DispaRisk: Assessing Fairness Through Usable Information

Jonathan Vasquez[1,2](✉), Carlotta Domeniconi[2], and Huzefa Rangwala[2]

[1] Universidad de Valparaiso, Chile
jonathan.vasquez@uv.cl
[2] George Mason University, Fairfax, USA
{jvasqu6,cdomenic,rangwala}@gmu.edu

Abstract. Machine Learning algorithms (ML) impact virtually every aspect of human lives and have found use across diverse sectors including healthcare, finance, and education. Often, ML algorithms have been found to exacerbate societal biases present in datasets leading to adversarial impacts on subsets/groups of individuals and in many cases on minority groups. To effectively mitigate these untoward effects, it is crucial that disparities/biases are identified early in a ML pipeline. This proactive approach facilitates timely interventions to prevent bias amplification and reduce complexity at later stages of model development. In this paper, we leverage recent advancements in usable information theory to introduce DispaRisk, a novel framework designed to proactively assess the potential risks of disparities in datasets during the initial stages of the ML pipeline. We evaluate DispaRisk's effectiveness by benchmarking it against commonly used datasets in fairness research. Our findings demonstrate DispaRisk's capabilities to identify datasets with a high risk of discrimination, detect model families prone to biases within an ML pipeline, and enhance the explainability of these bias risks. This work contributes to the development of fairer ML systems by providing a robust tool for early bias detection and mitigation. The code is available at https://github.com/jovasque156/disparisk.

Keywords: usable information · fairness · uncertainty · bias

1 Introduction

Extensive research on fairness in machine learning (ML) has shown that biased datasets can amplify historical and societal inequities [24, 19, 31, 8], harming minorities and disadvantaged groups in areas like criminal justice [25], healthcare [2], and education [30]. This underscores the need to detect biases throughout the ML pipeline—especially in its early stages [7, 12]. To address this, *data-* and *model-focused* metrics help identify potential discrimination risks, but they each have limitations.

Data-focused metrics are computed directly from the dataset and include Class Imbalance (CL) [11], Difference in Positive proportions in observed Labels (DPL) [11], and Mutual Information (MI) between the sensitive attribute and the

rest of the features [9]. While useful, these approaches do not account for model selection or preferences, making it difficult to determine which models are more likely to produce disparate outcomes. To address this gap, *model-focused* metrics analyze trained models directly. These methods, discussed by [11, 30], detect existing discrimination (e.g., Demographic Disparity (DEMP) and Equalized Opportunity (EQODD)) or provide explanations for disparate predictions (e.g., KernelSHAP [23, 27]). However, these evaluations occur late in the pipeline and are tied to specific models, limiting their generalizability. Moreover, they do not fully capture how the interaction between data characteristics and model capabilities affects fairness.

While fairness metrics help assess bias at different stages, they do not address how models interact with data properties in practice. Even when datasets appear balanced, models may process information differently across groups, leading to hidden disparities. For instance, in an ML pipeline classifying rural and urban loan applicants as approved or denied, a dataset may predict approvals equally across groups. However, simple models might effectively leverage credit scores for urban applicants while struggling with interaction-based features critical for rural ones. This disparity in information usability can lead to uneven model performance and potential discrimination, even with seemingly fair datasets. Moreover, increasing model complexity does not necessarily resolve these issues, as it depends on whether the model can effectively utilize nuanced information for different groups. Hence, key questions emerge: Can differences in usable information across groups be quantified to trace disparities? How does model choice influence these differences and outcomes? Addressing these requires a deeper assessment of ML pipelines, beyond dataset balance and final model evaluation. Specifically, an approach that enables early detection of disparity risks while accounting for model-specific characteristics is needed.

To operationalize these insights, we introduce DispaRisk, a framework designed to detect disparity risks early in the ML pipeline while considering the characteristics of the predictive models being used. Building on the *usable information* notions studied by Xu et al. [32], DispaRisk enables proactive fairness assessments by guiding the estimation of usable information-based metrics. Specifically, given a set of potential model choices, DispaRisk facilitates assessment analyses that: (1) can be conducted in the early stages of ML pipelines, (2) account for the predictive families selected, (3) correlate with data- and model-focused fairness metrics, and (4) explain why different model families generate disparate outcomes. This approach serves as an effective predictor of discrimination risks that may emerge later in the pipeline.

The key contributions of this study are threefold:

1. We introduce DispaRisk, a framework that leverages recent advances in *usable information* theory to detect early-stage disparities across predictive model families in ML pipelines. To this end, we develop:
   - Instance-level disparity scores (DispaRisk, DR) using pointwise $\mathcal{V}$-entropy to identify individuals at high risk (see Section 3).

  - Feature-level explanations (Uncertainty Reduction, UR) by quantifying how masking each feature alters model uncertainty (see Section 4.2).
2. We bridge the gap between *data*- and *model*-focused bias assessment approaches.
3. We demonstrate practical applications through experiments across diverse datasets, showcasing its ability to identify high-risk datasets, detect bias-prone model families, and improve bias explainability.

2 Basics and Preliminaries

2.1 ML Pipeline Basics

Let X, S, and Y be random variables in the space $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$, representing the input features, sensitive attributes, and target variable, respectively. An ML pipeline is given access to a dataset $\mathcal{D}_n = \{x_i, s_i, y_i\}_{i=1}^n \in \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$ of n instances to learn a mapping function $h : \mathcal{X} \mapsto \mathcal{Y}$ by employing a finite set of possible models $\mathcal{V}$. We assume that there is access to sufficient information about the ML pipeline to identify the set of possible models.

