# Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings

**Thomas Manzini**[†][*] , **Yao Chong Lim**[‡][*], **Yulia Tsvetkov**[‡], **Alan W Black**[‡]

Microsoft AI Development Acceleration Program[†], Carnegie Mellon University[‡]

`Thomas.Manzini@microsoft.com`, {`yaochonl,ytsvetko,awb`}`@cs.cmu.edu`

## Abstract

Online texts—across genres, registers, domains, and styles—are riddled with human stereotypes, expressed in overt or subtle ways. Word embeddings, trained on these texts, perpetuate and amplify these stereotypes, and propagate biases to machine learning models that use word embeddings as features. In this work, we propose a method to debias word embeddings in multiclass settings such as race and religion, extending the work of (Bolukbasi et al., 2016) from the binary setting, such as binary gender. Next, we propose a novel methodology for the evaluation of multiclass debiasing. We demonstrate that our multiclass debiasing is robust and maintains the efficacy in standard NLP tasks.

## 1 Introduction

In addition to possessing informative features useful for a variety of NLP tasks, word embeddings reflect and propagate social biases present in training corpora (Caliskan et al., 2017; Garg et al., 2018). Machine learning systems that use embeddings can further amplify biases (Barocas and Selbst, 2016; Zhao et al., 2017), discriminating against users, particularly those from disadvantaged social groups.

(Bolukbasi et al., 2016) introduced a method to *debias* embeddings by removing components that lie in stereotype-related embedding subspaces. They demonstrate the effectiveness of the approach by removing gender bias from word2vec embeddings (Mikolov et al., 2013), preserving the utility of embeddings and potentially alleviating biases in downstream tasks. However, this method was only for *binary* labels (e.g., male/female), whereas most real-world demographic attributes,

---

* Equal contributions

† Work done while at CMU and The Microsoft AI Development Acceleration Program

| Racial Analogies | |
|---|---|
| black → criminal | caucasian → lawful |
| caucasian → hillbilly | asian → yuppie |
| asian → engineer | black → killer |
| **Religious Analogies** | |
| christian → conservative | jew → liberal |
| muslim → terrorist | jew → journalist |
| christian → conservative | muslim → regressive |

Table 1: Examples of racial and religious biases in analogies generated from word embeddings trained on the Reddit data from users from the USA.

including gender, race, religion, are not binary but continuous or categorical, with more than two categories.

In this work, we show a generalization of Bolukbasi et al.'s (2016) which enables *multiclass* debiasing, while preserving utility of embeddings (§3). We train word2vec embeddings using the Reddit L2 corpus (Rabinovich et al., 2018) and apply multiclass debiasing using lexicons from studies on bias in NLP and social science (§4.2). We introduce a novel metric for evaluation of bias in collections of word embeddings (§5). Finally, we validate that the utility of debiased embeddings in the tasks of part-of-speech (POS) tagging, named entity recognition (NER), and POS chunking is on par with off-the-shelf embeddings.

## 2 Background

As defined by (Bolukbasi et al., 2016), debiasing word embeddings in a binary setting requires identifying the bias subspace of the embeddings. Components lying in that subspace are then removed from each embedding.

### 2.1 Identifying the bias subspace

(Bolukbasi et al., 2016) define the gender subspace using *defining sets* of words, where the words in

each set represent different ends of the bias. For example, in the case of gender, one defining set might be the gendered pronouns {*he*, *she*} and another set might be the gendered nouns {*man*, *woman*}. The gender subspace is then computed from these defining sets by 1) computing the vector differences of the word embeddings of words in each set from the set's mean, and 2) taking the most significant components of these vectors.

## 2.2 Removing bias components

Following the identification of the gender subspace, one can apply hard or soft debiasing (Bolukbasi et al., 2016) to completely or partially remove the subspace components from the embeddings.

