Investigating the Effect of Debiasing Methods on Intersectional Biases in Language Models

Stanford CS224N Custom Project

Ellie Talius
Department of Computer Science
Stanford University
etalius@stanford.edu

Ananya Karthik
Department of Computer Science
Stanford University
ananya23@stanford.edu

Abstract

Previous work in debiasing language models has focused on removing one form of bias, such as racial or gender bias, but no current work has attempted to reduce intersectional bias, which is bias resulting from being in more than one marginalized group. We propose three different methods to reduce intersectional bias in the BERT-Tiny model through fine-tuning in which we de-bias the contextual word embeddings of the BERT-Tiny model. We evaluate our models using intersectional embedding association tests across race, gender, and age. Our de-biasing methods are able to successfully reduce intersectional bias in all three intersectional identities tested. We find that Method 1 results in the best performance for less extreme intersectionality tests, while Methods 2 and 3 perform better for the more extreme intersectionality tests (such as European-American male vs. African-American female). Additionally, we see greater improvements in debiasing for the intersectional identity of African-American x Female, likely as a result of more data availability for this intersection.

1 Key Information to include

- Mentor: Ethan Chi
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

While work has been done to determine the extent of bias present in language models and develop de-biasing methods for these models, to our knowledge, no work has been done to develop and evaluate de-biasing methods for intersectional bias in language models.

We define intersectional bias as the bias towards someone who is part of 2 or more marginalized groups, such as a black woman. We also note that intersectional bias concerns the unique interaction between being part of more than one marginalized group. Concretely, intersectional bias focuses on how the experience of being a black female is different than the experience of being black and the experience of being female because of the interplay between the identities.

Specifically, in our work, we extend the method proposed in [1] for debiasing contextual embeddings in large scale language models to reducing intersectional bias in the BERT-Tiny model. The original method from [1] is only used to reduce gender bias from large models. Our original contribution is extending this method by adapting it to also reduce racial bias against African-Americans and Hispanic-Americans, as well as age bias, and then developing three ways to reduce intersectional bias that are based upon [1]'s method. We evaluate the effectiveness of our debiasing methods using

the intersectional bias evaluation metrics defined by [2], and also create two new metrics based on this method.

We note that [1] includes a GitHub repository containing an implementation of their debiasing method. However, we choose to reimplement this method as the original repository was difficult to build off of when implementing our intersectional debiasing methods. As such, **we implemented the entire intersectional debiasing fine-tuning of the BERT model, including data preprocessing and training.** We do use the code from [2] for the implementation of the intersectional bias evaluation metrics (https://github.com/tanyichern/social-biases-contextualized) but add two new intersectional embedding tests to the repository.

3 Related Work

Several studies have demonstrated how language models pick up on and amplify social biases [3].

Different debiasing techniques for single identities have been developed (such as gender independently), using either static and contextualized word embeddings. These techniques have included a modified GloVe objective [4], adversarial learning [5], and the pre-trained contextualized embeddings paper that informed our research [1]. To our knowledge, no debiasing techniques have been developed thus far for intersectional biases. In this project, we aim to contribute to designing intersectional debiasing techniques.

Different bias metrics have also been developed for single identities, including Word Embedding Association Tests (WEAT) that measure biases using semantic similarities between word embeddings [3], WinoBias that evaluates bias using the ability to predict gender pronouns with equal probabilities for gender neutral nouns [6], and Sentence Embedding Association Tests (SEAT) [7]. [2] is unique in considering contextual word representations in addition to WEAT and SEAT, using state-of-the-art models like BERT and GPT-2, and including intersectional bias tests. We use the intersectional bias tests from [2] for the intersection of European-American vs African-American and male vs female as our evaluation metrics. We also add two new tests–the intersection of European-American vs. Hispanic-American and male vs. female, and the intersection of age and gender.

4 Approach

4.1 Baselines

For our debiasing methods, our baseline is the results of the pretrained BERT-Tiny model on the intersectional evaluation metric presented by [2]. We will discuss these metrics and their interpretation more in the experiments section. We ran the metrics from[2] on BERT-Tiny ourselves and compared our debiasing methods to this baseline of BERT-Tiny with no debiasing.

