Demographic-Aware Language Model Fine-tuning as a Bias Mitigation Technique

Aparna Garimella
Adobe Research
garimell@adobe.com

Akhash Amarnath*
Amazon Search
akhashna@amazon.com

Rada Mihalcea
University of Michigan
mihalcea@umich.edu

Abstract

BERT-like language models (LMs), when exposed to large unstructured datasets, are known to learn and sometimes even amplify the biases present in such data. These biases generally reflect social stereotypes with respect to gender, race, age, and others. In this paper, we analyze the variations in gender and racial biases in BERT, a large pre-trained LM, when exposed to different demographic groups. Specifically, we investigate the effect of fine-tuning BERT on text authored by historically disadvantaged demographic groups in comparison to that by advantaged groups. We show that simply by fine-tuning BERT-like LMs on text authored by certain demographic groups can result in the mitigation of social biases in these LMs against various target groups.

1 Introduction

Bias is defined as any kind of preference or prejudice of an individual or group, towards another individual or group (Moss-Racusin et al., 2012; Sun et al., 2019). The underlying traits of one's demographic group shape one's thoughts and worldviews (Garimella et al., 2016), and therefore may surface in one's language preferences and biases in the day-to-day life. For example, the word *admit* is more often associated with *hospital* by Indian bloggers, whereas American bloggers associate it with *guilt* (Garimella et al., 2017).

Most prior work in bias mitigation has largely taken the "one-size-fits-all" approach, with most models being agnostic to the language of the speakers behind the language (Sun et al., 2019; Liang et al., 2020; Dinan et al., 2020; Garimella et al., 2021). In this paper, we draw inspiration from previous research that showed the effect of demographic information on NLP tasks, such as word embeddings (Bamman et al., 2014), word associations (Garimella et al., 2017; Welch et al., 2020), empathy prediction (Guda et al., 2021), varied model performance of demographic-aware models (Hovy, 2015). We hypothesize that biases toward or against a specific group vary based on the demographic lens through which the world is viewed, and analyzing the social biases of various demographic groups from their language use can help uncover their characteristics. We believe such an understanding can move us beyond "one-size-fits-all" models, while at the same time developing demographic-aware bias mitigation techniques.

The advent of large pre-trained Transformer-based (Vaswani et al., 2017) language models, such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2019), have revolutionized NLP techniques for several downstream tasks (Joshi et al., 2019; Liu and Lapata, 2019; Wang et al., 2019; Zhang et al., 2019). However, along with their high performances, they can also inherit the various social biases that may be present in the large unstructured datasets they are exposed to (Kurita et al., 2019; Sheng et al., 2019; Tan and Celis, 2019; Zhao et al., 2019). In this paper, we focus on gender (male, female) and racial (African American, European American) biases, and address two main research questions: **(1)** How do gender and racial biases encoded in BERT vary when exposed to language authored by different demographic groups? **(2)** How do biases in demographic-specific BERT models compare to those in vanilla BERT?

The paper makes two main contributions. First, we present an empirical analysis of gender and racial biases encoded in BERT when it is trained on datasets authored by various demographic groups, and show that the biases, as measured by the Sentence Encoder Association Test (May et al., 2019), vary across different speaker groups under consideration. Second, we compare the biases in demographic-specific BERT models with those in vanilla BERT, and examine the extent to which the biases are either amplified or reduced in BERT

*This work was done when the author was at Adobe Research.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 311–319
November 20–23, 2022.

upon exposure to language authored by specific demographic groups. To our knowledge, this is the first work that explores the effect of speaker demographic information on biases in BERT-like language models.

2 Effect of Speaker Demographics on Biases in BERT

To investigate the effect of speaker demographic information on biases encoded in BERT, we build variants of the pre-trained BERT model that are exposed to the language of various demographic groups. We consider the resulting change in the model biases to be a result of the underlying group's bias. Specifically, we use the pre-trained BERT model and fine-tune it on datasets authored by different demographic groups for masked language modelling (MLM).[1]