### Hard debiasing

Hard debiasing (also called "Neutralize and Equalize") involves two steps. First, bias components are removed from words that are not gendered and should not contain gender bias (e.g., *doctor*, *nurse*), and second, gendered word embeddings are centered and their bias components are equalized. For example, in the binary case, *man* and *woman* should have bias components in opposite directions, but of the same magnitude. Intuitively, this then ensures that any neutral words are equidistant to any biased words with respect to the bias subspace.

More formally, to *neutralize*, given a bias subspace $\mathcal{B}$ spanned by the vectors $\{b_1, b_2, ..., b_k\}$, we compute the component of each embedding in this subspace:

$$\mathbf{w}_\mathcal{B} = \sum_{i=1}^{k} \langle \mathbf{w}, b_i \rangle b_i \qquad (1)$$

We then remove this component from words that should be bias-neutral and normalize to get the debiased embedding:

$$\mathbf{w}' = \frac{\mathbf{w} - \mathbf{w}_\mathcal{B}}{\|\mathbf{w} - \mathbf{w}_\mathcal{B}\|} \qquad (2)$$

To *equalize* the embeddings of words in an equality set $E$, let $\boldsymbol{\mu} = \frac{1}{|E|} \sum_{\mathbf{w} \in E} \mathbf{w}$ be the mean embedding of the words in the set and $\boldsymbol{\mu}_\mathcal{B}$ be its component in the bias subspace as calculated in Equation 1. Then, for $\mathbf{w} \in E$,

$$\mathbf{w}' = (\boldsymbol{\mu} - \boldsymbol{\mu}_\mathcal{B}) + \sqrt{1 - \|\boldsymbol{\mu} - \boldsymbol{\mu}_\mathcal{B}\|^2} \frac{\mathbf{w}_\mathcal{B} - \boldsymbol{\mu}_\mathcal{B}}{\|\mathbf{w}_\mathcal{B} - \boldsymbol{\mu}_\mathcal{B}\|} \qquad (3)$$

Note that in both Equations 2 and 3, the new embedding has unit length.

### Soft debiasing

Soft debiasing involves learning a projection of the embedding matrix that preserves the inner product between biased and debiased embeddings while minimizing the projection onto the bias subspace of embeddings that should be neutral.

Given embeddings $\mathbf{W}$ and $\mathbf{N}$ which are embeddings for the whole vocabulary and the subset of bias-neutral words respectively, and the bias subspace $\mathcal{B}$ obtained in Section 2.1, soft debiasing seeks for a linear transformation $A$ that minimizes the following objective:

$$\|(\mathbf{AW})^\mathsf{T}(\mathbf{AW}) - \mathbf{W}^\mathsf{T}\mathbf{W}\|_F^2 \\ + \lambda\|(\mathbf{AN})^\mathsf{T}(\mathbf{A}\mathcal{B})\|_F^2 \qquad (4)$$

Minimizing the first term preserves the inner product after the linear transformation $\mathbf{A}$, and minimizing the second term minimizes the projection onto the bias subspace $\mathcal{B}$ of embeddings. $\lambda \in \mathbb{R}$ is a tunable parameter that balances the two objectives.

## 3 Methodology

We now discuss our proposed extension of word embedding debiasing to the multiclass setting.

## 3.1 Debiasing

As in the binary setting, debiasing consists of two steps: identifying the "bias subspace" and removing this component from the set of embeddings.

### Identifying the bias subspace

The core contribution of our work is in *identifying* the "bias subspace" in a multiclass setting; if we can identify the bias subspace then prior work can be used for multiclass debiasing.