2.2 Fairness Notion

We examine fairness through independence and separation notions [7, 12, 24, 30]. Independence requires the learned mapping function's outcomes to be independent of the sensitive attribute ($h(X) \perp S$), while separation also requires independence, but conditioned to the ground truth ($h(X) \perp S|Y$). Our analysis focuses on positive class of binary classifications, which typically signify favorable decisions with significant social implications. For example, in contexts such as university admissions or loan approvals, positive outcomes (e.g., being admitted or approved) directly influence individuals' opportunities. To evaluate these disparity kinds, we employ the metrics DEMP and EQOPP explained as follows:

Definition 1 (Demographic Disparity (DEMP)). *Difference in the* ***positive rate*** *of class $k \in Y$ between the advantaged (s) and disadvantaged (s') groups.*

$$\Delta_{DEMP}(h, S, Y_k) = P(h(X) = 1|S = s) - P(h(X) = 1|S = s')$$

Definition 2 (Equalized Opportunity (OPP)). *Difference in the* ***true positive rate*** *of class $k \in Y$ between advantaged (s) and disadvantaged (s') groups:*

$$\Delta_{OPP}(h, S, Y_k) = P(h(X) = 1|S = s, Y_k = 1) - P(h(X) = 1|S = s', Y_k = 1)$$

2.3 Usable information framework

The *usable information* framework [32] quantifies uncertainty differences across groups within a model family, highlighting the impact of model selection. We next replicate Xu et al.'s [32] metric formulations, propose a new metric, and outline their estimation within the $\mathcal{V}$-information framework. The next subsection introduces DispaRisk, a framework for improving fairness analysis in ML pipelines by assessing model class, usable information, and disparate outcomes.

$\mathcal{V}$-information framework. Xu et al. [32] introduces the $\mathcal{V}$-information framework to estimate *usable information* withim a family of models $\mathcal{V}$. A first formulated concept is the predictive conditional $\mathcal{V}$-entropy, which represents the minimum achievable expected negative log-likelihood to predict Y given X using models from the predictive family $\mathcal{V}$. Formally:

Definition 3 (Predictive conditional $\mathcal{V}$-*entropy*). *For a family $\mathcal{V}$ of models, the conditional $\mathcal{V}$-entropy[3] of Y given X is defined as:*

$$H_{\mathcal{V}}(Y|X) = \inf_{h \in \mathcal{V}} \mathbb{E}_{x,y \sim X,Y}[-\log_2 h[x](y)] \tag{1}$$

The infimum in Equation 1 is attained by finding the function $h \in \mathcal{V}$ that minimizes the expected negative log-likelihood,[4]. Measuring a model class's uncertainty in predicting Y from X requires identifying its best-performing model. Unlike Shannon entropy, $\mathcal{V}$-entropy depends on $\mathcal{V}$, providing distinct uncertainty measures across model classes, making it valuable for comparing predictive capacities – a key focus of our study.

While $\mathcal{V}$-entropy aggregates uncertainty over the dataset, bias assessment requires analyzing specific data slices, such as demographic differences. To address this, we propose Pointwise $\mathcal{V}$-entropy ($\texttt{P}\mathcal{V}\texttt{E}$) to quantify instance-level uncertainty within a model family $\mathcal{V}$. Formally:

Definition 4 (Pointwise $\mathcal{V}$-entropy ($\texttt{P}\mathcal{V}\texttt{E}$)). *For a family $\mathcal{V}$, and an instance represented by the tuple (x, y), the pointwise $\mathcal{V}$-entropy ($\texttt{P}\mathcal{V}\texttt{E}$) is defined as:*

$$\texttt{P}\mathcal{V}\texttt{E}(x \mapsto y) = -\log_2 h[x](y) \tag{2}$$

where $h \in \mathcal{V}$ such that $\mathbb{E}[-\log_2 h[X](Y)] = H_{\mathcal{V}}(Y|X)$.

Higher $\texttt{P}\mathcal{V}\texttt{E}$ values indicate greater uncertainty, meaning models within $\mathcal{V}$ struggle to predict the instance accurately. $\texttt{P}\mathcal{V}\texttt{E}$ complements $\texttt{PVI}$ [6], which estimates usable information by comparing predictions with and without x. In contrast, $\texttt{P}\mathcal{V}\texttt{E}$ focuses on the remaining uncertainty when x is given, simplifying estimation and reducing estimation costs.

Estimating $\mathcal{V}$-entropy and $\texttt{P}\mathcal{V}\texttt{E}$. The $\mathcal{V}$-entropy can be empirically estimated on a finite dataset $\mathcal{D}$ of n instances as:

$$\hat{H}_{\mathcal{V}}(Y|X;\mathcal{D}) = \inf_{h \in \mathcal{V}} \frac{1}{n} \sum_{x_i, y_i \in \mathcal{D}} -\log_2 h[x_i](y_i) \tag{3}$$

$$= \inf_{h \in \mathcal{V}} \frac{1}{n} \sum_{x_i, y_i \in \mathcal{D}} \texttt{P}\mathcal{V}\texttt{E}(x_i \mapsto y_i) \tag{4}$$

[3] In this article, conditional $\mathcal{V}$-entropy is referred to as $\mathcal{V}$-entropy.
[4] With $\log_2$, the measure is in bits; for nats, use $\log_e$.

Algorithm 1 $\mathcal{V}$-entropy and $\texttt{P}\mathcal{V}\texttt{E}$

Require: $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{k}$, $\mathcal{D}_{held-out} = \{(x_i, y_i)\}_{i=k+1}^{n}$, and family $\mathcal{V}$
Ensure: $\hat{H}_{\mathcal{V}}$ and $\texttt{P}\mathcal{V}\texttt{E}$ estimates.
1: $h \leftarrow$ fine-tune $\mathcal{V}$ on $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{k}$
2: $\hat{H}_{\mathcal{V}}(Y|X) \leftarrow 0$
3: **for** $(x_i, y_i) \in \mathcal{D}_{held-out}$ **do**
4: $\quad \hat{H}_{\mathcal{V}}(Y|X) \leftarrow \hat{H}_{\mathcal{V}}(Y|X) - \frac{1}{n-k} \log_2 h[x_i](y_i)$
5: $\quad \texttt{P}\mathcal{V}\texttt{E}(x_i \mapsto y_i) \leftarrow -\log_2 h[x_i](y_i)$
6: **end for**

where the infimum $h \in \mathcal{V}$ is approximated using cross-entropy loss to minimize the negative log-likelihood of Y given X [32, 6]. The approximation of $\hat{H}_{\mathcal{V}}$ is achieved by training or fine-tuning a pretrained model following Algorithm 1, which extends [6] to focus on $\mathcal{V}$-entropy and $\texttt{P}\mathcal{V}\texttt{E}$. The algorithm splits data into training and held-out sets, using the latter to estimate $\hat{H}_{\mathcal{V}}$ and $\texttt{P}\mathcal{V}\texttt{E}$. Since estimation is based on finite data, results may deviate from true $\mathcal{V}$-entropy. Xu et al. [32] provide Probably Approximately Correct (PAC) bounds, showing that larger datasets and simpler $\mathcal{V}$ yield tighter bounds.