Datasets. We use several datasets to fine-tune BERT. First, we use gender-aware datasets, to measure gender bias: *(i)* blogs collected from Google Blogger (Garimella et al., 2017), and *(ii)* Reddit comments (Welch et al., 2020). The former consists of a large set of blog posts collected from Google Blogger from 1999 to 2016, where the gender information of the bloggers is self-provided in their profiles. The latter consists of publicly available Reddit comments from 2007 to 2015; since Reddit users do not have profiles with personal information fields that could be scraped, Welch et al. (2020) extracted the demographic attributes of users from self-identification in their text. Gender was extracted by searching for statements referring to oneself as a 'boy', 'man', 'male', 'guy', for male (e.g., 'i am a male'), or 'girl', 'woman', 'female', 'gal', for female (e.g., 'i am a female'). We use 50K examples for each gender, randomly sampled from these datasets for fine-tuning BERT.

Second, we use location-aware datasets, covering Africa, Asia, Europe, North America, and Oceania, to measure the racial bias held in these regions: *(i)* GeoWAC, a geographically-balanced gigaword corpus, that consists of web pages from the Common Crawl (Dunn and Adams, 2020a), and *(ii)* Reddit comments (Welch et al., 2020).[2] In GeoWAC, the language samples are geo-located using country-specific top-level domains; *e.g.,* a web page under the *.ca* domain is assumed to have originated from Canada (Dunn and Adams, 2020b).[3] This dataset consists of gigaword corpora for 48 languages, with the English corpus spanning across more than 150 countries. We consider top three countries per region with the highest number of examples, and select around 94K examples (the minimum number of examples for any country) from each of them (Table 1 in Appendix) , resulting in around 283K examples for each region.

For the Reddit dataset, similar to the case in gender, Welch et al. (2020) segregated the comments region-wise based on the usage of phrases such as 'i am from' and 'i live in'. This dataset consists of eight regions, namely Africa, Asia, Canada, Europe, Mexico, Oceania, South America, United Kingdom, and United States (Table 5 in Appendix shows the number of examples present in each of the five regions in the Reddit dataset). For our experiments, we merge the comments from United States and Canada to obtain examples for the North America region, and merge United Kingdom with Europe. We do not consider Mexico and South America regions for our experiments. We use around 80K examples from each region from Reddit dataset (based on the minimum number of Reddit comments for any region).

We only consider examples with length > 20 and < 500 tokens. For each gender and region, we perform fine-tuning five times on each dataset, by randomly sampling the required number of examples for each fold, and use 90:10 ratio to obtain training and validation splits. We report the results averaged on the five folds. Further implementation details are provided in Appendix.

Evaluation. Social biases are typically measuring using the Word Embedding Association Test (WEAT) (Caliskan et al., 2016). WEAT imitates the human implicit association test (Greenwald et al., 1998) for word embeddings, by measuring bias via the association between two sets of target concepts and two sets of attributes. For example, to measure gender bias with respect to career/family, which is a common historical gender bias (Caliskan et al., 2016), it uses target words such as *female, woman* and *male, man*, and attributes such as *executive, office* and *home, children*. The bias is determined by the difference between the relative similarity of the target concepts to the two sets of attributes: greater

[1] https://github.com/huggingface/transformers

[2] While the Reddit comments from (Welch et al., 2020) are available from 2007-15 and are of size > 1 TB, we use data only from the latest years for time and memory constraints.

[3] While this may not imply that the language user is born in Canada, it is assumed that the user lives in the country.

	Bias against male/female			Bias against female			Bias against male		
Model	**6: M/F names**	**6b: M/F terms**	**Avg.**	**6: M/F names**	**6b: M/F terms**	**Avg.**	**6: M/F names**	**6b: M/F terms**	**Avg.**
BERT	0.48	0.11	0.29	0.48	0.11	0.29	0.00	0.00	0.00
Blogs									
BERT-Male	0.82	0.23	0.52	0.82	0.23	0.52	0.00	0.00	**0.00**
BERT-Female	0.37	0.15	**0.26**	0.37	0.01	**0.19**	0.00	0.14	0.07
Reddit comments									
BERT-Male	0.78	0.23	0.50	0.78	0.14	0.46	0.00	0.08	**0.04**
BERT-Female	0.57	0.08	**0.32**	0.57	0.00	**0.28**	0.00	0.08	**0.04**

Table 1: SEAT effect sizes for gender tests 6/6b with *career/family* attributes for BERT and its gender-specific variants. Least average scores among the variants are in **bold** for each test. Average scores lower than that of BERT are underlined.

the difference, higher is the bias. This difference is called the *effect size* in the WEAT metric.