Past work has shown that it is possible to linearly separate multiple social classes based on components of word embeddings (Garg et al., 2018). Based on this we hypothesize that there exists some component of these embeddings which can capture multiclass bias. While a multiclass problem is inherently not a linearly separable problem, a one versus rest classifier is. Following from this, the computation of a multiclass bias subspace does not have any linear constraints, though it does come with a loss of resolution. As a result we can compute the principal components

| Gender Debiasing | MAC | $p$-Value |
|---|---|---|
| Biased | 0.623 | N/A |
| Hard Debiased | 1.000 | 1.582e-14 |
| Soft Debiased ($\lambda = 0.2$) | 0.747 | 1.711e-12 |
| **Race Debiasing** | MAC | $p$-Value |
| Biased | 0.892 | N/A |
| Hard Debiased | 1.009 | 7.235e-04 |
| Soft Debiased ($\lambda = 0.2$) | 0.985 | 6.217e-05 |
| **Religion Debiasing** | MAC | $p$-Value |
| Biased | 0.859 | N/A |
| Hard Debiased | 1.004 | 3.006e-07 |
| Soft Debiased ($\lambda = 0.2$) | 0.894 | 0.007 |

Table 2: The associated mean average cosine similarity (MAC) (defined in Section 3.2) and $p$-Values for debiasing methods for gender, race, and religious bias.

required to compute the "bias subspace" by simply adding an additional term for each additional bias class to each defining set.

Formally, given defining sets of word embeddings $D_1, D_2, ..., D_n$, let the mean of the defining set $i$ be $\boldsymbol{\mu}_i = \frac{1}{|D_i|} \sum_{\mathbf{w} \in D_i} \mathbf{w}$, where $\mathbf{w}$ is the word embedding of $w$. Then the bias subspace $\mathcal{B}$ is given by the first $k$ components of the following principal component analysis (PCA) evaluation:

$$\mathbf{PCA}\left( \bigcup_{i=1}^{n} \bigcup_{\mathbf{w} \in D_i} \mathbf{w} - \boldsymbol{\mu}_i \right) \quad (5)$$

The number of components $k$ can be empirically determined by inspecting the eigenvalues of the PCA, or using a threshold. Also, note that the defining sets do not have to be the same size. We discuss the robustness of this method later.

**Removing Bias Components**

Following the identification of the bias subspace, we apply the hard Neutralize and Equalize debiasing and soft debiasing method presented in (Bolukbasi et al., 2016) and discussed in Section 2.2 to completely or partially remove the subspace components from the embeddings.

For equalization, we take the defining sets to be the equality sets as well.

### 3.2 Quantifying Bias Removal

We propose a new metric for the evaluation of bias in collections of words which is simply the mean average cosine similarity (MAC). This approach is motivated by the WEAT evaluation method proposed by (Caliskan et al., 2017) but modified for a multiclass setting. To compute this metric the following data is required: a set of target word embeddings $T$ containing terms that inherently contain some form of social bias (e.g. {*church*, *synagogue*, *mosque*}), and a set $A$ which contains sets of attributes $A_1, A_2, ..., A_N$ containing word embeddings that should not be associated with any word embeddings contained in the set $T$ (e.g. {*violent*, *liberal*, *conservative*}).

We define a function $S$ that computes the mean cosine distance between a particular target $T_i$ and all terms in a particular attribute set $A_j$:

$$S(\mathbf{t}, A_j) = \frac{1}{N} \sum_{\mathbf{a} \in A_j} cos(\mathbf{t}, \mathbf{a}), \quad (6)$$

where the cosine distance is:

$$cos(\mathbf{u}, \mathbf{v}) = 1 - \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}. \quad (7)$$

Finally, we define MAC as:

$$\text{MAC}(T, A) = \frac{1}{|T||A|} \sum_{T_i \in T} \sum_{A_j \in A} S(T_i, A_j) \quad (8)$$

We also perform a paired $t$-test on the distribution of average cosines used to calculate the MAC. Thus we can quantify the effect of debiasing on word embeddings in $T$ and sets in $A$.

### 3.3 Measuring Downstream Utility

To measure the utility of the debiased word embeddings, we use the tasks of NER, POS tagging, and POS chunking. This is to ensure that the debiasing procedure has not destroyed the utility of the word embeddings. We evaluate test sentences that contain at least one word affected by debiasing. Additionally, we measure the change in performance after replacing the biased embedding matrix by a debiased one, and retraining the model on debiased embeddings.

