FineDeb: A Debiased Finetuning Approach for Language Models

Anonymous ACL submission

Abstract

As language models are increasing included in human-facing machine learning tools, bias against demographic subgroups has gained attention. We consider the problem of debiasing in language models. Rather than modifying a model's already learned representations, we focus on modifying them during model training itself. We propose a two-phase methodology (FineDeb) that starts with contextual debiasing of embeddings learned by the language models during training, then finetunes the model on the original language modelling objective. We apply our method to debias for demographics with multiple classes, demonstrating its effectiveness through extensive experiments and comparing with state of the art techniques, and on three metrics.

1 Introduction & Related Work

Machine learning tools that rely on natural language processing (NLP) are increasingly being developed for scenarios with immediate impact on individuals, such as healthcare (Velupillai et al., 2018), conversational agents (Zhang et al., 2020), and legal systems (Dale, 2019). While the language models here vary in the type of embeddings used, they rely on representations that may reflect or exhibit bias (Manzini et al., 2019; Bolukbasi et al., 2016). When used in downstream tasks such as prediction, health care diagnoses, or other decision-making, such representations can amplify bias and result in discriminatory actions against individuals in disadvantaged demographic subgroups. Our focus in the present paper is on such representational harms (Blodgett et al., 2020).

There is prior work that relies on word embeddings to analyze bias and proposes debiasing techniques for NLP methods. Debiasing on word embeddings was first introduced by Bolukbasi et al. (2016) and further refined to enable debiasing on multiple classes by Manzini et al. (2019). However, recent advances in NLP have been focused on large pretrained transformer-based language models (LM) such as BERT and GPT. Such models differ in that rather than considering individual word embeddings, they create representations that take into account large and connected components such as sentences and context. For comprehensive bias mitigation we must consider bias in the context of sentences, beyond mere word embeddings. We therefore focus on bias in transformer-based contextual LMs.

There has been some recent work on bias in such models (Liang et al., 2021; Zhang et al., 2022), which, like much of prior work focus on debiasing of representations. Previous work (Bolukbasi et al., 2016; Liang et al., 2020, 2021) apply debiasing techniques on embeddings after bias subspaces in these representations have been detected. While such techniques attempt to create debiased representations, downstream tasks may not necessarily reflect debiased language. Furthermore, good performance on those downstream tasks, such as language generation that makes sense, must still be an objective. Bordia and Bowman (2019) thus act on the training objective by adding a regularizer for debiasing. Similar to this but instead of a regularizer, our method uses an entirely new training objective to minimize distance between relevant embeddings.

A further limitation of state of the art debiasing techniques is that they largely consider demographics with binary classes or when they consider multiple classes, they focus on the three largest subgroups (Manzini et al., 2019). In real scenarios, social disadvantage is represented through more than the binary or majority/minority dichotomy, with demographic groups containing many classes. We apply our debiasing methodology on demographics with multiple classes.

In this work we propose a debiasing method - FineDeb, for large pretrained LMs, with two phases of fine tuning: one for debiasing and one for improving LM performance. In the debiasing phase we modify representations by using a training objective that minimizes the distance between embeddings of target words while considering their sentence contexts. In a second phase, we further finetune the model in an attempt to restore the LM performance. We thus extend prior work in several ways: by debiasing on LM models rather than word embeddings, including the debiasing objective in the training itself, including multiple classes in the debiasing phase, and further fine tuning for LM performance.

2 Data

We use two types of data - one for the debiasing phase and one for the LM finetuning phase.

Debiasing Data Our methodology starts by first finetuning the LM with a new objective function on a debiasing dataset, which consists of examples of debiased sentences. The debiased sentence examples are created using word lists crafted with multiple classes per demographic (in english language). Our final debiased model is thus trained by debiasing for multiple classes. Note that this is a novel contribution beyond word lists with pairs created previously.

