All Names Matter: Mitigating Racial Bias in Word Embeddings with Last Name-Based Counterfactual Data Substitution

Aditya Mengani
School of Information
University of California, Berkeley

Gabriela May Lagunes
School of Information
University of California, Berkeley

Abstract

In this project implemented Counterfactual Data Substitution with Name Intervention (Maudslay et al., 2019) to a Wikipedia corpus in order to mitigate racial bias in the resulting word embeddings. We used word2vect to train the embeddings and used USA census data to find the most common last names in the country per racial or ethnic group. We found that it was possible to take this methodology to tackle racial bias. This is relevant because it was previously used to mitigate only gender bias. Moreover, we showed that this was a successful process even when framed as a multi-class problem for race. Finally, we proved that the performance of the debiased embeddings remained the same before and after debasing for NLP tasks such as sentiment analysis and word analogy. All the code, data sets, embeddings, models and analysis generated and used in this project can be found in the Git repository of the project.

1 Introduction

Word embeddings are distributed representations of words as vectors, and they are an important element of several natural language processing (NLP) models [1]. Nevertheless, it has been proven that embeddings carry significant human-like semantic biases, such as gender and racial discrimination [2]. Such biases propagate NLP tasks and products that use them, therefore it is important work to mitigate these biases.

Important efforts have been done to reduce gender bias, and each methodology is built around the way bias is defined [3]. This dependency on the operationalisation of bias is a challenge because certain definitions can limit the way we identify bias within an embedding. Gonen et al. (2019)[3] showed this by evaluating some of the main gender debiasing methodologies and proving that they were successful at diminishing direct bias, i.e. the distance of a non gendered word (like *nurse*) to both male and female subspaces, but kept indirect bias, i.e. the clustering of non gendered words in groups that are stereotypically female (like *nurse* or *delicate*) or male (like *business* or *engineer*).

In the case of race this challenge becomes even more complex. Addressing gender bias can be framed as a binary problem. In addition, there are specific parts of English speech that can help us identify, and therefore work with, gender, e.g. pronouns (*he*, *she*), gendered word pairs (*king*, *queen*) and even names (*John*, *Mary*). Race, on the other hand, is a multi-class problem and the relevant classes depend on the language and the region of the world that we are looking at [4]. In addition, unlike the case of familial or occupational gendered word pairs (*mom-dad*, *actor-actress*), there are no obvious ways to differentiate these roles based on race (e.g. you do not have a different word for professor when you are being thought by a White, Black, Asian or Hispanic person).

In this project we aim to modify and implement a successful methodology for gender bias mitigation in order to tackle racial bias.

2 Background

After reviewing the literature, we found two main trends of gender bias operationalisation.

2.1 Word Embedding Debiasing (WED): The Race and Gender Directions

The first trend was started by the work of Bulokbasi et al. [5], where gender bias was defined as the projection of each word on the gender direction (the difference between the *he* and *she* vectors). Using an already trained embedding, Bolukbasi modified the values of word vectors to reduce this projection in words that are not gendered. Also, he took gendered word pairs and got all neutral words

to have the same distance to each word of these pairs [5].

Zhao et al. (2018) took this definition of bias and implemented a different methodology. They changed the loss of the GloVe model concentrating most of the gender information in the last coordinate of each vector. This allowed them to later use the word representation without the gender coordinate [3]. Finally, they made neutral words being orthogonal to the gender direction [6]. The most relevant contribution of this work is the proposition that bias is better handled when tacked before training the embedding, and not after. Kaneko [7] took the idea of relying on orthogonality even further and proposed to separate female, male, neutral and stereotypical words in orthogonal subspaces. They used feed forward neural networks and seed words to divide the whole vocabulary in those categories.

These methodologies are successful at mitigating direct gender bias while keeping the features that make embeddings useful for NLP tasks. Nevertheless, Gonen et al. showed that Bolukbasi's and Zhao's experiments were not capable of addressing indirect bias [3]. We reproduced Gonen's evaluation on Kaneko's *debiased* embedding and found similar results. Moreover, Manzini et al. successfully applied Bolukbasi's framework and transformed it into a multi-class problem with the aim of mitigating race bias in their word embeddings. They were able to diminish direct racial bias, but indirect bias remain intact. Manzini's group showed this by reproducing Gonen's experiments in their own results [8]. The main results of Gonen's indirect bias experiments in Bolukbasi's and Zhao's embeddings, our results in Kaneko's embedding and Manzini's results in their own embedding can be found in the appendix section.