To measure the bias in BERT, we use the Sentence Encoder Association Test (SEAT) (May et al., 2019), which is an extension of WEAT to measure the bias between contextual representations obtained using BERT. The word-level test is extended to sentence contexts by using semantically bleached sentence templates, such as "This is a <word>", "The <word> is here", which convey very little meaning beyond that of the term inserted in them. We use the tests 6/6b and 3/3b to measure gender and racial bias respectively. 6/6b use attributes *family* and *career* for {male, female} groups, and 3/3b use *pleasant* and *unpleasant* for {African American, European American} groups. The effect size for each test $\in \{-\infty, \infty\}$, with sizes of larger magnitude indicating more severe bias toward or against a group. A score > 0 (or < 0) for gender tests (6/6b) indicates that male is more (or less) associated to career than to family, in comparison to female. Similarly, a score > 0 (or < 0) for race tests (3/3b) indicates that European American is more (or less) associated to pleasant than to unpleasant, in comparison to African American. Thus, an effect size > 0 indicates that the model is biased against female and African American groups (or toward male and European American groups) for gender and race respectively, and an effect size < 0 indicates the the model is biased toward them.[4]

3 Results and Discussion

Tables 1 and 2 show the SEAT effect sizes of BERT, and its fine-tuned gender-specific and race-specific variants respectively, averaged over five folds (with absolute values taken for BERT and each fold of its variants). While the effect sizes can be positive or negative, we present their absolute values as they indicate the severity of the models' bias toward or against any given group.

3.1 Bias Variation Across Groups

In the case of gender (Table 1), with blogs as training data, the effect sizes of BERT-F model for both tests (0.37, 0.15) and their average (0.26) are lower in magnitude than those of BERT-M model (0.82, 0.23, 0.52). Similar trend is seen with Reddit comments as training data (0.57, 0.08, 0.32 compared to 0.78, 0.23, 0.50 respectively). The gender bias in BERT (according to tests 6/6b) stems from the high association of male terms with career than family in comparison to female terms; the decrease in bias for BERT-F indicates that such associations are lower in female language than in male language, in both blogs and Reddit comments.

To examine biases against a specific gender, we consider the direction of the SEAT effect size. If the effect size is d, we consider bias against female (or African American for race) as $\begin{cases} d, & \text{if } d \geq 0 \\ 0, & \text{otherwise} \end{cases}$, and bias against male (or European American) as $\begin{cases} |d|, & \text{if } d \leq 0 \\ 0, & \text{otherwise} \end{cases}$. With this formulation, we observe that bias against females is lower for BERT-F (0.19) than BERT-M (0.52) for both the datasets, while that against males is lower for BERT-M when trained with blogs (0.00 compared to 0.07), and more or less the same for both the BERT variants when trained with Reddit comments (0.04). In other words, bias against a specific gender in BERT is lower when the model is trained with data authored by that gender, as per the SEAT tests 6/6b.

For racial bias against African American (AA) or European American (EA) groups (Table 2), with

[4]This is assuming the historical bias against women to not give as much importance to their career as they do for family (Caliskan et al., 2016).

	Bias against EA/AA			Bias against AA			Bias against EA		
Model	**3: EA/AA names**	**3b: EA/AA terms**	**Avg.**	**3: EA/AA names**	**3b: EA/AA terms**	**Avg.**	**3: EA/AA names**	**3b: EA/AA terms**	**Avg.**
BERT	0.10	0.37	0.23	0.00	0.37	0.18	0.10	0.00	0.05
GeoWAC									
BERT-Africa	0.10	0.19	**0.14**	0.08	0.10	**0.09**	0.02	0.09	0.05
BERT-Asia	0.32	0.32	0.32	0.32	0.32	0.32	0.00	0.00	**0.00**
BERT-Europe	0.15	0.18	0.17	0.09	0.16	0.12	0.06	0.03	0.04
BERT-NA	0.41	0.21	0.31	0.41	0.17	0.29	0.00	0.05	0.02
BERT-Oceania	0.25	0.23	0.24	0.16	0.18	0.17	0.06	0.02	0.04
Reddit comments									
BERT-Africa	0.42	0.30	0.36	0.30	0.23	0.27	0.11	0.07	0.09
BERT-Asia	0.32	0.25	0.29	0.07	0.15	**0.11**	0.24	0.11	0.18
BERT-Europe	0.33	0.21	0.27	0.29	0.21	0.25	0.05	0.00	**0.02**
BERT-NA	0.20	0.14	**0.17**	0.17	0.14	0.16	0.03	0.00	**0.02**
BERT-Oceania	0.29	0.27	0.28	0.25	0.22	0.24	0.05	0.05	0.05

