representative group-robust algorithm, i.e., Adversarial Removal (Elazar and Goldberg, 2018), Group DRO (Sagawa et al., 2020), IRM (Arjovsky et al., 2020) and REx (Krueger et al., 2020).

We standardize and put together four datasets: ECtHR (Chalkidis et al., 2021), SCOTUS of (Spaeth et al., 2020), FSCS (Niklaus et al., 2021), and SCP (Xiao et al., 2018; Wang et al., 2021b) that are already publicly available. ECtHR cases are partially annonymized by the court. Its data is processed and made public in accordance with the European Data Protection Law. SCOTUS cases may also contain personal information and the data is processed and made available by the US Supreme Court, whose proceedings are public. In FSCS, the names of the parties have been redacted by the court according to its official guidelines. The same applies for SPC.

We note that some protected attributes within our datasets are extracted automatically, i.e., the gender and the age of the ECtHR dataset, by means of Regular Expressions, or manually clustered by the authors, such as the defendant state in the ECtHR dataset and the respondent attribute in the SCOTUS dataset. Those assumptions and simplifications can hold in an experimental setting only and by no means should be used in real-world applications where some simplifications, e.g., binary gender, would not be appropriate. By no means, we endorse the law standards or framework of the examined datasets, to any degree rather than the publication and use of the data.

We believe that this work can help practitioners to build assisting technology for legal professionals - with respect to the legal framework (jurisdiction) they operate -; technology that does not only rely on performance on majority groups, but also considering minorities and the robustness of the developed models across them. We believe that this is an important application field, where more research should be conducted (Tsarapatsanis and Aletras, 2021) in order to improve legal services and democratize law, but more importantly highlight (inform the audience on) the various multi-aspect shortcomings seeking a responsible and ethical (fair) deployment of technology.

References

Nikolaos Aletras, Elliott Ash, Leslie Barrett, Daniel Chen, Adam Meyers, Daniel Preotiuc-Pietro, David Rosenberg, and Amanda Stent, editors. 2019. *Proceedings of the 1st Natural Legal Language Processing Workshop at NAACL 2019*. Minneapolis, Minnesota.

Elizabeth Anderson. 1999. What is the point of equality? *Ethics*, 109(2).

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization.

Woodrow Barfield. 2020. *The Cambridge Handbook of the Law of Algorithms*. Cambridge Law Handbooks. Cambridge University Press.

Kristen Bell, Jenny Hong, Nick McKeown, and Catalin Voss. 2021. The Recon Approach: A New Direction for Machine Learning in Criminal Law. *Berkeley Technology Law Journal*, 37.

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

Ilias Chalkidis and Dimitrios Kampas. 2019. Deep learning in law: Early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2):171–198.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Council of European Union. 2000. Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. *Publications Office of the EU*.

Council of European Union. 2006. Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation. *Publications Office of the EU*.

Council of European Union. 2021. Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *Publications Office of the EU*.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(10).

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Sandra Fredman. 2016. Substantive equality revisited. *I-CON Oxford Legal Studies*, 14:712—-773.

Janneke Gerards and Raphaële Xenedis. 2020. Algorithmic discrimination in europe: challenges and opportunities for gender equality and non-discrimination law.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. In *Advances in Neural Information Processing Systems*.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938, Stockholmsmässan, Stockholm Sweden. PMLR.

Daniel Martin Katz. 2012. Quantitative legal prediction-or-how I learned to stop worrying and start preparing for the data-driven future of the legal services industry. *Emory Law Journal*, 62:909.

Daniel Martin Katz, Corinna Coupette, Janis Beckedorf, and Dirk Hartung. 2020. Complex societies and the growth of the law. *Scientific Reports*, 10:18737.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Rémi Le Priol, and Aaron C. Courville. 2020. Out-of-Distribution Generalization via Risk Extrapolation (REx). *CoRR*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Alan Lundgard. 2020. Measuring justice in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 680, New York, NY, USA. Association for Computing Machinery.

Laura Macula. 2019. *The Potential to Secure a Fair Trial Through Evidence Exclusion: A Swiss Perspective*, pages 15–60. Springer International Publishing, Cham.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Court-Predict: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the 2022 NLLP Workshop*, Online.

Anya E.R. Prince and Daniel Schwarcz. 2019. Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review*, 105.

John Rawls. 1971. *A Theory of Justice*, 1 edition. Belknap Press of Harvard University Press, Cambridge, Massachussets.

Ingrid Robeyns. 2009. Justice as fairness and the capability approach. *Arguments for a Better World. Essays for Amartya Sen's*, 75:397–413.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*.

Sebastian Felix Schwemer, Letizia Tomada, and Tommaso Pasini. 2021. Legal ai systems in the eu's proposed artificial intelligence act. In *In Joint Proceedings of the Workshops on Automated Semantic Analysis of Information in Legal Text (ASAIL 2021) and AI and Intelligent Assistance for Legal Professionals in the Digital, Workplace (LegalAIIA 2021)*.