2.2 Counterfactual Data Augmentation and Substitution

Counterfactual Data Augmentation (CDA) and Substitution (CDS) are techniques in which a transformation to the original corpus is designed and implemented in order to invert the bias that is being targeted. In CDA the text transformation is appended to the corpus [9]. This means that the original text is duplicated, therefore there can be unknown statistical properties in the resulting embedding [10]. For this reason, Maudslay et al.[10] propose a change to apply the text transformation probabilistically, i.e. they apply half the possible transformations over the original corpus instead of duplicating it, and hence define CDS.

Maudslay et al.[10] applied Bulokbasi's WED, Lu's CDA and their proposed CDS to two different corpora to evaluate the mitigation of direct and indirect gender bias, as well as the performance of the resulting embeddings in different NLP tasks [10]. In the CDA and CDS implementation they included a *Name Intervention*. In Lu et al.[9]'s methodology a list of gendered pair words were swapped every time they appeared in the text. Nevertheless, when the words were linked to a first name, they remained untouched. Similarly, where the only gendered word of the sentence was a first name, it would also remained unchanged. Maudslay et al.[10] constructed first names pairs based on the number of gender occurrences in the United States Social Security Administration (SSA) dataset from 1879 to date, and on its frequency in the original corpus. They showed that resulting embeddings from CDA and CDS with name interventions were the only embeddings that were free from both direct and indirect bias, and CDS with name intervention performed better at NLP tasks [10].

Maudslay proposed at the end of the paper that CDS with a name intervention could be implemented to tackle racial bias given than names have been used as a psychological proxy for race [10]. Nevertheless, this implementation was expected to be challenging. To our knowledge, this has not been done to date. Therefore, we decided to modify Maudslay et al.[10] methodology to tackle racial bias. The followed process and experiments were as follows.

3 Methods

3.1 Counterfactual Data Substitution for Racial Bias Mitigation

We decided to implement CDS with name intervention to a Wikipedia corpus to try to tackle racial bias. We got the latest Wikipedia dump and used articles with at least 50 words on them, giving us a total of 5,009,203 articles as corpus [11].

The first step was to define the word pairs that were going to be swap during the CDS corpus transformation. We used last names instead of first names because it was possible to find census data which correlated a list of last names with the amount of occurrences of self-identification for racial and ethnic groups. We considered the groups labeled as White, Black, Hispanic and Asian for

our experiments. We use a 2010 census dataset that contained over 150,000 surnames and the times a person self-identified as one of the used racial groups [12]. In addition to the quantity and quality of last names data available, by using last names we did not had to consider gender at the moment of implementing the transformation in the corpus.

Cleaning the data was a challenging. Maudslay's group used Name Entity Recognition (NER) in order to identify first names that were also common nouns and discard them [10]. We found that last names could be not only nouns, but verbs, adjectives, nationalities, languages and names of places like countries, cities and regions. Therefore, even with NER, the dataset needed to be manually revised.

In addition to last names, it was necessary to form word pairs that could be used to identify the 4 racial groups we were working with. This was another challenge to address, because the English language does not have components that can be obviously used to identify people of different races. For instance, there are not different pronouns that can be used or different ways to call an actor or a family member depending on the race of the person we are talking about, so gendered pairs such as *he-she*, *actor-actress* or *dad-mom* do not have an equivalent in the case of race. We first thought of using things like nationalities, languages and places that represented the different groups, but we were concerned about the effect swapping these pairs could have in embedding capabilities on NLP tasks, such as analogies like *Paris is to France* like *Tokyo is to Japan*. To tackle this we decided to use ethnic slur, which are derogatory terms targeted to an specific ethnic or racial group in different contexts [13]. We thought that by swapping the derogatory or stereotypical terms related to specific racial groups, then the negative connotations targeted to one group would be closer to the other. Once we had the final list of words (last names and slur) for CDS, we created 7 different sets of word pairs. The first three were *asian-white*, *black-white* and *hispanic-white*. We formed the pairs of surnames using all of those with a minimum 50% racial self-identification occurrences. Then we matched last names by finding the pairs with a minimum Euclidean distance in the racial percentage - corpus appearance frequency plane. The slur pairs were made based only in their frequency on the corpus because each of them target a single racial group. The resulting word-pairs lists had between 7,000 and 5,000 unique pairs.

