Benchmarking Intersectional Biases in NLP

John P. Lalor,[1,2] **Yi Yang,**[3]**, Kendall Smith,**[1,2] **Nicole Forsgren,**[4] **Ahmed Abbasi**[1,2]
[1] Human-centered Analytics Lab, University of Notre Dame
[2] Department of IT, Analytics, and Operations, University of Notre Dame
[3] Department of Information Systems and Operations Management, HKUST
[4] Microsoft Research
{john.lalor, ksmith77, aabbasi}@nd.edu, imyiyang@ust.hk, nicole.forsgren@microsoft.com

Abstract

There has been a recent wave of work assessing the fairness of machine learning models in general, and more specifically, on natural language processing (NLP) models built using machine learning techniques. While much work has highlighted biases embedded in state-of-the-art language models, and more recent efforts have focused on how to debias, research assessing the fairness and performance of biased/debiased models on downstream prediction tasks has been limited. Moreover, most prior work has emphasized bias along a single dimension such as gender or race. In this work, we benchmark multiple NLP models with regards to their fairness and predictive performance across a variety of NLP tasks. In particular, we assess intersectional bias - fairness across multiple demographic dimensions. The results show that while current debiasing strategies fare well in terms of the fairness-accuracy trade-off (generally preserving predictive power in debiased models), they are unable to effectively alleviate bias in downstream tasks. Furthermore, this bias is often amplified across demographic dimensions. We conclude with implications for future NLP debiasing research.

1 Introduction

As state-of-the-art natural language processing (NLP) language models become increasingly powerful and pervasive, recent progress in NLP has underscored the need for deeper analyses of how such models perform with respect to underrepresented groups. Research on fairness in NLP has shown that distributed representations of words often encode stereotypes - particularly towards different demographic groups (Blodgett et al., 2020; Bender et al., 2021). There is a growing stream of research that looks at mitigating these biases, especially when it manifests in the learned embedding state (Bolukbasi et al., 2016; Zmigrod et al., 2019; Kaneko and Bollegala, 2021). While prior work has undoubtedly moved the needle, recent surveys and research articles have identified several important gaps and issues (Blodgett et al., 2020; Tan and Celis, 2019). First, much of the current work on examining NLP bias (and proposing debiasing strategies) has focused on representational harm - how a model describes certain groups, including stereotyping and other misrepresentations (Blodgett et al., 2020; Suresh and Guttag, 2019). Conversely, there has been far less work exploring *allocational harm in downstream NLP prediction tasks* - when a system distributes resources or opportunities differently (Blodgett et al., 2020; Suresh and Guttag, 2019). Downstream tasks, such as sequence classification, also affect underrepresented groups, as these models show disparate impact on various demographic subsets, including women, African Americans, and the elderly (Blodgett et al., 2020; Bender et al., 2021; Shah et al., 2020).

Second, there has been limited work that examines *intersectional bias across a wide array of relevant charactersitics*, including several demographic dimensions, for a variety of non-debiased and debiased embeddings,[1] on a multitude of downstream tasks. Some work has studied demographic intersections such as young men and old women from a theoretical perspective (e.g., Kearns et al., 2018). Other recent studies have empirically shown that the biases inherent in language models for gender and race intersections might exceed those observed for gender and race alone (Tan and Celis, 2019), and that only debiasing along a single dimension can be problematic (Subramanian et al., 2021). Based on these two gaps, there is a need for a more systematic analysis of how current state-of-the-art language models and mitigation strategies perform with regards to intersectional bias in down-

[1] In this work, our scope is debiasing embeddings, not debiasing classifiers. While there is much work in the area of debiasing classifiers, here we restrict our focus to the debiasing of embeddings.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3598 - 3609

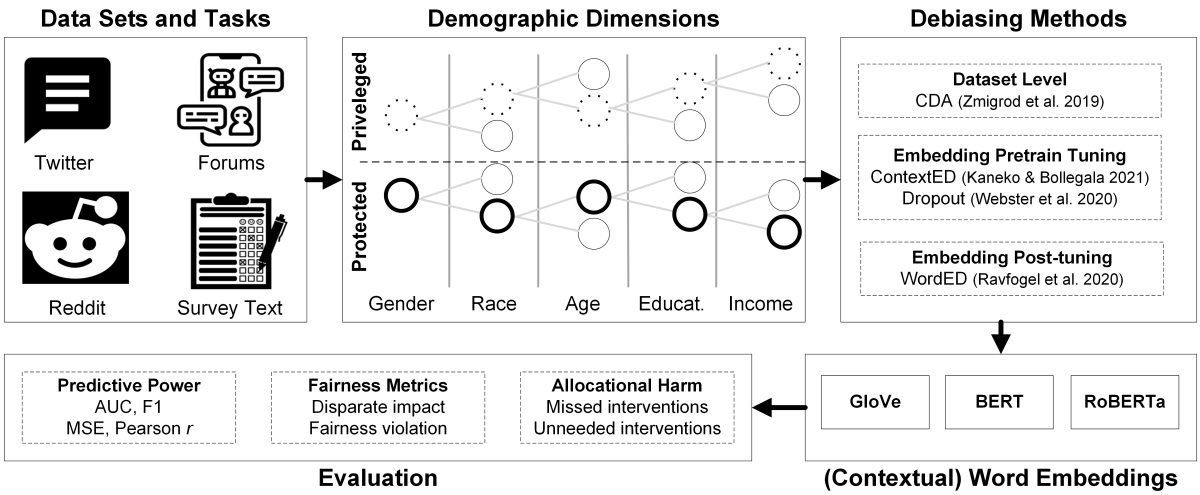

Figure 1: Overview of our fairness benchmarking analyses. We benchmark performance across datasets, models, and debiasing methods for tasks involving multiple demographic variables.

stream tasks.