3 DispaRisk

3.1 $\mathcal{V}$-information in disparity assessment

We propose to assess unfairness by comparing uncertainty for predicting Y from X across slices of the dataset that we are interested, arguing that these differences are expected to align with fairness metrics in later stages of the ML pipeline. To this end, we introduce DispaRisk (DR), a framework for computing differences over P$\mathcal{V}$E averages of data slices and analyzing their relationship with disparities. In the following paragraphs, we first explain the rationale of using average of P$\mathcal{V}$E,[5] and then, introduce how to compute DR for *independence* and *separation* notions of fairness.

Rationale of using average P$\mathcal{V}$E. For disparity metrics to be meaningful, uncertainty estimates across groups must be comparable—that is, all derived under the same reference mapping function. Computing separate infimum h for advantaged and disadvantaged slices produces inherently incomparable $\mathcal{V}$-entropy values, since each is defined over a different input–output space. In contrast, using a single global infimum ensures every average P$\mathcal{V}$E is evaluated against the same baseline, isolating uncertainty differences that arise solely from the groups' data distributions rather than from variations in the learnable mapping function. Indeed, for fairness criteria defined over a single class (such as EQOPP,

[5] It is worthy to note that $\mathcal{V}$-entropy coincides with the expected P$\mathcal{V}$E only if the model used to estimate P$\mathcal{V}$E has been trained on the entire dataset; accordingly, the mean P$\mathcal{V}$E of any subset does not equal to its own $\mathcal{V}$-entropy [32, 6].

discussed later), estimating group-specific $\mathcal{V}$-entropy is ill-posed: taking an infimum over slices of only one target class yields zero $\mathcal{V}$-entropy. Nevertheless, we leave exploration of group-specific training regimes to future work.

DR and DEMP: to analyze fairness under *independence* notions (in this study, measured through DEMP), we propose to compute the average of $\mathtt{P\mathcal{V}E}$ on slices comprised by instances belonging to the advantaged ($\mathcal{D}_a$) and disadvantaged ($\mathcal{D}_d$) group. Formally:

$$\mathtt{DR}(\mathcal{D}_a, \mathcal{D}_d|\mathcal{V}) = \frac{1}{|\mathcal{D}_a|}\sum_{x,y\in\mathcal{D}_a} \mathtt{P\mathcal{V}E}(x \mapsto y) - \frac{1}{|\mathcal{D}_d|}\sum_{x,y\in\mathcal{D}_d} \mathtt{P\mathcal{V}E}(x \mapsto y) \tag{5}$$

where DR quantifies the difference in P$\mathcal{V}$E among groups in the family $\mathcal{V}$ when forecasting Y from X. Since $\mathcal{V}$ is defined over the entire dataset, DR serves as a computationally efficient alternative, requiring only one model per group instead of a full dataset estimation.

DR and EQOPP: Under EQOPP in *separation*, where disparity is assessed only for the positive class, DR is computed over $Y = 1$ as follows:

$$\begin{aligned}\mathtt{DR}(\mathcal{D}_{a,y=1}, \mathcal{D}_{d,y=1}|\mathcal{V}) = &\frac{1}{|\mathcal{D}_{a,y=1}|}\sum_{x,y\in\mathcal{D}_{a,y=1}} \mathtt{P\mathcal{V}E}(x \mapsto y) \\ &-\frac{1}{|\mathcal{D}_{d,y=1}|}\sum_{x,y\in\mathcal{D}_{d,y=1}} \mathtt{P\mathcal{V}E}(x \mapsto y)\end{aligned} \tag{6}$$

where $\mathcal{D}_{a,y=1}$ and $\mathcal{D}_{d,y=1}$ represent dataset slices for advantaged and disadvantaged groups with target label $y = 1$. Thus, when comparing DR with DEMP, we use Equation (5), and for EQOPP, we use Equation (6).

3.2 The relationship between DR and fairness notions

We now examine the relationship between DR and the fairness notions of *separation* and *independence*. Higher uncertainty implies that models in $\mathcal{V}$ are less confident in predicting Y from X. When $\mathcal{V}$-entropy is high, models in $\mathcal{V}$ tend to rely on guessing, favoring the majority class in Y. Consequently, instances from the group with the highest average $\mathtt{P\mathcal{V}E}$ are more likely to be predicted as the majority class. How does this affect disparities in $\mathcal{V}$? In the following, we outline rules of thumb to address this question.

DR and *separation* through EQOPP. EQOPP measures the difference in true positive rates between advantaged and disadvantaged groups. By the definition of DR in Equation (6), higher absolute DR values should positively correlate with EQOPP. The reasoning is that greater DR differences indicate that the group with higher average $\mathtt{P\mathcal{V}E}$ experiences greater uncertainty in predicting Y from X, leading to more inaccurate predictions and a lower true positive rate. This

results in higher EQOPP values, reflecting greater disparities under *separation*. Based on this analysis, we establish the following rules of thumb, demonstrated in the experiments: *For **higher** absolute values of `DR`, **higher** levels of disparities under EQOPP are expected.*

`DR` and *independence* through DEMP. DEMP measures the difference in positive ratios between advantaged and disadvantaged groups. To determine whether higher `DR` values from Equation (5) correspond to higher or lower positive ratios for the group with greater uncertainty—and thus the expected DEMP levels—we identify two key dataset characteristics.