We also wanted to explore the trade off between bias mitigation and the performance of the embedding. We thought that a higher number of transformations over the corpus could lead to a higher level of bias mitigation, but the changes in embedding performance for NLP tasks could be more noticeable, and not necessarily in a good way. Therefore, we created other three sets of word pairs, one for each racial pair, but using only the 2500 more frequent last names per race on the corpus. We kept all racial slur words for both cases. Finally, we wanted to see if it was possible to use this technique to tackle racial bias against more than one racial group. Therefore, we created a last set of word pairs by combining each of the 2500 last names for Asian, Black and Hispanic people and matching it with an extended list of 7500 White last names.

3.2 Experiment Set Up

With the resulting word pairs we applied probabilistic CDS (with probability of 0.5) and created 8 embeddings using the Wikipedia corpus described above and word2vec. We labeled the resulting embeddings as **baseline**, the embedding resulting from the Wikipedia corpus with no pre-training transformations, **asian-white**, **black-white**, **hispanic-white**, the embeddings resulting of applying CDS with each of the first 3 sets of word pairs, **asian-white-2500**, **black-white-2500**, **hispanic-white-2500**, the embeddings resulting from CDS with the second, shorter sets of word pairs, and **all-white**, the last embedding with the extended list of the three racial categories against White last names. With each of these embeddings we implemented experiments to evaluate the change on *direct bias* and *indirect bias* against the baseline, as well as the change in performance on *sentiment analysis* and *analogies* experiments in order to evaluate the embeddings performance in different NLP tasks compared to the baseline.

All the code, data sets, embeddings, models and analysis generated and used in this project can be found in the Git repository of the project.

3.3 Direct Racial Bias

We implemented a modified version of Caliskan et al.[2]'s Word Embedding Association Test (WEAT). The test computes Cohen's d, a measurement of the difference of relative similarities between two sets of target words X and Y, and

two sets of attribute words *A* and *B* (a higher Cohen's d value means a higher bias). Maudslay used three tests proposed by Nosek et al. [14] which measure the strength of various gender stereotypes: art–maths, arts–sciences, and careers–family. In this case, the target words *X* and *Y* were male and female first names, and the attribute words *A* and *B* were words related to each of the fields or subjects that each test was evaluating [10].

In our version of the test we recreated the target words *X* and *Y* with the most frequent last names per racial group, and the attribute word sets *A* and *B* to measure stereotypes against racial minorities in the United States. According to social science's literature racial minorities in the United States are more likely to be linked to negative or disadvantaged concepts, such as blue collar work, negative attributions such as laziness, aggressiveness or unreliability, or illegal activity, such as drug dealing or abuse, prostitution, illegal immigration and homelessness [15][16]. Therefore we created three new attribute word groups: *white collar professions - blue collar professions* with words like 'doctor' and 'janitor', *positive adjectives - negative adjectives* with words like 'educated' and 'uneducated', *lawful status - criminal status* with words like 'patriot' and 'prostitute'.

3.4 Indirect Racial Bias

To evaluate the change in indirect racial bias we reproduced Gonen's cluster experiments (2019). For each embedding we got the 500 most biased words, clustered them in two clusters using k-means and compared with the baseline. In addition, we also plotted the number of white neighbors for a list of professions as a function of its original racial bias before and after debasing, and calculated its Pearson coefficient [3].

3.5 Sentiment Analysis

To evaluate sentiment analysis performance in the different embeddings, we did a smaller version of the Maudslay's experiment [10]. We took a sample of 6323 words with sentiment score from -1 to 1 taken from Speer (2017). We used 90% of the words to train a logistic regression classifier, and the rest to were used as a test set to evaluate the accuracy of the models trained with the different embeddings.

3.6 Analogies

...

Figure 1: Direct racial bias experiments results. The final average Cohen's d values marked in blue are those where all d values were lower than the baseline for all categories, and those marked in red are averages that are lower than the baseline average, but that got one or two categories with greater d than the baseline.

Race Pairing	Embedding	Cohen's d			Average d
		Professions	Adjectives	Status	
Asian - White	Baseline	**0.385**	**0.969**	**0.962**	**0.772**
	Debiased with all names	0.758	0.941	0.634	0.778
	Debiased with 2500 names	0.423	0.495	0.282	0.400
Black - White	Baseline	**0.022**	**0.567**	**0.976**	**0.522**
	Debiased with all names	0.712	0.065	0.114	0.297
	Debiased with 2500 names	0.004	0.994	0.026	0.341
Hispanic - White	Baseline	**0.631**	**0.759**	**0.974**	**0.788**
	Debiased with all names	0.469	0.279	0.630	0.459
	Debiased with 2500 names	0.185	0.151	0.152	0.163
All - White	Baseline	**0.529**	**0.994**	**0.999**	**0.841**
	Debiased	0.535	0.516	0.440	0.497

