# Evaluating Debiasing Techniques for Intersectional Biases

**Shivashankar Subramanian**    **Xudong Han**
**Timothy Baldwin    Trevor Cohn    Lea Frermann**
School of Computing and Information Systems
University of Melbourne, Australia

shivashankarrs@gmail.com    xudongh1@student.unimelb.edu.au

{tbaldwin,t.cohn,lfrermann}@unimelb.edu.au

## Abstract

Bias is pervasive in NLP models, motivating the development of automatic debiasing techniques. Evaluation of NLP debiasing methods has largely been limited to binary attributes in isolation, e.g., debiasing with respect to binary gender or race, however many corpora involve multiple such attributes, possibly with higher cardinality. In this paper we argue that a truly fair model must consider 'gerrymandering' groups which comprise not only single attributes, but also intersectional groups. We evaluate a form of bias-constrained model which is new to NLP, as well an extension of the iterative nullspace projection technique which can handle multiple protected attributes.

## 1 Introduction

Text data reflects the social and cultural biases in the world, and NLP models and applications trained on such data have been shown to reproduce and amplify those biases. Discrimination has been identified across diverse sensitive attributes including gender, disability, race, and religion (Caliskan et al., 2017; May et al., 2019; Garimella et al., 2019; Nangia et al., 2020; Li et al., 2020). While early work focused on debiasing typically binarized protected attributes in isolation (e.g., age, gender, or race; Caliskan et al. (2017)), more recent work has adopted a more realistic scenario with multiple sensitive attributes (Li et al., 2018) or attributes covering several classes (Manzini et al., 2019).

In the context of multiple protected attributes, *gerrymandering* refers to the phenomenon where an attempt to make a model fairer towards some group results in increased unfairness towards another group (Buolamwini and Gebru, 2018; Kearns et al., 2018; Yang et al., 2020). Notably, algorithms can be fair towards independent groups, but not towards all intersectional groups. Despite this, debiasing approaches within NLP have so far been evaluated only using *independent group fairness*
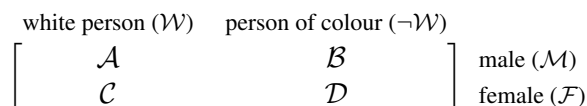


Figure 1: Group intersection and gerrymandering: $\mathcal{A}$=white male, $\mathcal{B}$=male person of colour, $\mathcal{C}$=white female, $\mathcal{D}$=female person of colour.

when modelling datasets with multiple attributes, disregarding intersectional subgroups defined by combinations of sensitive attributes (see Figure 1).

The primary goal of this work is to evaluate independent and intersectional identity debiasing approaches in relation to fairness gerrymandering for text classification tasks. To this end, we evaluate bias-constrained models (Cotter et al., 2019b) and iterative nullspace projection (INLP; Ravfogel et al. (2020)), a post-hoc debiasing method which we extend to handle intersectional groups. The constrained model jointly optimizes model performance and model fairness, while INLP seeks to learn a hidden representation which is independent of the protected attributes. INLP does not consider the trade-off between accuracy and fairness, but rather it iteratively maximizes fairness in an unconstrained fashion.

In this work, we address the following questions:
- Are debiasing approaches based on independent groups more prone to fairness gerrymandering than methods using intersectional groups?
- How do INLP and bias-constrained approaches, and their extensions to handle intersectional groups, fare compare in terms of both predictive accuracy and fairness?

## 2 Background

Debiasing with respect to more than a single protected attribute ($|Z| > 1$) requires grouping data

2492

points according to their associated protected values. In this work, debiasing approaches are trained *wrt* the three settings detailed below (Kearns et al., 2018; Yang et al., 2020), evaluated based on aggregated violations over $G_{\text{GERRY}}$, which is a commonly used evaluation strategy for intersectional identities (Yang et al., 2020).

**Independent groups ($G_{\text{INDEP}}$)** Each private attribute is treated as an independent group. In the case of all binary attributes, this gives $2|Z|$ groups. As illustrated in Figure 1, for binary gender and race,[1] $\mathcal{F}$, $\mathcal{M}$, $\mathcal{W}$, and $\neg\mathcal{W}$ are four independent groups, which are overlapping, e.g., with $\mathcal{F}$ referring to both white females and female persons of colour ($\mathcal{C}$ and $\mathcal{D}$).