The first is the **majority class in the target**, which helps predict whether the higher-uncertainty group will receive a higher or lower positive rate. Since `DR` implies that the group with higher average $\mathtt{P}\mathcal{V}\mathtt{E}$ is more likely to be predicted as the majority class, the relationship between `DR` and DEMP depends on whether the majority class is positive or negative. If the majority class is positive, the higher-uncertainty group is expected to receive higher positive ratios, leading to greater disparities under DEMP. Thus, a second rules of thumb is: *For **higher** absolute values of `DR`, **higher** levels of disparities under DEMP are expected.*

Conversely, if the **majority class is negative**, the relationship is reversed. The group with higher uncertainty is now less likely to be predicted as positive, reducing the difference in positive ratios. Therefore, the third rules of thumb is defined as: *For **higher** absolute values of `DR`, **lower** levels of disparities under DEMP are expected.*

3.3 Benefits of `DR`

Our simple yet effective approach offers two key benefits. First, it aligns with fairness notions by accounting for the dependency between labels and sensitive attributes. `DR` translates fairness concepts into the $\mathcal{V}$-entropy framework, where fairness implies uncertainty differences close to zero, ensuring equal *usable information* across groups and reducing disparities. However, as we will show, this holds only under certain conditions and disparity notions.

Second, $\mathcal{V}$-entropy enables pipeline-dependent metrics for model selection. Since it is defined over $\mathcal{V}$, this set can be tailored to the models used in the ML pipeline, making `DR` context-specific rather than dataset- or model-specific. A more granular approach could involve multiple $\mathcal{V}$ sets, each representing different model families, allowing for comparable metrics across model types. The following sections demonstrate how DispaRisk enhances disparity risk assessment through a thorough analysis.

4 Experiments

4.1 Machine Learning Pipelines

We assess disparity risks in three ML pipelines using datasets KDD, FACET, and Hate Speech, denoted as $\mathcal{D}^{kdd}$, $\mathcal{D}^{facet}$, and $\mathcal{D}^{hs}$. Each dataset $\mathcal{D}$ includes

Table 1. Disadvantaged group and positive class from the sensitive attribute and target variable for each ML pipeline.

ML Pipeline	Sensitive	Disadvantaged	Target	Positive class
$\mathcal{D}^{kdd}$	sex	female	income	$> 50K$
$\mathcal{D}^{facet}$	gender	non-masculine	person-related	lawman nurse
$\mathcal{D}^{hs}$	dialectal	african-american	harrasement	non-harrasement

input features X, sensitive attribute S, and target Y for learning $h : X \mapsto Y$. While S is excluded from mapping, it remains available for fairness analysis. We now describe each dataset.

The KDD Census-Income dataset ($\mathcal{D}^{kdd}$) originates from the 1994–1995 U.S. Census Bureau surveys, containing 41 demographic and employment-related variables for 299,285 individuals. It is used to classify whether a person earns more than $50K$ per year, with sensitive attributes such as age, sex, and race.

The FACET dataset ($\mathcal{D}^{facet}$) is a benchmark from Meta AI for evaluating vision model fairness [10]. It includes 32,000 images labeled with demographic (e.g., perceived gender presentation) and person-related attributes (e.g., *lawman*, *nurse*), covering 50,000 people. We extract a dataset of 50,000 images (one per person) using provided bounding boxes, along with a binary masculine gender attribute and person-related class labels.

The Hate Speech dataset ($\mathcal{D}^{hs}$) by Davidson et al. [3] contains 24,802 tweets labeled as *hate speech*, *offensive*, or *neither*. We augment it with demographic dialect predictions from Blodgett et al. [1], estimating dialect proportions for African-American, Hispanic, White, and other groups per tweet.

Following fairness conventions for binary classification, we define disadvantaged group membership using sensitive attributes and the positive class based on target variables. Table 1 summarizes these criteria. We transform the sensitive attribute, assigning 1 to disadvantaged groups and 0 otherwise. Likewise, the target variable is set to 1 for positive class instances and 0 for all others.

4.2 Disparity Risk Assessments

We conduct two approaches to evaluate disparity risks in each ML pipeline: (1) a *baseline* using popular *dataset-focused* metrics from literature, and (2) an approach using `DR` and comparing with popular *model-focused* metrics.

Baseline. We use *data-focused* metrics—CIm, DPL, KL, r_ϕ, and Matthews Correlation Coefficient—to assess bias [11, 18]. Table 2 presents the results, highlighting varying bias levels across datasets. For $\mathcal{D}^{kdd}$, we observe slight overrepresentation of the disadvantaged group, with a moderate negative correlation ($r_\phi = -0.159$) between sensitive attributes and labels and a higher positive rate for the male socio-demographic group (DPL> 0). In $\mathcal{D}^{hs}$, bias is strongest for

Table 2. *Data-focused* metrics computed from datasets. Higher values indicate stronger relationships between sensitivities and labels.

Pipeline	Class	CIm	DPL	r_ϕ	KL
$\mathcal{D}^{kdd}$	$> 50k$	-0.04	0.08	-0.16	0.07
$\mathcal{D}^{hs}$	*no_harassment*	-0.06	0.25	-0.34	0.33
$\mathcal{D}^{facet}$	*lawman*	0.34	0.06	-0.10	0.03
	nurse		-0.04	0.12	0.02

the positive class (*no_harassment*), where DPL, r_ϕ, and KL reach the highest absolute values, indicating that the advantaged group has more tweets labeled as *no_harassment* than disadvantaged one. $\mathcal{D}^{facet}$ shows a strong overrepresentation of the advantaged group. Based on r_ϕ and DPL, the *lawman* class has a higher positive rate for the advantaged group, while *nurse* exhibits the opposite trend.

While these results offer valuable insights, they provide a global perspective, lacking the granularity needed to analyze specific model types within each ML pipeline. The following section applies *DispaRisk* to enable a more nuanced assessment of potential biases in ML pipelines.