## 4 Data

In this section we discuss the different data sources we used for our initial word embeddings, the social bias data used for evaluating bias, and the linguistic data used for evaluating the debiasing process.

### 4.1 Embedding Language Corpus

We used the L2-Reddit corpus (Rabinovich et al., 2018), a collection of Reddit posts and comments

| | Embedding Matrix Replacement | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Hard Gender Debiasing | | | Hard Racial Debiasing | | | Hard Religious Debiasing | | |
| | NER Tagging | POS Tagging | POS Chunking | NER Tagging | POS Tagging | POS Chunking | NER Tagging | POS Tagging | POS Chunking |
| Biased F1 | 0.9954 | 0.9657 | 0.9958 | 0.9948 | 0.9668 | 0.9958 | 0.9971 | 0.9665 | 0.9968 |
| $\Delta$ F1 | +0.0045 | -0.0098 | +0.0041 | +0.0051 | -0.0117 | +0.0041 | +0.0103 | -0.0345 | +0.0120 |
| $\Delta$ Precision | 0.0 | -0.0177 | 0.0 | 0.0 | -0.0208 | 0.0 | 0.0 | -0.0337 | 0.0 |
| $\Delta$ Recall | +0.0165 | -0.0208 | +0.0156 | +0.0186 | -0.0250 | +0.0155 | +0.00286 | -0.0174 | +0.0031 |
| | Soft Gender Debiasing | | | Soft Racial Debiasing | | | Soft Religious Debiasing | | |
| | NER Tagging | POS Tagging | POS Chunking | NER Tagging | POS Tagging | POS Chunking | NER Tagging | POS Tagging | POS Chunking |
| Biased F1 | 0.9952 | 0.9614 | 0.9950 | 0.9946 | 0.9612 | 0.9946 | 0.9964 | 0.9616 | 0.9961 |
| $\Delta$ F1 | +0.0047 | -0.0102 | +0.0049 | +0.0053 | -0.0107 | +0.0053 | +0.0128 | -0.0242 | +0.0148 |
| $\Delta$ Precision | 0.0 | -0.0202 | 0.0 | 0.0 | -0.0223 | 0.0 | 0.0 | -0.0199 | 0.0 |
| $\Delta$ Recall | +0.0169 | -0.0198 | +0.0187 | +0.0193 | -0.0197 | +0.0202 | +0.0035 | -0.0112 | +0.0038 |
| | Model Retraining | | | | | | | | |
| | Hard Gender Debiasing | | | Hard Racial Debiasing | | | Hard Religious Debiasing | | |
| | NER Tagging | POS Tagging | POS Chunking | NER Tagging | POS Tagging | POS Chunking | NER Tagging | POS Tagging | POS Chunking |
| Biased F1 | 0.9954 | 0.9657 | 0.9958 | 0.9948 | 0.9668 | 0.9958 | 0.9971 | 0.9665 | 0.9968 |
| $\Delta$ F1 | +0.0045 | -0.0137 | +0.0041 | +0.0051 | -0.0165 | +0.0041 | +0.0103 | -0.0344 | +0.0120 |
| $\Delta$ Precision | 0.0 | -0.0259 | 0.0 | 0.0 | -0.0339 | 0.0 | 0.0 | -0.0287 | 0.0 |
| $\Delta$ Recall | +0.0165 | -0.0278 | +0.0156 | +0.0186 | -0.0306 | +0.0156 | +0.00286 | -0.0161 | +0.0031 |
| | Soft Gender Debiasing | | | Soft Racial Debiasing | | | Soft Religious Debiasing | | |
| | NER Tagging | POS Tagging | POS Chunking | NER Tagging | POS Tagging | POS Chunking | NER Tagging | POS Tagging | POS Chunking |
| Biased F1 | 0.9952 | 0.9614 | 0.9950 | 0.9946 | 0.9612 | 0.9946 | 0.9964 | 0.9616 | 0.9961 |
| $\Delta$ F1 | +0.0047 | +0.00178 | +0.0049 | +0.0053 | -0.00119 | +0.0053 | +0.0128 | -0.0098 | +0.0148 |
| $\Delta$ Precision | 0.0 | +0.0048 | 0.0 | 0.0 | -0.00187 | 0.0 | 0.0 | -0.0125 | 0.0 |
| $\Delta$ Recall | +0.0169 | +0.00206 | +0.0187 | +0.0193 | -0.00264 | +0.0202 | +0.0035 | -0.0057 | +0.0038 |