Accordingly, in this study we perform a broad benchmark analysis of intersectional bias (Figure 1) encompassing the following key characteristics:

- Benchmark analysis on ten downstream sequence classification tasks related to five datasets that span common modes of user-generated content: Twitter, forums, Reddit, and survey responses. For these tasks, we also note the allocational harm implications of disparate impact, namely the harm associated with biased NLP-guided interventions.

- Inclusion of five demographic dimensions: gender, race, age, education, and income. Having three or more dimensions on many of the tasks affords opportunities to examine bias for various demographic intersection subgroups in a more in-depth manner. On four of the datasets, these demographics are self-reported as opposed to being algorithmically or heuristically inferred - an important consideration for debiasing research.

- Evaluation of three prominent word embeddings, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GloVe (Pennington et al., 2014), and four state-of-the-art model debiasing methods (Ravfogel et al., 2020; Kaneko and Bollegala, 2021; Zmigrod et al., 2019; Webster et al., 2020). This allows us to draw empirical insights regarding the effectiveness of mitigation strategies for downstream tasks.

Our results show that existing debiasing methods are generally very adept at preserving predictive power in downstream tasks. However, their ability to mitigate intersectional bias in such tasks is limited. In general, debiasing BERT/RoBERTa only incrementally alleviates disparate impact of model classifications. Further, while gender bias alone has disparate impact rates of 5-10% or less on most tasks, the range of bias is amplified for intersections - with unfairness rates often being 20 to 50% higher. On tasks such as inferring personality traits, literacy, or numeracy of users, these debiased models are still outside the fairness ranges recommended by governing bodies (Barocas and Selbst, 2016). Interestingly, these biases are more pronounced in models using GloVe, suggesting that debiased transformer-based models generally have better predictive power, and are fairer.

Our main contributions are two-fold. First, we perform a large-scale examination of intersectional bias across an array of downstream tasks. Our benchmark evaluation offers empirical evidence that the concerns voiced in recent critical surveys about too much emphasis on representational debiasing devoid of explicit normative goals (Blodgett et al., 2020), relative to mitigation of downstream allocational harm, are well-founded. Second, we quantify the size and scope of the intersectional bias problem, and the risks it can introduce for select underprivileged sub-groups when deploying NLP models for sequence prediction tasks. We are hopeful our work will spur future research that further sheds light on intersectional biases in down-

stream tasks, as well as mitigation strategies for alleviating allocational harm. Towards this goal, the code and data used in this work is publicly available via GitHub.[2]

2 Related Work

2.1 Allocational and Representational Harms

In their survey on bias in NLP, Blodgett et al. (2020) drew a distinction between allocational and representational harms. They found that most papers in NLP describe methods for measuring and mitigating representational harms - when "a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether" (Blodgett et al., 2020). One well-known example are stereotypes in word embeddings, such as certain ethnic groups being more closely associated with "housekeeper" (Garg et al., 2018).

In contrast, (Blodgett et al., 2020) only found four papers in their survey that were classified as having techniques for measuring/mitigating allocational harms - these "arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups." Allocational harm is often aligned with downstream tasks/interventions guided by the NLP model. For instance, all four of the aforementioned allocational harm papers measure and/or mitigate gender bias with respect to an NLP-based occupation classifier (De-Arteaga et al., 2019; Prost et al., 2019; Romanov et al., 2019; Zhao et al., 2020). More specifically, these studies examine the allocational harm of biased occupation classification predictions on decisions that affect humans, specifically whether an HR NLP system scraping web bios classifies individuals as relevant or not for a position. Our work builds on the nascent allocational harm literature by examining ten downstream tasks related to five data sets spanning Twitter, Reddit, forum, and survey response text.

2.2 Intersectional Biases

Intersectional biases arising as a result of interacting demographics have been studied in the broader machine learning literature, either from a theoretical perspective (Kearns et al., 2018; Yang et al., 2020), or in the context of facial recognition (Buolamwini and Gebru, 2018). In NLP, Tan and Celis (2019) evaluate and reveal important intersectional biases in contextualized word embedding models such as BERT and GPT-2. However, in their study, intersectional biases are evaluated using the word association test with an emphasis on representational harm - it remains unclear how intersectional biases affect allocational harm in downstream NLP tasks. Subramanian et al. (2021) looked at intersectional biases of classification models specifically designed for unbiased prediction, but do not evaluate embedding debiasing techniques. We build on the emergent literature on intersectional biases by assessing datasets encompassing up to five demographic dimensions, in conjunction with state-of-the-art word embeddings and debiasing methods, on downstream tasks where biased predictions can lead to allocational harm (§3.1).

2.3 Debiasing

Pretrained word embeddings, including static word embeddings such as GloVe and contextualized word embeddings such as BERT, contain human-like biases and stereotypical associations (Caliskan et al., 2017; Garg et al., 2018; May et al., 2019). A burgeoning body of NLP work has explored debiasing techniques to mitigate biases in pretrained word embeddings. One body of work has focused on debiasing static word embeddings (Bolukbasi et al., 2016; Zhao et al., 2020, 2018; Kaneko and Bollegala, 2019; Ravfogel et al., 2020).

Given the wide adoption of transformer-based contextualized embedding models, recent research has investigated bias mitigation in models such as BERT and RoBERTa (Zmigrod et al., 2019; Webster et al., 2020; Garimella et al., 2021; Kaneko and Bollegala, 2021; Guo et al., 2022). Existing methods for debiasing static and contextualized embeddings have alleviated representational harm along demographic dimensions such as gender. However, Gonen and Goldberg (2019) raised the concern that some debiasing strategies geared towards static word embeddings simply cover up the biases - which can resurface. Moreover, the seemingly debiased static embeddings often do not alleviate biases in downstream NLP prediction tasks (Goldfarb-Tarrant et al., 2021). The extent to which state-of-the-art debiasing methods can mitigate downstream intersectional biases remains unclear. This is precisely one of the gaps our study attempts to shed light on.