DispaRisk in Practice. DispaRisk is applied in three steps: (1) constructing model families based on the intended models for the ML pipeline, (2) estimating `DR`, and (3) analyzing results to generate insights.

Construction of families $\mathcal{V}$. For each hypothetical ML pipeline, we define model families based on assumed preferences. For $\mathcal{D}^{kdd}$, we construct five Feedforward Neural Network (FNN) families with different activation functions: no activation (linear), ReLU [33], LeakyReLU [13], Sigmoid, and GELU [15]. For $\mathcal{D}^{hs}$, we analyze transformer-based families: BERT [4], RoBERTa [21], GPT2 [26], BART [20], and DeBERTa [14]. For $\mathcal{D}^{facet}$, we employ popular vision model families: VGG [28], Inception [29], DenseNet [17], MobileNet [16], and Vision-Transformer [5]. Model families are identified using activation functions or model architecture names as subscripts. For example, $\mathcal{V}leaky_relu$ represents FNNs with LeakyReLU, and $\mathcal{V}gpt2$ denotes the GPT2 family.

`DR` Estimates. To estimate uncertainty differences via `DR` (Section 3), we follow this protocol for each ML pipeline using dataset $\mathcal{D}^{(p)}$ and family $\mathcal{V}_i$:

(1) Split $\mathcal{D}^{(p)}$ into $\mathcal{D}^{(p)}_{train}$ and $\mathcal{D}^{(p)}_{held-out}$ sets at 80/20 ratio.
(2) Approximate the infimum $h \in \mathcal{V}_i$ by training or fine-tuning a pretrained model using cross-entropy loss (Step 1 of Algorithm 1).
(3) Estimate $H_{\mathcal{V}_i}$ and `PVE` following Steps 2–6 of Algorithm 1.

Since the most computationally powerful model in $\mathcal{V}$ often attains the infimum (Definition 3), this weakens the PAC bound [32], requiring overfitting

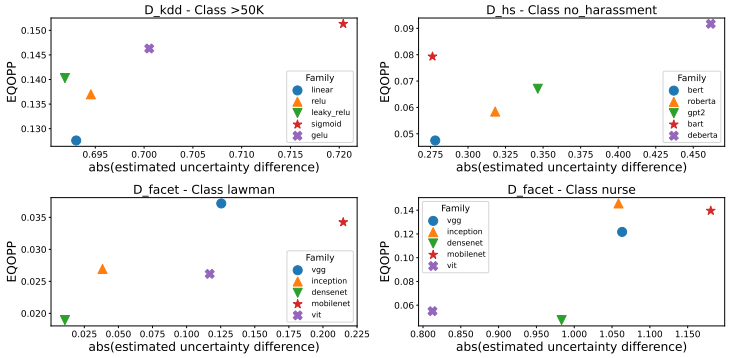

Fig. 1. Average EQOPP disparity in family versus estimated uncertainty difference through $\texttt{DR}(\mathcal{D}_{a,y=1}, \mathcal{D}_{d,y=1})$ for each family $\mathcal{V}_i$.

prevention. To mitigate this, we create a validation set $\mathcal{D}^{(p)}_{val}$ by sampling 10% of $\mathcal{D}^{(p)}_{train}$ and evaluate estimates per epoch. Models are trained/fine-tuned for 5 epochs with a learning rate of $5e-5$ and batch size 32. If overfitting arises, we lower the learning rate to $5e-6$, halve the batch size, and rerun Algorithm 1. We use the AdamW optimizer [22] with a linear scheduler[6] for all experiments.

Assessing Fairness Through Usable Information. We use the estimates to address two key questions that *data-focused* metrics alone cannot answer:

(Q.1) Which model families in the ML pipeline are more likely to replicate or exacerbate biases? To investigate this, we simulate later ML pipeline stages and compare estimated DR with observed disparities, identifying model families prone to higher bias reproduction. For example, in the $\mathcal{D}^{kdd}$ pipeline, we first estimate DR for each family. Next, we train FNNs with varying hidden layers from each family and compute average disparity levels using EQOPP and DEMP. In parallel, we estimate DR for each family. Finally, we compare uncertainty difference estimates with observed disparities to evaluate whether DR effectively signals model families more prone to exacerbating biases, as inferred in Section 3. Applying this protocol across all ML pipelines, we obtain the results shown in Figures 1 and 2, which depicts the relationship between DR estimates and EQOPP/DEMP across different pipelines and model families.

Figure 1 shows the relationship between absolute $\texttt{DR}(\mathcal{D}_{a,y=1}, \mathcal{D}_{d,y=1})$ values and average EQOPP. The observed trend confirms the rule of thumb from Section 3 across all ML pipelines, validating DR as a predictor of disparity risks for future models in downstream tasks. Given this, we derive the first insight for **(Q.1)**: *the model families most prone to higher disparities under separation notions are* $\mathcal{V}_{sigmoid}$, $\mathcal{V}_{deberta}$, *and* $\mathcal{V}_{mobilenet}$ *for the* $\mathcal{D}^{kdd}$, $\mathcal{D}^{hs}$, *and* $\mathcal{D}^{facet}$ *pipelines, respectively.*

[6] Minimum learning rate set to 0.

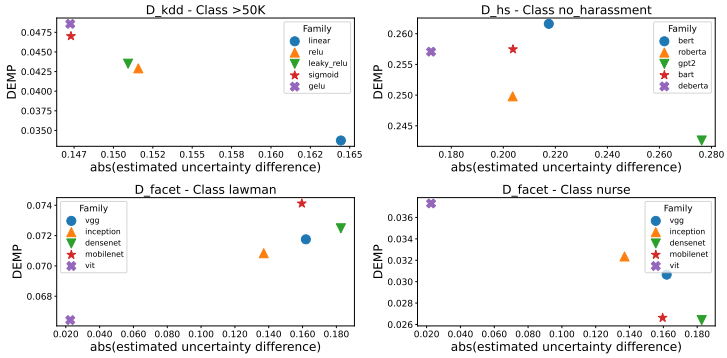

Fig. 2. Average DEMP in predictive families versus estimated uncertainty difference through $\texttt{DR}(\mathcal{D}_a, \mathcal{D}_d)$ for each family $\mathcal{V}_i$.