**Intersectional groups ($G_{\text{INTER}}$)** Non-overlapping, exhaustive combinations of groups, which in the binary attribute case gives $2^{|Z|}$ combinations. For our example in Figure 1, the intersectional groups are $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$.

**Gerrymandering intersectional groups ($G_{\text{GERRY}}$)** Overlapping subgroups are defined by group intersections of any subset of private attributes ($3^{|Z|}$ for binary attributes, where the 3rd value denotes a wildcard). For the example in Figure 1, the $G_{\text{GERRY}}$ groups are $\mathcal{F}$, $\mathcal{M}$, $\mathcal{W}$, $\neg\mathcal{W}$, $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$.

## 3 Proposed Approaches

We aim to learn a model to predict main task labels (e.g., sentiment) from input data $X$ (e.g., reviews) by means of a learnt hidden representation $X_h$. $X_h$ will generally encode implicit biases, and our goal is to predict the output without the influence of social identities encoded in $X_h$. We assume access to the protected attributes for each instance. Below, we detail the two primary models targeted in this work: INLP, and the bias-constrained model.

### 3.1 INLP

INLP is applied to frozen hidden representations, where we first learn a linear classifier $\mathbf{W}$ using $X_h$ as the independent variables to predict a protected attribute (binary- or multi-valued). Next, $X_h$ is projected onto the nullspace of $\mathbf{W}$ (denoted $P_{N(\mathbf{W})}X_h$) to remove the protected information, which forms the input for learning a classifier of the main task label. In real-world scenarios, bias may arise due to a combination of multiple protected attributes, hence we extend the single attribute setting

of Ravfogel et al. (2020) to this situation. Assuming $k$ protected attributes, and classifiers trained to predict those denoted as $\{\mathbf{W}^i\}_{i=1}^k$, we compute the intersection of nullspaces (Strang, 2014) across identities,

$$\bigcap_{i=1}^{k} N(\mathbf{W}^i) = N\left(\begin{bmatrix}\mathbf{W}^1 \\ \mathbf{W}^2 \\ \vdots \\ \mathbf{W}^k\end{bmatrix}\right). \quad (1)$$

No special properties are imposed over $\mathbf{W}^i$ (e.g., orthogonality), which allows for parallel training.

Even after debiasing by projecting $X_h$ onto the nullspace of classifiers, hidden representations can still retain protected information (Ravfogel et al., 2020). Hence we follow Ravfogel et al. (2020) and resort to iteratively removing bias, rather than doing it in a single step. Nullspace projection will reduce the rank of data rapidly with intersectional identities[2] (see Figure 2). To circumvent this issue, we project onto the nullspace of the *principal component* of the bias subspace ($N_p$). This is equivalent to projecting onto the nullspace of some of the intersectional identities in each iteration; similar strategies have been employed for multi-class supervised principal component analysis (Piironen and Vehtari, 2018).

Following Ravfogel et al. (2020), while incorporating our principal component-based nullspace projection as described above, the intersection of nullspaces across iterations is computed as

$$\bigcap_{i=0}^{m} N_p(\mathbf{W}_i) = N\left(\sum_{i=0}^{m} P_{R'(\mathbf{W}_i)}\right), \quad (2)$$

where $N_p$ denotes the nullspace based on the principal component of the bias subspace, $\mathbf{W}_i$ is the set of classifiers trained in each iteration to predict the protected attributes, and $R'$ is the principal component of the rowspace computed by stacking classifier weight vectors across multiple protected attributes. Unlike Eq (1), the formulation in Eq (2) requires orthogonality of subspaces across each iteration.

---

[1] Higher cardinality attributes are also supported.

[2] The rank reduces by 1 for each protected attribute.

