Measuring Fairness of Text Classifiers via Prediction Sensitivity

Anonymous ACL submission

Abstract

With the rapid growth in language processing applications, fairness has emerged as an important consideration in data-driven solutions. Although various fairness definitions have been explored in the recent literature, there is lack of consensus on which metrics most accurately reflect the fairness of a system. In this work, we propose a new formulation – *accumulated prediction sensitivity*, which measures fairness in machine learning models based on the model's prediction sensitivity to perturbations in input features. The metric attempts to quantify the extent to which a single prediction depends on a protected attribute, where the protected attribute encodes the membership status of an individual in a protected group. We show that the metric can be theoretically linked with a specific notion of group fairness (statistical parity) and individual fairness. It also correlates well with humans' perception of fairness. We conduct experiments on two text classification datasets – Jigsaw Toxicity, and Bias in Bios, and evaluate the correlations between metrics and manual annotations on whether the model produced a fair outcome. We observe that the proposed fairness metric based on prediction sensitivity is statistically significantly more correlated with human annotation than the existing counterfactual fairness metric.

1 Introduction

Ongoing research is increasingly emphasizing the development of methods which detect and mitigate unfair social bias present in machine learning-based language processing models. These methods come under the umbrella of algorithmic fairness which has been quantitatively expressed with numerous definitions (Mehrabi et al., 2019b; Jacobs and Wallach, 2021). These fairness definitions are broadly categorized into two types, i.e, individual fairness and group fairness. Individual fairness (e.g., counter-factual fairness (Kusner et al., 2017)) is aimed at evaluating whether a model gives similar predictions for individuals with similar personal attributes (e.g., age or race). On the other hand, group fairness (e.g., statistical parity (Dwork et al., 2012)) evaluates fairness across cohorts with same protected attributes instead of individuals (Mehrabi et al., 2019b). Although these two broad categories of fairness define valid notions of fairness, human understanding of fairness is also used to measure fairness in machine learning models (Dhamala et al., 2021). Existing studies often consider only one or two these verticals of measuring fairness.

In our work, we propose a formulation based on models sensitivity to input features – the *accumulated prediction sensitivity*, to measure fairness of model predictions. We establish its theoretical relationship with statistical parity (group fairness) and individual fairness (Dwork et al., 2012) metrics. We then demonstrate the correlation between the proposed metric and human perception of fairness using empirical experiments.

Researchers have proposed metrics to quantify fairness based on a model's sensitivity to input features. Specifically, Maughan and Near (2020); Ngong et al. (2020) propose a *prediction sensitivity* metric that attempts to quantify the extent to which a single prediction depends on a protected attribute. The protected attribute encodes the membership status of an individual in a protected group. Prediction sensitivity can be seen as a form of feature attribution, but specialized to the protected attribute. In our work, we extend their concept of prediction sensitivity to propose *accumulated prediction sensitivity*. Akin to the metric proposed by (Maughan and Near, 2020; Ngong et al., 2020), our metric also relies on model output's sensitivity to changes in input features. Our metric generalizes their notion of sensitivity, where the model sensitivity to various input features can be weighted non-uniformly. We show that the formulation follows certain properties for the chosen definitions of group and individual fairness and also present several methodologies to select weights assigned to sensitivity of model's output to input features. For each selection, we present the correlation between the *accumulated prediction sensitivity* and human assessment of the model-output fairness.

We define our metric in the Section 3 and present bounds on it (under settings when a classifier follows the selected group fairness or individual fairness constraints) in Sections 4 and 5, respectively. Next, given that the human perception of fairness is not theoretically defined, we present an empirical study on two text classification tasks in Section 6. We request a group of annotators to annotate whether they think that model output is biased against a specific gender and observe that the proposed metric correlates positively with more biased outcomes. We then observe correlations between our metric and the stated human understanding of fairness. We find that not only the proposed accumulated prediction sensitivity metric correlates positively with human perception of bias, but also beats an existing baseline based on counterfactual fairness.

2 Related Work

Over the past decade multiple efforts have been made on defining, measuring, and mitigating biases in natural language understanding and generation models (Sun et al., 2019; Mehrabi et al., 2019a; Sheng et al., 2021). Dwork et al. (2012) and Kusner et al. (2017) focus on individual fairness and propose novel classification approaches to ensure that a classification decision is fair towards an individual. Another set of works focus on group fairness. Corbett-Davies et al. (2017) present fair classification to ensure population from different race groups receive similar treatment. Hardt et al. (2016) focus on shifting the cost of incorrect classification from disadvantaged groups for group fairness. Zhao and Chang (2020) propose an approach to measure group fairness in local regions. Finally, Kearns et al. (2019) combine the best properties of the group and individual notions of fairness.