Figure 2 illustrates the relationship between $\texttt{DR}(\mathcal{D}_a, \mathcal{D}_d)$ and DEMP. To analyze these results and address **Q.1**, we follow a structured approach: (1) identify the majority class, (2) evaluate whether the observed trends align with the expected DR-DEMP relationship from Section 3, and (3) synthesize insights to answer **Q.1**. Applying this approach, we find that for $\mathcal{D}^{kdd}$ the negative class (income $< 50k$) is the majority, suggesting an inverse relationship between absolute DR and DEMP (Section 3). Figure 2 confirms this, with higher absolute DR values corresponding to lower DEMP. Models in $\mathcal{V}_{gelu}$ and $\mathcal{V}_{sigmoid}$ show higher disparity risks. Additionally, in $\mathcal{D}^{hs}$ the majority class is *harassment* (negative class), indicating a similar inverse DR-DEMP relationship as in $\mathcal{D}^{kdd}$. Consistent with Section 3, models in $\mathcal{V}_{deberta}$ are more prone to disparities under *independence*. Finally, for $\mathcal{D}^{facet}$ the majority class is *lawman*. From Figure 2, the rule of thumb (Section 3) that an inverse DR-DEMP relationship is expected for the *nurse* class and a direct one for *lawman* is confirmed. Models in $\mathcal{V}_{densenet}$ are more prone to disparities for *lawman*, while $\mathcal{V}vit$ shows similar tendencies for *nurse*.

(Q.2) Why might these model types produce disparate outcomes? The rules of thumb not only help identify high-risk model families but also explain why these families contribute to disparities in later pipeline stages. For example, in $\mathcal{V}_{sigmoid}$, EQOPP is higher because the group with a higher average $\texttt{P}\mathcal{V}\texttt{E}$ is less likely to be correctly predicted, leading to a lower true positive rate. Similarly, DEMP is higher as the lower average $\texttt{P}\mathcal{V}\texttt{E}$ differences shows that models are reflecting dataset biases seen in Table 2. This pattern generalizes across model families and ML pipelines.

To further explore these disparities, we analyze which features contribute to the computed average $\texttt{P}\mathcal{V}\texttt{E}$ in each model family. We select the riskiest families per ML pipeline and identify key features by measuring uncertainty reduction when a feature is added to the input space. Specifically, we compare $\texttt{P}\mathcal{V}\texttt{E}$ when feature i is masked using transformation τ_i versus when x is complete. The

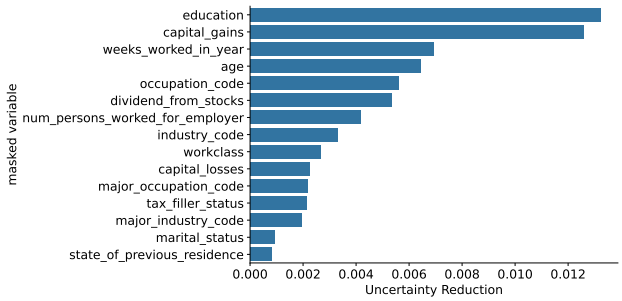

Fig. 3. Top 15 UR of features over the advantaged group (male) for the $\mathcal{V}_{sigmoid}$ family in the $\mathcal{D}^{kdd}$ ML pipeline.

transformations applied in each pipeline are: setting feature i to 0 in $\mathcal{D}^{kdd}$, replacing word i with a blank space in $\mathcal{D}^{hs}$, and setting specific pixel sets to 0 in $\mathcal{D}^{facet}$. Formally, this uncertainty reduction is measured as follows:

$$\mathtt{UR}(\mathcal{D}|\mathcal{V},\tau_i) = \frac{1}{|\mathcal{D}|}\sum_{x,y\in\mathcal{D}} \mathtt{P\mathcal{V}E}(\tau_i(x)\mapsto y) - \mathtt{P\mathcal{V}E}(x\mapsto y) \tag{7}$$

$$= \frac{1}{|\mathcal{D}|}\sum_{x,y\in\mathcal{D}} -\log_2 h[x_{\neg i}](y) + \log_2 h[x](y) \tag{8}$$

where UR represents Uncertainty Reduction, $\tau_i(x)$ denotes the transformation process that masks feature i, and $x_{\neg i}$ is the resulting output. We use $h \in \mathcal{V}$ such that $\mathbb{E}[-\log_2 h[X](Y)] = H_{\mathcal{V}}(Y|X)$. Higher UR values for feature i indicate its importance for models in $\mathcal{V}$ to accurately predict the target variable. Notably, we use the same infimum of $\mathcal{V}$ for P$\mathcal{V}$E with both $\tau_i(x)$ and the unmasked input to avoid the computational overhead of determining a separate infimum for each masked feature. While this simplification has limitations, which will be discussed in Section 5, the primary goal here is to demonstrate how *DispaRisk* extends beyond identifying risky models to offer deeper insights into potential disparities. In the following paragraphs we apply this approach to all ML pipelines.

In $\mathcal{D}^{kdd}$, the positive DPL indicates higher labeled positive rates for the advantaged group, a disparity replicated in $\mathcal{V}_{sigmoid}$ due to its lower DR. Thus, we compute UR on $\mathcal{D}_a$. Figure 3 shows the top 15 most relevant features by UR for $\mathcal{V}_{sigmoid}$, the highest-risk family. The main contributors to disparity risks are `education`, `capital_gains`, `weeks_worked_in_year`, `age`, and `occupation`, with `education` and `capital_gains` providing the greatest uncertainty reduction. Thus, complementing **(Q.1)**, uncertainty differences are largely attributed to these variables in $\mathcal{V}_{sigmoid}$. This might suggest insights such as careful preprocessing of these features to mitigate bias or further considerations on these variables during model constructions.