Table 2: SEAT effect sizes for race tests 3/3b with *pleasant*/*unpleasant* attributes for BERT and its region-specific variants. Least scores among the variants are in **bold** for each test. Average scores lower than that of BERT are underlined.

GeoWAC training data, the average effect size of BERT-Africa is the least (0.14) compared to the other region-specific BERT variants. For bias against AA, the BERT-Africa model has the least score of 0.09, while the highest scores are seen for BERT-Asia and BERT-North America models, both for the case of racial bias against either EA or AA (0.32, 0.31 respectively), and only AA (0.32, 0.29 respectively). For the bias against EA, BERT-Asia and BERT-North America achieve least scores, while BERT-Africa has the highest score. Thus, similar to the case in gender, bias against a specific race is lower when the model is trained with data authored by that racial group, with GeoWAC data.

It is interesting to note that with Reddit comments as the training data, BERT-North America model achieves the least average effect size against EA/AA (0.17), while BERT-Africa has the highest bias score (0.36). While BERT-NA and BERT-Europe models have least bias scores against EA similar to the case with GeoWAC data, BERT-Asia has the least score for bias against AA and highest score against EA, and BERT-Africa has the highest bias score against its own group. We suspect the Reddit comments authored by AA group are particularly stereotypical, and further investigation is needed to more concretely understand this.

3.2 Bias Variation Between BERT and its Demographic-Specific Variants

Here, we address the second research question of how the biases in the demographic-specific variants of BERT compare to those in BERT. We observe from Table 1 that BERT-F obtained using

Model	**6: M/F names**	**6b: M/F terms**	**Avg.**
Reddit comments			
50K Training Examples			
BERT-Male	0.78	0.23	0.50
BERT-Female	0.57	0.08	**0.32**
30K Training Examples			
BERT-Male	0.63	0.13	0.38
BERT-Female	0.48	0.15	**0.31**
10K Training Examples			
BERT-Male	0.54	0.19	0.37
BERT-Female	0.44	0.16	**0.30**
Blogs			
50K Training Examples			
BERT-Male	0.82	0.23	0.52
BERT-Female	0.37	0.15	**0.26**
30K Training Examples			
BERT-Male	0.80	0.24	0.52
BERT-Female	0.20	0.20	**0.20**
10K Training Examples			
BERT-Male	0.72	0.10	0.41
BERT-Female	0.22	0.34	**0.28**

Table 3: SEAT scores for tests 6/6b with Reddit and blog datasets for BERT variants with varying training sizes.

blogs achieves lower bias score (0.26) compared to not only BERT-M (0.52) but also to BERT itself (0.29). In other words, a small degree of bias mitigation is achieved in BERT by only exposing it to female language. However, the bias score of BERT-F increases when it is trained with Reddit data (0.32), (though the increase is much higher for BERT-M); this may be due to the possible biased nature of the Reddit data itself. To further examine this, we fine-tune BERT on Reddit data with smaller sizes of 30K and 10K examples (Table 3).

We see that the average bias scores decrease upon reducing the training size, hinting at the possible biased nature of the Reddit data. In the case of blogs, however, while fine-tuning BERT with 30K female-authored examples results in a decreased score (0.20), it increases slightly (0.28) with 10K examples. We also note that such decrease in model bias of BERT-F compared to BERT is also seen for bias against female, more so with 30K training examples (0.19 and 0.07 with 50K and 30K examples respectively with blogs, 0.28 and 0.24 examples respectively with Reddit data; complete results with bias against specific groups are provided in Table 4 in Appendix).