4.2 Intersectional De-biasing Methods

4.2.1 Original Method

To explain our three intersectional debiasing methods, we will first explain the gender debiasing method from [1] and then our methods which build upon it, depicted in Figure 1. This method uses 2 types of words, attribute words V_A and target words V_T. There is one set of masculine attribute words and one set of feminine attribute words, where each set contains words that indicate that group, such as she, her, he, etc. There is also one set of target words that consists of words that should not have any bias associated with them but often do due to bias in society (i.e. nurse, teacher, etc).

The method then extracts sentences that contain attribute words or target words. Let A be the set of sentences extracted for containing an attribute word, T be the set of sentences extracted for containing a target word and $\Omega(w)$ be the set of sentences extracted because they contain word w. Additionally, Let $E_i(w, x; \theta_e)$ be the embedding of word w in sentence x in model E at layer i (where there are N layers) parameterized by θ_e.

The loss used minimizes the similarity between the non-contextualized embeddings of each attribute word a, denoted as $v_i(a)$ and the contextualized embeddings of each target word, denoted as

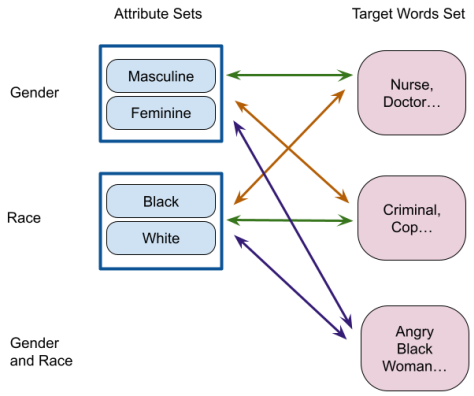


Figure 1: Arrows represent the sets of word embeddings between which the dot product is minimized in the different intersectional debiasing methods. Method 1's loss: green; Method 2's loss: green and orange; Method 3's loss: green and purple.

$E_i(t, x; \theta_e)$ The loss is then defined as

$$L_{bias} = \sum_{i=1}^{N} \sum_{t \in V_T} \sum_{x \in \Omega(t)} \sum_{a \in V_A} (v_i(a)^T E_i(t, x; \theta_e))^2$$

where $v_i(a) = \frac{1}{\Omega(a)} \sum_{x \in \Omega(a)} E_i(a, x; \theta_e)$. Additionally, the method tries to retain the semantic information in the attribute word embeddings by adding an L2 regularization term that minimizes the change in word embeddings in sentences containing an attribute word, expressed as

$$L_{reg} = \sum_{x \in A} \sum_{w \in x} \sum_{i=1}^{N} \|E_i(w, x; \theta_e) - E_i(w, x; \theta_{pre})\|$$

The final loss is then

$$L = \alpha L_{bias} + (1 - \alpha) L_{reg}$$

4.2.2 Intersectional Debiasing Method 1

We then extend this method to be used to reduce intersectional bias instead of only gender bias by creating three new methods for reducing intersectional bias. Now, let V_{Tj} and V_{Aj} be the target and attribute words associated with bias of type j, and A_j and T_j be the set of sentences extracted for containing attribute and target words relating to bias type j when there are M types of bias being debiased for. In method one, we minimize the difference between the target and attribute words for each type of bias individually. Thus, in this method we do not yet consider the interactions between the different types of bias. We use the the loss $L = \sum_{j=1}^{M} \alpha L_{bias_j} + (1 - \alpha) L_{reg_j}$ where

$$L_{bias_j} = \sum_{i=1}^{N} \sum_{t \in V_{Tj}} \sum_{x \in \Omega(t)} \sum_{a \in V_{Aj}} (v_i(a)^T E_i(t, x; \theta_e))^2$$

$$L_{reg_j} = \sum_{x \in A_j} \sum_{w \in x} \sum_{i=1}^{N} \|E_i(w, x; \theta_e) - E_i(w, x; \theta_{pre})\|$$

In Figure 1, each green arrow is one L_{bias_j} term, and the sum of the green arrows gives the loss for this method.