4 Results and Discussion

4.1 Direct Racial Bias

Figure 3 shows the resulting values of Cohen's d for each of the direct bias experiments for each of the embeddings. We can see that the average value of Cohen's d is lower than the baseline for all cases except for **asian-white**. A possible explanation could be that this embedding was the one with the largest number of swapped word pairs for a single race during corpus pre-processing, therefore, there could have been some added noise that stopped the corpus transformation from tackling the bias as expected. This is supported by the results of **asian-white-2500** where the average d diminishes significantly and that the result for professions improves in comparison to the **asian-white** embedding.

In the case of **black-white** embeddings, we can see that the average d improves for both cases, but that the value of d increases with respect to the baseline for Professions in the case of **black-white** and for Adjectives in **black-white-2500**. A potential cause for this is that some of the most popular Black last names were occupations, like *Barker* or *Cooper*, and negative adjectives, like *Awkward*. It is possible that some of these names had stayed in the word-pairs formed to apply CDS, which indicates more systematic efforts need to be done to clean the data before applying CDS. It is important to note that the d values for the Status category decreased significantly for both cases, which means

there is a much weaker relationship between Black last names and criminal concepts after debiasing.

For **hispanic-white** embeddings we can see that both of the average d values, and all d values for all categories decreased. We believe this was because the Hispanic set of last names had the lesser occurrences of nouns, adjectives, verbs or other kind of words that were not desirable to have in the word pairs lists. In addition, in the cases were names of places were also used as last names, those places were mostly written in Spanish, and were cities or regions from Latin American countries or from Spain, like *Apodaca* or *Salamanca*. If examples of such last names stayed in the set they probably had a less negative effect than their English counterparts. Finally, we can see that the average d value of **hispanic-white-2500** is lower than **hispanic-white**, supporting the idea that a smaller amount word pairs can have better effects in the corpus transformation.

Finally, we can see that the average d value of the **all-white** embedding decreased, which indicates that it is possible to use CDS with last name intervention to tackle multi-class racial bias. It is interesting to note that the average d value of the baseline for this experiment is greater than all the other baseline values, which indicate a higher initial bias. This is probably because by using the top word pairs per race we got more occurrences of transformations with a smaller number of examples.

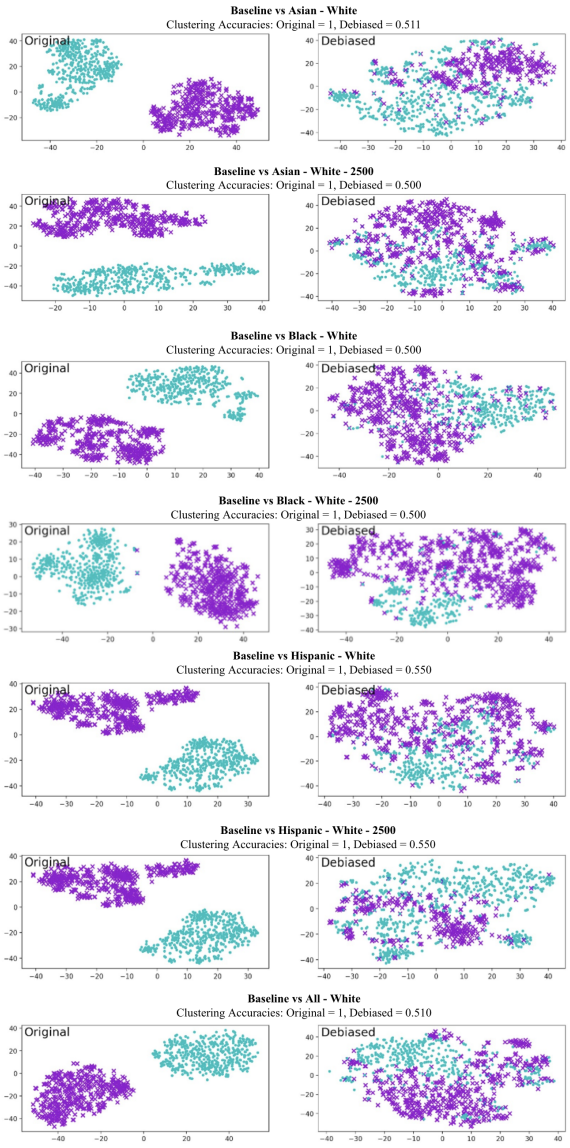
4.2 Indirect Racial Bias

Figure 2 shows the results of the indirect racial bias experiments based on the K-mean clustering of the most biased words before and after debiasing for each embedding. For all cases the accuracy of clustering went from 1 to between 0.5 and 0.55. Similarly, figure 3 shows that the correlation between professions and race significantly lowers after debiasing for all cases. This means that it was possible to tackle indirect racial bias with our methodology, and that it is possible to do this as a multi-class problem.