[2] https://github.com/nd-hal/naacl-2022

Dataset	Task	Demographics	Data source	N
Psychometrics	Anxiety, Literacy, Numeracy, Trust	Gender, Race, Age, Income, Education	Survey	8,395
Multilingual Twitter Corpus (MTC)	Hate Speech Identification	Gender, Race, Age	Twitter	83,078
Five Item Personality Inventory (FIPI)	Extraverted, Stable	Gender, Race, Age, Income, Education	Survey	6,805
AskAPatient	Sentiment	Gender, Age	Forums	20,000
Myers–Briggs Type Indicator (MBTI)	Perceiving, Thinking	Gender, Age	Reddit	7,406 (1,584)

Table 1: Details of the datasets used for benchmarking. For MBTI, users were able to provide multiple texts, we report unique users in parentheses.

3 Data, Models, Experiments

As previously depicted in Figure 1, our experimental setup is as follows. We assess predictive performance and fairness across five datasets spanning ten dependent variables/tasks and five demographic dimensions. We train three models (GloVe, BERT, and RoBERTa) as our prediction and fairness baselines. We then debias the input embeddings for these models (Ravfogel et al., 2020; Zmigrod et al., 2019; Kaneko and Bollegala, 2021) and re-train them to compare the performance. Details of the data, models, and evaluation metrics are below.

3.1 Data

We examine five datasets (Table 1) across several NLP tasks: psychometric dimension prediction, hate speech identification, personality detection, and sentiment analysis. The psychometric data set (Abbasi et al., 2021) consists of free-text responses on four psychometric dimensions: subjective health literacy, numeracy, anxiety, and trust in doctors. These free-text responses were then linked to survey-based psychometric scores also provided by the participants (serving as gold-standard numeric response labels). The data also includes self-reported demographics for each individual: age, race, gender, income, and education level. This data set was collected using crowd workers from Amazon Mechanical Turk and Qualtrics.

Similarly, the Five Item Personality Inventory (FIPI) and Myers–Briggs Type Indicator (MBTI) datasets include free text responses to estimate one of the FIPI or MBTI personality traits (Gjurković et al., 2021). In particular, due to space constraints, we focus on the MBTI traits of perceiving and thinking, and the FIPI traits of extraverted and stable. For FIPI, available demographics are gender, race, age, income, and education. For MBTI, self-reported gender and age are available. The AskAPatient dataset (Limsopatham and Collier, 2016) is taken from web forums and has labeled sentiment, along with gender and age information.

The Multilingual Twitter Corpus (MTC) hate-speech dataset contains labeled Twitter messages for the task of hate speech detection (Huang et al., 2020). The dataset also contains inferred author demographic factors. We use three demographics: gender, race, and age.

The Psychometrics, FIPI, AskAPatient, and MBTI tasks are all relevant from an allocational harms perspective. Biases in predictions for healthcare-related variables (Psychometrics), or personality type variables (MBTI, FIPI) can affect an individual's health care plan, personalized interventions, job prospects, etc. Biased predictions for drug rating sentiment can affect which drugs a future user chooses to take.

3.2 Models and Debiasing Methods

In the experiments, we considered several different text classification models. We used a word convolutional neural network (CNN) model, initialized with GloVe embeddings. We also considered two transformer-based contexualized embedding models: BERT and RoBERTa.

CNN We trained a word convolutional neural network (CNN) model, initialized with GloVE embeddings. The model consists of 3 concatenated CNN layers with kernel size of 1, 2 and 3 respectively. Each layer has a filter size of 256, rectified linear unit (ReLU) activation, L2 regularization (0.001), and global max pooling. The models were trained for 35 epochs with a batch size of 32 and learning rate of $1e^{-4}$.

Debiased-CNN We debiased the GloVe model using (Ravfogel et al., 2020). We kept all parameters the same as in the original paper based on

their publically available implementation.[3] The projection matrix was learned over 50 epochs.

BERT and RoBERTa We fine-tuned BERT and RoBERTa on downstream prediction tasks. We used `BERT-base-uncased` and `RoBERTa-base` model loaded from the transformers library. We fine-tuned BERT and RoBERTa model for five epochs using the following hyperparameters: a batch size to 32, learning rate of $1e^{-5}$, weight decay of 0.01. We saved the final model that achieves the lowest loss on validation set.

Debiased-BERT and Debiased-RoBERTa We debiased BERT and RoBERTa using (Kaneko and Bollegala, 2021). We obtained the gender word lists and stereotype word lists[4]. We used News-commentary-v15 corpus[4] as the external corpus to locate sentences where the gender and stereotype words occur and then debias. All BERT or RoBERTa layers are debiased at the token level, and the debiasing loss weight is set to 0.8. The model is fine-tuned for three epochs used the following hyperparameters: a batch size of 32 and learning rate of $5e^{-5}$.

Training Details For each dataset we trained using five-fold cross validation, so that for each example in each dataset, we could generate predictions as unseen test data.Each test fold was then concatenated for a given model for fairness calculations. All models were trained on the same data with hyperparameter tuning. All prediction models, debiasing models are trained on a NVIDIA GeForce RTX 3090 GPU card, with 11.2 CUDA version.

Debiasing Strategy Static word embeddings (GloVe, Pennington et al., 2014) were debiased using WordED (Ravfogel et al., 2020)[5]. This method iteratively learns a projection of embeddings that removes the bias information with minimal impact on embedding distances.

Contextualized word embedding models BERT and RoBERTa were debiased using ContextED (Kaneko and Bollegala, 2021)[6], which has been shown to work well at removing gender-bias encoded in embeddings. This method uses predefined word lists to identify sentences that contain the gendered or stereotype words, and then fine-tunes the pretrained model parameters by encouraging gendered and stereotype words to have orthogonal representations.