Table 3: The performance of embeddings the downstream tasks of NER, POS Tagging, and POS Chunking.

by both native and non-native English speakers. The native countries of post authors are determined based on their posts in country- and region-specific subreddits (such as r/Europe and r/UnitedKingdom), and other metadata such as user flairs, which serve as self-identification of the user's country of origin.

In this work, we exclusively explore data collected from the United States. This was done to leverage extensive studies of social bias done in the United States. To obtain the initial biased word embeddings, we trained word2vec embeddings (Mikolov et al., 2013) using approximately 56 million sentences.

### 4.2 Social Bias Data

We used the following vocabularies and studies to compile lexicons for bias detection and removal.[1]

For gender, we used vocabularies created by (Bolukbasi et al., 2016) and (Caliskan et al., 2017).

For race we consulted a number of different sources for each race: Caucasians (Chung-Herrera and Lankau, 2005; Goad, 1998); African Americans (Punyanunt-Carter, 2008; Brown Givens and Monahan, 2005; Chung-Herrera and Lankau, 2005; Hakanen, 1995; Welch, 2007; Kawai, 2005); and Asian Americans (Leong and Hayes, 1990; Lin et al., 2005; Chung-Herrera and Lankau, 2005; Osajima, 2005; Garg et al., 2018).

Finally, for religion we used the following sources and labels: Christians (Rios et al., 2015; Zuckerman, 2009; Unnever et al., 2005); Jews (Dundes, 1971; Fetzer, 2000); and Muslims (Shryock, 2010; Alsultany, 2012; Shaheen, 1997).

### 4.3 Downstream Tasks

We evaluate biased and debiased word embeddings on several downstream tasks. Specifically, the CoNLL 2003 shared task (Tjong Kim Sang and De Meulder, 2003) which provides evaluation data for NER, POS tagging, and POS chunking.

## 5 Results and Discussion

In this section we review the results of our experiments and discuss what those results mean in the context of this work.

### 5.1 Observations of Bias

We use the analogy task from (Bolukbasi et al., 2016) to demonstrate that bias exists in these word embeddings. In order to construct our analogies we trained five word2vec (Mikolov et al., 2013) embedding spaces on the same data. We then constructed a set of analogies for each embedding space taking the intersection of these to form a working set of analogies. We then reversed each analogy to predict the source word from the generated word (e.g. attempting to generate the analogy "caucasian is to lawful as black is to criminal" having observed the analogy "black is to criminal as caucasian is to lawful"). We refined our working

set by removing analogies that failed to generate the source term in the top ten candidates. We performed these additional tests in order to ensure the analogies were robust to perturbations in the embedding space, and the analogies themselves were robust in both directions. Following this analysis we observe that bias is present in generated analogies by viewing them directly. A small subset of these analogies are in Table 1 to highlight our findings.

## 5.2 Removal of Bias

We perform our debiasing in the same manner as described in Section 3.1 and calculate the MAC scores and $p$-values to measure the effects of debiasing. Results are presented in Table 2.

