Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics

Paula Czarnowska[♠]
University of Cambridge, UK
pjc211@cam.ac.uk

Yogarshi Vyas
Amazon AI, USA
yogarshi@amazon.com

Kashif Shah
Amazon AI, USA
shahkas@amazon.com

Abstract

Measuring bias is key for better understanding and addressing unfairness in NLP/ML models. This is often done via fairness metrics, which quantify the differences in a model's behaviour across a range of demographic groups. In this work, we shed more light on the differences and similarities between the fairness metrics used in NLP. First, we unify a broad range of existing metrics under three generalized fairness metrics, revealing the connections between them. Next, we carry out an extensive empirical comparison of existing metrics and demonstrate that the observed differences in bias measurement can be systematically explained via differences in parameter choices for our generalized metrics.

1 Introduction

The prevalence of unintended social biases in NLP models has been recently identified as a major concern for the field. A number of papers have published evidence of uneven treatment of different demographics (Dixon et al., 2018; Zhao et al., 2018; Rudinger et al., 2018; Garg et al., 2019; Borkan et al., 2019; Stanovsky et al., 2019; Gonen and Webster, 2020; Huang et al., 2020a; Nangia et al., 2020), which can reportedly cause a variety of serious harms, like unfair allocation of opportunities or unfavorable representation of particular social groups (Blodgett et al., 2020).

Measuring bias in NLP models is key for better understanding and addressing unfairness. This is often done via **fairness metrics**, which quantify the differences in a model's behavior across a range of social groups. The community has proposed a multitude of such metrics (Dixon et al., 2018; Garg et al., 2019; Huang et al., 2020a; Borkan et al., 2019; Gaut et al., 2020). In this paper, we aim to shed more light on how those varied means of quantifying bias differ and what facets of bias they capture. Developing such understanding is crucial for drawing reliable conclusions and actionable recommendations regarding bias. We focus on bias measurement for downstream tasks, as Goldfarb-Tarrant et al. (2021) have recently shown that there is no reliable correlation between bias measured intrinsically on, for example, word embeddings, and bias measured extrinsically on a downstream task. We narrow down the scope of this paper to tasks that do not involve prediction of a sensitive attribute.

We survey 146 papers on social bias in NLP and unify the multitude of disparate metrics we find under three **generalized fairness metrics**. Through this unification we reveal the key connections between a wide range of existing metrics—we show that they are simply *different parametrizations* of our generalized metrics. Next, we empirically investigate the role of different metrics in detecting the systemic differences in performance for different demographic groups, namely, differences in *quality of service* (Jacobs et al., 2020). We experiment on three transformer-based models—two models for sentiment analysis and one for named entity recognition (NER)—which we evaluate for fairness with respect to seven different sensitive attributes, qualified for protection under the United States federal anti-discrimination law:[1] *Gender*, *Sexual Orientation*, *Religion*, *Nationality*, *Race*, *Age*, and *Disability*. Our results highlight the differences in bias measurements across the metrics and we discuss how these variations can be systematically explained via different parameter choices of our generalized metrics. Our proposed unification and observations can

[♠] Work done during an internship at Amazon AI.

[1] https://www.ftc.gov/site-information/no-fear-act/protections-against-discrimination.

Transactions of the Association for Computational Linguistics, vol. 9, pp. 1249–1267, 2021. https://doi.org/10.1162/tacl_a_00425
Action Editor: Dirk Hovy. Submission batch: 2/2021; Revision batch: 5/2021; Published 11/2021.

guide decisions about which metrics (and parameters) to use, allowing researchers to focus on the pressing matter of bias mitigation, rather than reinventing parametric variants of the same metrics. While we focus our experiments on English, the metrics we study are language-agnostic and our methodology can be trivially applied to other languages.

We release our code with implementations of all metrics discussed in this paper.[2] Our implementation mirrors our generalized formulation (Section 3), which simplifies the creation of new metrics. We build our code on top of CHECKLIST[3] (Ribeiro et al., 2020), making it compatible with the CHECKLIST testing functionalities; that is, one can evaluate the model using the fairness metrics, as well as the CHECKLIST-style tests, like *invariance*, under a single **bias evaluation framework**.

2 Background

2.1 Terminology

We use the term **sensitive attribute** to refer to a category by which people are qualified for protection (e.g., *Religion* or *Gender*). For each sensitive attribute we define a set of **protected groups** T (e.g., for *Gender*, T could be set to {female, male, non-binary}). Next, each protected group can be expressed through one of its **identity terms**, I (e.g., for the protected group *female* those terms could be {woman, female, girl} or a set of typically female names).

2.2 Definitions of Fairness in NLP

The metrics proposed to quantify bias in NLP models across a range of social groups can be categorized based on whether they operationalize notions of group or counterfactual fairness. In this section we give a brief overview of both and encourage the reader to consult Hutchinson and Mitchell (2019) for a broader scope of literature on fairness, dating back to the 1960s.

Group fairness requires parity of some statistical measure across a small set of protected groups (Chouldechova and Roth, 2020). Some prominent examples are *demographic parity* (Dwork et al., 2012), which requires equal positive classification rate across different groups, or *equalized odds* (Hardt et al., 2016) which for binary classification requires equal true positive and false negative rates. In NLP, group fairness metrics are based on performance comparisons for different sets of examples, for example, the comparison of two F1 scores: one for examples mentioning female names and one for examples with male names.

Source Example	Female	Male
I like {person}.	I like Anna. I like Mary. I like Liz.	I like Adam. I like Mark. I like Chris.
{Person} has friends.	Anna has friends. Mary has friends. Liz has friends.	Adam has friends. Mark has friends. Chris has friends.

Table 1: Example of counterfactual fairness data. $T = \{\text{female, male}\}$ and $|I| = 3$ for both groups.

Counterfactual fairness requires parity for two or more versions of an individual, one from the actual world and others from counterfactual worlds in which the individual belongs to a *different protected group*; that is, it requires invariance to the change of the protected group (Kusner et al., 2017). Counterfactual fairness is often viewed as a type of individual fairness, which asks for similar individuals to be treated similarly (Dwork et al., 2012). In NLP, counterfactual fairness metrics are based on comparisons of performance for variations *of the same sentence*, which differ in mentioned identity terms. Such data can be created through perturbing real-world sentences or creating synthetic sentences from templates.

In this work, we require that for each protected group there exists *at least one* sentence variation for every source example (pre-perturbation sentence or a template). In practice, the number of variations for each protected group will depend on the cardinality of I (Table 1). In contrast to most NLP works (Dixon et al., 2018; Garg et al., 2019; Sheng et al., 2020), we allow for a protected group to be realized as more than one identity term. To allow for this, we separate the variations for each source example into $|T|$ *sets*, each of which can be viewed as a separate counterfactual world.

[2] https://github.com/amazon-research/generalized-fairness-metrics.

[3] https://github.com/marcotcr/checklist.

3 Generalized Fairness Metrics

We introduce three **generalized fairness metrics** that are based on different comparisons between protected groups and are model and task agnostic. They are defined in terms of two parameters:

(i) A scoring function, ϕ, which calculates the *score* on a subset of examples. The *score* is a base measurement used to calculate the metric and can be either a scalar or a set (see Table 2 for examples).