4.2.3 Intersectional Debiasing Method 2

In method 2, we debias for all types of bias in one fine-tuning run and minimize the similarity between all target and attribute words, even if they are not relating to the same type of bias. In this method, we consider the interactions between the different forms of bias by debiasing all attribute and target words. Let $V_{T_{all}} = \cup_{j=1}^{M} V_{Tj}$ and $V_{A_{all}} = \cup_{j=1}^{M} V_{Aj}$. Also let $A_{all} = \cup_{j=1}^{M} A_j$ and $T_{all} = \cup_{j=1}^{M} T_j$. Then, we perform one run of fine-tuning with the loss function $L = \alpha L_{bias} + (1-\alpha)L_{reg}$ where

$$L_{bias} = \sum_{i=1}^{N} \sum_{t \in V_{T_{all}}} \sum_{x \in \Omega(t)} \sum_{a \in V_{A_{all}}} (v_i(a)^T E_i(t, x; \theta_e))^2$$

$$L_{reg} = \sum_{x \in A_{all}} \sum_{w \in x} \sum_{i=1}^{N} \|E_i(w, x; \theta_e) - E_i(w, x; \theta_{pre})\|$$

In Figure 1, the orange and green arrows are encapsulated by the L_{bias} term since it combines the attribute word sets into one set and target words sets into one set and reduces the similarity between all words in the two sets.

4.2.4 Intersectional Debiasing Method 3

In the final method we present, we add a term to the loss function used in method 1 specifically aimed at reducing the intersectional bias resulting from all the types of bias being reduced. This extends method 1 to consider the interactions between the different forms of bias being reduced. Let $V_{Tintersect}$ be a set of target words that are stereotypes for the intersectional bias being reduced by the fine-tuning(such as words representing the angry black women stereotype).

We use the loss $L = \sum_{j=1}^{M} (\alpha L_{bias_j} + (1-\alpha) L_{reg_j}) + \beta L_{intersect}$ where

$$L_{intersect} = \sum_{i=1}^{N} \sum_{t \in V_{T_{intersect}}} \sum_{x \in \Omega(t)} \sum_{a \in V_{A_{all}}} (v_i(a)^T E_i(t, x; \theta_e))^2$$

In Figure 1, each green arrow is one L_{bias_j} term, as in method 1, and the purple arrows represent the $L_{intersect}$ term added in this method.

5 Experiments

5.1 Data

As per [1], we use the monolingual English section of the News-commentary-v15 corpus [8]. As described above, we pre-process the data by extracting sentences that contain an attribute or target word that is in one of the attribute or target word sets. Sentences that contain words from more than one set will not be used (as per [1]) in order to maintain clarity. For all the methods described above, we pre-process the data by extracting sentences from the corpus based on the sets of attribute and target words used in the method.

For the attribute and target word lists, we use the lists provided by [1] for words related to gender bias, and develop the words related to racial and age bias ourselves. We generated word lists with 2539 words in total.

5.2 Evaluation method

We measure intersectional bias by using the intersectional embedding association tests proposed by [2]. The WEAT and SEAT embedding association tests measure the association between two target concepts and two attributes. If X and Y are equal-size sets of target concept embeddings and A and B are sets of attribute embeddings, the tests measure the effect size of the association between a concept X with attribute A and concept Y with attribute B, as opposed to concept X with attribute B and concept Y with attribute A, where the associations are calculated using the mean of cosine

Gender - African-American Tests	
I1	EA F, AA F (least extreme)
I2	AA M, AA F
I3	EA M, AA M
I4	EA M, EA F
I5	EA M, AA F (most extreme)
Gender - Age Test*	
I6	Y M, E F (most extreme)
Gender - Hispanic-American Test*	
I7	EA M, HA F (most extreme)

Table 1: 7 intersectional embedding association tests (*: self-designed). EA: European-American, AA: African-American, M: Male, F: Female, Y: Young, E: Elderly, HA: Hispanic-American. Note: In this project, we use binary gender, but we understand that gender identity is fluid and hope to incorporate different gender identities into future experiments.

similarities. To investigate bias at the contextual word level, the authors modify SEAT to use the contextual word representation of the token of interest (representation of the word before pooling) instead of the sentence encoding.