Multiple recent works also focus on developing new dataset and associated metrics to capture various types of biases in specific application domains. For example, Dhamala et al. (2021) and Nangia et al. (2020) propose dataset and metrics to measure social biases and stereotypes in language model generations, Bolukbasi et al. (2016); Caliskan et al. (2017); Manzini et al. (2019) define metrics to access gender and race biases in word vector representations, and Wang et al. (2019) define metrics to quantify and mitigate biases in visual recognition task. Ethayarajh (2020) propose Bernstein bounds to represent uncertainty about the bias. Majority of these bias metrics are automatically computed, for example, using a regard classifier (Sheng et al., 2019), sentiment classifier (Dhamala et al., 2021), toxicity classifier (Dixon et al., 2018) or true positive rate difference between privileged and underprivileged groups (De-Arteaga et al., 2019b). A few works additionally validate the alignment of these automatically computed bias metrics with human understanding of biases by collecting annotations of biases on a subset of test data from crowd-workers (Sheng et al., 2019; Dhamala et al., 2021). Blodgett et al. (2021, 2020) discuss the limitations of several these bias datasets and measurements.

However, the majority of existing bias metrics are specific to the type of the model and the application domain used, they may not be tested for correlation with human judgement of biases, and their relationship to existing definitions of fairness has not been explored. Additionally, metrics such as true positive or error difference between groups requires ground truth labels, thereby making their computation in real-time systems difficult. Speicher et al. (2018) have attempted to present unified approach to measuring group and individual fairness via inequality indices, however we note that such metrics are non-trivial to extend to unstructured data such as text. For example, gender information in a text may be subtle (e.g. mention of softball) and it is unclear whether presence of this word should be considered to impact the genderness of the text. *Accumulated prediction sensitivity* metric, presented in this paper, attempts to address all the above limitations of existing bias metrics. We acknowledge that the proposed metric is yet to be associated with other notions of fairness (e.g. preference based notion of fairness (Zafar et al., 2017)).

3 Accumulated Prediction Sensitivity

Below, we define *accumulated prediction sensitivity*, a metric that capture the sensitivity of a model to protected attributes.

Definition 1 (Accumulated Prediction sensitivity)**.**

Let $\boldsymbol{x} \in \boldsymbol{X}$ be a feature vector drawn from the input space $\boldsymbol{X}$. Let $\boldsymbol{w}, \boldsymbol{v}$ be stochastic vectors whose entries are non-negative values that sum to one. Given $\boldsymbol{x}$, let $\boldsymbol{f}$ be a K-class classifier, such that $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), .., f_k(\boldsymbol{x}), .., f_K(\boldsymbol{x})]$ denotes the K-dimensional probability output generated by the classifier. We define accumulated prediction sensitivity P as:

$$P = \boldsymbol{w}^T \boldsymbol{J} \boldsymbol{v}; \quad where \;\; \boldsymbol{J}(k,i) = \left| \frac{\partial f_k(\boldsymbol{x})}{\partial x_i} \right|. \quad (1)$$

$\boldsymbol{J}$ is a matrix such that the $(k,i)^{\text{th}}$ entry is $\left|\frac{\partial f_k(\boldsymbol{x})}{\partial x_i}\right|$, where x_i is the i^{th} entry in $\boldsymbol{x}$. The product $\boldsymbol{w}^T \boldsymbol{J}$ sums the absolute derivatives $|\frac{\partial f_k(\boldsymbol{x})}{\partial x_i}|$ across $f_k, k = 1, .., K$ and returns a vector of summed derivatives with respect to each $x_i \in \boldsymbol{x}$. The product of $\boldsymbol{v}$ with $\boldsymbol{w}^T \boldsymbol{J}$ further averages the derivatives across all the features $x_i \in \boldsymbol{x}$ to yield the scalar P.

The value $\frac{\partial f_k(\boldsymbol{x})}{\partial x_i}$ captures the expected change in model output for the k^{th} class given a perturbation in x_i. If x_i is a protected feature, arguably a smaller value of $\frac{\partial f_k(\boldsymbol{x})}{\partial x_i}$ implies a fairer model; as then the model's outcome does not change sharply with changes in x_i. In order to capture the sensitivity of the model with respect to the protected features, one also needs to choose $\boldsymbol{v}$ judiciously. For example, given the explicit set of protected features in $\boldsymbol{x}$, one can select $\boldsymbol{v}$ such that only entries corresponding those features are assigned a non-zero value, while the rest are set to zero. Given this heuristics, we expect the value P to be smaller for fairer models. In the next sections, we connect the accumulated prediction sensitivity to two known notions of fairness and human perception of fairness. Note that we use the following notation scheme in this paper – bold capital letters for matrices, bold small letters for vectors and un-bolded letters for scalars.