## 3.2 Bias-constrained Model

A bias-constrained model combines the main task objective and fairness constraints,

$$\min_{\theta} \quad \ell(\mathcal{F}(x, \theta), y)$$
$$\text{s.t.} \quad \forall g \in [G] \; \gamma_g |\phi_g - \phi| \le \nu,$$

where $\ell(\mathcal{F}(x, \theta), y)$ denotes the primary objective (e.g., error rate); $G$ are the groups defined in Section 2; $\phi_g$ denotes the group-wise performance; and $\phi$ denotes overall performance of the model (e.g., TPR). $\nu$ is a slack variable, which controls the maximum deviation allowed. $\gamma_g$ is the inverse proportion of positive examples in each group, $\frac{|g|}{|g_{y=1}|}$, meaning that the more underrepresented a group is *wrt* some target class (say *toxicity*), the smaller its accepted deviation from the overall performance.

Each group-wise constraint is denoted as $\psi_g$. The constraints involve a linear combination of indicator variables, which is not differentiable *wrt* $\theta$. A common approach to handle this constrained optimization problems is using the Lagrangian,

$$\mathcal{L}(\theta, \lambda) = \ell(\mathcal{F}(x, \theta), y) + \sum_{g=1}^{|G|} \lambda_g \psi_g, \quad (3)$$

which is minimized over $\theta$ and maximized over $\lambda$. Similar formulations have been used for learning fair models with structured data (Cotter et al., 2019b; Yang et al., 2020; Zafar et al., 2019). In this work, we apply this method to NLP tasks and use the two-player zero-sum game approach for optimization, where the first player chooses $\theta$ to minimize $\mathcal{L}(\theta, \lambda)$, and the second player enforces fairness constraints by maximizing $\lambda$ (Kearns et al., 2018; Cotter et al., 2019b; Yang et al., 2020). Specifically we use the implementations available in TensorFlow constrained optimization (Cotter et al., 2019a,b).[3]

The approach iterates over T iterations, where the primal player updates $\theta$ with a fixed $\lambda$, and dual player updates $\lambda$ with a fixed $\theta$. This results in multiple models, $\{\theta_1, \theta_2, ...\theta_T\}$ and $\{\lambda_1, \lambda_2, ...\lambda_T\}$, which trade off performance and fairness. We use the Adam optimizer (Kingma and Ba, 2014) with learning rate $10^{-3}$.

## 4 Experimental Results

We evaluate using two binary classification datasets: (1) Twitter hate speech detection; and

---

[3] https://github.com/google-research/tensorflow_constrained_optimization
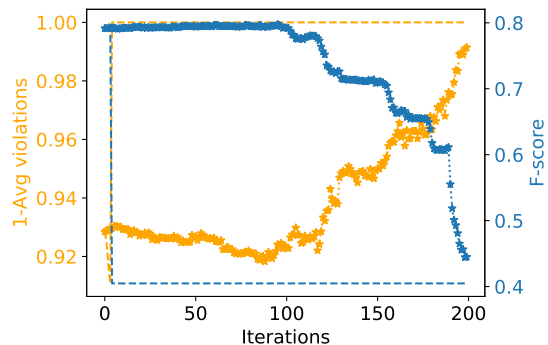


Figure 2: F-score vs. fairness results for INLP (hate-speech dev set). Projecting onto the null space of the principal bias component (dotted line and stars) results in a better trade-off than directly projecting onto the nullspace of the entire bias sub-space (dashed line).

(2) occupation classification using biography text. For INLP we use logistic regression to construct **W**. All the approaches have hyperparameters controlling the performance–fairness trade-off, and as no choice of hyperparameter will optimize both objectives simultaneously, we adopt the concept of *Pareto optimality* (Godfrey et al., 2007).