We also assessed two alternative debiasing methods for the contextualized word embedding models: counterfactual data augmentation (CDA) (Zmigrod et al., 2019) and Dropout (Webster et al., 2020).[7] CDA augments the training corpora with counterfactual data so that the language model is pretrained on gender-balanced text. Dropout mitigates gender biases by increasing the dropout rate in the pretrained models. Therefore, the debiasing methods in our experiments represent different ways of mitigating biases: dataset level (CDA), debiasing during pretraining (ContextED and Dropout), and post-tuning debiasing (WordED).

3.3 Evaluation

There are several definitions of fairness in the literature (Mehrabi et al., 2021), each with corresponding methods of assessment. In this work we rely on two prior metrics from the literature, and also present a new metric, adjusted disparate impact, to account for base rates in the dataset.

Disparate Impact One of the most common fairness assessments is disparate impact (DI, Friedler et al., 2019). DI measures the inequality of positive cases between privileged and non-privileged groups for a particular demographic. DI comes from from the legal field, where certain regulations require DI be above a threshold of 0.8 (or below 1.2 in the inverse case). For true labels y, predicted labels $\hat{y}$, and relevant demographic group A:

$$DI = \frac{p(\hat{y} = 1|A = 0)}{p(\hat{y} = 1|A = 1)} \quad (1)$$

Where $A = 0$ refers to the protected group and $A = 1$ refers to the privileged group. A DI ratio of 1 indicates *demographic parity*, where the rates of positive predictions are consistent across demographic classes: $P(\hat{y} = 1|A = 0) = P(\hat{y} = 1|A = 1)$ (Mehrabi et al., 2021).

Statistical Parity (SP) Subgroup Fairness Recent theoretical work on intersectional biases also assesses demographic parity, where the score compares group-specific rates to the global rate in the dataset instead of a comparison between privileged and protected classes (Kearns et al., 2018):

$$p(A = g) \times |p(\hat{y} = 1) - p(\hat{y} = 1|A = g)| \quad (2)$$

[3] https://github.com/shauli-ravfogel/nullspace_projection
[4] https://github.com/kanekomasahiro/context-debias/
[5] https://github.com/shauli-ravfogel/nullspace_projection
[6] https://github.com/kanekomasahiro/context-debias
[7] https://github.com/google-research-datasets/zari

This value is compared to an acceptability parameter λ to assess fairness. As this method was proposed for the intersectional case, it gives a way to identify the upper-bound of the *fairness violation* in a dataset (Yang et al., 2020):

$$FV = \max_{g \in G_f} |TPR_g - TPR_D| \quad (3)$$

Where G_f is the set of demographic groups under consideration for analysis, TPR_g is the true positive rate of the classifier on the instances in g, and TPR_D is the overall true positive rate for the classifier on the dataset. Prior work considered the average violation across groups (Subramanian et al., 2021), but for the purposes of this study we are interested in a worst case analysis.

Adjusted Disparate Impact We propose reweighting DI to account for differences in base rates. Adjusted DI (ADI) divides DI by the base rate ratio for the protected and privileged groups: $DI^* = \frac{p(y=1|A=0)}{p(y=1|A=1)}$, $ADI = \frac{DI}{DI^*}$

Note that the disparate impact metrics are not defined for cases where there are no positive instances for either the protected or privilege classes in the data, or when there are no positive predictions for the privileged class (due to zero division). Therefore, we use additive smoothing when calculating DI and adjusted DI (Zhai and Lafferty, 2004).

Intersectional Fairness To assess intersectional fairness we enumerated all combinations for each n-demographic scenario (e.g., 2-demographic, 3-demographic, etc.). We set a reference demographic, specifically gender, because of the prior work on debiasing word embeddings for gender (Bolukbasi et al., 2016; Gonen and Goldberg, 2019; Kaneko and Bollegala, 2021). For intersectional cases, we calculated DI and FV for all possible combinations of demographics that included gender. For example, the 2-demographic case for the psychometrics dataset involves calculating DI and FV for the following protected groups: older women, lower education women, lower income women, and non-white women. Our privileged groups are the negations of the protected groups, e.g., for the above case they are younger men, higher education men, higher income men, and white men. By considering disjoint demographic groups, we avoid cumulative effects of merging fairness results from individual demographics during the intersectional phase. We follow the same procedure for enumerating protected groups for the 3- and 4-demographic cases. For 5-demographics we consider all demographics together. For all models and datasets, we calculated fairness and performance metrics. For performance, we report mean squared error (MSE), Pearson's r,[8] F1, and area under the receiver operating curve (AUC). For fairness, we report adjusted DI and fairness violation (FV, §3.3).[9]

4 Results and Discussion

Figure 2 shows the ADI results for BERT and GloVe using ContextED and WordED for debiasing, respectively. In most cases, particularly for BERT, disparate impact scores for gender alone are in a reasonable range (within 10%). For GloVe, we do observe high gender ADI on Anxiety and Thinking. However, as the number of demographics under consideration grows, the range of ADI scores widens. While debiasing the word embeddings typically helps to reduce the unfairness for the target demographic (e.g., gender), in the intersectional cases the model still performs poorly. There are similar trends in FV scores as the number of demographics increases, with the extent of violations often increasing by a factor of 3x to 10x as intersections increase (Table 2).

In some cases the intersectional disparities are extreme. On the BERT models, the ratio of positive Numeracy predictions for the protected class is three-to-one compared to the privileged class. In the other direction, for 3-demographics, hate-speech detection positive predictions are significantly less likely for the protected group than the privileged group. This is consistent with prior hate-speech detection work that has shown large (absolute value) fairness gaps between protected and privileged groups (e.g., Liu et al., 2021).