4 Relation to Group Fairness: Statistical Parity

Given a set of protected features (e.g. gender), a model satisfies statistical parity if model outcome is independent of the protected features (we note that identifying protected features may not always be feasible in the real world). We represent the feature vector $\boldsymbol{x} = [\boldsymbol{x}_p, \boldsymbol{x}_l]$, where $\boldsymbol{x}_p$ is the set of protected features and $\boldsymbol{x}_l$ is the remainder. Accordingly, we choose $\boldsymbol{v}$ to be a vector such that the entries that sum $|\frac{\partial f_k(x_p)}{\partial x_i}| \forall x_p \in \boldsymbol{x}_p$ in $\boldsymbol{J}$ are non-zero; and zero otherwise. This choice is intuitive as then we sum the gradients in $\boldsymbol{J}$ that correspond to protected features and measure model's sensitivity to them. The predictor $\boldsymbol{f}(\boldsymbol{x})$ will satisfy statistical parity if $\boldsymbol{f}(\boldsymbol{x}_p, \boldsymbol{x}_l) = \boldsymbol{f}(\boldsymbol{x}'_p, \boldsymbol{x}_l) \forall \boldsymbol{x}_p \neq \boldsymbol{x}'_p$. Given this, we state the following theorem.

Theorem 1. *Given a vector $\boldsymbol{v}$ with non-zero entries corresponding to $\boldsymbol{x}_p$ and zero entries for $\boldsymbol{x}_l$, if the predictor $\boldsymbol{f}(\boldsymbol{x})$ satisfies statistical parity with respect to $\boldsymbol{x}_p$, accumulated prediction sensitivity will be zero.*

Proof: If $\boldsymbol{f}(\boldsymbol{x})$ satisfies statistical parity with respect to $\boldsymbol{x}_p$, the values $\frac{\partial f_k(\boldsymbol{x})}{\partial x_p} \forall x_p \in \boldsymbol{x}_p$ will be all zeros. This is due to the fact that the function $f_k(\boldsymbol{x})$ can not be defined based on entries $x_p \in \boldsymbol{x}_p$ for it to be independent of them. Therefore, for every multiplication in the product $\boldsymbol{J}\boldsymbol{v}$, either the entry $\frac{\partial f_k(\boldsymbol{x})}{\partial x_p}$ will be 0 or the entry in $\boldsymbol{v}$ corresponding to $\boldsymbol{x}_l$ will be 0. Hence, P will be 0.

5 Relation to Individual Fairness

Dwork et al. (2012) state the notion of individual based fairness as: "*We interpret the goal of mapping similar people similarly to mean that the distributions assigned to similar people are similar*". They propose adding a Lipschitz property constraint during the classifier optimization. Given a loss function $\mathcal{L}$ defined to optimize the parameters $\boldsymbol{\theta}$ of the classifier $\boldsymbol{f}(\boldsymbol{x})$, a distance function $d(\boldsymbol{x}, \boldsymbol{x}')$ that computes distance between data-points $\boldsymbol{x}, \boldsymbol{x}'$, another distance function $\mathcal{D}(\boldsymbol{f}(\boldsymbol{x})), \boldsymbol{f}(\boldsymbol{x}'))$ that computes distance between classifier predictions on $\boldsymbol{x}, \boldsymbol{x}'$ and a constant L, Dwork et al. (2012) propose the following constrained optimization.

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \mathcal{L}; \quad \text{such that} \\ &\mathcal{D}(\boldsymbol{f}(\boldsymbol{x})), \boldsymbol{f}(\boldsymbol{x}')) < Ld(\boldsymbol{x}, \boldsymbol{x}'); \forall \boldsymbol{x}, \boldsymbol{x}' \in \boldsymbol{X}. \end{aligned} \quad (2)$$

It is natural to choose an Lp norm (Bourbaki, 1987) for d and $\mathcal{D}$. For a classifier $\boldsymbol{f}$ that is trained with the above constrained optimization and the choice of distance metrics $\mathcal{D}, d$ is an Lp norm, we state the following.