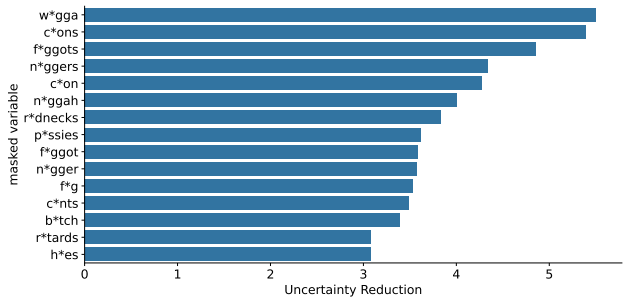

Fig. 4. Top 15 `UR` of words over the advantaged group (Not-African-American dialect) for the $\mathcal{V}_{deberta}$ family in the $\mathcal{D}^{hs}$ ML pipeline. Words are modified to avoid exposition of inappropriate text.

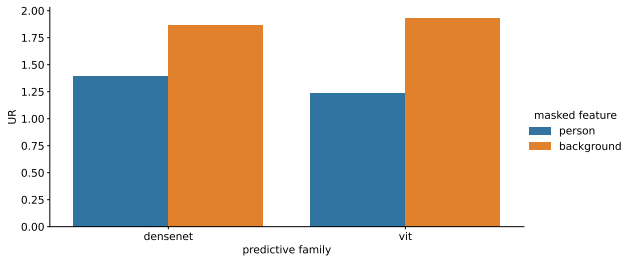

Fig. 5. `UR` of feature `person` and `background` over the disadvantaged group (Female) for the $\mathcal{V}_{densenet}$ and $\mathcal{V}_{vit}$ families in the $\mathcal{D}^{facet}$ ML pipeline.

For the $\mathcal{D}^{hs}$ pipeline, we analyze two word sets: (1) the most relevant words for each target class identified by Ethayarajh et al. [6] and (2) a manually curated list of problematic words for $\mathcal{D}^{hs}$. For each word i, we compute `UR` using a subsample of texts containing i, capturing its impact on uncertainty reduction when present versus absent. Given the `DR` values and following the approach for $\mathcal{D}^{kdd}$, we compute `UR` on the advantaged group due to the positive DPL. Figure 4 presents the top 15 words in $\mathcal{V}_{deberta}$, identified as the highest-risk family in **(Q.1)**. The analysis shows that racial and homophobic slurs most significantly reduce uncertainty in $\mathcal{V}_{deberta}$ for the advantaged group. This indicates that a specific set of biased terms largely drives uncertainty differences and, consequently, disparity risks in this predictive family.

Finally, for the $\mathcal{D}^{facet}$ dataset, we focus on the disadvantaged group, which had the highest average `PVE` in the estimated `DR`. We analyze $\mathcal{V}_{densenet}$ and $\mathcal{V}_{vit}$, identified as the highest-risk families for the `lawman` and `nurse` classes, respectively. We examine the image background and the person to determine which contributes more to uncertainty reduction, explaining the disparity risks identified earlier. Figure 5 shows that in both families, the background has the highest uncertainty reduction. This reinforces that image backgrounds significantly im-

pact the elevated uncertainty of the disadvantaged group, driving disparity risks. These findings help narrow the focus on key features when analyzing biases.

5 Discussion

DispaRisk bridges the gap between *data-focused* and *model-focused* bias detection methods. While it operates on datasets like *data-focused* techniques, it also provides insights that are both *data-centric* and *model-family-aware*, making it distinct from traditional approaches. A key difference from *model-focused* metrics is its broader applicability. Model-focused metrics evaluate bias based on a specific model's output, limiting generalizability. In contrast, *DispaRisk* estimates metrics over an entire model family $\mathcal{V}$ (Definition 3), enabling a more generalizable assessment earlier in the ML pipeline. Another advantage of *DispaRisk* is its scalability to high-dimensional data. Unlike mutual information, which struggles with high-dimensional variables, empirical $\mathcal{V}$-metrics remain tractable as dimensionality increases [32]. This is especially useful in modern ML, where high-dimensional data is common.

The utility of DispaRisk depends on choosing a model family $\mathcal{V}$ that meaningfully spans the hypothesis space of interest. A narrowly defined $\mathcal{V}$ may miss key biases, while an overly broad one can dilute signal. We therefore recommend: (i) to select a variety of families by using criteria such as Vapnik-Chervonenkis (VC) dimension or model capacity tiers, (ii) remove models that under- or overfit relative to targets, ensuring $\mathcal{V}$ reflects realistic candidates, and (iii) ensure $\mathcal{V}$ matches anticipated production pipelines.

As we described, the accuracy of $\mathcal{V}$-entropy estimates depends on the size and representativeness of held-out data: small datasets yield higher variance in pointwise entropy measurements. We recommend using an enough size of dataset in line with PAC-style bounds on entropy estimation, or employing techniques like bootstrap resampling to quantify confidence intervals. For Uncertainty Reduction (`UR`) in high-dimensional settings (e.g. large vocabularies), computing average P$\mathcal{V}$E for every feature can be prohibitive. In such cases, practitioners can apply pre-filtering strategies such as: (i) grouping similar words or (ii) performing expert-domain-driven pre-selection of interpretable feature subsets. These approaches maintain `UR`'s explanatory power while keeping computation tractable.

However, *DispaRisk* has limitations. Computing pointwise $\mathcal{V}$-entropy requires training or fine-tuning each candidate in $\mathcal{V}$. While this cost exceeds that of simpler data-centric metrics, it can be substantially reduced in practice by (i) sampling $\mathcal{V}$ by capacity tier, selecting one representative per group of similar architectures, or (ii) applying early-stopping on a small held-out slice, aborting models whose P$\mathcal{V}$E curves fall below a baseline threshold. These strategies retain the ability to detect relative uncertainty disparities while cutting training iterations.

6 Conclusions

DispaRisk enhances bias assessment in ML pipelines by addressing veracity and value challenges in large datasets. Using the $\mathcal{V}$-entropy framework, it reveals how different model families may amplify societal biases, bridging *data-focused* and *model-focused* metrics while considering computational constraints. Illustrative experiments across diverse datasets demonstrate *DispaRisk*'s effectiveness in identifying disparity sources and explaining bias propagation. By pinpointing areas where biases are amplified, it helps improve dataset quality and fairness in ML applications. Its context-specific assessments make it a valuable tool for regulatory compliance and internal use, ensuring fairness in ML-based systems.