(ii) A comparison function, d, which takes a range of different scores—computed for different subsets of examples—and outputs a single scalar value.

Each of the three metrics is conceptually different and is most suitable in different scenarios; the choice of the most appropriate one depends on the scientific question being asked. Through different choices for ϕ and d, we can systematically formulate a broad range of different fairness metrics, targeting different types of questions. We demonstrate this in Section 4 and Table 2, where we show that many metrics from the NLP literature can be viewed as parametrizations of the metrics we propose here. To account for the differences between group and counterfactual fairness (Section 2.2) we define *two different versions of each metric*.

Notation Let $T = \{t_1, t_2, \ldots, t_{|T|}\}$ be a set of all protected groups for a given sensitive attribute, for example, *Gender*, and $\phi(A)$ be the *score* for some set of examples A. This score can be either a set or a scalar, depending on the parametrization of ϕ. For group fairness, let S be the set of all evaluation examples. We denote a subset of examples associated with a protected group t_i as S^{t_i}. For counterfactual fairness, let $X = \{x_1, x_2, \ldots, x_{|X|}\}$ be a set of *source examples*, e.g., sentences pre-perturbation, and $S' = \{S'_1, S'_2, \ldots, S'_{|S|}\}$ be a *set of sets* of evaluation examples, where S'_j is a set of all variations of a source example x_j, i.e., there is a one-to-one correspondence between S' and X. We use $S'^{t_i}_j$ to denote a subset of S'_j associated with a protected group t_i. For example, if $T = \{\text{female}, \text{male}\}$ and the templates were defined as in Table 1, then $S'^{\text{female}}_1 = \{$'I like Anna.', 'I like Mary.', 'I like Liz.'$\}$.

3.1 Pairwise Comparison Metric

Pairwise Comparison Metric (PCM) quantifies how distant, on average, the scores for two different, randomly selected groups are. It is suitable for examining whether and to what extent the chosen protected groups differ from one another. For example, for the sensitive attribute *Disability*, are there any performance differences for cognitive vs mobility vs no disability? We define Group (1) and Counterfactual (2) PCM as follows:

$$\frac{1}{N} \sum_{t_i, t_j \in \binom{T}{2}} d\left(\phi(S^{t_i}), \phi(S^{t_j})\right) \tag{1}$$

$$\frac{1}{|S'|\, N} \sum_{S'_j \in S'} \sum_{t_i, t_k \in \binom{T}{2}} d\left(\phi(S'^{t_i}_j), \phi(S'^{t_k}_j)\right) \tag{2}$$

where N is a normalizing factor, for example, $\binom{|T|}{2}$.

3.2 Background Comparison Metric

Background Comparison Metric (BCM) relies on a comparison between the score for a protected group and the score of its **background**. The definition of the background depends on the task at hand and the investigated question. For example, if the aim is to answer whether the performance of a model for the group differs from the model's *general* performance, the background can be a set of *all* evaluation examples. Alternatively, if the question of interest is whether the groups considered disadvantaged are treated differently than some privileged group, the background can be a set of examples associated with that privileged group. In such a case, T should be narrowed down to the disadvantaged groups only. For counterfactual fairness the background could be the unperturbed example, allowing us to answer whether a model's behavior differs for any of the counterfactual versions of the world. Formally, we define Group (3) and Counterfactual (4) BCM as follows:

$$\frac{1}{N} \sum_{t_i \in T} d\left(\phi(\beta^{t_i, S}), \phi(S^{t_i})\right) \tag{3}$$

$$\frac{1}{|S'|\, N} \sum_{S'_j \in S'} \sum_{t_i \in T} d\left(\phi(\beta^{t_i, S'_j}), \phi(S'^{t_i}_j)\right) \tag{4}$$

where N is a normalizing factor and $\beta^{t_i,S}$ is the background for group t_i for the set of examples S.

Vector-valued BCM In its basic form BCM aggregates the results obtained for different protected groups in order to return a single scalar value. Such aggregation provides a concise signal about the presence and magnitude of bias, but it does so at the cost of losing information. Often, it is important to understand how different protected groups contribute to the resulting outcome. This requires the individual group results not to be accumulated; that is, dropping the $\frac{1}{N}\sum_{t_i \in T}$ term from equations (3) and (4). We call this version of BCM, the vector-valued BCM (VBCM).

3.3 Multi-group Comparison Metric

Multi-group Comparison Metric (MCM) differs from the other two in that the comparison function d takes as arguments the scores for *all protected groups*. This metric can quantify the global effect that a sensitive attribute has on a model's performance; for example, whether the change of *Gender* has any effect on model's scores. It can provide a useful initial insight, but further inspection is required to develop better understanding of the underlying bias, if it is detected.

Group (5) and Counterfactual (6) MCM are defined as:

$$d(\phi(S^{t_1}), \phi(S^{t_2}), \ldots, \phi(S^{t_{|T|}})) \tag{5}$$

$$\frac{1}{|S'|}\sum_{S'_j \in S'} d(\phi(S'^{t_1}_j), \phi(S'^{t_2}_j), \ldots, \phi(S'_j t_{|T|})) \tag{6}$$

4 Classifying Existing Fairness Metrics Within the Generalized Metrics

Table 2 expresses 22 metrics from the literature as instances of our generalized metrics from Section 3. The presented metrics span a number of NLP tasks, including text classification (Dixon et al., 2018; Kiritchenko and Mohammad, 2018; Garg et al., 2019; Borkan et al., 2019; Prabhakaran et al., 2019), relation extraction (Gaut et al., 2020), text generation (Huang et al., 2020a) and dependency parsing (Blodgett et al., 2018).

We arrive at this list by reviewing 146 papers that study bias from the survey of Blodgett et al. (2020) and selecting metrics that meet three criteria: (i) the metric is extrinsic; that is, it is applied to at least one downstream NLP task,[4] (ii) it quantifies the difference in performance across two or more groups, and (iii) it is not based on the *prediction* of a sensitive attribute—metrics based on a model's predictions of sensitive attributes, for example, in image captioning or text generation, constitute a specialized sub-type of fairness metrics. Out of the 26 metrics we find, only four do not fit within our framework: BPSN and BNSP (Borkan et al., 2019), the $\prod$ metric (De-Arteaga et al., 2019), and Perturbation Label Distance (Prabhakaran et al., 2019).[5]

Importantly, many of the metrics we find are PCMs defined for only two protected groups, typically for male and female genders or white and non-white races. Only those that use commutative d can be straightforwardly adjusted to more groups. Those that cannot be adjusted are marked with gray circles in Table 2.

Prediction vs. Probability Based Metrics Beyond the categorization into PCM, BCM, and MCM, as well as group and counterfactual fairness, the metrics can be further categorized into *prediction* or *probability* based. The former calculate the score based on a model's predictions, while the latter use the probabilities assigned to a particular class or label (we found no metrics that make use of both probabilities and predictions). Thirteen out of 16 group fairness metrics are prediction based, while *all* counterfactual metrics are probability based. Since the majority of metrics in Table 2 are defined for *binary* classification, the prevalent scores for prediction based metrics include false positive/negative rates (FPR/FNR) and true positive/negative rates (TPR/TNR). Most of the probability-based metrics are based on the probability associated with the positive/toxic class (class 1 in binary classification). The exception are the metrics of Prabhakaran et al. (2019), which utilize the probability of the *target* class ⑱⑲㉑.