*Does multiclass debiasing decrease bias?* We see that this debiasing procedure categorically moves MAC scores closer to 1.0. This indicates an increase in cosine distance. Further, the associated P-values indicate these changes are statistically significant. This demonstrates that our approach for multiclass debiasing decreases bias.

## 5.3 Downstream Effects of Bias Removal

The effects of debiasing on downstream tasks are shown in Table 3. Debiasing can either help or harm performance. For POS tagging there is almost always a decrease in performance. However, for NER and POS chunking, there is a consistent increase. We conclude that these models have learned to depend on some bias subspaces differently. Note that many performance changes are of questionable statistical significance.

*Does multiclass debiasing preserve semantic utility?* We argue the minor changes in Table 3 support the preservation of semantic utility in the multiclass setting, especially compared to gender debiasing which is known to preserve utility (Bolukbasi et al., 2016).

*Is the calculated bias subspace robust?* The bias subspace is at least robust enough to support the above debiasing operations. This is shown by statistically significant changes in MAC scores.

## 6 Limitations & Future Work

Calculating multiclass bias subspace using our proposed approach has drawbacks. For example, in the binary gender case, the extremes of bias subspace reflect extreme male and female terms. However, this is not possible when projecting multiple classes into a linear space. Thus, while we can calculate the magnitude of the bias components, we cannot measure extremes of each class.

Additionally, the methods presented here rely on words that represent biases (defining sets) and words that should or should not contain biases (equality sets). These lists are based on data collected specifically from the US. Thus, they may not translate to other countries or cultures. Further, some of these vocabulary terms, while peer reviewed, may be subjective and may not fully capture the bias subspace.

Recent work by Gonen and Goldberg (2019) suggests that debiasing methods based on bias component removal are insufficient to completely remove bias in the embeddings, since embeddings with similar biases are still clustered together after bias component removal. Following Gonen and Goldberg's (2019) procedure, we plot the number of neighbors of a particular bias class as a function of the original bias, before and after debiasing in Figure 1 and 2 in the Appendix. In line with Gonen and Goldberg's (2019) findings, simply removing the bias component is insufficient to remove multiclass "cluster bias". However, increasing the size of the bias subspace reduces the correlation of the two variables (Table 4 in the Appendix).

## 7 Conclusion

We showed that word embeddings trained on `www.reddit.com` data contain multiclass biases. We presented a novel metric for evaluating debiasing procedures for word embeddings. We robustly removed multiclass bias using a generalization of existing techniques. Finally, we showed that this multiclass generalization preserves the utility of embeddings for different NLP tasks.

## Acknowledgments

# References

Evelyn Alsultany. 2012. *Arabs and Muslims in the Media: Race and Representation after 9/11*. nyu Press.

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

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. In *Proc. of NIPS*, pages 4349–4357.

Sonja M Brown Givens and Jennifer L Monahan. 2005. Priming mammies, jezebels, and other controlling images: An examination of the influence of mediated stereotypes on perceptions of an african american woman. *Media Psychology*, 7(1):87–106.

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

Beth G Chung-Herrera and Melenie J Lankau. 2005. Are we there yet? an assessment of fit between stereotypes of minority managers and the successful-manager prototype. *Journal of Applied Social Psychology*, 35(10):2029–2056.

Alan Dundes. 1971. A study of ethnic slurs: The jew and the polack in the united states. *The Journal of American Folklore*, 84(332):186–203.

Joel S Fetzer. 2000. *Public attitudes toward immigration in the United States, France, and Germany*. Cambridge University Press.

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

Jim Goad. 1998. *The Redneck Manifesto: How Hillbillies Hicks and White Trash Becames America's Scapegoats*. Simon and Schuster.

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.

Ernest A Hakanen. 1995. Emotional use of music by african american adolescents. *Howard Journal of Communications*, 5(3):214–222.

Yuko Kawai. 2005. Stereotyping asian americans: The dialectic of the model minority and the yellow peril. *The Howard Journal of Communications*, 16(2):109–130.