Theorem 2. *If the predictor $\boldsymbol{f}(\boldsymbol{x})$ is trained with the constrained optimization stated in Eq. (2), the accumulated prediction sensitivity will be upper bounded by L.*

Proof: We restate the constraint in Eq. (2) as (note that the inequality sign does not change as distance metrics $\mathcal{D}, d$ are required to be positive for $\boldsymbol{x} \neq \boldsymbol{x}'$)

$$\forall \boldsymbol{x} \neq \boldsymbol{x}', \quad L > \frac{\mathcal{D}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}'))}{d(\boldsymbol{x}, \boldsymbol{x}')}. \qquad (3)$$

Given the inequality holds for any pair of $\boldsymbol{x}, \boldsymbol{x}'$, it must also hold true for an $\boldsymbol{x}'$ of the following choice.

$$\boldsymbol{x}' = \boldsymbol{x} + [0, 0, \Delta x_i, 0, 0];$$

where Δx_i is a scalar perturbation in the i^{th} entry in $\boldsymbol{x}$. For a chosen Lp norm, Eq (3) becomes

$$\begin{aligned} L &> \frac{[\sum_{k=1}^{K} |f_k(\boldsymbol{x}) - f_k(\boldsymbol{x}')|^p]^{\frac{1}{p}}}{|\Delta x_i|} \\ &> \frac{[|f_k(\boldsymbol{x}) - f_k(\boldsymbol{x}')|^p]^{\frac{1}{p}}}{|\Delta x_i|}. \qquad (4) \end{aligned}$$

Since each entry $|f_k(\boldsymbol{x}) - f_k(\boldsymbol{x}')|^p, k = 1, ..K$ is expected to be non-zero and zeroing out all such entries (but one) will yield a lower value than the summation $\sum_{k=1}^{K} |f_k(\boldsymbol{x}) - f_k(\boldsymbol{x}')|^p$. We can re-write Eq. (4) as:

$$\frac{|f_k(\boldsymbol{x}) - f_k(\boldsymbol{x} + [0, 0, \Delta x_i, 0, 0])|}{|\Delta x_i|}.$$

We can further chose Δx_i such that it is small perturbation, leading to the following.

$$\begin{aligned} L > \lim_{\Delta x_i \to 0} & \frac{|f_k(\boldsymbol{x}) - f_k(\boldsymbol{x} + [0, 0, \Delta x_i, 0, 0])|}{|\Delta x_i|} \\ = & \left| \frac{\partial f_k(\boldsymbol{x})}{\partial x_i} \right|. \end{aligned}$$

Therefore, each entry in $\boldsymbol{J}$ is upper bounded by L. As vectors $\boldsymbol{v}, \boldsymbol{w}$ are stochastic and they compute weighted averages of bounded entries in $\boldsymbol{J}$, P (defined in Eq. (1)) must be less than or equal to L.

We also note that as L becomes larger, the constraint in the Eq. (2) becomes looser. Therefore, a higher value of L during optimization is expected to loosen the fairness constraint as well as the bound on fairness sensitivity. This aligns with our intuition of lower values of P for fairer models.

6 Correlations with Human Perception of Fairness

While the conditional statistical parity and individual fairness establish theoretical constraints on the model behaviour (e.g. independence from protected features and similarity in prediction outcomes for similar data-points), humans may carry a different notion of fairness for model outcomes on individual data-points. This notion may be based on their understanding of cultural norms, which in turn effect their decisions in identifying which model outputs could be considered biased. In this section, we present experiments that correlate accumulated prediction sensitivity with human perception of fairness.

6.1 Human Perception of Fairness

Given a data-point $\boldsymbol{x}$ and model prediction $\boldsymbol{f}(\boldsymbol{x})$, we assign one of the K classes to the data-point. In order to evaluate the human perception of fairness on the data-point, we request a group of annotators to evaluate the model prediction (taken as the arg-max of the model output) and assess whether they believe the output is biased. For instance, given the social/cultural norms, a profession classifier assigning a data-point "she worked in a hospital" to nurse instead of doctor can be perceived as biased. To correlate the accumulated prediction sensitivity P with the human understanding of fairness, we conduct experiments on two text classification datasets. We describe the datasets below, followed by our choices for $\boldsymbol{w}$ and $\boldsymbol{v}$.

6.2 Datasets

We experiment with our proposed metric on two classification tasks, i.e, occupation classification on *Bias in Bios* dataset (De-Arteaga et al., 2019a)[1] and toxicity classification with *Jigsaw Toxicity* dataset[2]. We focus on these two datasets as they have been investigated in several previous studies (Pruksachatkun et al., 2021) and have been reported to carry significant presence of bias. *Bias in bios* data (De-Arteaga et al., 2019a) is purposed to train occupation classifier which predicts occupation given the biography of an individual. We split the data to have 107,171 train samples, 71,447 validation samples and 91,917 test examples. For this data, the task classifier is an occupation classification model which is composed of a standard LSTM-based encoder combined with the output layer of 28 nodes, i.e, number of occupation classes. *Jigsaw Toxicity* dataset is commonly used to train

[1]The data is available at https://github.com/microsoft/biosbias

[2]The data is available at https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

toxic classifier which is tasked to predict if an input sentence is toxic or not. This dataset has input sentences as the comments from Wikipedia's talk page edits labeled with the degree of toxicity. We split the dataset such that we have 1,443,900 training, 360,974 validation samples and 97,320 test samples. In this dataset, the task classifier is a binary classifier trained to predict whether a comment is toxic or not. We labeled the samples with >0.5 toxicity score as *toxic* and others as *non-toxic* to train the task classifier. The task classifier trained with *Jigsaw Toxicity* dataset achieved an AUC of 0.957.