[4] We do not consider language modeling to be a downstream task.

[5] BPSN and BNSP can be defined as Group VBCM if we relax the definition and allow for a separate ϕ function for the background—they require returning different confidence scores for the protected group and the background. The metrics of Prabhakaran et al. (2019) ⑱⑲㉑ originally have not been defined in terms of protected groups. In their paper, T is a set of different names, both male and female.

	Metric	Gen. Metric	$\phi(A)$	d	N	$\beta^{t_i,S}$
			GROUP METRICS			
①	False Positive Equality Difference (FPED)	BCM	False Positive Rate	$\lvert x-y\rvert$	1	S
②	False Negative Equality Difference (FNED)	BCM	False Negative Rate	$\lvert x-y\rvert$	1	S
③	Average Group Fairness (AvgGF)	BCM	$\{f(x,1) \mid x \in A\}$	$W_1(X,Y)$	$\lvert T\rvert$	S
④	FPR Ratio	VBCM	False Positive Rate	$\frac{y}{x}$	–	$S \setminus S^{t_i}$
⑤	Positive Average Equality Gap (PosAvgEG)	VBCM	$\{f(x,1) \mid x \in A, y(x)=1\}$	$\frac{1}{2} - \frac{MWU(X,Y)}{\lvert X\rvert\lvert Y\rvert}$	–	$S \setminus S^{t_i}$
⑥	Negative Average Equality Gap (NegAvgEG)	VBCM	$\{f(x,1) \mid x \in A, y(x)=0\}$	$\frac{1}{2} - \frac{MWU(X,Y)}{\lvert X\rvert\lvert Y\rvert}$	–	$S \setminus S^{t_i}$
⑦	Disparity Score	PCM	F1	$\lvert x-y\rvert$	$\lvert T\rvert$	–
⑧	*TPR Gap	PCM	True Positive Rate	$\lvert x-y\rvert$	$\binom{\lvert T\rvert}{2}$	–
⑨	*TNR Gap	PCM	True Negative Rate	$\lvert x-y\rvert$	$\binom{\lvert T\rvert}{2}$	–
⑩	*Parity Gap	PCM	$\frac{\lvert\{x \mid x\in A, \hat{y}(x)=y(x)\}\rvert}{\lvert A\rvert}$	$\lvert x-y\rvert$	$\binom{\lvert T\rvert}{2}$	–
⑪	*Accuracy Difference	PCM	Accuracy	$x-y$	1	–
⑫	*TPR Difference	PCM	True Positive Rate	$x-y$	1	–
⑬	*F1 Difference	PCM	F1	$x-y$	1	–
⑭	*LAS Difference	PCM	LAS	$x-y$	1	–
⑮	*Recall Difference	PCM	Recall	$x-y$	1	–
⑯	*F1 Ratio	PCM	Recall	$\frac{x}{y}$	1	–
			COUNTERFACTUAL METRICS			
⑰	Counterfactual Token Fairness Gap (CFGap)	BCM	$f(x,1),\ A=\{x\}$	$\lvert x-y\rvert$	$\lvert T\rvert$	$\{x_j\}$
⑱	Perturbation Score Sensitivity (PertSS)	VBCM	$f(x,\ y(x)),\ A=\{x\}$	$\lvert x-y\rvert$	$\lvert T\rvert$	$\{x_j\}$
⑲	Perturbation Score Deviation (PertSD)	MCM	$f(x,\ y(x)),\ A=\{x\}$	$\text{std}(X)$	–	–
⑳	Perturbation Score Range (PertSR)	MCM	$f(x,\ y(x)),\ A=\{x\}$	$\max(X)-\min(X)$	–	–
㉑	Average Individual Fairness (AvgIF)	PCM	$\{f(x,1) \mid x \in A\}$	$W_1(X,Y)$	$\binom{\lvert T\rvert}{2}$	–
㉒	*Average Score Difference	PCM	$\text{mean}(\{f(x,1) \mid x \in A\})$	$x-y$	$\binom{\lvert T\rvert}{2}$	–

Table 2: Existing fairness metrics and how they fit in our generalized metrics. $f(x,c)$, $y(x)$ and $\hat{y}(x)$ are the probability associated with a class c, the gold class and the predicted class for example x, respectively. MWU is the Mann-Whitney U test statistic and W_1 is the Wasserstein-1 distance between the distributions of X and Y. Metrics marked with * have been defined in the context of only two protected groups and do not define the normalizing factor. The metrics associated with gray circles cannot be applied to more than two groups (see Section 4). ① ② (Dixon et al., 2018), ③ ㉑ (Huang et al., 2020a), ④ (Beutel et al., 2019), ⑤ ⑥ (Borkan et al., 2019), ⑦ (Gaut et al., 2020), ⑧ (Beutel et al., 2017; Prost et al., 2019), ⑨ (Prost et al., 2019), ⑩ (Beutel et al., 2017), ⑪ (Blodgett and O'Connor, 2017; Bhaskaran and Bhallamudi, 2019), ⑫ (De-Arteaga et al., 2019), ⑬ (Stanovsky et al., 2019; Saunders and Byrne, 2020), ⑭ (Blodgett et al., 2018), ⑮ (Bamman et al., 2019), ⑯ (Webster et al., 2018), ⑰ (Garg et al., 2019), ⑱ ⑲ ⑳ (Prabhakaran et al., 2019), ㉒ (Kiritchenko and Mohammad, 2018; Popović et al., 2020).

Choice of ϕ and d For scalar-valued ϕ the most common bivariate comparison function is the (absolute) difference between two scores. As outliers, Beutel et al. (2019) ④ use the ratio of the group score to the background score and Webster et al. (2018) ⑯ use the ratio between the first and the second group. Prabhakaran et al.'s (2019) MCM metrics use multivariate d. Their Perturbation Score Deviation metric ⑲ uses the standard deviation of the scores, while their Perturbation

Score Range metric ⑳ uses the range of the scores (difference between the maximum and minimum score). For set-valued ϕ, Huang et al. (2020a) choose Wasserstein-1 distance (Jiang et al., 2020) ③ ㉑, while Borkan et al. (2019) define their comparison function using the Mann-Whitney U test statistic (Mann and Whitney, 1947).

5 Experimental Details

Having introduced our generalized framework and classified the existing metrics, we now *empirically* investigate their role in detecting the systemic performance difference across the demographic groups. We first discuss the relevant experimental details before presenting our results and analyses (Section 6).

Models We experiment on three RoBERTa (Liu et al., 2019) based models:[6] (i) a binary classifier trained on SemEval-2018 valence classification shared task data (Mohammad et al., 2018) processed for binary classification (SemEval-2)[7] (ii) a 3-class classifier trained on SemEval-3, and (iii) a NER model trained on the CoNLL 2003 Shared Task data (Tjong Kim Sang and De Meulder, 2003) which uses RoBERTa to encode a text sequence and a Conditional Random Field (Lafferty et al., 2001) to predict the tags. In NER experiments we use the BILOU labeling scheme (Ratinov and Roth, 2009) and, for the probability-based metrics, we use the probabilities from the encoder's output. Table 5 reports the performance on the official dev splits for the datasets the models were trained on.