### 4.1 Settings

For INLP we consider different definitions of the bias subspace, which we illustrate *wrt* Figure 1: (a) two linear classifiers, trained to discriminate between independent groups $G_{\text{INDEP}}$: $\mathcal{F}$ and $\mathcal{M}$, $\mathcal{W}$ and $\neg\mathcal{W}$; (b) four linear classifiers to discriminate between intersectional groups $G_{\text{INTER}}$: $\mathcal{A}, \mathcal{B}, \mathcal{C}$, and $\mathcal{D}$ (one vs. rest); and (c) eight linear classifiers to discriminate between $G_{\text{GERRY}}$ groups. We refer to the three approaches as $\text{INLP}_{\text{INDEP}}$, $\text{INLP}_{\text{INTER}}$, and $\text{INLP}_{\text{GERRY}}$ respectively. For the constrained model, constraints capturing the performance deviation of $G_{\text{INDEP}}$, $G_{\text{INTER}}$, and $G_{\text{GERRY}}$ are used in Eq (3), which we refer to as $\text{CON}_{\text{INDEP}}$, $\text{CON}_{\text{INTER}}$, and $\text{CON}_{\text{GERRY}}$, respectively.

**Metrics** We assess a model's predictive performance using F-score, and fairness based on average-violations (Yang et al., 2020) of true positive rate (*equality of opportunity*), across all the $G_{\text{GERRY}}$ groups: $\frac{1}{|G_{\text{GERRY}}|} \sum_{g=1}^{G_{\text{GERRY}}} |\text{tpr}_g - \text{tpr}|$.

### 4.2 Hate speech detection

We use the English Twitter hate speech detection dataset of Huang et al. (2020), where each tweet is labeled with a binary hate speech label, as well as binary demographic group identity indicators for
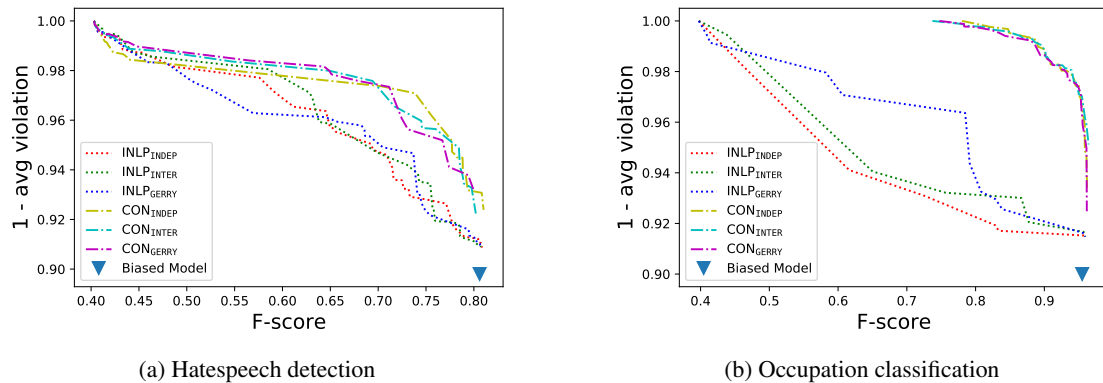
(a) Hatespeech detection      (b) Occupation classification

Figure 3: Pareto frontier model test results for the trade-off between $F_1$ score and fairness, measured as $1-$ the average TPR violation across all $G_{\text{GERRY}}$ groups. The top-right represents ideal performance.

the tweet authors: gender (f or m), age ($\leq$ or $>$ median), location (US or not US), and race (white or other). We use the data splits from Huang et al. (2020), but further filter each partition to the subset of tweets for which all demographic author labels are available, resulting in 23K (train), 5K (dev), and 5K (test) tweets.

We use a biGRU as our base model, following Huang et al. (2020). In Figure 2, we show the benefits of our proposed INLP approach for debiasing gerrymandering intersectional identities, compared to the naive extension of INLP for multiple protected attributes (which projects onto the entire bias subspace). With 80 $G_{\text{GERRY}}$ groups, the naive extension leads to a drastic drop in performance after 4 iterations (the dashed blue line), as in each iteration the rank is reduced by 80. On the other hand, the proposed approach achieves a much more nuanced trade-off between accuracy and fairness (the blue and yellow dotted line + stars, resp.).

We present the Pareto frontier for the different models on the test-set in Figure 3a, by varying the hyper-parameters of the approaches[4] each under of the three different debiasing settings, i.e. $G_{\text{INDEP}}$, $G_{\text{INTER}}$, and $G_{\text{GERRY}}$. From the results, for F-score close to the base biGRU (= biased model), the independent model provides the best trade-off, while debiasing for intersectional identities provides a better trade-off overall, and constrained models perform better than INLP.