In most cases, trends are consistent between the BERT and GloVe models (e.g., Extraverted, Numeracy, Perceiving). Some counterexamples are the Trust and Anxiety tasks. Here model choice impacts the direction of bias. As more demographics are considered, the GloVe model skews more unfair against the protected group, while the BERT model remains mostly fair, skewing slightly unfair against the privileged group. Higher trust in physicians is associated with better well-being and lower anxiety when visiting a doctor (Netemeyer et al.,

[8] MSE and Pearson's r were calculated for datasets where continuous gold standard values were available

[9] Standard disparate impact results were consistent with ADI and are not included due to space considerations.

Task	ModelName	MSE	Pearson's r	F1	AUC	DI	DI+	DI++	FV	FV+	FV++
Anxiety	**BERT**	**0.04**	0.53	0.68	0.74	1.04	0.89	1.04	**0.03**	**0.06**	0.09
	BERT-D	**0.04**	0.53	0.67	0.74	**1.06**	1.11	1.08	**0.03**	0.05	0.08
	RoBERTa	**0.04**	**0.55**	**0.69**	**0.75**	1.03	1.08	1.05	**0.03**	**0.06**	**0.12**
	RoBERTa-D	**0.04**	**0.55**	**0.69**	**0.75**	1.04	0.93	1.06	**0.03**	**0.06**	0.1
	word2vec	**0.04**	0.45	0.53	0.71	1.05	**0.83**	0.9	0.02	0.05	0.09
	word2vec-D	**0.04**	0.44	0.58	0.7	1.03	**0.83**	**0.89**	0.02	**0.06**	**0.12**
Extraverted	**BERT**	0.24	0.26	0.43	0.65	0.93	1.46	1.58	-	0.09	0.21
	BERT-D	0.24	0.27	0.41	**0.67**	0.94	1.46	1.65	-	0.07	0.21
	RoBERTa	0.22	**0.3**	**0.5**	**0.67**	0.93	1.42	1.56	-	**0.1**	**0.26**
	RoBERTa-D	0.24	0.23	0.48	0.63	0.94	1.35	1.5	-	0.08	0.22
	word2vec	**0.1**	-0.02	-	0.51	**0.44**	**3.12**	**2.96**	-	-	0.01
	word2vec-D	**0.1**	-0.01	-	0.51	**0.44**	**3.12**	**2.96**	-	-	0.01
Hatespeech	**BERT**	-	-	0.94	0.94	0.99	0.98	0.97	**0.04**	**0.11**	**0.17**
	BERT-D	-	-	0.94	0.94	1	0.98	0.97	**0.04**	**0.11**	**0.17**
	RoBERTa	-	-	**0.95**	**0.95**	1	0.98	0.96	**0.04**	**0.11**	**0.17**
	RoBERTa-D	-	-	**0.95**	**0.95**	0.99	0.98	0.96	**0.04**	**0.11**	**0.17**
	word2vec	-	-	0.76	0.81	**0.97**	0.8	0.82	0.01	0.04	0.07
	word2vec-D	-	-	0.75	0.81	0.98	**0.79**	**0.79**	0.01	0.04	0.07
Literacy	**BERT**	**0.01**	0.61	0.7	0.78	1.01	0.8	0.61	**0.02**	0.05	**0.06**
	BERT-D	**0.01**	0.6	0.68	0.78	1	0.82	0.64	**0.02**	**0.06**	0.02
	RoBERTa	**0.01**	**0.62**	**0.74**	**0.79**	0.95	0.74	0.63	0.01	0.05	0.04
	RoBERTa-D	**0.01**	**0.62**	0.73	**0.79**	0.98	0.76	0.65	**0.02**	0.05	0.02
	word2vec	0.02	0.46	-	0.72	**0.44**	**3.07**	**2.92**	-	-	0.01
	word2vec-D	**0.01**	0.49	0.04	0.73	1.17	0.41	0.32	-	-	-
Numeracy	**BERT**	**0.03**	0.55	0.69	0.75	1.21	2.46	3.23	**0.03**	**0.15**	0.3
	BERT-D	0.04	0.56	0.71	0.75	1.19	2.5	3.04	**0.03**	**0.15**	0.32
	RoBERTa	**0.03**	**0.58**	**0.72**	**0.77**	1.24	2.9	3.91	0.02	0.14	0.3
	RoBERTa-D	**0.03**	**0.58**	**0.72**	0.76	1.25	2.7	3.24	0.02	**0.15**	**0.34**
	word2vec	0.05	0.36	-	0.67	**0.71**	**26.93**	**45.11**	-	-	0.01
	word2vec-D	0.05	0.38	-	0.67	**0.71**	**26.93**	**45.11**	-	-	0.01
Perceiving	**BERT**	-	-	**0.37**	0.53	1.01	1.38	-	0.03	0.2	-
	BERT-D	-	-	0.36	0.54	0.91	1.44	-	**0.05**	0.23	-
	RoBERTa	-	-	0.34	**0.67**	0.9	2.45	-	0.03	0.34	-
	RoBERTa-D	-	-	0.25	0.55	**0.83**	**2.51**	-	0.03	**0.35**	-
	word2vec	-	-	0.29	0.54	0.91	2.13	-	0.03	0.32	-
	word2vec-D	-	-	0.29	0.53	0.98	2.09	-	0.03	0.31	-
Sentiment	**BERT**	**0.03**	0.82	0.84	0.93	0.95	**0.97**	-	**0.02**	**0.14**	-
	BERT-D	**0.03**	0.82	0.85	0.93	0.95	0.98	-	**0.02**	**0.14**	-
	RoBERTa	**0.03**	**0.84**	**0.86**	**0.94**	0.95	0.99	-	**0.02**	**0.14**	-
	RoBERTa-D	**0.03**	**0.84**	**0.86**	**0.94**	**0.94**	0.98	-	**0.02**	**0.14**	-
	word2vec	**0.03**	0.82	0.84	0.93	0.95	**0.97**	-	**0.02**	**0.14**	-
	word2vec-D	**0.03**	0.82	0.85	0.93	0.95	**0.97**	-	**0.02**	**0.14**	-
Stable	**BERT**	0.22	0.36	**0.6**	**0.71**	1.11	1.24	1.31	**0.02**	**0.09**	0.19
	BERT-D	0.23	0.32	0.57	0.68	1.1	1.29	1.37	**0.02**	**0.09**	**0.21**
	RoBERTa	0.23	0.34	0.49	0.69	1.08	1.55	1.39	0.01	0.07	0.15
	RoBERTa-D	0.22	**0.37**	0.58	**0.71**	1.1	1.36	1.39	**0.02**	**0.09**	**0.21**
	word2vec	**0.04**	0.18	-	0.59	**0.44**	**3.12**	**2.96**	-	-	0.01
	word2vec-D	**0.04**	0.18	-	0.6	**0.44**	**3.12**	**2.96**	-	-	0.01
Thinking	**BERT**	-	-	0.49	0.59	0.86	1.12	-	**0.07**	0.11	-
	BERT-D	-	-	0.51	0.58	0.92	1.02	-	0.06	0.06	-
	RoBERTa	-	-	**0.54**	**0.74**	**0.81**	**1.49**	-	0.06	**0.21**	-
	RoBERTa-D	-	-	0.44	0.58	0.87	1.33	-	0.05	0.17	-
	word2vec	-	-	0.47	0.58	0.82	1.2	-	**0.07**	0.14	-
	word2vec-D	-	-	0.43	0.56	0.96	1.4	-	0.04	0.17	-
Trust	**BERT**	**0.01**	0.73	0.83	0.87	1.03	1.09	1.09	**0.01**	**0.04**	**0.06**
	BERT-D	0.02	0.72	0.83	0.87	1.03	1.11	1.06	**0.01**	**0.04**	0.05
	RoBERTa	**0.01**	**0.74**	**0.84**	**0.88**	1.03	1.1	1.05	**0.01**	**0.04**	0.04
	RoBERTa-D	**0.01**	**0.74**	**0.84**	0.87	1.02	0.93	1.02	**0.01**	**0.04**	0.05
	word2vec	0.02	0.6	0.5	0.82	1.1	**1.19**	**0.84**	-	0.01	0.04
	word2vec-D	0.02	0.59	0.62	0.82	**1.11**	1.14	0.9	**0.01**	**0.04**	0.04