Our word lists are compiled from various sources, both online[1] and existing work (Bolukbasi et al., 2016; Zhao et al., 2018), to create a list of target word tuples for each demographic (race, religion and gender). The word list contains 2 classes for gender[2], 5 classes for race, and 7 classes for religion. For each demographic, the word list consists of several tuples of target words, where within each tuple the words are comparable. For example, *("Muslim", "Christian", "Jewish", "Hindu", "Buddhist", "Confucianist", "Taoist")* is one tuple within the religion word list. We compile 10 word tuples for Race, 32 for Religion and 158 for Gender. We make the word lists freely available in our codebase, providing samples in Appendix D.

In order to generate our final debiasing dataset, we first craft sentence templates from the RedditBias (Barikeri et al., 2021) dataset. RedditBias is a dataset of human conversation data from Reddit across four demographics: *gender, race, religion,* and *queerness*. We convert all sentences containing the former three demographics into sentence templates by removing the target words. An example for race would be *"all ____ are criminals"*. During the training process we choose a relevant word tuple and sample target words from different classes such as *Black* and *White* to generate sentence pairs that differ in only the target word. We generate such sentence sets pairwise among all classes.

LM Finetuning Data Taking inspiration from prior work (Qian et al., 2019; Bordia and Bowman, 2019), we use CNN-DailyMail (Hermann et al., 2015) for our language model finetuning objective. It consists of $300,000+$ English news articles.

3 FineDeb

Our method, FineDeb, adopts a two phase approach for training. In the first phase, we debias the model by modifying the embeddings learned by the language model, and in the second phase, we finetune the debiased model on the original language modeling objective. Our method is demonstrated on a BERT model (Devlin et al., 2019), specifically *bert-base-uncased*. All hyperparameters are listed in Appendix B.

3.1 Debiasing Phase

In the debiasing phase, we train our model using the sentence pairs generated in Section 2. Our training objective is inspired by the traditional method of determining relationships between pairs of words by computing the distance between their embeddings (Mikolov et al., 2013; Bolukbasi et al., 2016). However, when we consider that most modern language models are contextual in nature (Devlin et al., 2019), the same word may have different meanings based on the context in which the word is used. For instance, the word *"temple"* could refer to a building or a part of the human body. Thus it becomes important to perform our debiasing only in those contexts where the bias may exist. Given two near-identical sentences that differ only by a target word (or phrase), we first compute the difference between the embeddings of the two sentences (using their [CLS] token embedding as in Devlin et al. (2019)) and the difference between the embeddings of the target words in each sentence. Our training objective then minimizes the difference between these two quantities to debias the model. In other words, the difference between the embeddings of the two sentences should be equal to the difference between the embeddings of the two target words as the sentences are otherwise identical. Formally, our loss function is as follows:

[1] Wiki: Religious Groups, Lumen: Religions, Wiki: Race

[2] NLP datasets have limited data for other gender classes

$L(S_1, S_2, W_1, W_2) = \mathcal{D}(S_1 - S_2, W_1 - W_2)$, where S_i is the embedding for sentence i, W_i is the embedding for the target word in sentence i, and $\mathcal{D}(\cdot)$ is the distance between the two quantities (we use Mean Squared Error). Taking an example from Nangia et al. (2020), if we had two sentences *"The crafty Jews made a plan to steal the money."* and *"The crafty Christians made a plan to steal the money."*, W_1 and W_2 would be the embeddings of *"Jews"* and *"Christians"* respectively, while S_1 and S_2 would respectively be the sentence embeddings.