Note that SENT-DEBIAS, proposed for debiasing sentence representations via a post-training technique (Liang et al., 2020), achieves an average absolute SEAT score of 0.27 for tests 6/6b. It is interesting that BERT-F trained with blogs achieves a comparable score of 0.26 with 50K examples, and a much lower score of 0.20 with 30K examples, just by exposing BERT to female language.

In the case of racial bias (Table 2) as well, we note a reduced bias of BERT-Africa model (0.14) compared to that of BERT (0.23) in the GeoWAC setting, and of BERT-NA model (0.17) in the Reddit setting. Similar drops can be seen for bias against AA and EA groups using both the datasets. These results indicate that not only biases encoded in BERT vary across speaker demographics of the language BERT is exposed to, but also that such exposure via simple fine-tuning can sometimes also result in bias mitigation of the pre-trained LM. The results obtained using blogs (for gender) and GeoWAC data (for race) further hint at the possibility of gender or racial bias mitigation in BERT against a specific target group by fine-tuning it with language authored by that very group (female for gender and African American for race).[5]

4 Conclusions

In this paper, we analyzed gender and racial bias in BERT when it is fine-tuned on datasets authored by different demographic groups. We found that BERT when exposed to female language exhibits lower gender bias than when it is exposed to male language as measured by the SEAT effect size with respect to career/family attributes. For European American/African American racial bias, we observed that with one dataset, BERT exposed to African language exhibits lower bias, while on another dataset, BERT exposed to North American language results in lower bias. We also found that simply fine-tuning BERT on MLM tasks with data authored by specific demographic groups can result in bias mitigation in BERT, indicating that depending on the lens through which the world is viewed, biases can be lowered in large pre-trained LMs.

Based on these initial findings, we believe further research is warranted in this direction of bias mitigation using demographic data and demographic-aware bias mitigation methods.

[5]Given that BERT fine-tuning can be unstable due to the randomness in data shuffling and initialization (Devlin et al., 2019), there may be slight variations in some of the results if the same experiments are re-run with the same set of hyper-parameters and data splits. Our aim in this paper is to only highlight the variations in the biases in BERT when exposed to language authored by different demographic groups, and bring to attention that sometimes this could lead to bias mitigation in it.

Acknowledgments

This material is based in part upon work supported by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. We thank Ian Stewart and Yiting Shen for all the feedback they provided on earlier iterations of this work.

References

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL 2014)*, pages 828–834.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in

dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online.

Jonathan Dunn and Ben Adams. 2020a. Geographically-balanced Gigaword corpora for 50 language varieties. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2528–2536, Marseille, France. European Language Resources Association.

Jonathan Dunn and Ben Adams. 2020b. Mapping languages and demographics with georeferenced corpora. *arXiv preprint arXiv:2004.00809*.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu N, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545, Online.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2285–2295, Copenhagen, Denmark.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. EmpathBERT: A BERT-based framework for demographic-aware empathy prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3072–3079, Online. Association for Computational Linguistics.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. 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, Minneapolis, Minnesota.

Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41):16474–16479.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, volume 32, pages 13230–13241.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, volume 30, pages 5998–6008.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5878–5882, Hong Kong, China.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089, Online. Association for Computational Linguistics.

Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 789–797, Hong Kong, China.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized 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 629–634, Minneapolis, Minnesota.

A Appendix

Region	Country	# Examples
Africa	Nigeria	3,153,761
Africa	Mali	660,916
Africa	Gabon	645,769
Asia	India	12,327,494
Asia	Singapore	6,130,047
Asia	Philippines	3,166,971
Europe	Ireland	8,689,752
Europe	United Kingdom	7,044,434
Europe	Spain	465,780
North America	Canada	7,965,736
North America	United States	8,521,094
North America	Bermuda	244,500
Oceania	New Zealand	94,476
Oceania	Palau	486,437
Oceania	Vanuatu	165,355

Table 4: Country-specific details in GeoWAC dataset.

Blog dataset. We use a subset of the blogs described in (Garimella et al., 2017), which consists of around 211K female blog posts and 121K male blog posts. We randomly sample 50K examples from each of these two genders for our experiments (for five folds).

Region	# Examples
Africa	3,153,761
Asia	660,916
Europe	645,769
North America	12,327,494
Oceania	6,130,047

Table 5: Region-specific details in Reddit dataset.