Table 2: Benchmarking results. For Psychometrics and FIPI, + and ++ indicate 3- and 5-way demographics, respectively. For MTC, AskAPatient, and MBTI, + and ++ indicate 2- and 3-way demographics, respectively. Best performance metrics (lowest for MSE, highest for Pearson r, F1, and AUC) and least fair for fairness metrics (furthest from 1 for DI, highest for FV) are **bolded**.

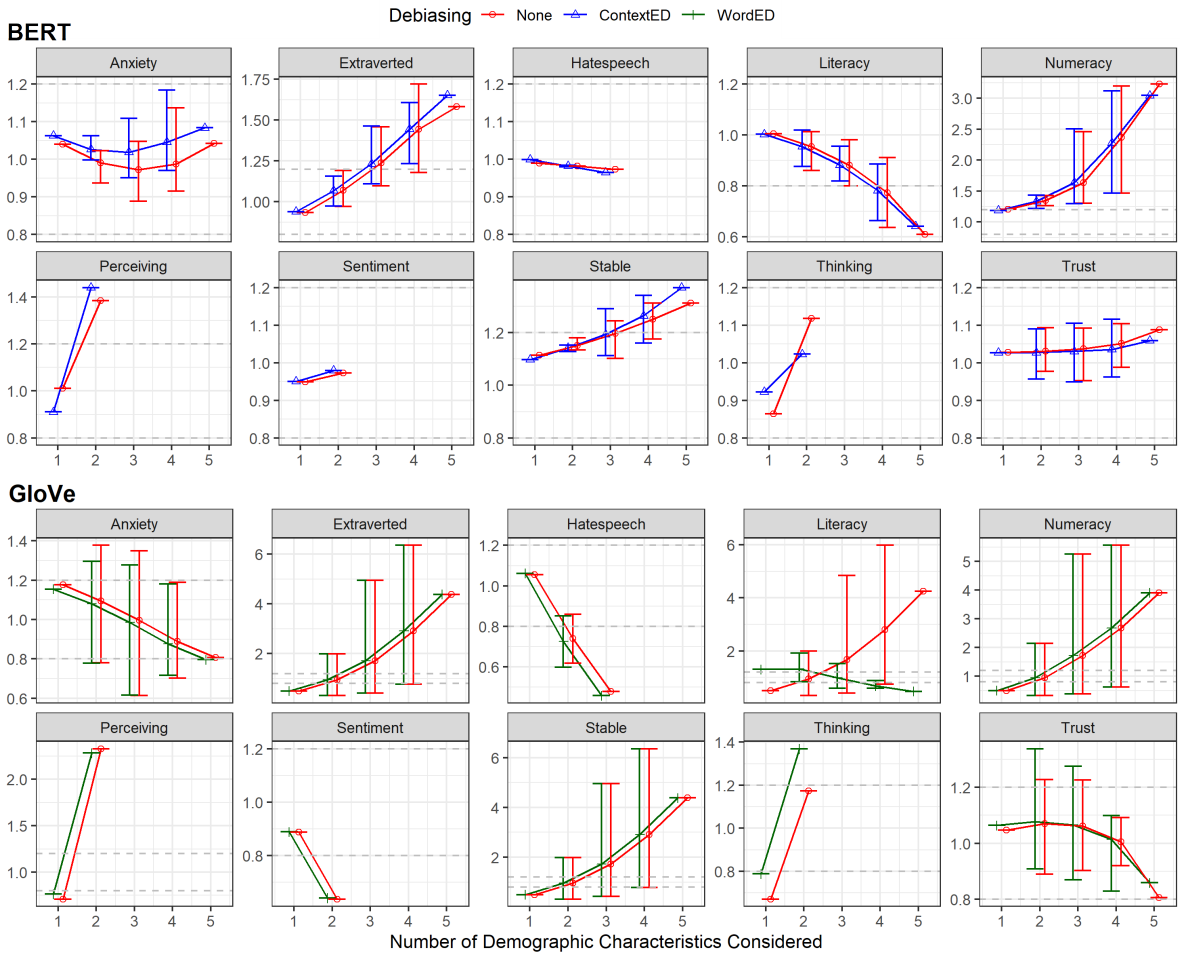

Figure 2: Effect of intersectionality on adjusted disparate impact for BERT and GloVe models. For x-axes with more than one demographic characteristic under consideration, we report the mean ADI and 95% confidence intervals.