3.2 LM Finetuning Phase

Pre-trained language models generally consist of a model which generates embeddings for words and a language modeling head (LM head) which gives probabilities for each of the words based on these embeddings. In the case of our debiasing strategy, we only update the embedding-generating model and not the LM head. Thus the weights of the LM head become incompatible with the new embeddings, leading to a poor language modeling ability. To remedy this, we finetune the entire model (debiased model + LM head) on the standard BERT language modeling objective (masked word prediction) using the CNN-DailyMail dataset. During this finetuning process, we freeze the debiased model so that the embeddings do not change, and only the weights in the LM head are updated. This results in an LM head compatible with the debiased model. While this method of improving the language modeling ability of our model may re-introduce some biases that exist in the CNN-DailyMail dataset, it does not eliminate the effects of our debiasing as we show in our results.

4 Metrics

We evaluate FineDeb on the three demographics of gender, race, and religion using three metrics: StereoSet, SEAT, and Crow-S Pairs. These metrics differ in how bias is evaluated, the data used in evaluating bias, and whether the language modeling performance is considered.

StereoSet Following recent work (Meade et al., 2021; Zhang et al., 2022), we use StereoSet (Nadeem et al., 2021) to evaluate our work. StereoSet measures the Stereotype Score (SS) which gives a measure of bias and the Language Modeling Score (LMS) which determines performace at language modeling tasks. There is a trade-off here, as a model could be perfectly unbiased but be a poor language model (or vice versa). Thus the authors provide a combined measure named ICAT. Following prior work (Meade et al., 2021), we use the intrasentence variant of StereoSet.

Crow-S Pairs The Crowdsourced Stereotype Pairs (Crow-S Pairs) (Nangia et al., 2020) uses crowdsourced pairs of sentences that differ only by a small number of tokens such that one sentence reflects a stereotype while the other violates that stereotype. Under this metric, a perfect model is equally likely to pick the stereotypical sentence as it is to pick the anti-stereotypical sentence. This metric does not test the language modeling ability of the model but covers a wide variety of biases.

SEAT Sentence Encoder Association Test (SEAT) (May et al., 2019) is a sentence level extension of the WEAT metric (Caliskan et al., 2017) which measures bias between two sets of attribute words and two sets of target words. Specifically, SEAT uses sentence templates to obtain representations of words. The metric is measured in terms of the average effect size across several tests, where a value closer to 0 indicates a lower degree of bias but it does not test the language modeling ability of the model.

5 Results and Discussion

We compare FineDeb on 3 demographics against 2 baselines and 5 prior works. Our baselines are pre-trained BERT and a pretrained BERT model where only the LM finetuning phase is applied. We include results on this latter model to show that the LM finetuning phase does not significantly alter bias in the model compared to the base BERT model, and any change in bias in our model is strictly due to our debiasing phase. This is evident in the results which show that BERT with LM finetuning is on par with or slightly better than the base BERT model for all listed metrics. We also compare with state of the art methods - CDA, Dropout, INLP, Self-Debias and, Sentence Debias, the results of which we cite from Meade et al. (2021).

We first present results[3] on the StereoSet metric in Table 1. Under the SS measure, FineDeb outperforms all techniques for all three demographics. Under the LMS measure, Self-Debias, Sentence Debias, and the baselines have the best or second best results, across different demographics. It is

[3] Evaluation code taken from Nadeem et al. (2021) (StereoSet) and Meade et al. (2021) (SEAT and Crow-S)

Table 1: StereoSet evaluation. LMS indicates Language Modeling Score, SS is the Stereotype Score, ICAT is the overall score. Higher is better for LMS & ICAT, closer to 50 is better for SS (**Best**; Next Best).