GeoWAC dataset. From the GeoWAC dataset, we consider three countries for each of the five regions, as shown in Table 4. Table 4 also shows the number of examples from each of these countries. Note that these numbers includes all the very short or long examples as well; we discard those with $<$ 20 and $>$ 500 tokens while selecting examples for fine-tuning BERT.

Reddit dataset. Welch et al. (2020) segregated the Reddit comments gender-wise based on the usage of phrases such as 'i am a male' and 'i am a female', and region-wise based on the usage of phrases such as 'i am from' and 'i live in'. We consider Reddit comments authored by males and females from the years 2014-15 from (Welch et al., 2020) for the case of gender. These amount to more than 49M female and male examples; we randomly sample 50K examples (five folds) for our BERT fine-tuning experiments. Table 5 shows the number of examples present in each of the five regions in the Reddit dataset. Note that this dataset (spanning 2013-15) consists of eight regions, namely Africa, Asia, Canada, Europe, Mexico, Oceania, South America, United Kingdom, and United States. For our experiments, we merge the comments from United States and Canada to obtain examples for the North America region, and merge United Kingdom with Europe. We do not consider Mexico and South America regions for our experiments.

The overall dataset statistics for the finetuning experiments are provided in Table 6.

Evaluation. Word Embedding Association Test (WEAT) imitates the human implicit association test (Greenwald et al., 1998) for word embeddings. Specifically, it measures the association between two sets of target concepts and two sets of attributes.

Implementation details. BERT is fine-tuned for 3 epochs with every dataset for each of the five folds.

GENDER	
Dataset	**# Examples**
Blogs (Garimella et al., 2017)	50K
Reddit comments (Welch et al., 2020)	50K
RACE	
GeoWAC (Dunn and Adams, 2020a)	285K
Reddit comments (Welch et al., 2020)	80K

Table 6: Sizes of datasets used for finetuning BERT.

The region-specific fine-tuning experiment with GeoWAC dataset are run on single Tesla T4 GPU (22 GB memory), and the rest other experiments (region-specific fine-tuning with Reddit dataset, and gender-specific fine-tuning experiments) are run on single Tesla V100 GPU (52 GB). All of them use BERT-base-uncased model, with batch size 8, learning rate 1e-4, and maximum sequence length 512. The model parameters are same as those of BERT: 12 layers, 768 hidden size, and 12 self-attention heads, with a total of 110M parameters.

Results. Table 7 shows the SEAT effect sizes of gender-specific variants of BERT training with 50K and 30K examples, from blog and Reddit datasets.

	Bias against male/female			Bias against female			Bias against male		
Model	**6: M/F names**	**6b: M/F terms**	**Avg.**	**6: M/F names**	**6b: M/F terms**	**Avg.**	**6: M/F names**	**6b: M/F terms**	**Avg.**
BERT	0.48	0.11	0.29	0.48	0.11	0.29	0.00	0.00	0.00
Blogs									
50K Training Examples									
BERT-Male	0.82	0.23	0.52	0.82	0.23	0.52	0.00	0.00	**0.00**
BERT-Female	0.37	0.15	**0.26**	0.37	0.01	**0.19**	0.00	0.14	0.07
30K Training Examples									
BERT-Male	0.80	0.24	0.52	0.80	0.24	0.52	0.00	0.00	**0.00**
BERT-Female	0.20	0.20	**0.20**	0.14	0.00	**0.07**	0.06	0.20	0.13
Reddit comments									
50K Training Examples									
BERT-Male	0.78	0.23	0.50	0.78	0.14	0.46	0.00	0.08	**0.04**
BERT-Female	0.57	0.08	**0.32**	0.57	0.00	**0.28**	0.00	0.08	**0.04**
30K Training Examples									
BERT-Male	0.63	0.13	0.38	0.63	0.07	0.35	0.00	0.06	**0.03**
BERT-Female	0.48	0.15	**0.31**	0.48	0.00	**0.24**	0.00	0.15	0.07

Table 7: SEAT effect sizes (absolute values) for gender tests 6 and 6b with *career/family* attributes for BERT and its gender-specific variants, and their averages. Least scores among the variants are in **bold** for each test. Average scores lower than that of BERT are underlined.