Evaluation Data For classification, we experiment on seven sensitive attributes, and for each attribute we devise a number of protected groups (Table 3).[8] We analyze bias within each attribute

Sensitive attribute	Protected groups (T)
Gender	aab, female, male, cis, many-genders, no-gender, non-binary, trans
Sexual Orientation	asexual, homosexual, heterosexual, bisexual, other
Religion	atheism, buddhism, baha'i-faith, christianity, hinduism, islam, judaism, mormonism, sikhism, taoism
Race	african american, american indian, asian, hispanic, pacific islander, white
Age	young, adult, old
Disability	cerebral palsy, chronic illness, cognitive, down syndrome, epilepsy, hearing, mental health, mobility, physical, short stature, sight, unspecified, without
Nationality	We define 6 groups by categorizing countries based on their GDP.

Table 3: The list of sensitive attributes and protected groups used in our experiments.

Protected group	Identity terms (I)
aab	AMAB, AFAB, DFAB, DMAB, female-assigned, male-assigned
female	female (adj), female (n), woman
male	male (adj), male (n), man
many genders	ambigender, ambigendered, androgynous, bigender, bigendered, intersex, intersexual, pangender, pangendered, polygender, androgyne, hermaphrodite
no-gender	agender, agendered, genderless

Table 4: Examples of explicit identity terms for the selected protected groups of *Gender*.

independently and focus on *explicit* mentions of each identity. This is reflected in our choice of identity terms, which we have gathered from Wikipedia, Wiktionary, as well as Dixon et al. (2018) and Hutchinson et al. (2020) (see Table 4 for an example). Additionally, for the *Gender* attribute we also investigate implicit mentions—female and male groups represented with names typically associated with these genders. We experiment on synthetic data created using handcrafted templates, as is common in the literature (Dixon et al., 2018; Kiritchenko and Mohammad, 2018; Kurita et al., 2019; Huang et al., 2020a). For each sensitive attribute we use 60 templates

[6] Our preliminary experiments also used models based on Electra (Clark et al., 2020) as well as those trained on SST-2 and SST-3 datasets (Socher et al., 2013). For all models, we observed similar trends in differences between the metrics. Due to space constraints we omit those results and leave a detailed cross-model bias comparison for future work.

[7] We process the SemEval data as is commonly done for SST (Socher et al., 2013). For binary classification, we filter out the neutral class and compress the multiple fine-grained positive/negative classes into a single positive/negative class. For 3-class classification we do not filter out the neutral class.

[8] For Disability and Race we used the groups from Hutchinson et al. (2020) and from the Racial and Ethnic Categories and Definitions for NIH Diversity Programs (`https://grants.nih.gov/grants/guide/notice-files/not-od-15-089.html`), respectively. For the remaining attributes, we rely on Wikipedia and Wiktionary, among other sources.

SemEval-2	SemEval-3	CoNLL 2003
Accuracy		F1
0.90	0.73	0.94

Table 5: RoBERTa performance on the official development splits for the three tasks.

with balanced classes: 20 negative, 20 neutral and 20 positive templates. For each attribute we use 30 generic templates—with adjective and noun phrase slots to be filled with identity terms—and 30 attribute-specific templates.[9] In Table 6 we present examples of both generic templates and attribute-specific templates for *Nationality*. Note that the slots of generic templates are designed to be filled with terms that explicitly reference an identity (Table 4), and are unsuitable for experiments on female/male names. For this reason, for names we design additional 30 name-specific templates (60 in total). We present examples of those templates in Table 6.

For NER, we only experiment on *Nationality* and generate the evaluation data from 22 templates with a missing {*country*} slot for which we manually assign a BILOU tag to each token. The {*country*} slot is initially labeled as U-LOC and is later automatically adjusted to a *sequence* of labels if a country name filling the slot spans more than one token, for example, B-LOC L-LOC for *New Zeland*.

Metrics We experiment on metrics that support more than two protected groups (i.e., the *white-circled* metrics in Table 2). As described in Section 2.2, for each source example we allow for a number of variations for each group. Hence, for counterfactual metrics that require only one example per group (all counterfactual metrics but Average Individual Fairness (21)) we evaluate on the $|T|$-ary Cartesian products over the sets of variations for all groups. For groups with large $|I|$ we sample 100 elements from the Cartesian product, without replacement. We convert Counterfactual Token Fairness Gap (17) and Perturbation Score Sensitivity (18) into PCMs because for templated-data there is no single *real-world* example.

Average Group Fairness (3), Counterfacutal Token Fairnes Gap (17), and Average Individual Fairness (21) calculate bias based on the probability

[9]The templates can be found with the code.