- **Word**: Single word embedding association (*Alice* vs. *Doctor*)
- **Sent**: Sentence embedding association (*This is a doctor* vs. *Alice is here*)
- **C-word**: Single contextual word embedding association (This is a *doctor* vs. *Alice* is here)

Intersectional bias tests were developed by [2] by matching names with attribute words of pleasantness/unpleasantness. As shown in Table 1, we use seven intersectional bias tests as our metrics, where I1 through 15 are from [2] and I6 and I7 are self-designed. We run Word Embedding (Word), Sentence Encoder (Sent), and Contextual Word Representation (C-word) Association Tests for each of I1 through I7.

5.3 Experimental details

We use the Bert-Tiny model. (We had initially tried Bert-Base but obtained very few statistically significant results, so we followed the suggestion from [2] to decrease the size of the model.) We use the hyperparameters set by [1] for our debiasing fine-tuning. Specifically, for all hyperparameters besides learning rate and batch size, we use the default values from the HuggingFace library. We use a learning rate of $5e-5$ and batch size of 32. In our loss functions, we use $\alpha = 0.2$. We use $\beta = 0.5$ for method 3.

We run our 3 de-biasing methods for the intersections of African-American Race x Gender, Age x Gender and Hispanic American x Gender, as well as reporting results on a model with no-debiasing (Bert-Tiny), and models debiased only for Gender (Gender-Only), European-American v. African-American (EA-AA), Age (Age-Only), and European-American v. Hispanic-American (EA-HA). We refer to the models debiased using the intersectional methods as AA-G for African-American x Gender, A-G for Age x Gender, and HA-G for Hispanic-American x Gender followed by the method number (1, 2, or 3) used to debias the model.

5.4 Results

As shown in Table 2, 3, and 4, we see promise for all three intersectional debiasing methods in reducing the intersectional bias of the original Bert-Tiny model. We tend to see more significant results for the AA-G models compared to the A-G or H-G models, most probably due to less available data on the age / gender intersection and the Hispanic-American vs European-American / gender intersection. Debiasing for the African-American x Female bias shows more improvements over the baseline than debiasing for the Elderly x Female bias, likely because there is higher racial bias than age bias in the baseline model, and more training data related to race than age. Debiasing for the Hispanic-American x Female bias had the fewest significant results, likely due to a lack of data on this intersection.

Test	Encoding	Bert-Tiny	EA-AA Only	Gender Only	AA-G 1	AA-G 2	AA-G 3
	word	**1.425**	**1.457**	**1.453**	**1.322**	**1.355**	**1.365**
I1	sent	**1.362**	**1.38**	**1.361**	**1.273**	**1.334**	**1.275**
	c-word	-0.465	-0.269	**0.447**	**0.447**	**0.517**	**0.462**
	word	**1.407**	**1.152**	**1.256**	**1.403**	**1.254**	**1.381**
I2	sent	**0.838**	**0.643**	**0.749**	**0.857**	**0.667**	**0.864**
	c-word	-0.173	-0.065	0.118	0.152	0.224	0.13
	word	-0.21	0.391	0.36	0.365	-0.072	0.482
I3	sent	**0.437**	0.5	0.318	**0.397**	0.209	**0.414**
	c-word	-0.296	-0.377	0.254	0.143	0.203	0.279
	word	-0.75	-0.539	-0.176	0.176	-0.398	0.215
I4	sent	-0.822	-0.749	-0.518	-0.205	-0.665	-0.143
	c-word	0	-0.175	-0.08	-0.156	-0.098	-0.058
	word	1.502	**1.242**	**1.339**	**1.402**	**1.061**	**1.436**
I5	sent	**1.326**	**0.959**	**1.051**	**1.139**	**0.918**	**1.169**
	c-word	-0.465	-0.441	**0.47**	0.294	**0.423**	**0.406**

Table 2: Effect sizes for intersectional tests involving **EA-AA and and gender** run on different models. Bolded values represent significant tests ($p < 0.01$). Positive values represent pro-stereotypical bias, negative values represent anti-stereotypical bias.