	Gender			**Race**			**Religion**		
StereoSet	LMS	SS	ICAT	LMS	SS	ICAT	LMS	SS	ICAT
BERT	84.17	60.28	66.86	84.17	57.03	72.34	84.17	59.70	67.84
BERT (LM finetuning)	**85.01**	59.01	69.69	83.78	56.38	73.08	83.07	61.47	64.01
FineDeb	77.70	**53.27**	**72.62**	75.37	**50.82**	74.13	74.10	**50.39**	**73.52**
CDA	83.08	59.61	67.11	83.41	56.73	72.18	83.24	58.37	69.31
Dropout	83.04	60.66	65.34	83.04	57.07	71.30	83.04	59.13	67.88
INLP	80.63	57.25	68.94	83.12	57.29	71.00	83.36	60.31	66.17
Self-Debias	84.09	59.34	68.38	**84.24**	54.30	**77.00**	84.23	57.26	72.00
Sentence Debias	84.20	59.37	68.42	83.95	57.78	70.89	**84.26**	58.73	69.55

Table 2: SEAT: Average effect size for each demographic. BERT (LM finetuning) & Self-Debias do not modify internal representations and so have the same SEAT score as BERT. Lower is better (**Best**; Next Best).

	Gender	**Race**	**Religion**
BERT	0.62	0.62	0.49
FineDeb	0.36	0.62	0.67
CDA	0.72	0.57	**0.34**
Dropout	0.77	**0.55**	0.38
INLP	**0.20**	0.64	0.46
Sentence Debias	0.43	0.61	0.44

Table 3: Crow-S Pairs: Metric scores for each demographic. Closer to 50 is better (**Best**; Next Best).

	Gender	**Race**	**Religion**
BERT	57.25	62.33	62.86
BERT (LM finetuning)	57.63	62.91	58.10
FineDeb	54.58	65.24	**44.76**
CDA	56.11	**56.70**	60.00
Dropout	55.34	59.03	**55.24**
INLP	**51.15**	67.96	60.95
Self-Debias	52.29	**56.70**	56.19
Sentence Debias	52.29	62.72	63.81

expected that our method not perform best on the LMS measure since we focus first and foremost on debiasing. However the reduction in LMS performance is not too severe since under the combined measure of ICAT our model performs best for gender and religion, and second best for race.

For SEAT (Table 2), our method performs second best for gender, while for race and religion we do better than one method and no other methods respectively. We note that SEAT measures distances between embeddings, whereas the StereoSet metric is based on final word outputs. Further, SEAT does not measure LM performance. We reason that since discriminatory harm is manifested due to contextual outputs from language models and not just their representations, a metric such as StereoSet, which considers both a measure of final LM performance, as well as a measure of the bias in contextual outputs, more accurately represents real-world impact of bias.

The results for Crow-S (Table 3) show that for religion, FineDeb outperforms all others, with similar performance as Dropout. For gender, our method beats the baseline and two other methods, but performs poorly for race. Crow-S measures bias similarly to StereoSet by considering whether a sterotypical sentence is preferred among sentence pairs. The sentences in a pair, however, differ on attributes rather than target words. We cannot be certain if this difference accounts for the differing results, but we note that Crow-S does not measure LM ability, an important component of any model.

Considering results over the three demographics, all benchmarks, and across all three metrics, FineDeb contributes significantly to both debiasing and attempting to maintain LM performance. Our debiasing technique can be considered stronger than other methods - the training objective itself is to minimize embedding distance, which may to some extent even lead to a distortion in the embeddings. This is seen in the StereoSet SS measures that are nearly perfect and in the somewhat poorer scores for LMS and SEAT. For the ICAT measure which combines debiasing and LM performance our method performs the best or second best, suggesting a better overall performance for our method.

There are a few avenues for future work in this area. Namely, expanding our work to include more classes (such as in gender) or to other demographics; a more comprehensive analysis of our model on downstream tasks; replacing our current two-phase training approach with a single phase of interleaved debiasing and finetuning. Further, current metrics for bias rely on comparisons of either target words or attribute words, resulting in varying performance across the different techniques. This suggests the need for a more comprehensive metric on bias that is agnostic to the debiasing technique.

References

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *ACL*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Robert Dale. 2019. Law and word order: NLP in legal tech. *Nat. Lang. Eng.*, 25(1):211–217.

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. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

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, Online. Association for Computational Linguistics.