Harold J. Spaeth, Lee Epstein, Jeffrey A. Segal Andrew D. Martin, Theodore J. Ruger, and Sara C. Benesh. 2020. Supreme Court Database, Version 2020 Release 01. Washington University Law.

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

V. Vapnik. 1992. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann.

Sandra Wachter, Brent Daniel Mittelstadt, and Chris Russell. 2021. Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law. *West Virginia Law Review*, 123.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021a. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Yuzhong Wang, Chaojun Xiao, Shirong Ma, Haoxi Zhong, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2021b. Equality before the law: Legal judgment consistency analysis for fairness. *Science China - Information Sciences*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

A Discrimination and Fairness in Law

The legal notion of *discrimination* has a different scope and semantics in comparison to the notions of *fairness and bias* used in the context of machine learning (Gerards and Xenedis, 2020), where the aim usually is to achieve *equal odds*, e.g. that a court shall rule the same decision for both men and woman based on similar facts, or to have 50/50 favourable decisions for both man and woman, but *equal opportunities* (Rawls, 1971).

In particular, EU non-discrimination law (Council of European Union, 2000, 2006) prohibits both *direct and indirect discrimination*. *Direct discrimination* occurs when one person is treated "*less favourably than another is, has been or would be treated in a comparable situation*" on grounds of sex, racial or ethnic origin, disability, sexual orientation, religion or belief and age in the context of a protected sector (e.g. the workplace and provision of goods and services) (Wachter et al., 2021). Prohibiting direct discrimination allows to provide people with equal access to opportunities (i.e. formal equality). This however does not suffice, nor guarantee to create equality of opportunity (i.e. substantive equality), which can instead be achieved only by accounting for protected attributes and for social and historical realities and by taking positive measures to level the playing field (Fredman, 2016). The notion of *indirect discrimination* is grounded on achieving substantive equality in practice. Indirect discrimination refers to situations in which an apparently neutral provision, criterion or practice would put persons with a protected characteristic at disadvantage in comparison to other persons, unless 'that provision, criterion or practice is "*justified by a legitimate aim and the means of achieving that aim are appropriate and necessary*".

Nevertheless, the current EU non-discrimination law framework suffers from limitations, both as regards its personal (i.e. it only protects six characteristics) and material scope (i.e. the prohibition on discrimination is limited only to certain fields) (Gerards and Xenedis, 2020). These limitations pose problems in connection to algorithmic discrimination. For example, as algorithmic bias often creates seemingly neutral distinctions which

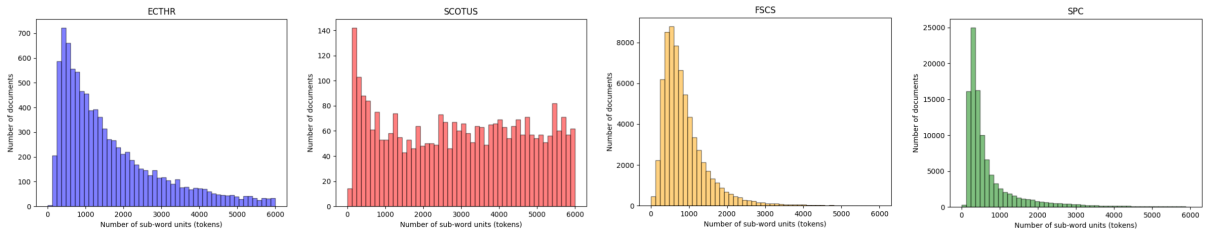

Figure 2: Distribution of sequence (document) length across FairLex datasets (ECtHR, SCOTUS, FSCS).

however often correlate to a protected group (i.e. proxy discrimination), the limited list of protected grounds renders difficult to tackle the effects of algorithmic bias through the concept of direct discrimination (Prince and Schwarcz, 2019). Indirect discrimination can help address those cases. but its application in this context poses several challenges.

In April 2021 the European Commission presented a proposal for a Regulation laying down harmonized rules on artificial intelligence (AI Act / AIA) (Council of European Union, 2021). The proposal aims at avoiding "*significant risks to the health and safety or fundamental rights of persons*" and would, once adopted, complement the currently applicable legal framework for tackling algorithmic discrimination, thereby overcoming some of its existing limitations. The envisaged implementation of the proposed AI Act highlights the importance that the legislator poses in preventing and mitigating discrimination and biases arising from the development and use of AI systems in several areas of application, including in the legal sector (Schwemer et al., 2021). AI systems used for the administration of justice and democratic processes are proposed to be deemed high-risk in order "*to address the risks of potential biases, errors, and opacity*" (recital 40 AIA). The consequence is that such systems would be subject to a variety of design and development requirements, e.g. related to the training, validation and testing data sets which would have to be examined *inter alia* in relation to possible biases (art. 10(2) lit. f AIA) or related to human oversight of such AI system with a view to remain aware of automation bias (art. 14(4) lit. b AIA).