4.3 Sentiment Analysis

Figure 4 shows the results of the sentiment analysis experiments for all embeddings. We can see that the accuracy of the models train with all embeddings are above 90% and are close to the baseline. This means that all embeddings maintain their capability of performing well when applied to sentiment analysis tasks.

Figure 2: Indirect racial bias experiments results - K-means Clusters for most biased words.



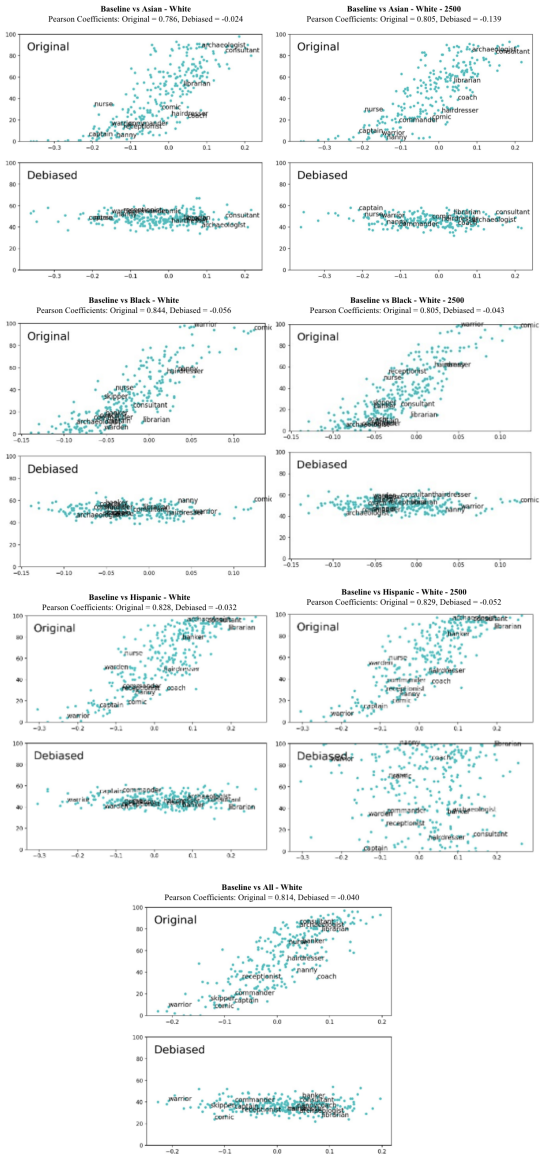
4.4 Analogies

...

5 Conclusion

In this project we modified Conterfactual Data Substitution with name intervention, a methodology presented by Maudslay et al.[10], to tackle gender bias within word embeddings, and implemented our resulting methodology to tackle racial bias, both as a binary and a multi-class problem. We saw that using last names instead of first names allowed us to tackle both direct and indirect racial bias while keeping the performance of the debi-

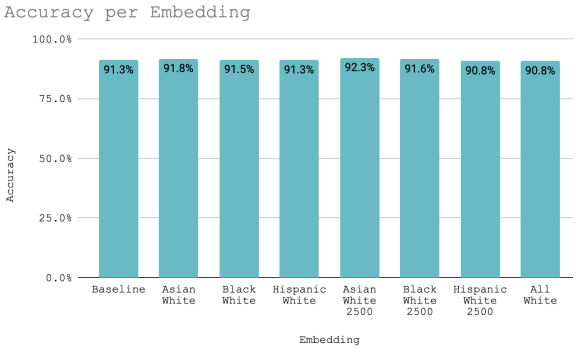
Figure 3: Indirect racial bias experiments results - Number of white neighbors for each profession as a function of its original bias.



ased embeddings in NLP tasks such as sentiment analysis and word analogies.