6.3 Selecting the vectors $\boldsymbol{w}$

The vector $\boldsymbol{w}$ sums up the absolute partial derivatives of $f_k(\boldsymbol{x})$ with respect to a given feature $x_i, \forall k = 1, .., K$. In our setup, we consider input features to be the word embeddings and the matrix $\boldsymbol{J}$ is computed over the same. Given a D-dimensional word embedding, K classes and N words in $\boldsymbol{x}$, $\boldsymbol{J}$ will be a matrix of size $(K) \times (DN)$. In all our experiments, we choose $\boldsymbol{w}$ to be a uniform vector with entries $1/K$. Such a choice assigns equal weight to the partial derivatives computed over each class. One may chose to put a higher weight on derivatives computed over a specific class, if there is a reason to believe that the accumulated prediction sensitivity should be informed more with respect to that class. For instance, for a classifier that stratifies medical images into various diseases (Agrawal et al., 2019), disparity in model performance with respect to malicious diseases can be considered more costly. Therefore, derivatives for classes that represent more malicious disease can be weighted higher.

6.4 Selecting the vectors $\boldsymbol{v}$

Through the vector $\boldsymbol{v}$, we aim to select words in $\boldsymbol{x}$ that carry gendered information. We use two formulations for the the vector $\boldsymbol{v}$ as discussed below.

6.4.1 Using a list of gendered words

In this setup, we use the set of gendered words from (Bolukbasi et al., 2016) and assign entries in $\boldsymbol{v}$ corresponding to those words as $1/(N_g \times D)$, where N_g is the count of gendered words in the data-point.

6.4.2 Using a Protected Status Model (PSM)

While prior work has used word matching to a pre-defined corpus of tokens describing various demographic cohorts (Bolukbasi et al., 2016), these corpus do not contain words that stereotypically are associated with a particular cohort but may not be explicitly tied to that cohort. For example, the word "volleyball" is associated with females in the analysis presented by (Dinan et al., 2020).

To capture this nuance, we propose using another classifier (that acts on the same dataset as used to train the original classifier, for which we aim to compute P) and using it to identify tokens containing information about the protected attribute (e.g. gender). We discuss the model training below.

Protected Status Model: To extend accumulated prediction sensitivity to settings with no explicit protected attribute, we train a *protected status model* $\boldsymbol{g}$. Given the data-point $\boldsymbol{x}$, goal of the PSM model $\boldsymbol{g}(\boldsymbol{x})$ is to predict the protected attributes. Given a trained $\boldsymbol{g}(\boldsymbol{x})$, we then compute another matrix $\boldsymbol{J}_g$, where the $(j,i)^{\text{th}}$ entry is $|\frac{\partial \boldsymbol{g}_m(\boldsymbol{x})}{\boldsymbol{x}_i}|$ ($\boldsymbol{g}_m$ is the probability outcomes corresponding to the m^{th} protected attribute class; e.g. male in a gender classifier). We then define an entry $v_i \in \boldsymbol{v}$ as $\sum_j \boldsymbol{J}_g(m,i)$ (the vector $\boldsymbol{v}$ is normalized to be stochastic). Intuitively, the sum $\sum_j \boldsymbol{J}_g(m,i)$ captures the model output sensitivity with respect to the input features x_i and is expected to higher if x_i carries more gendered information.

In our experiments, we train separate PSM models for gender sensitivity computation on Bias-in-bios and Jigsaw data-sets, as each data-point in these data-sets is additionally labeled with a binary gender class (male/female)[3]. Gender PSMs predicts the associated gender given the datapoint $\boldsymbol{x}$. Training PSM on the same datasets used to train the task classifier $\boldsymbol{f}$ helps capture the gender stereotypes present in the respective datasets. For instance, in a given dataset, if the word "volleyball" appears more often in the data-points that correspond to the female gender, the gender classifier's sensitivity to this word is expected to be high as the classifier may pay higher emphasis to this word for gender classification. We use the same model architecture as the task classifier models for PSM training. PSM models for gender classification achieve an accuracy of 98.79% and 95.39% for *Bias in bios* and *Jigsaw Toxicity* datasets, respectively. These accuracies are computed over the same train/test split as the task classifier.

[3] We note that this is a limitation of this work as gender can be non-binary.