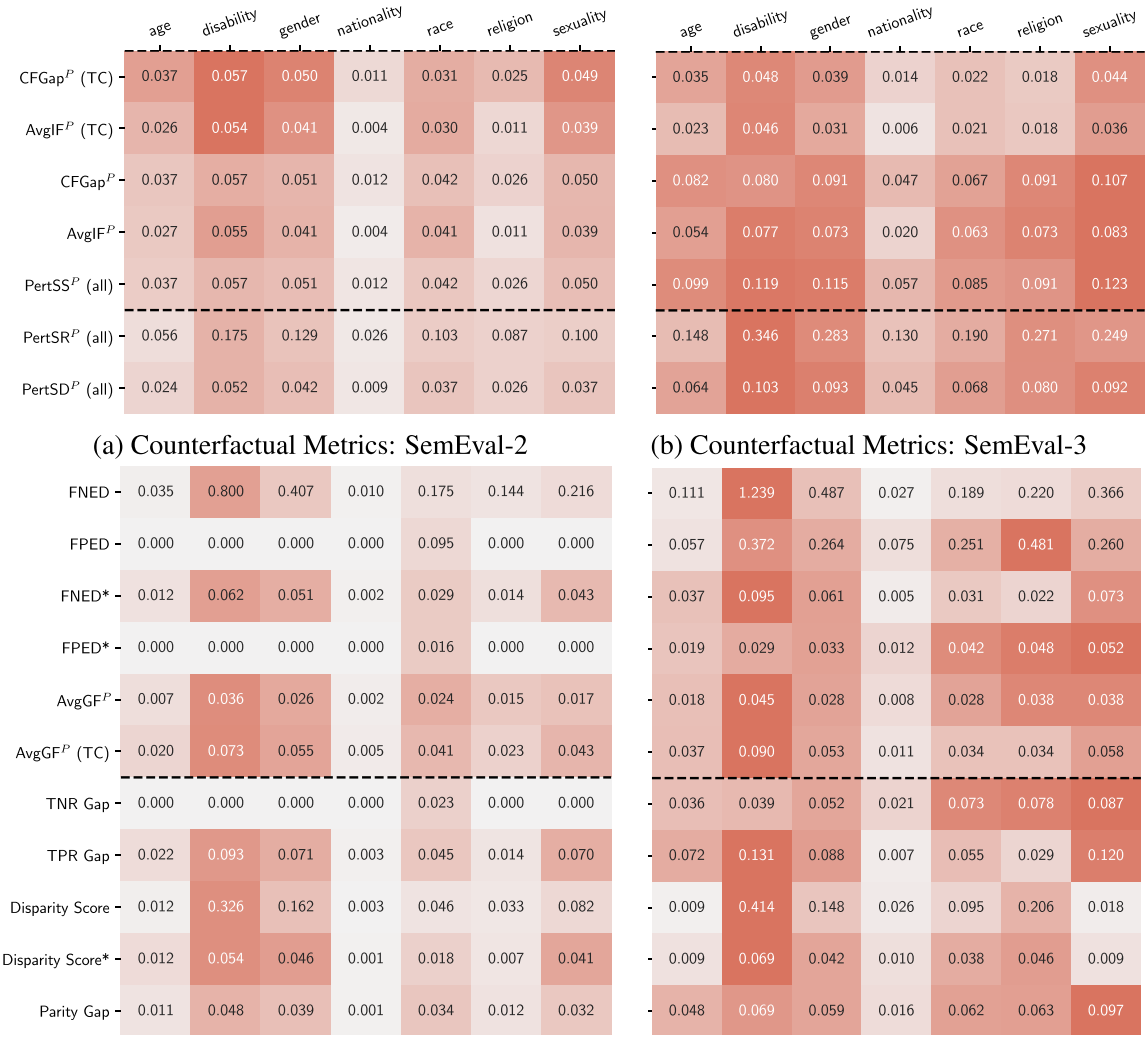
	Generic templates
POS	I liked that {identity_adj} writer.
POS	As {a:identity_np}, I feel hopeful.
POS	This summit is a wonderful event to celebrate {identity_adj} folk.
NEUT	I recently found out that Alex is {a:identity_np}.
NEUT	The main character was {identity_adj}.
NEUT	I watched a documentary about {identity_adj} writers.
NEG	What a sad day to be {identity_adj}.
NEG	It's scary how they treat {identity_adj} people.
NEG	Being {identity_adj} makes life really hard.
	Nationality templates
POS	The next ACL will be in the beautiful {country}!
POS	The outstanding festival celebrated the best artists from {country}.
POS	{Country} has done a great job in fighting economic inequality.
NEUT	Alex always wanted to live in {country}.
NEUT	Next speaker travelled all the way from {country}.
NEUT	{Country} is UK's most important trading partner.
NEG	{Country} is the worst holiday destination.
NEG	{Country_adj} Music Awards were a disaster.
NEG	The designs of this {country_adj} based firm are distressing.
	Name templates
POS	{Person} teared up and it was heartwarming.
POS	{Person} feels fulfilled as a parent.
POS	You should read this ground-breaking paper by {person}.
NEUT	As a child, {person} had big dreams.
NEUT	{Person} received their high-school diploma in 1990.
NEUT	{Person} came out with new strategies for managing and financing that program.
NEG	{Person} sadly didn't get the job.
NEG	Nothing special, {person} thought so too.
NEG	It's unfair that {person} got that reputable position without prior experience.

Table 6: Template examples.

of positive/toxic class on *all* examples. We introduce alternative versions of these metrics which calculate bias *only* on examples with gold label c, which we mark with a *(TC)* (for true class) suffix. The original versions target *demographic parity* (Dwork et al., 2012), while the TC versions target *equality of opportunity* (Hardt et al., 2016) and can pinpoint the existence of bias more precisely, as we show later (Section 6).

5.1 Moving Beyond Binary Classification

Fourteen out of 15 *white-circled* metrics from Table 2 are inherently classification metrics, 11 of which are defined exclusively for binary classification. We adapt binary classification metrics



(a) Counterfactual Metrics: SemEval-2

(b) Counterfactual Metrics: SemEval-3

(c) Group Metrics: SemEval-2

(d) Group Metrics: SemEval-3

Figure 1: BCM, PCM, and MCM metrics calculated for different sensitive attributes, for the positive class. Metrics marked with (all) are inherently multiclass and are calculated for all classes. Superscripts ^{P} and * mark the probability-based and correctly normalized metrics, respectively. We row-normalize the heatmap coloring, across the whole figure, using maximum absolute value scaling.

to (i) multiclass classification and (ii) sequence labeling to support a broader range of NLP tasks.

Multiclass Classification Probability-based metrics that use the probability of the target class (⑱ ⑲ ⑳) do not require any adaptations for multiclass classification. For other metrics, we measure bias independently for each class c, using a one-vs-rest strategy for prediction-based metrics and the probability of class c for the scores of probability-based metrics (③ ⑤ ⑥ ⑰ ㉑).

Sequence Labeling We view sequence labeling as a case of multiclass classification, with each token being a separate classification decision. As for multiclass classification, we compute the bias measurements for each class independently. For prediction-based metrics, we use one-vs-rest strategy and base the F1 and FNR scores on exact span matching.[10] For probability-based metrics, *for each token* we accumulate the probability scores for different labels of the same class. For example, with the BILOU labeling scheme, the probabilities for B-PER, I-PER, L-PER, and U-PER are summed to obtain the probability for the class PER. Further, for counterfactual metrics, to account for different identity terms yielding different number of tokens, we average the probability scores for all tokens of multi-token identity terms.

[10]We do not compute FPR based metrics, because false positives are unlikely to occur for our synthetic data and are less meaningful if they occur.