Test	Encoding	Bert-Tiny	Age Only	Gender Only	A-G 1	A-G 2	A-G 3
	word	0.153	0.513	0.222	0.388	0.043	0.505
I6	sent	0.283	**0.85**	**0.967**	**0.905**	0.527	**0.758**
	c-word	0.252	0.25	-0.331	0.392	0.339	0.33

Table 3: Effect sizes for the intersectional test involving **age and and gender** run on different models. Bolded values represent significant tests ($p < 0.01$). Positive values represent pro-stereotypical bias, negative values represent anti-stereotypical bias.

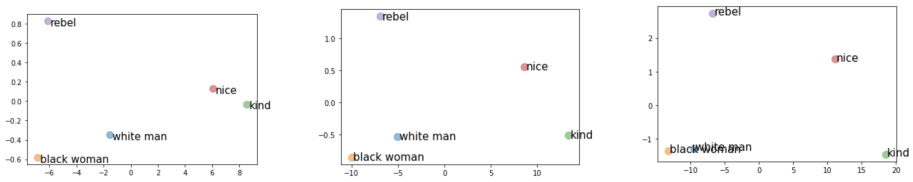
6 Analysis

We perform further analysis of our debiased models by creating visualizations of the models' embeddings of different words related to the intersectionalities we are trying to de-bias for. To do so, we generate the embeddings and then reduce them to 2 dimensions using PCA to plot them. For each of AA-G, A-G and H-G, we visualize word embeddings of related words in the base Bert-Tiny model, Gender debiased only model, and model debiased for the intersectionaity using method 3. We see that in general, our method tends to debias models by making the embeddings for different identities closer together, rather than changing the stereotypes they are associated with. We will explain this result further when discussing the first set of plots. Additionally, models that perform better on the intersectional embedding tests have more separation between embeddings for the identities and the stereotypes being debiased for, which we expect as our evaluation metric measures associations between the embeddings of identities and stereotypes.

Firstly, in Figure 2, we compare the embeddings of "black woman", "white man", "nice", "kind" and "rebel". In the baseline model, white man is much closer than "black woman" to the positive stereotype terms of "nice" and "kind", the model demonstrating that the model has bias related to these words. In the gender de-biased model, the identity terms of "white man" and "black woman" have moved closer together, while also becoming more distant from all the stereotype terms. This

Test	Encoding	Bert-Tiny	EA-HA Only	Gender Only	H-G 1	H-G 2	H-G 3
	word	-0.614	-0.486	0.730	0.770	-0.176	0.594
I7	sent	-0.413	-0.387	0.039	**0.021**	-0.358	**-0.161**
	c-word	-0.401	-0.319	0.299	0.280	**0.346**	**0.31**

Table 4: Effect sizes for the intersectional test involving **European-American vs Hispanic-American / gender** run on different models. Bolded values represent significant tests ($p < 0.01$). Positive values represent pro-stereotypical bias, negative values represent anti-stereotypical bias.

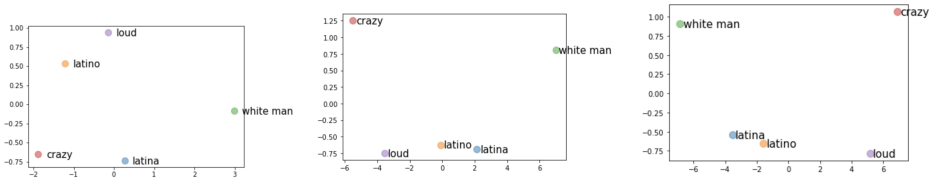


(a) Base Bert-Tiny (b) Gender De-biased (c) AA-G Method 3

Figure 2: Visualization of African-American x Female word embeddings in 3 models

indicates that the method is performing as desired since there is less association between the identities and stereotypes, and less of a difference between the embeddings that could indicate bias.