Individual Fairness Metrics	Bias in Bios		Jigsaw Toxicity	
	Corr.	MI	Corr.	MI
P1 (uniform $\boldsymbol{w}, \boldsymbol{v}$)	0.206	0.013	0.117	0.007
CF (Garg et al., 2019)	0.326	0.025	0.214	0.022
P4 ($\boldsymbol{v}$ set using gendered words)	0.34	0.037	0.227	0.054
P5 ($\boldsymbol{v}$ set using gendered words and embedding vectors)	**0.363**	**0.098**	0.295	0.061
P2 ($\boldsymbol{v}$ set using PSM)	**0.397**	**0.102**	0.358	**0.097**
P3 ($\boldsymbol{v}$ set using PSM and embedding vectors)	**0.441**	**0.105**	**0.374**	**0.101**

Table 1: Point bi-serial correlations (Corr.) and Mutual Information (MI) between different individual fairness metrics with human annotations on Bios in Bias and Jigsaw toxicity datasets. Bold numbers are the correlations where we see statistically significant increase over CF baseline. The metric variants are sorted based on the correlation values. We use the bootstrap method to compute statistical significance (Koehn, 2004) at p-value<0.05.

6.4.3 Using Word Embedding Vectors

In addition to using the list of gendered words and PSM, we also test with a setting where we multiply the word embedding vectors to the proposed formulations of $\boldsymbol{v}$. We stack the word embedding vectors for each word $x_i \in \boldsymbol{x}$ to obtain a vector of embeddings $\boldsymbol{e}_i$. We perform an element-wise multiplication of the embedding vectors $\boldsymbol{e}_i$ with the vector with entries $1/(N_g \times D)$ for gendered words or $\sum_j \boldsymbol{J}_g(j,i)$ obtained using PSM. This choice is motivated based upon the findings in (Han et al., 2020). They leverage the magnitude of embedding vectors in determining saliency of the input words for the classification task at hand. Their proposed methodology computes saliency maps over the features $x_i \in \boldsymbol{x}$ by multiplying embedding vectors with partial derivatives of the class probabilities with respect to embedding vectors themselves.

6.5 Fairness Metrics

We experiment with six fairness metrics. Out of the six, one metric is a baseline based on counter-factual fairness and the rest are variants of the accumulated prediction sensitivity P.

Counter-factual Fairness (CF) : We use the counter-factual fairness definition mentioned in Garg et al. (2019) and compute the metric as the difference in model predictions between the original sample $\boldsymbol{f}(\boldsymbol{x})$ and its corresponding counter-factual gendered sample $\boldsymbol{f}(\hat{\boldsymbol{x}})$. We take the L1 norm of the vector $\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\hat{\boldsymbol{x}})$. For example, we take the difference in predictions between the sample "She practices dentistry" and "He practices dentistry", which is the corresponding counter-factual sample. We use the definitional gender token substitutions from Bolukbasi et al. (2016) to create counter-factual samples.

P1: Uniformly weighted prediction sensitivity : In this setting, the values of $\boldsymbol{w}$ and $\boldsymbol{v}$ are set to uniform values $\frac{1}{K}$ and $\frac{1}{DN}$, respectively. This is a weak baseline as the choice of $\boldsymbol{v}$ does not provide any information regarding the gender-ness of the input words.

P2: Weighted Prediction Sensitivity based on PSM : In this setting, $\boldsymbol{w}$ is chosen to be a uniform vector, while $\boldsymbol{v}$ is chosen based on the PSM model.

P3: Weighted Prediction sensitivity + Embedding weights : In this setting, $\boldsymbol{v}$ is chosen based on the PSM model (akin to the metric in P2) which is further multiplied element-wise with the word embedding vectors.

P4: Hard gender weights based Prediction sensitivity : In this metric, we use the list of gendered words described in section 6.4.1 to determine $\boldsymbol{v}$. The value of entries in $\boldsymbol{v}$ is set to $\frac{1}{DN_g}$.

P5: Hard gender weights based prediction sensitivity + Embeddings: This setting is same as above, except entries in $\boldsymbol{v}$ are further multiplied element-wise with the word embedding vectors.