### 4.3 Occupation classification

For our second experiment, we use the occupation classification dataset of De-Arteaga et al. (2019).

This comprises short web biographies, annotated by gender and profession. We augment the dataset with economic status (wealthy vs. rest, using World Bank data) based on the country the individual is based in, labelled based on the first sentence of each bio, which we perform entity linking on using AIDA-light (Nguyen et al., 2014), and map locations and institutions to countries based on their Wikipedia articles. We computed the accuracy of the economic status mapping by manually analyzing the output of over 200 biographies, resulting in a 93% accuracy and 90% agreement (Jaccard coefficient) between three annotators. We use a subset of two highly stereotyped classes of the dataset — nurse vs. surgeon — resulting in 13K/1.7K/5.4K instances in train/dev/test.[5]

We use the BERT (base uncased) (Devlin et al., 2019) [CLS]-encoding followed by an MLP (300-d with ReLU activations) for classification. The Pareto results[6] on the test set are shown in Figure 3b. INLP$_{\text{INTER}}$ and INLP$_{\text{GERRY}}$ perform better than INLP$_{\text{INDEP}}$, and the constrained models once again perform best overall, with intersectional identities providing no significant gains over CON$_{\text{INDEP}}$. Even in a highly stereotyped setting, constrained models achieve high predictive performance, in contrast to INLP which improves fairness but greatly reduces predictive performance, consistent with the single protected attribute results in Ravfogel et al. (2020).

### 4.4 Constrained analysis

In addition to Pareto curves (Figure 3), we compare the fairest models under a minimum performance

---

[4]For INLP, # iterations = 1 to $\dim(X_h)$; for the constrained models, # iterations (T) = 1 to 100, $\nu$ = 1e-4 to 1, $\gamma_g$ = 1 or $\frac{|g|}{|g_{y=1}|}$.

[5]90% of surgeons are male and 87% of nurses are female; 97.5% of nurses are from wealthy economies.

[6]Constrained models are run for T = 1 to 500 iterations.

|  | Trade-off 5% | | | Trade-off 10% | | |
|---|---|---|---|---|---|---|
| **Approach** | **$F_1$** | **Max violation** | **Avg violation** | **$F_1$** | **Max violation** | **Avg violation** |
| INLP$_{\text{INDEP}}$ | 0.763 | 0.333 | 0.079 | 0.732 | 0.242 | 0.076 |
| INLP$_{\text{INTER}}$ | 0.785 | 0.312 | 0.083 | 0.736 | 0.258 | 0.064 |
| INLP$_{\text{GERRY}}$ | 0.778 | 0.453 | 0.086 | 0.737 | 0.199 | 0.054 |
| CON$_{\text{INDEP}}$ | 0.777 | 0.230 | **0.048** | 0.724 | 0.274 | **0.031** |
| CON$_{\text{INTER}}$ | 0.772 | 0.254 | 0.063 | 0.746 | 0.172 | 0.043 |
| CON$_{\text{GERRY}}$ | 0.759 | **0.135** | 0.059 | 0.715 | **0.150** | 0.047 |
| Biased model | 0.806 | 0.338 | 0.102 | 0.806 | 0.338 | 0.102 |

(a) Hatespeech classification

|  | Trade-off 5% | | | Trade-off 10% | | |
|---|---|---|---|---|---|---|
| **Approach** | **$F_1$** | **Max violation** | **Avg violation** | **$F_1$** | **Max violation** | **Avg violation** |
| INLP$_{\text{INDEP}}$ | 0.949 | 0.291 | 0.087 | 0.949 | 0.291 | 0.087 |
| INLP$_{\text{INTER}}$ | 0.956 | 0.313 | 0.090 | 0.956 | 0.313 | 0.090 |
| INLP$_{\text{GERRY}}$ | 0.917 | 0.242 | 0.090 | 0.904 | 0.224 | 0.088 |
| CON$_{\text{INDEP}}$ | 0.953 | **0.134** | **0.041** | 0.940 | **0.102** | **0.028** |
| CON$_{\text{INTER}}$ | 0.949 | 0.225 | 0.065 | 0.908 | 0.110 | **0.028** |
| CON$_{\text{GERRY}}$ | 0.951 | 0.170 | 0.049 | 0.905 | 0.137 | 0.035 |
| Biased model | 0.955 | 0.350 | 0.100 | 0.955 | 0.350 | 0.100 |