Since our work showed that it is indeed possible to diminish racial bias within word embeddings with this methodology, it is important that future work aims to have a more comprehensive approach to last name data cleaning, to explore even further the balance between the number of applied transformations to the corpus and the resulting bias mitigation, and to evaluate in a more systematic way the resulting embedding's performance in NLP tasks. This can allow future work to develop an efficient embedding with racial bias mitigated for several groups.

Figure 4: Sentiment analysis experiments results



Finally, it is important to note that here we just addressed racial bias against minorities in the United States, and that similar efforts should be done to tackle racial bias in the rest of the world.

References

[1] Thanapon Noraset, Chen Liang, Larry Birnbaum, Doug Downey. *Definition Modeling: Learning to Define Word Embeddings in Natural Language*. AAAI Conference, 2017.

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

[3] Hila Gonen and Yoav Goldberg. *Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them*. Proceedings of NAACL-HLT 2019, pages 609–614. 2019

[4] Kamelia Angelova. *MAPS: A Complete Guide To National Stereotypes All Around The World*. Business Insider Australia. 2011.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai. *Man is to Computer Programmer as Woman is to Homemaker?Debiasing Word Embeddings*. arXiv:1607.06520. 2016.

[6] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai- Wei Chang. *Learning gender-neutral word embeddings*. In Proceedings of EMNLP, pages 4847–4853. 2018.

[7] Masahiro Kaneko, Danushka Bollegala. *Learning gender-neutral word embeddings*. arXiv:1906.00742v1 [cs.CL]. 2019.

[8] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, Alan W Black. *Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings*. arXiv:1904.04047v3 [cs.CL]. 2019.

[9] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Aman- charla, and Anupam Datta. *Gender bias in neural natural language processing*. CoRR, abs/1807.11714. 2018.

[10] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, Simone Teufel. *It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution*. arXiv:1909.00871v3 [cs.CL]. 2019.

[11] Wikipedia. *Wikipedia Data Dumps*. https://dumps.wikimedia.org/enwiki/latest/

[12] United States Census Bureau. *2010 Census Data*. https://www.census.gov/data/developers/datasets/surnames.html

[13] Wikipedia. *List of Ethnic Slurs*. https://en.wikipedia.org/wiki/List_of_ethnic_slurs

[14] Brian A. Nosek, Mahzarin R. Banaji, and Anthony G Greenwald. *Harvesting implicit group attitudes and beliefs from a demonstration web site*. Group Dynamics: Theory, Research, and Practice, 6 1:101–115. 2002.

[15] Taylor, E., Guy-Walls, P., Wilkerson, P. et al. *The Historical Perspectives of Stereotypes on African-American Males*. J. Hum. Rights Soc. Work 4, 213–225. 2019. https://doi.org/10.1007/s41134-019-00096-y

[16] Naomi Priest ,Natalie Slopen, Susan Woolford, Jeny Tony Philip, Dianne Singer, Anna Daly Kauffman, Kathryn Mosely, Matthew Davis, Yusuf Ransome, David Williams. *Stereotyping across intersections of race and age: Racial stereotyping among White adults working with children*.2018. https://doi.org/10.1371/journal.pone.0201696

A Appendices

Figure 5: From top to bottom: Resulting K-means clusters of the 1000 most gendered biased words of Bulokbasi's and Zhao's embeddings before and after debiasing taken from (Gonen et al. 2019), and from Kaneko's embedding generated by us using Gonen's code (Gonen et al. 2019)(Kaneko et al., 2019)

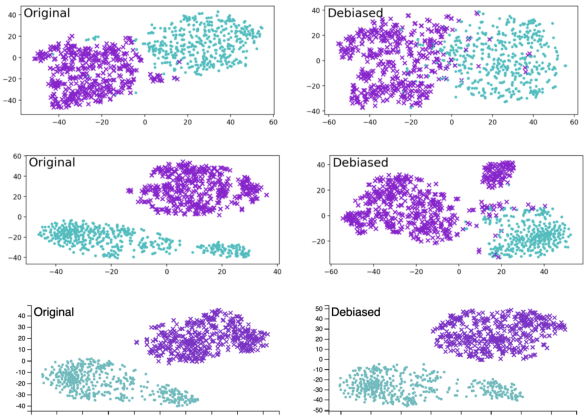


Figure 6: From top to bottom: The number of male neighbors for each profession as a function of its original bias, before and after debiasing for Bulokbasi's and Zhao's embeddings taken from (Gonen et al. 2019), for Kaneko's embeddings generated by us, and the number of Jewish neighbors for each profession as a function of its original bias, before and after debiasing, taken from (Manzini et al., 2019)