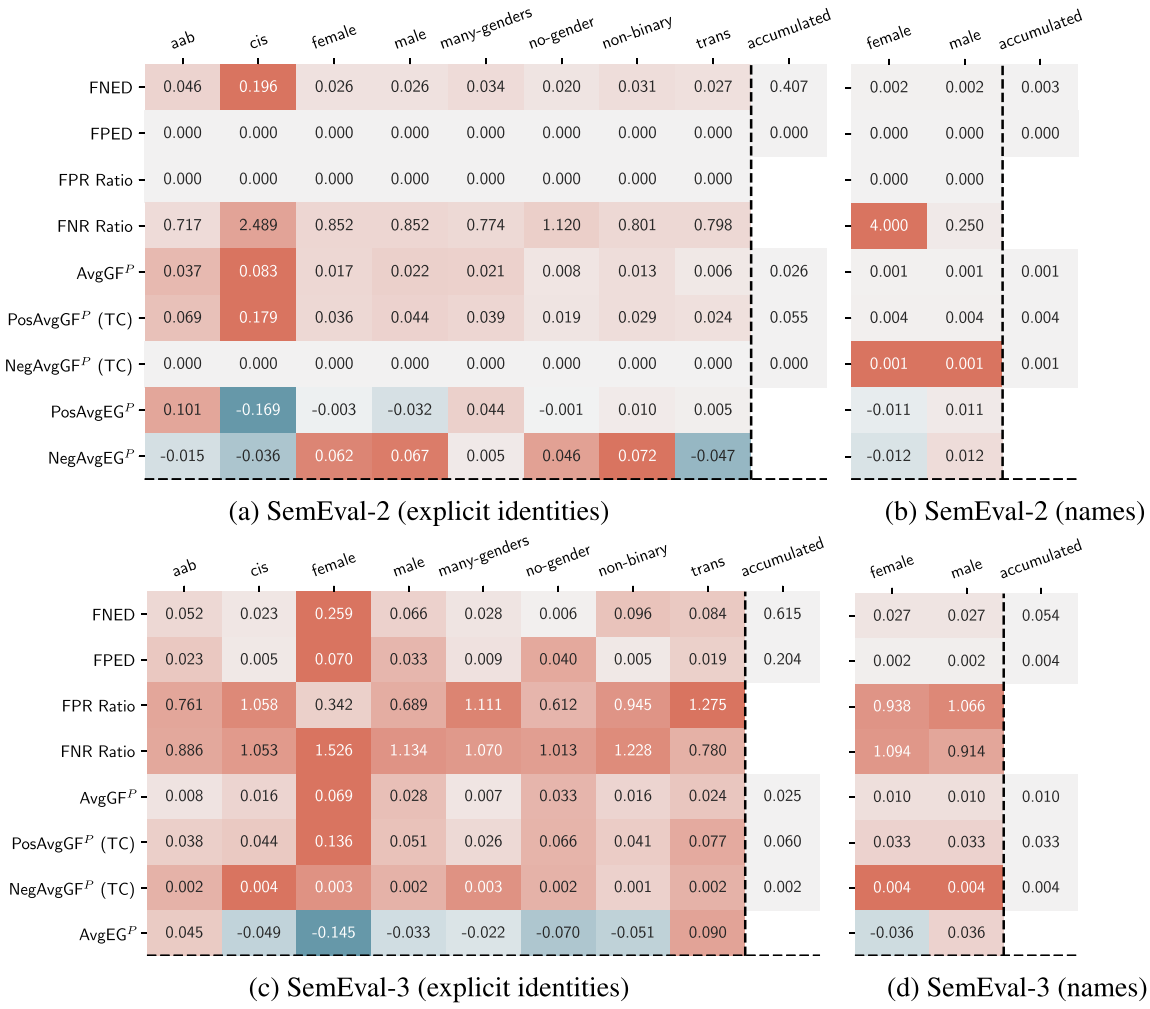


Figure 2: Results for BCM and VBCM metrics on the positive class on *Gender* for explicit (left) and implicit identities, signaled through names (right).

6 Empirical Metric Comparison

Figure 1 shows the results for sentiment analysis for all attributes on BCM, PCM, and MCM metrics. In each table we report the original bias measurements and row-normalize the *heatmap coloring* using maximum absolute value scaling to allow for some cross-metric comparison.[11] Figure 1 gives evidence of unintended bias for most of the attributes we consider, with *Disability* and *Nationality* being the most and least affected attributes, respectively. We highlight that because we evaluate on simple synthetic data in which the expressed sentiment is evident, even small performance differences can be concerning. Figure 1 also gives an initial insight into how the bias measurements vary across the metrics.

In Figure 2 we present the per-group results for VBCM and BCM metrics for the example *Gender* attribute.[12] Similarly, in Figure 3 we show results for NER for the relevant LOC class. The first set of results indicates that the most problematic *Gender* group is *cis*. For NER we observe a big gap in the model's performance between the most affluent countries and countries with lower GDP. In the context of those empirical results we now discuss how different parameter choices affect the observed bias measurement.

Key Role of the Base Measurement Perhaps the most important difference between the metrics lies in the parametrization of the scoring function ϕ. The choice of ϕ determines what

[11] Even after normalization, bias measurements across metrics are not *fully* comparable—different metrics use different base measurements (TPR, TNR, etc.) and hence measure different aspects of bias.

[12] We omit the per-group results for the remaining attributes due to the lack of space. For BCM, we do not include *accumulated* values in the normalization.

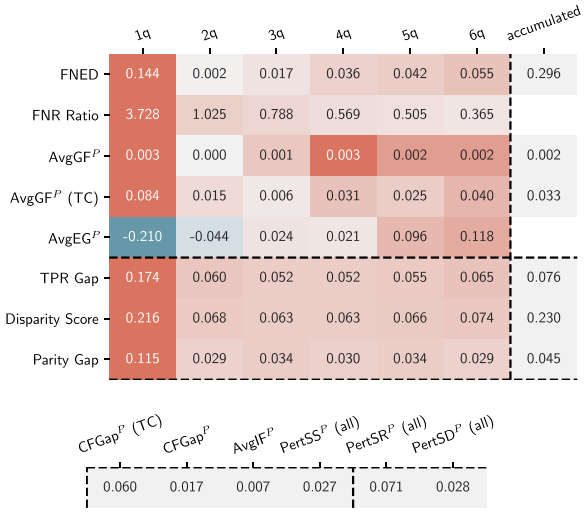


Figure 3: Results for the NER model on *Nationality* attribute for six groups defined by categorizing countries based on their GDP (six quantiles) for the (most relevant) LOC class. We present group metrics at the top and the counterfactual metrics at the bottom. The probability-based metrics not marked with (TC) use probability scores for LOC for *all* tokens, including O; hence they are less meaningful than their TC alternatives.

type and aspect of bias is being measured, making the metrics *conceptually* different. Consider, for example ϕ of Average Group Fairness (3) —$\{f(x,1) \mid x \in A\}$—and Positive Average Equality Gap (5)—$\{f(x,1) \mid x \in A, y(x) = 1\}$. They are both based on the probabilities associated with class 1, but the former is computed on *all* examples in A, while the latter is computed on only those examples that belong to the positive class (i.e., have gold label 1). This difference causes them to measure different types of bias—the first targets *demographic parity*, the second *equality of opportunity*.

Further, consider FPED (1) and FNED (2), which use FPR and FNR for their score, respectively. This difference alone can lead to entirely different results. For example, in Figure 2a FNED reveals prominent bias for the *cis* group while FPED shows none. Taken together, these results signal that the model's behavior for this group *is* notably different from the other groups but this difference manifests itself *only* on the positive examples.

(In)Correct Normalization Next, we highlight the importance of correct normalization. We argue that fairness metrics should be invariant to the number of considered protected groups, otherwise the bias measurements are incomparable and can be misleadingly elevated. The latter is the case for three metrics—FPED (1), FNED (2), and Disparity Score (7). The first two lack any kind of normalization, while Disparity Score is incorrectly normalized—N is set to the number of groups, rather than group pairs. In Figure 1 we present the results on the original versions of those metrics and for their correctly normalized versions, marked with *. The latter result in much lower bias measurements. This is all the more important for FPED and FNED, as they have been very influential, with many works relying *exclusively* on these metrics (Rios, 2020; Huang et al., 2020b; Gencoglu, 2021; Rios and Lwowski, 2020).

Relative vs Absolute Comparison Next, we argue that the results of metrics based on the relative comparison, for example, FPR Ratio (4), can be misleading and hard to interpret if the original scores are not reported. In particular, the relative comparison can amplify bias in cases when both scores are low; in such scenarios even a very small absolute difference can be relatively large. Such amplification is evident in the FNR Ratio metric (FNR equivalent of FPR Ratio) on female vs male names for RoBERTa fine-funed on SemEval-2 (Figure 2b). Similarly, when both scores are very high, the bias can be underestimated—a significant difference between the scores can seem relatively small if both scores are large. Indeed, such effects have also been widely discussed in the context of reporting health risks (Forrow et al., 1992; Stegenga, 2015; Noordzij et al., 2017). In contrast, the results of metrics based on absolute comparison can be meaningfully interpreted, even without the original scores, if the range of the scoring function is known and interpretable (which is the case for all metrics we review).