6.6 Evaluation

To evaluate whether the proposed prediction sensitivity correlates with human perception of fairness, we collect annotations from crowd workers using the Amazon Mechanical Turk platform. Crowd workers are asked to annotate if a model prediction appears to be a biased prediction or not. For *Bias in Bios* dataset, each sample presented to the annotators has the biography and occupation predicted by the model. We collect annotations on a random sample of the test set. For each biography and a predicted occupation, we ask annotators to label if the prediction is indicative of bias or if it is unbiased. Bias refers to a situation where an occupation is incorrectly predicted based on the gender associated with the biography. For instance, if the input biography is "she studied at Harvard Medical School and practices dentistry." and is

Model	Heat Map
Examples from the Bias in Bios dataset	
TC	And she serves on the executive board of san francisco bay area physicians for social responsibility.
PSM	And she serves on the executive board of san francisco bay area physicians for social responsibility.
TC	he obtained his master's in architecture from the university of tehran and phd in architecture and landscape history at georgia institute of technology
PSM	he obtained his master's in architecture from the university of tehran and phd in architecture and landscape history at georgia institute of technology
Example from the Jigsaw toxicity dataset	
TC	gee maybe she shouldn't have cheated on her dead husband. how about that? nasty woman, it seems like.
PSM	gee maybe she shouldn't have cheated on her dead husband. how about that? nasty woman, it seems like.

Table 2: Color coded representations for the vectors $\boldsymbol{w}^T J$ (top entry in each row) and $\boldsymbol{v}$ (bottom entry in each row) per input word x_i. Darker the color, the higher the magnitude of each of these vectors. These vectors are multiplied to compute accumulated prediction sensitivity. TC: task classifier, PSM: Protected Status Model.

predicted as nurse, then we call this prediction biased since the biography fits better for a doctor. In case of unbiased predictions, the prediction is not expected to be influenced by the gender content in the biography.

Figure 1 presents a sample of examples provided to the annotators for annotating the *Bias in bios* dataset. Each page in the annotation task consisted of ten biography-profession pairs. We collect annotations for each biography-profession pair from at least three annotators and pick the label with majority vote. Similarly for *Jigsaw Toxicity* dataset, each sample presented to the annotators contains the text and associated toxicity predicted by the model. We restrict the set of annotators to be master annotators and the location of annotators to be Unites States. Based on the initial pilot studies conducted in the Amazon Mechanical Turk platform, we setup a payment rate to ensure a fair compensation of at least $15\$/hour$ for all annotators that work at an average pace.

We annotated 900 test data-points from each dataset. We note that these test data-points were misclassified by the classifiers $\boldsymbol{f}$ trained for each dataset. While such a sampling may not conform to the true distribution of biased/unbiased model outcomes on the overall test set, we expect to get more biased samples amongst the misclassified samples. The distribution between biased and unbiased outputs was about 55:45 for *Bias in Bios* and 50:50 for *Jigsaw Toxicity*. For the *Bias in Bios* and *Jigsaw Toxicity* datsets, we obtained a Fliess' kappa of 0.43 and 0.47, respectively, amongst the three annotators. This is considered a moderate level of agreement, which we believe is expected for an relatively ambiguous task to identify model outcomes influenced by gender. We compute mutual information and bi-serial correlations as the primary measures of association between the human annotations and the *accumulated model sensitivity*.

7 Results

Table 1 lists the bi-serial correlations and mutual information between manual annotations and the different fairness metrics. First, we observe that correlations of the baseline with human judgement are mediocre (0.326 and 0.214) compared to the human judgement. We attributed this to the fact that the metric attempts to quantify a fairly subjective assessment of bias that may have different interpretation (as also pointed out by the moderate level of annotation agreement across annotators). However, the proposed variants of P have stronger correlations compared to the counter-factual baseline (except the method P1). As expected, we see the smallest correlation for P1, since this metric does not account for gender-ness in $\boldsymbol{v}$. However, metrics that determine $\boldsymbol{v}$ based on PSM prediction sensitivity and gendered words get higher correlations over P1 and the CF baseline. Variant of P with $\boldsymbol{v}$ informed using the embedding vectors further lead to improved correlations. We also observe weaker statistical significance in the case of *Jigsaw Toxicity* due to a weaker PSM. We attribute this to the noise present in gender annotations for *Jigsaw Toxicity* dataset. Hence, the performance of PSM in predicting the protected status is crucial for accurately measuring fairness.

7.1 Discussion

In order to further analyse the effect of PSM, we look into heat-maps capturing $\boldsymbol{w}^T \boldsymbol{J}$ and $\boldsymbol{v}$ separately. As a reminder, the first quantity captures the weighted average of partial derivatives of class

Examples of unbiased samples:
The predicted profession is unrelated to gender stereotypes about professions.
1. Bio: she received a master's degree in computer science from the university of north Carolina at chapel hill. Predicted profession: **computer scientist**
2. Bio: he received a master's degree in computer science from the university of north Carolina at chapel hill. Predicted profession: **computer scientist**

Examples of biased samples:
Strongly biased predictions are based on associating a specific gender to a specific profession even when there are evidences against it in the biography.
1. Bio: Mary has 25 years of experience in data analytics, business intelligence and information governance with fortune 100 companies. **Predicted profession: nurse**
2. Bio: He achieved a masters degree in nursing from the university of north Carolina at chapel hill. **Predicted profession: computer scientist**