(b) Occupation classification

Table 1: Test performance of salient models on hatespeech and occupation classification. Least biased models within a given performance tradeoff thresholds are chosen from development set.

constraint, where the most fair models (based on average violations) which exceed a performance threshold on the development set are chosen, and evaluated on the test set. This allows us to measure the generalizability of models occupying optimal Pareto region for the development set, to an unseen test set. We provide results under two constrained scenarios — models with least bias on development set trading off 5% and 10% of predictive performances. In addition to average violation, we also report maximum violation as $\max_{g \in \text{G}_{\text{GERRY}}} |\text{tpr}_g - \text{tpr}|$, which measures the upper bound of unfairness towards any subgroup (Yang et al., 2020).

For hatespeech classification, results are provided in Table 1a. With a slack of up to 10% trade-off in predictive performance, INLP$_{\text{GERRY}}$ shows better predictive performance and less violations than both INLP$_{\text{INTER}}$ and INLP$_{\text{INDEP}}$. For the closer 5% tradeoff, INLP$_{\text{INTER}}$ has higher F-score with a slightly worse avg violation and better max violation. The constrained models debiasing for G$_{\text{GERRY}}$ are fairest (lowest max and avg violation). Similarly, for occupation classification (Table 1b), INLP$_{\text{GERRY}}$ achieves lowest maximum violation among INLP models, while overall the constrained models, specifically CON$_{\text{INDEP}}$, perform best. This confirms the conclusion based on Pareto plots that adding intersectional identities for INLP-based de-

biasing is beneficial, whereas for constrained models which directly regulate the expected performance *wrt* different subgroups, they only provide mild gains.

## 5 Conclusion and Future Work

We examined the impact of independent vs. intersectional groups on classifier fairness. We proposed an extended version of INLP, which we compared against bias-constrained models. INLP is a post-hoc debiasing method, while the constrained model requires full joint training. Debiasing in INLP happens in an unconstrained way, where sensitive information is removed from the latent representations independent of predictive performances, leading to fairer models at the cost of overall accuracy (Han et al., 2021). From a practical perspective, post-hoc debiasing may be the only option if it is impossible or undesirable to re-train a model from scratch with a regularized objective. Our paper shows that lighter-weight post-hoc debiasing is indeed possible in the case of overlapping groups, and quantifies the additional advantage of joint training and debiasing in the context of two data sets. The constrained models especially are more robust to stereotyped settings, and using intersectional identities benefits INLP in particular.

## Acknowledgements

## References

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR.

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

Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2019a. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pages 1397–1405. PMLR.

Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. 2019b. Two-player games for efficient non-convex constrained optimization. In *Algorithmic Learning Theory*, pages 300–332. PMLR.

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.

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.

Parke Godfrey, Ryan Shipley, and Jarek Gryz. 2007. Algorithms and analyses for maximal vector computation. *The VLDB Journal*, 16(1):5–28.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765.

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.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. Unqovering stereotyping biases via underspecified questions. *arXiv preprint arXiv:2010.02428*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628.

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

Dat Ba Nguyen, Johannes Hoffart, Martin Theobald, and Gerhard Weikum. 2014. AIDA-light: High-throughput named-entity disambiguation. In *LDOW*.

Juho Piironen and Aki Vehtari. 2018. Iterative supervised principal components. In *International Conference on Artificial Intelligence and Statistics*, pages 106–114. PMLR.

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.

G. Strang. 2014. *Linear Algebra and Its Applications*. Elsevier Science.

Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. 2020. Fairness with overlapping groups; a probabilistic perspective. In *Advances in Neural Information Processing Systems*, volume 33, pages 4067–4078.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *J. Mach. Learn. Res.*, 20(1):2737–2778.