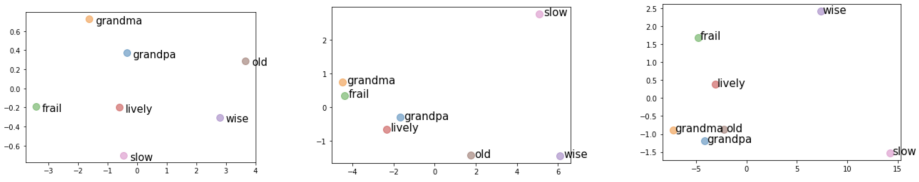
In the embeddings for the AA-G Method 3 model, the embeddings for "white man" and "black woman" have changed to be almost the exact same. We did not initially expect this change, since we would have expected all embeddings to shift to reduce bias in the model between them, but upon further analysis, this result is to be expected with our method. Without the regularization term, the de-biasing loss treats all of the identity words the same: it minimizes the similarity between the identity and all words in the stereotype target set. Thus, the model would learn one optimal embedding for all identity words if there was no regularization term. However, since there is a regularization term, the model learns similar, but not the exact same embeddings for the identities it is being de-biased for. This does successfully reduce bias in the model, since the identities are all about the same distance from the stereotype embeddings, but it does mean that it is likely that information contained in the identity embedding is lost.



(a) Base Bert-Tiny (b) Gender De-biased (c) H-G Method 3

Figure 3: Visualization of Hispanic-American x Female word embeddings in 3 models

Next, in Figure 3, we see that a similar change in the embeddings for "latino" and "latina" occurs as the model becomes less biased. The original model has the embeddings quite far apart, but they get closer together with the gender de-biased model, and even closer with the H-G 3 model. We also note that "white man" doesn't get closer to the embeddings of "latino and latina" in any of the models, which may explain why it the improvement of the de-biased models over the base line is smaller than in the African-American x Female case above. We do also see movement of the embeddings of the stereotypes in these models, with them moving further away from the identity embeddings, which also indicates that the de-biasing is reducing association.



(a) Base Bert-Tiny (b) Gender De-biased (c) H-G Method 3

Figure 4: Visualization of Age x Gender word embeddings in 3 models

Lastly, in Figure 4, we see an interesting trend in the gender de-biased model which explains why the results of the model during evaluation are worse than the baseline of Tiny-Bert and the intersectional de-biasing methods. We note that while the identities of "grandma" and "grandpa" have moved closer

together in this model, "grandma" and "frail" and "grandpa" and "lively" have also become very close to each other. We hypothesize that removing gender bias makes the age bias more pronounced in the model since it has not been removed and the embeddings have shifted to be closely aligned with age bias during the debiasing since it was not considered. This aligns with the results of [9] which also demonstrate that doing one form of debiasing can increase another form of bias.

From our visualizations, we come to understand that our method de-biases models by making embeddings of different identities more similar, since it treats all identities almost the same, and the model is most successful when there is a large distance between all identity and stereotype embeddings. Additionally, we see that reducing one form of bias can unintentionally increase another form, as in the case of the gender de-biased model.

7 Conclusion

In conclusion, while this effort is definitely in its early stages, we were nevertheless able to decrease intersectional bias found in language models using three intersectional debiasing methods, all of which perform better than single identity debiasing. Interestingly, Method 1 performs the best for less extreme intersectionality tests (where we define "less extreme" as a smaller difference between intertwined identities, such as European-American Female vs. African-American Female [I1], since gender is the same for both). However, debiasing methods 2 and 3 perform the best for more extreme intersectionality tests, such as European-American Male vs. African-American Female [I5], Young Male vs. Elderly Female [I6], and European-American Male vs. Hispanic-American Female [I7], all of which represent the largest power differential among their respective groups. This result aligns with our intuitions since Methods 2 and 3 take interacting identities into account more comprehensively than Method 1, although additional work needs to be done to check our intuitions. So far, these results may indicate that a one-size-fits-all approach for intersectional debiasing is not ideal, and different methods may be needed for different scenarios, depending on the groups involved and varying degrees of vulnerability.

In this project, we used Bert-Tiny in order to obtain statistically significant results with which to evaluate our debiasing techniques. In the future, we hope to experiment with larger models and different SOTA model types. We also aim to expand to different intersectionalities, taking into account disability, sexuality, and other aspects of identity. However, our techniques will need to adapt to the type of identity–our current approach of using names (traditionally European-American names, or traditionally female names, for example) in our word lists will not work for other facets of identity like disability or sexuality.

References

[1] Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings, 2021.

[2] Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations, 2019.

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

[4] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[5] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[6] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[7] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, et al. Proceedings of the fifth conference on machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, 2020.

[9] Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. Gender bias in multilingual embeddings and cross-lingual transfer. *CoRR*, abs/2005.00699, 2020.