Frederick TL Leong and Thomas J Hayes. 1990. Occupational stereotyping of asian americans. *The Career Development Quarterly*, 39(2):143–154.

Monica H Lin, Virginia SY Kwan, Anna Cheung, and Susan T Fiske. 2005. Stereotype content model explains prejudice for an envied outgroup: Scale of anti-asian american stereotypes. *Personality and Social Psychology Bulletin*, 31(1):34–47.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Keith Osajima. 2005. Asian americans as the model minority: An analysis of the popular press image in the 1960s and 1980s. *A companion to Asian American studies*, pages 215–225.

Narissra M Punyanunt-Carter. 2008. The perceived realism of african american portrayals on television. *The Howard Journal of Communications*, 19(3):241–257.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. In *Proc. of the Transactions of Association for Computational Linguistics*.

Kimberly Rios, Zhen Hadassah Cheng, Rebecca R Totton, and Azim F Shariff. 2015. Negative stereotypes cause christians to underperform in and disidentify with science. *Social Psychological and Personality Science*, 6(8):959–967.

Jack G Shaheen. 1997. *Arab and Muslim stereotyping in American popular culture*. Center for Muslim-Christian Understanding, History and International Affairs .

Andrew Shryock. 2010. *Islamophobia/Islamophilia: Beyond the politics of enemy and friend*. Indiana University Press.

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-Volume 4*, pages 142–147. Association for Computational Linguistics.

James D Unnever, Francis T Cullen, and Brandon K Applegate. 2005. Turning the other cheek: Reassessing the impact of religion on punitive ideology. *Justice Quarterly*, 22(3):304–339.

Kelly Welch. 2007. Black criminal stereotypes and racial profiling. *Journal of Contemporary Criminal Justice*, 23(3):276–288.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proc. of EMNLP*, pages 2979–2989.

Phil Zuckerman. 2009. Atheism, secularity, and well-being: How the findings of social science counter negative stereotypes and assumptions. *Sociology Compass*, 3(6):949–971.

| Target | $k$ | $r$ | $\rho$ |
|---|---|---|---|
| *jew* | 0 | 0.767 | 0.875 |
| | 1 | 0.795 | 0.891 |
| | 2 | 0.718 | 0.756 |
| | 3 | 0.736 | 0.772 |
| *christian* | 0 | 0.925 | 0.947 |
| | 1 | 0.835 | 0.841 |
| | 2 | 0.825 | 0.831 |
| | 3 | 0.832 | 0.839 |
| *muslim* | 0 | 0.858 | 0.894 |
| | 1 | 0.774 | 0.812 |
| | 2 | 0.715 | 0.721 |
| | 3 | 0.712 | 0.718 |

Table 4: Pearson's $r$ and Spearman's $\rho$ correlation coefficients between the number of biased neighbors and the original bias of professions with respect to target classes for religion. $k$ is the dimension of the bias subspace used ($k = 0$ represents the original embedding). All correlation coefficients have $p$-values $< 10^{-30}$.

## A   Addressing Cluster Bias

To visualize the degree of cluster bias before and after our debiasing procedure, we follow a similar procedure to Gonen and Goldberg (2019). For a defining set $D$ for the target task (e.g. religion, race, gender), we compute the mean embedding $\boldsymbol{\mu} = \frac{1}{|D|}\sum_{\mathbf{c}\in D}\mathbf{c}$. Then, for each class $\mathbf{c}$ in the defining set, we define the bias direction as $\mathbf{b} = \frac{\mathbf{c}-\boldsymbol{\mu}}{\|\mathbf{c}-\boldsymbol{\mu}\|}$. Using this, we find the 500 most biased words in each direction in the whole vocabulary based on their component in the bias direction: $\langle\mathbf{w}, \mathbf{b}\rangle$.