2020); disparate predictions can lead to missed interventions for trust-increase and anxiety reduction across demographic groups. Though not depicted in the main paper, plots for RoBERTa show similar trends to those observed for BERT while debiasing with ContextED (see Appendix A).

Results are similar when looking at alternate BERT debiasing methods beyond ContextED, namely CDA and Dropout (Figure 3). These findings on the Anxiety, Literacy, Numeracy, and Trust tasks suggest that debiasing at the dataset, embedding pretraining, and post-tuning levels leads to similar increases in unfairness as the number of demographic intersections considered increases.

Collectively, the results underscore the allocational harm implications of NLP models on several downstream tasks - ones that even well-designed and well-intentioned debiasing strategies cannot overcome. This can be problematic in the era of personalized marketing and precision health, with NLP-based persona-generation playing a bigger role. For tasks like numeracy and literacy, this can affect how a patient is treated by a medical staff during a hospital visit (i.e., a false positive high literacy prediction for a person who has trouble understanding his or her medical record). For the personality indicators, inconsistent predictions may lead to biased decisions in the workplace (e.g., a manager looking to form a team of extroverts).

5 Conclusion

In this work we present a comprehensive benchmarking analysis of fairness for sequence prediction models. We also look at known debiasing methods for these models and show that while the debiased versions maintain predictive performance (as expected), they do not help with mitigating biases. While most models are relatively fair when looking at a single demographic characteristic, accounting for intersectional groups leads to less fair models and wider ranges of bias because of the

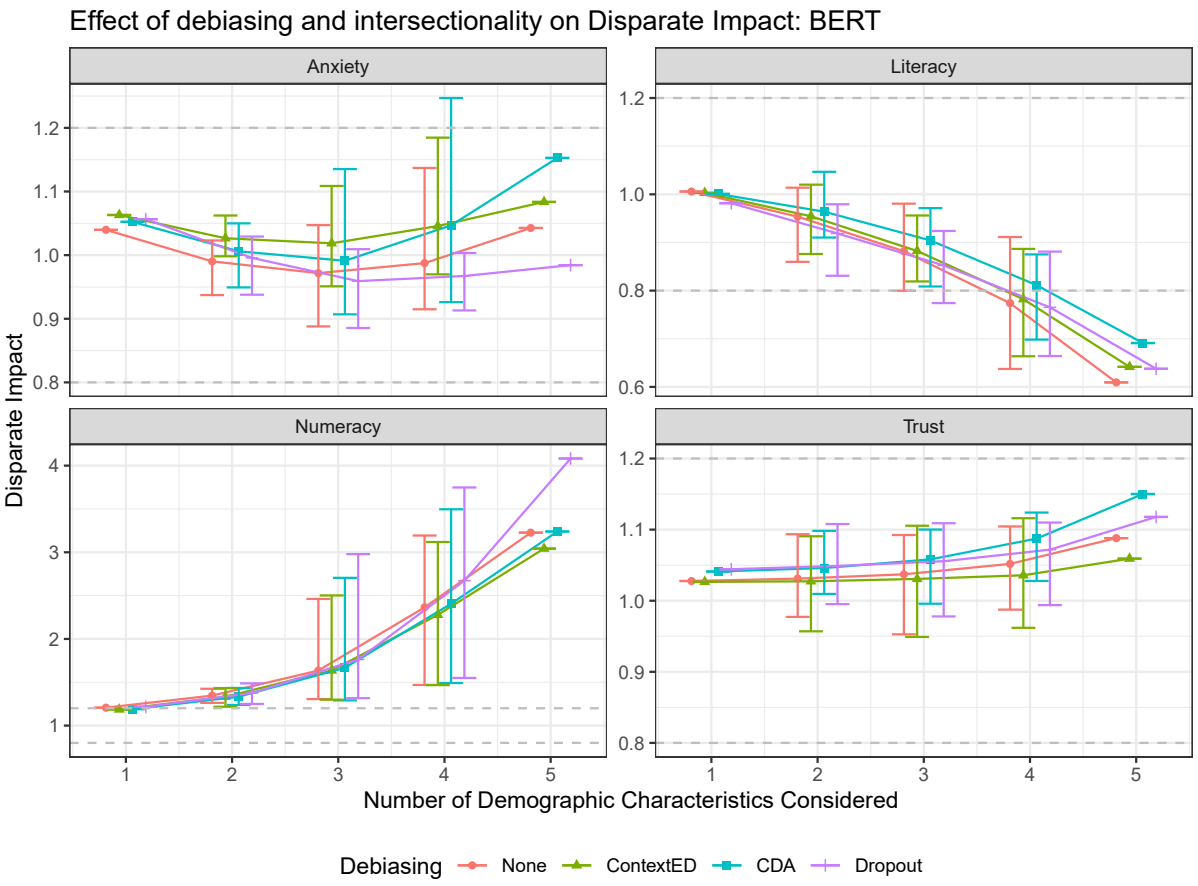

Figure 3: Effect of different debiasing methods on adjusted disparate impact.

combinatorial considerations of the intersectional groups. It is our hope that this benchmarking encourages future work into mitigating intersectional biases, and also to collect more demographic information when creating new datasets.