Importance of Per-Group Results Most group metrics accumulate the results obtained for different groups. Such accumulation leads to diluted bias measurements in situations where the performance differs only for a small proportion of all groups. This is evident in, for example, the per-group NER results for correctly normalized metrics (Figure 3). We emphasize the importance of reporting per-group results whenever possible.

Prediction vs Probability Based In contrast to prediction-based metrics, probability-based

metrics also capture more subtle performance differences that do not lead to different predictions. This difference can be seen, for example, for *aab Gender* group results for SemEval-2 (Figure 2a) and the results for female/male names for SemEval-3 (Figure 2d). We contend that it is beneficial to use both types of metrics to understand the effect of behavior differences on predictions and to allow for detection of more subtle differences.

Signed vs Unsigned Out of the 15 *white-circled* metrics only two are signed; Positive and Negative Average Equality Gap (AvgEG) ⑤⑥. Using at least one signed metric allows for quick identification of the bias direction. For example, results for Average Equality Gap reveal that examples mentioning the *cis Gender* group are considered less positive than examples mentioning other groups and that, for NER, the probability of LOC is *lower* for the richest countries (first and second quantiles have negative signs).

True Class Evaluation We observe that the TC versions of probability-metrics allow for better understanding of bias location, compared with their non-TC alternatives. Consider Average Group Fairness ③ and its TC versions evaluated on the positive class (PosAvgGF) and negative class (NegAvgGF) for binary classification (Figure 2a). The latter two reveal that the differences in behavior apply solely to the positive examples.

6.1 Fairness Metrics vs Significance Tests

Just like fairness metrics, statistical significance tests can also detect the presence of systematic differences in the behavior of a model, and hence are often used as alternative means to quantify bias (Mohammad et al., 2018; Davidson et al., 2019; Zhiltsova et al., 2019). However, in contrast to fairness metrics, significance tests *do not* capture the magnitude of the differences. Rather, they quantify the likelihood of observing given differences under the null hypothesis. This is an important distinction with clear empirical consequences, as even very subtle differences between the scores can be statistically significant.

To demonstrate this, we present p-values for significance tests for which we use the probability of the positive class as a dependent variable (Table 7). Following Kiritchenko and Mohammad (2018), we obtain a single probability score for each template by averaging the results across all identity terms per group. Because we evaluate on synthetic data, which is balanced across all groups, we use the scores for all templates regardless of their gold class. We use the Friedman test for all attributes with more than two protected groups. For *Gender* with male/female names as identity terms we use the Wilcoxon signed-rank test. We observe that, despite the low absolute values of the metrics obtained for the *Nationality* attribute (Figure 1), the behavior of the models across the groups is unlikely to be equal. The same applies to the results for female vs male names for SemEval-3 (Figure 2d). Utilizing a test for statistical significance can capture such nuanced presence of bias.

Attribute	SemEval-2	SemEval-3
Gender (names)	8.72×10^{-1}	3.05×10^{-6}
Gender	1.41×10^{-8}	3.80×10^{-24}
Sexual Orientation	2.76×10^{-9}	9.49×10^{-24}
Religion	1.14×10^{-23}	8.24×10^{-36}
Nationality	1.61×10^{-2}	1.45×10^{-14}
Race	2.18×10^{-5}	8.44×10^{-5}
Age	4.86×10^{-2}	4.81×10^{-8}
Disability	9.67×10^{-31}	2.89×10^{-44}

Table 7: P-values for the Wilcoxon signed-rank test (attribute *Gender, (names)*) and the Friedman test (all other attributes).

Notably, Average Equality Gap metrics ⑤⑥ occupy an atypical middle ground between being a fairness metric and a significance test. In contrast to other metrics from Table 2, they *do not quantify the magnitude* of the differences, but the likelihood of a group being considered less positive than the background.

7 Which Metrics to Choose?

In the previous section we highlighted important differences between the metrics which stem from different parameter choices. In particular, we emphasized the difference between prediction and probability-based metrics, in regards to their *sensitivity* to bias, as well as the conceptual distinction between the fairness metrics and significance tests. We also stressed the importance of correct normalization of metrics and reporting per-group results whenever possible. However, one important question still remains unanswered: Out of the many

different metrics that can be used, which ones are the most appropriate? Unfortunately, there is no easy answer. The choice of the metrics depends on many factors, including the task, the particulars of how and where the system is deployed, as well as the goals of the researcher.

In line with the recommendations of Olteanu et al. (2017) and Blodgett et al. (2020), we assert that fairness metrics need to be grounded in the application domain and carefully matched to the type of studied bias to offer meaningful insights. While we cannot be too prescriptive about the exact metrics to choose, we advise against reporting results for all the metrics presented in this paper. Instead, we suggest a three-step process that helps to narrow down the full range of metrics to those that are the most applicable.

Step 1. Identifying the type of question to ask and choosing the appropriate generalized metric to answer it. As discussed in Section 3, each generalized metric is most suitable in different scenarios; for example, MCM metrics can be used to investigate whether the attribute has any overall effect on the model's performance and (V)BCM allows us to investigate how the performance for particular groups differs with respect to model's general performance.

Step 2. Identifying scoring functions that target the studied type and aspect of bias. At this stage it is important to consider practical consequences behind potential base measurements. For example, for sentiment classification, misclassyfing positive sentences mentioning a specific demographic as negative can be more harmful than misclassyfing negative sentences as positive, as it can perpetuate negative stereotypes. Consequently, the most appropriate ϕ would be based on FNR or the probability of the negative class. In contrast, in the context of convicting low-level crimes, a false positive has more serious practical consequences than a false negative, since it may have a long-term detrimental effect on a person's life. Further, the parametrization of ϕ should be carefully matched to the motivation of the study and the assumed type/conceptualization of bias.

Step 3. Making the remaining parameter choices. In particular, deciding on the comparison function most suitable for the selected ϕ and the targeted bias; for example, absolute difference if ϕ is scalar-valued ϕ or Wasserstein-1 distance for set-valued ϕ.

The above three steps can identify the most relevant metrics, which can be further filtered down to the minimal set sufficient to identify studied bias. To get a complete understanding of a model's (un)fairness, our general suggestion is to consider at least one prediction-based metric and one probability-based metric. Those can be further complemented with a test for statistical significance. Finally, it is essential that the results of each metric are interpreted in the context of the score employed by that metric (see Section 6). It is also universally good practice to report the results from all selected metrics, regardless of whether they do or do not give evidence of bias.

8 Related Work

To our knowledge, we are the first to review and empirically compare fairness metrics used within NLP. Close to our endeavor are surveys that discuss types, sources, and mitigation of bias in NLP or AI in general. Surveys of Mehrabi et al. (2019), Hutchinson and Mitchell (2019), and Chouldechova and Roth (2020) cover a broad scope of literature on algorithmic fairness. Shah et al. (2020) offer both a survey of bias in NLP as well as a conceptual framework for studying bias. Sun et al. (2019) provide a comprehensive overview of addressing gender bias in NLP. There are also many task specific surveys, for example, for language generation (Sheng et al., 2021) or machine translation (Savoldi et al., 2021). Finally, Blodgett et al. (2020) outline a number of methodological issues, such as providing vague motivations, which are common for papers on bias in NLP.