Then, using the list of professions from Bolukbasi et al. (2016)[2], we find the 100 closest neighbors for each profession. We then plot the number of neighbors with positive bias against the original bias of the profession word, as shown in Figures 1 and 2. The plots suggest that while the correlation between the bias component and the number of positively-biased neighbors might decrease slightly as the number of bias subspace dimensions increase, the cluster bias is still not fully removed. As Table 4 shows, while the correlation between the two quantities decreases as the number of subspace dimensions increase to 2 or 3, its magnitude is still high.
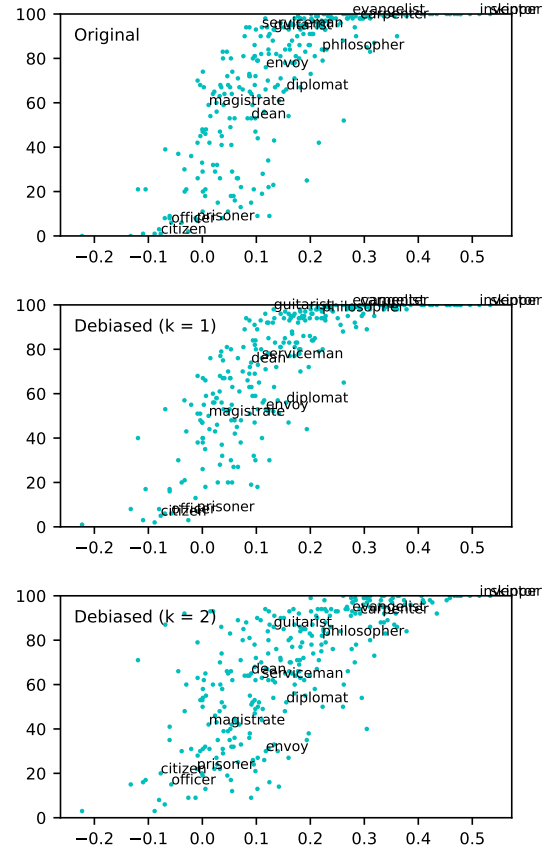


Figure 1: Plots of number of neighbors to *jew* for each profession as a function of its original bias with respect to *jew*, before and after debiasing, for different subspace dimensionalities $k$.
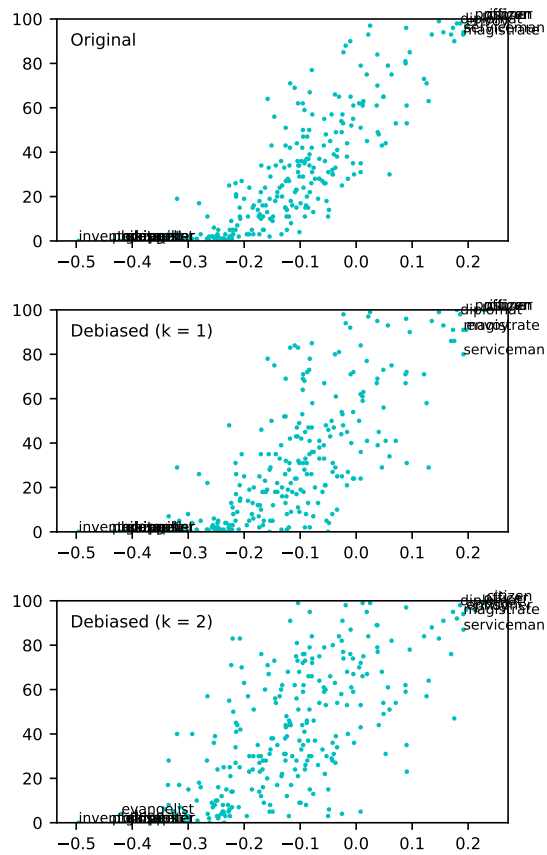
Figure 2: Plots of number of neighbors to *muslim* for each profession as a function of its original bias with respect to *muslim*, before and after debiasing, for different subspace dimensionalities $k$.