The topic deserves great attention because AI systems learning from historical data pose the risk of transporting biases previously encumbered in the data in future decision-making, thereby exponentially increasing their effect. For example, criminal justice is already often strongly influenced by racial bias, with people of colour being more likely to be arrested and receive higher punishments than others, both in both in the USA[16] and in the UK.[17]

B Train and Evaluation Details

We fine-tune all models using the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 3e-5. We use a batch size of 16 and train models for up to 20 epochs using early stopping on validation performance.[18] Across datasets and attributes, we run five repetitions with different random seeds and report averaged scores.

C Statistics

C.1 Distribution of Document Length

In Figure 2 we report the distribution of sequence (document) length across FairLex datasets (ECtHR, SCOTUS, FSCS). We observe that the documents are extremely long (3,000-6,000+ words) across datasets.

C.2 Group Distribution by Attribute

In Tables 5 and 6 we report the group distribution per examined attribute under consideration. In some cases, the extraction of the specific attribute, e.g., gender or age in ECtHR, was not possible, i.e., the applied rules would no suffice, possibly because the information is intentionally missing. During training, the groups of unidentified samples is included, but we report test scores excluding those, i.e., $\overline{\text{mF1}}$ and *GD* do not take into account the F1 of these groups.

D Label Distribution KL Divergences

In Tables 7, 8, 9, and 10, we report the Jensen-Shannon divergences between train-test, train-dev and test-test distribution of labels separately for each protrected attribute values and for each dataset in our framework.

[16] https://tinyurl.com/4cse552t
[17] https://tinyurl.com/hkff3zcb
[18] We train all models in a mixed-precision (fp16) setting to use the maximum available batch size.

ECtHR								
Applicant Age				*Applicant Gender*			*Defendant State*	
N/A	≤ 35	≤ 65	> 65	N/A	Male	Female	East	West
2,794	839	4,246	1,121	3,306	4,407	1,287	7,224	1,776

Table 5: Group distribution in training set for each attribute of ECtHR dataset. 'N/A' (Not Answered) refers to samples, where the respected attribute could not be extracted.

SCOTUS						
Defendant					*Direction*	
Other	Facility	Organization	Person	Public Entity	Conservative	Liberal
957	140	741	1847	2796	3146	3335

Table 6: Group distribution in training set for each attribute of SCOTUS dataset.

	Applicant Age			*Applicant Gender*		*Defendant State*	
	≤ 35	≤ 65	> 65	Male	Female	East	West
Train-Test	0.19	0.18	0.32	0.17	0.26	0.17	0.28
Train-Dev	0.18	0.19	0.22	0.17	0.22	0.18	0.17
Dev-Test	0.20	0.08	0.19	0.09	0.10	0.09	0.16

Table 7: Jensen-Shannon Divergence of label distribution between training, test and development sets of ECtHR by protected attribute values. The lower the values the more similar the distributions.

	Defendant					*Direction*	
	Facility	Organization	Other	Person	Pub. Entity	Conservative	Liberal
Train-Test	0.26	0.11	0.09	0.05	0.07	0.05	0.04
Train-Dev	0.28	0.11	0.11	0.07	0.03	0.06	0.05
Dev-Test	0.22	0.17	0.13	0.10	0.07	0.09	0.07

Table 8: Jensen-Shannon Divergence of label distribution between training, test and development set in Scotus by protected attribute values. The lower the values the more similar the distributions.

		Train-Test	Train-Dev	Dev-Test
Language	DE	0.0336	0.0275	0.0061
	FR	0.0517	0.0301	0.0216
	IT	0.0145	0.0405	0.0261
Legal Area	Other	0.1000	—	—
	Public Law	0.0007	0.0090	0.0083
	Penal Law	0.0018	0.0118	0.0136
	Civil Law	0.0248	0.0046	0.0202
	Social Law	0.0624	0.0570	0.0054
Region	Région lémanique	0.0447	0.0259	0.0188
	Zürich	0.0447	0.0345	0.0028
	Espace Mittelland	0.0765	0.0435	0.0331
	NW Switzerland	0.0280	0.0127	0.0407
	E Switzerland	0.0197	0.0394	0.0198
	C Switzerland	0.0267	0.0304	0.0036
	Ticino	0.0023	0.0284	0.0307
	Federation	0.0018	0.0385	0.0404

Table 9: Jensen-Shannon Divergence of label distribution between training, test and development set in FSCS by protected attribute values. The lower the values the more similar the distributions.

	Region							*Gender*	
	Beijing	Liaoning	Hunan	Guangdong	Sichuan	Guangxi	Zhejiang	Male	Female
Train-Test	0.0516	0.0458	0.0495	0.0524	0.0559	0.0696	0.0687	0.0345	0.0766
Train-Dev	0.0239	0.0270	0.0406	0.0584	0.0484	0.0426	0.0338	0.0164	0.0318
Dev-Test	0.0469	0.0296	0.0799	0.0431	0.0554	0.0496	0.0633	0.0307	0.0986

Table 10: Jensen-Shannon Divergence of label distribution between training, test and development set in SPC by protected attribute values. The lower the values the more similar the distributions.