We focus on measuring bias exhibited on classification and sequence labeling downstream tasks. A related line of research measures bias present in sentence or word representations (Bolukbasi et al., 2016; Caliskan et al., 2017; Kurita et al., 2019; Sedoc and Ungar, 2019; Chaloner and Maldonado, 2019; Dev and Phillips, 2019; Gonen and Goldberg, 2019; Hall Maudslay et al., 2019; Liang et al., 2020; Shin et al., 2020; Liang et al., 2020; Papakyriakopoulos et al., 2020). However, such intrinsic metrics have been recently shown not to correlate with application bias (Goldfarb-Tarrant et al., 2021). In yet another

line of research, Badjatiya et al. (2019) detect bias through identifying *bias sensitive words*.

Beyond the fairness metrics and significance tests, some works quantify bias through calculating a standard evaluation metric, for example, F1 or accuracy, or a more elaborate measure *independently* for each protected group or for each split of a challenge dataset (Hovy and Søgaard, 2015; Rudinger et al., 2018; Zhao et al., 2018; Garimella et al., 2019; Sap et al., 2019; Bagdasaryan et al., 2019; Stafanovičs et al., 2020; Tan et al., 2020; Mehrabi et al., 2020; Nadeem et al., 2020; Cao and Daumé III, 2020).

9 Conclusion

We conduct a thorough review of existing fairness metrics and demonstrate that they are simply parametric variants of the three generalized fairness metrics we propose, each suited to a different type of a scientific question. Further, we empirically demonstrate that the differences in parameter choices for our generalized metrics have direct impact on the bias measurement. In light of our results, we provide a range of concrete suggestions to guide NLP practitioners in their metric choices.

We hope that our work will facilitate further research in the bias domain and allow the researchers to direct their efforts towards bias mitigation. Because our framework is language and model agnostic, in the future we plan to experiment on more languages and use our framework as principled means of comparing different models with respect to bias.

Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful comments and suggestions. We also thank the members of Amazon AI for many useful discussions and feedback.

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pages 49–59. https://doi.org/10.1145/3308558.3313504

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, pages 15479–15488.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. 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 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Beutel, J. Chen, Zhe Zhao, and Ed Huai hsin Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017)*.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 453–459. https://doi.org/10.1145/3306618.3314234

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? Occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3809

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, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Su Lin Blodgett and Brendan T. O’Connor. 2017. Racial Disparity in Natural Language Processing: A case study of social media African-

American English. *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017)*.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter universal dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1131

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. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500, San Francisco USA. ACM. https://doi.org/10.1145/3308560.3317593

Aylin Caliskan, Joanna Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186. https://doi.org/10.1126/science.aal4230, PubMed: 28408601

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3804

Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89. https://doi.org/10.1145/3376898

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. *The International Conference on Learning Representations (ICLR)*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3504

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. 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*, FAT* '19, pages 120–128, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3287560.3287572

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 879–887. PMLR.

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*, AIES '18, pages 67–73, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3278721.3278729

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*, ITCS '12, pages 214–226, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/2090236.2090255

Lachlan Forrow, William C. Taylor, and Robert M. Arnold. 1992. Absolutely relative: How research results are summarized can affect treatment decisions. *The American Journal of Medicine*, 92(2):121–124. https://doi.org/10.1016/0002-9343(92)90100-P

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*, AIES '19, pages 219–226, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3306618.3317950

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1339

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.265

Oguzhan Gencoglu. 2021. Cyberbullying detection with fairness constraints. *IEEE Internet Computing*, 25(01):20–29. https://doi.org/10.1109/MIC.2020.3032461

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.150

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.180

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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 5267–5275, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1530

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference*

on Natural Language Processing (Volume 2: Short Papers), pages 483–488, Beijing, China. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020a. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.7

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020b. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58. https://doi.org/10.1145/3287560.3287600

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.487

Abigail Z. Jacobs, Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna Wallach. 2020. The meaning and measurement of bias: Lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 706, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3351095.3375671

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2020. Wasserstein fair classification. Ryan P. Adams and Vibhav Gogate, editors, In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 862–872. PMLR.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-2005

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3823

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.488

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin

Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692. Version 1.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18(1):50–60. https://doi.org/10.1214/aoms/1177730491

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, pages 231–232, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3372923.3404804

Ninareh Mehrabi, Fred Morstatter, N. Saxena, Kristina Lerman, and A. Galstyan. 2019. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635. Version 2.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1001

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *CoRR*, abs/2004.09456. Version 1. https://doi.org/10.18653/v1/2021.acl-long.416

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. 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, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.154

M. Noordzij, M. van Diepen, F. Caskey, and K. Jager. 2017. Relative risk versus absolute risk: One cannot be interpreted without the other: Clinical epidemiology in nephrology. *Nephrology Dialysis Transplantation*, 32:ii13–ii18. https://doi.org/10.1093/ndt/gfw465, PubMed: 28339913

Alexandra Olteanu, Kartik Talamadupula, and Kush R. Varshney. 2017. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 405–406. https://doi.org/10.1145/3091478.3098871

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 446–457, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3351095.3372843

Radomir Popović, Florian Lemmerich, and Markus Strohmaier. 2020. Joint multiclass debiasing of word embeddings. In *Foundations of Intelligent Systems*, pages 79–89, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-59491-6_8

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. 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 5740–5745, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1578

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3810

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity

recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics. https://doi.org/10.3115/1596374.1596399

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with Checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.442

Anthony Rios. 2020. FuzzE: Fuzzy fairness evaluation of offensive language classifiers on African-American English. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 881–889. https://doi.org/10.1609/aaai.v34i01.5434

Anthony Rios and Brandon Lwowski. 2020. An empirical study of the downstream reliability of pre-trained word embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3371–3388. International Committee on Computational Linguistics, Barcelona, Spain (Online). https://doi.org/10.18653/v1/2020.coling-main.299

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2002

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.690

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*.

João Sedoc and Lyle Ungar. 2019. The role of protected class word lists in bias identification of contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 55–61, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3808

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.291

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Seungjae Shin, Kyungwoo Song, JoonHo Jang, Hyemi Kim, Weonyoung Joo, and Il-Chul Moon. 2020. Neutralizing gender bias in word embeddings with latent disentanglement and counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3126–3140, Online.

Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.280

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the 5th Conference on Machine Translation (WMT)*, pages 629–638. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1164

Jacob Stegenga. 2015. Measuring effectiveness. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 54:62–71. https://doi.org/10.1016/j.shpsc.2015.06.003, PubMed: 26199055

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, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1159

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147. https://doi.org/10.3115/1119176.1119195

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617. https://doi.org/10.1162/tacl_a_00240

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2003

A. Zhiltsova, S. Caton, and Catherine Mulway. 2019. Mitigation of unintended biases against non-native english texts in sentiment analysis. In *Proceedings for the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*.