Acknowledgement

Members of Notre Dame's Human-centered Analytics Lab (HAL) were funded in part through U.S. NSF grant IIS-2039915.

References

Ahmed Abbasi, David Dobolyi, John P. Lalor, Richard Netemeyer, Kendall Smith, and Yi Yang. 2021. Constructing a psychometric testbed for fair natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.*, 104:671.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

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.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from lan-

guage corpora contain human-like biases. *Science*, 356(6334):183–186.

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*, pages 120–128.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *Findings of ACL*, pages 4534–4545.

Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. PANDORA talks: Personality and demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online. Association for Computational Linguistics.

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 ACL*.

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.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael Paul. 2020. 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.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of ACL*, pages 1641–1650.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023.

Haochen Liu, Wei Jin, Hamid Karimi, Zitao Liu, and Jiliang Tang. 2021. The authors matter: Understanding and mitigating implicit bias in deep text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 74–85, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of NAACL)*, pages 622–628.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Richard G Netemeyer, David G Dobolyi, Ahmed Abbasi, Gari Clifford, and Herman Taylor. 2020. Health literacy, health numeracy, and trust in doctor: effects on key patient health outcomes. *Journal of Consumer Affairs*, 54(1):3–42.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference*

on empirical methods in natural language processing (EMNLP), pages 1532–1543.

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.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? reducing bias in bios without access to protected attributes. 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 4187–4195.

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.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2498.

Harini Suresh and John V. Guttag. 2019. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Forest Yang, Mouhamadou Cisse, and Oluwasanmi O Koyejo. 2020. Fairness with overlapping groups; a probabilistic perspective. *Advances in Neural Information Processing Systems*, 33.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of EMNLP*.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

A Appendix: RoBERTa Results

Figure 4 shows results of our benchmarking experiments for RoBERTa. The trends of degrading performance are consistent with the results in BERT.

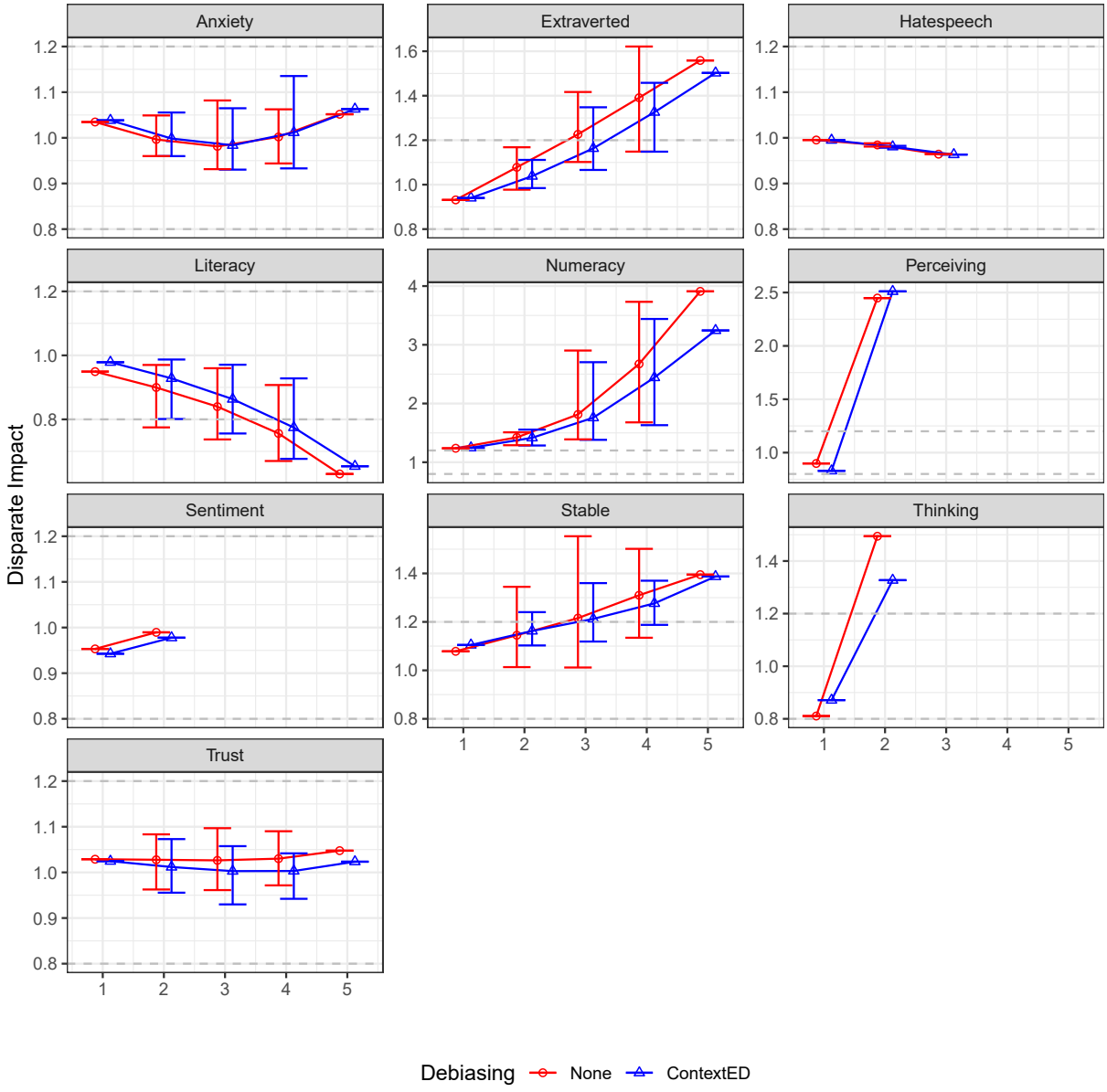

Figure 4: ADI results for RoBERTa on our benchmark datasets.