Future research on *DispaRisk* can explore several key directions to enhance its applicability and impact. One promising avenue is refining its estimation methods to improve efficiency, enabling faster assessments in large-scale ML pipelines. Another important direction is adapting *DispaRisk* for evolving ML architectures, ensuring its relevance as models become more complex and diverse. Finally, investigating its role in AI governance and regulatory compliance can help establish standardized fairness auditing practices, fostering greater transparency and accountability in machine learning systems. While our *rules of thumb* in Section 3.2 provide intuitive guidance, we do not yet offer a full formal proof under minimal assumptions. Developing rigorous guarantees (e.g. via PAC-Bayesian bounds or VC-dimension arguments) to bound the error of disparity predictions remains an important direction for future work. We defer a complete theoretical treatment to a follow-up study. Finally, DispaRisk's main novelty lies in integrating pointwise $\mathcal{V}$-entropy gaps with fairness evaluation and demonstrating its empirical utility across modalities. We acknowledge, however, that the core metric is an adaptation of existing usable-information measures. Deriving new theoretical insights—such as generalization bounds for disparity estimates, tighter links to information-bottleneck principles, or capacity-based criteria for model-family selection—would substantially strengthen the conceptual contribution. We plan to explore these avenues in future research.

Acknowledgments. We gratefully thank the anonymous reviewers for their insightful feedback and constructive suggestions, which have substantially improved the quality and clarity of this manuscript. Additionally, this study was supported by the National Agency for Research and Development (ANID - Agencia Nacional de Investigación y Desarrollo/Subdirección de Capital Humano), "Becas Chile" Doctoral Fellowship 2020 program; Grant No. 72210492 to Jonathan Patricio Vasquez Verdugo. Finally, Jonathan Vasquez was funded by the program "Magíster en Planificación y Control de gestión" of the "Ingeniería en Información y Control de Gestión", at Universidad de Valparaíso, Chile.

References

1. Blodgett, S.L., Green, L., O'Connor, B.: Demographic dialectal variation in social media: A case study of African-American English. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1119–1130. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1120, https://aclanthology.org/D16-1120
2. Chen, R.J., Wang, J.J., Williamson, D.F., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., Mahmood, F.: Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering **7**(6), 719–742 (2023)
3. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media **11**(1), 512–515 (May 2017). https://doi.org/10.1609/icwsm.v11i1.14955
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
6. Ethayarajh, K., Choi, Y., Swayamdipta, S.: Understanding dataset difficulty with $\mathcal{V}$-usable information. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 5988–6008. PMLR (17–23 Jul 2022)
7. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 259–268. KDD '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2783258.2783311, https://doi.org/10.1145/2783258.2783311
8. Fu, R., Huang, Y., Singh, P.V.: Crowds, lending, machine, and bias. Information Systems Research **32**(1), 72–92 (2021)
9. Gupta, U., Ferber, A.M., Dilkina, B., Ver Steeg, G.: Controllable guarantees for fair outcomes via contrastive information estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 7610–7619 (2021)
10. Gustafson, L., Rolland, C., Ravi, N., Duval, Q., Adcock, A., Fu, C.Y., Hall, M., Ross, C.: Facet: Fairness in computer vision evaluation benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20370–20382 (2023)
11. Hardt, M., Chen, X., Cheng, X., Donini, M., Gelman, J., Gollaprolu, S., He, J., Larroy, P., Liu, X., McCarthy, N., et al.: Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud. In: Proceedings of the 27th

ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 2974–2983 (2021)
12. Hardt, M., Price, E., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016)
13. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
14. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. In: 2021 International Conference on Learning Representations (May 2021), https://www.microsoft.com/en-us/research/publication/deberta-decoding-enhanced-bert-with-disentangled-attention-2/, under review
15. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
16. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
17. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
18. Jurman, G., Riccadonna, S., Furlanello, C.: A comparison of mcc and cen error measures in multi-class prediction. PLOS ONE **7**(8), 1–8 (08 2012). https://doi.org/10.1371/journal.pone.0041882
19. Kizilcec, R.F., Lee, H.: Algorithmic fairness in education. In: Holmes, W., Porayska-Pomsta, K. (eds.) The Ethics of Artificial Intelligence in Education, pp. 174–202. Taylor & Francis (2022)
20. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.703, https://aclanthology.org/2020.acl-main.703
21. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019), https://arxiv.org/abs/1907.11692
22. Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
24. Pessach, D., Shmueli, E.: Algorithmic Fairness, pp. 867–886. Springer International Publishing, Cham (2023). https://doi.org/10.1007/978-3-031-24628-9_37
25. ProPublica: How we analyzed the compas recidivism algorithm. ProPublica (2016)
26. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

27. Shapley, L.: 7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307-317., pp. 69–79. Princeton University Press, Princeton (1997). https://doi.org/doi:10.1515/9781400829156-012
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015). pp. 1–14. Computational and Biological Learning Society (2015)
29. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
30. Vasquez, J., Gitiaux, X., Ortega, C., Rangwala, H.: Faired: A systematic fairness analysis approach applied in a higher educational context. In: LAK22: 12th International Learning Analytics and Knowledge Conference. pp. 271–281 (2022)
31. Wen, M., Bastani, O., Topcu, U.: Algorithms for fairness in sequential decision making. In: International Conference on Artificial Intelligence and Statistics. pp. 1144–1152. PMLR (2021)
32. Xu, Y., Zhao, S., Song, J., Stewart, R., Ermon, S.: A theory of usable information under computational constraints. In: International Conference on Learning Representations (ICLR) (2020)
33. Zeiler, M., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., Hinton, G.: On rectified linear units for speech processing. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3517–3521 (2013). https://doi.org/10.1109/ICASSP.2013.6638312