Figure 1: Examples of biased/unbiased outcomes shown to the M-turk annotators

probabilites with respect to the input features, while the second quantity computes the weights assigned to sum up the aforementioned averages. Table 2 shows while $\boldsymbol{v}$ mostly captures gendered words such as “she”, “her” and “woman”, it also captures words such as “social”, “architecture” and “cheated” to carry more gendered information compared to other words. While these words conventionally are not gendered, for the datasets at hand, they seem to provide information whether the input data-point belongs to male/female gender. We also note that $\boldsymbol{w}^T \boldsymbol{J}$ weighs on occupation specific tokens such as "physician", "executive", etc.

This finding supports our motivations to compute $\boldsymbol{v}$ based on PSM and capturing feature attributions assigned to tokens that are implicitly related to a specific gender (instead of the definitional gender tokens only). Hence, by incorporating PSM in computing P, we can capture bias present in non-trivial gendered tokens.

8 Conclusion

Evaluating fairness is a challenging task as it requires selecting a notion of fairness (e.g. group or individual fairness) and then identifying metrics that can capture these notions of fairness while evaluating a classifier. Additionally, certain notions of fairness may not be well defined and can change based upon social norms (e.g. “volleyball” being closely associated with females); that may seep into the dataset at hand. In this work, we define an accumulated prediction sensitivity metric that relies on the partial derivatives of model's class probabilities with respect to input features. We establish properties of this metric with respect to the three verticals of fairness metrics: group, individual and human-perception based. We provide bounds on the metric's value when a predictor is expected to carry statistical parity or is trained with individual fairness. We also evaluate this metric with fairness as perceived through human evaluation of model outputs. We test variants of the proposed metric against an existing baseline derived from counter-factual fairness and observe better mutual information and correlation. Specifically, a variant of the metric that relies on a Protected Status Model (that identifies tokens that carry gender information but may not conventionally be considered gendered) yields the best correlation with the human evaluation.

In the future, one can associate the proposed formulation with other categories of group and individual fairness (Mehrabi et al., 2019a). We also aim to test the metric on other datasets with other protected attributes (e.g. race, nationality). Finally, we can compare the metric across these datasets to compare trends across protected groups.

9 Broader Impact

This work can be used to evaluate bias in models, and thus used to evaluate models serving human consumers. As with all metrics, the metric does not capture all notions of bias, and thus should not be the only consideration for serving models. While this is a valid risk, this is one that is not specific to prediction sensitivity. Good use of this metric requires users to be cognizant of these strengths and weaknesses. We also note that the metric requires defining protected attributes (e.g. gender) and our work carries the limitation that the selected datasets contain binary gender annotations. Defining protected attributes may not always be possible and when possible, the protected attribute classes may not be comprehensive.

References

Taruna Agrawal, Rahul Gupta, and Shrikanth Narayanan. 2019. On evaluating cnn representations for low resource medical image classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1363–1367. IEEE.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Nicolas Bourbaki. 1987. Topological vector spaces, elements of mathematics.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806.

Maria De-Arteaga, Alexey Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, Sahin Cem Geyik, K. Kenthapadi, and A. Kalai. 2019a. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019b. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Kawin Ethayarajh. 2020. Is your classifier actually biased? measuring fairness under uncertainty with bernstein bounds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2914–2919.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. *arXiv preprint arXiv:2005.06676*.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. FAccT ’21, page 375–385, New York, NY, USA. Association for Computing Machinery.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *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)*, pages 615–621.

Krystal Maughan and Joseph P Near. 2020. Towards a measure of individual fairness for deep learning. *arXiv preprint arXiv:2009.13650*.

Ninareh Mehrabi, Fred Morstatter, N. Saxena, Kristina Lerman, and A. Galstyan. 2019a. A survey on bias and fairness in machine learning. *ArXiv*, abs/1908.09635.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019b. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

Ivoline C Ngong, Krystal Maughan, and Joseph P Near. 2020. Towards auditability for fairness in deep learning. *arXiv preprint arXiv:2012.00106*.

Yada Pruksachatkun, Satyapriya Krishna, Jwala Dhamala, Rahul Gupta, and Kai-Wei Chang. 2021. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. *arXiv preprint arXiv:2106.10826*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3398–3403.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *ACL*.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P Gummadi, and Adrian Weller. 2017. From parity to preference-based notions of fairness in classification. *arXiv preprint arXiv:1707.00010*.

Jieyu Zhao and Kai-Wei Chang. 2020. LOGAN: Local group bias detection by clustering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1968–1977.