Yunlong Liang, Chulun Zhou, Fandong Meng, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2021. Towards making the most of dialogue characteristics for neural chat translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 67–79, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2021. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

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

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D. Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, Wendy Chapman, and Rina Dutta. 2018. Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*, 88:11–19.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

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

A Limitations

We believe that our work has the following limitations:

- Our results (Section 5) indicate that, our debiasing approach results in a strongly debiased model at the cost of LM ability. This remains true even after finetuning to improve our LM performance. Thus our approach is more suited towards cases, such as story telling and recommendation systems, where it is important to have as little bias as possible, while slightly compromising on the language modeling ability of the model.

- There may be cases where we want a bias to exist. For example, in the sentence "The ___ man went to the mosque.", the probability of "Muslim" should be higher than the probability of "Christian" or "Jew". While people of any religion could go to a mosque, a person who follows Islam is far more likely. This falls under explainable bias (Mehrabi et al., 2021).

- While our method is effective, it relies heavily on the word lists that we have compiled to the best of our knowledge. These are by no means fully representative of all bias targets for a given demographic and there is still scope for expansion.

B Hyperparameters

Considering the trade-off between bias reduction and LM performance discussed in this paper, we experiment with different dataset sizes for both the debiasing phase and LM finetuning phase. Intuitively, a larger debiasing dataset size would lead to a higher bias reduction but poorer LM performance. On the other hand, a large LM finetuning dataset would lead to higher bias but also better LM performance. The debiasing dataset sampling size is varied as - 500, 1000, 2000, 4000, and continues upto the debiasing dataset's size (different for different demographics). The LM finetuning dataset size is varied as a percentage of the total CNN dataset - 1%, 2%, 4%, 8%, 16% and 32%.

For training our language model (3.1, 3.2) we use a batch size of 64 to maximise GPU usage, maximum source length of 64. For language model finetuning, we use 100 epochs to allow for convergence, whereas for debiasing we find 30 epochs to be sufficient, with both training processes utilizing an early stopping mechanism. We use a learning rate of 1e-4 based on what we find in the original BERT paper (Devlin et al., 2019).

C Training Details

For each two phase training, we run a job with a time limit of 41 hours, 4 CPUs per job, 4GB of memory per CPU and 1 NVIDIA V100 GPU.

D Word Lists

D.1 Gender

In Table 4, we provide a sample of word sets in the Gender word list.

Table 4: Example word sets for Gender

Male	Female
countryman	countrywoman
fraternal	sororal
manservant	maidservant
divo	diva
actor	actress
bachelor	spinster
papa	mama
busboy	busgirl

D.2 Race

In Table 5, we provide a sample of word sets in the Race word list.

D.3 Religion

In Table 6, we provide a sample of word sets in the Religion word list.

E Additional Results

In this section, we provide the complete evaluation results for SEAT in Tables 7, 8 and 9.

Table 5: Example word sets for Race

African-American	Anglo-American	Hispanic	Asian	Native-American
Black	White	Latino	Brown	light-brown
Negroid	Caucasian	Latino	Brown	Native-American
African-American	Anglo-American	Hispanic	Asian	Native-American
Afro-American	Anglo-American	Hispanic	Asian	Native-American
African	American	Hispanic	Asian	Native-American
Afroamerican	Angloamerican	Hispanic	Asian	Native-American
Negro	Caucasian	Hispanic	Brown	Native-American
dark-skin	light-skin	white-latino	gray-skin	reddish-brown
dark-skin	light-skin	black-latino	gray-skin	reddish-brown

Table 6: Example word sets for Religion

Islam	Christianity	Judaism	Hinduism	Buddhism	Confucianism	Taoism
Islam	Christianity	Judaism	Hinduism	Buddhism	Ruism	Daoism
Quran	Bible	Torah	Gita	Tripitaka	Analects	Tao-Te-Ching
Koran	Bible	Tanakh	Veda	Tripitaka	Analects	Tao-Te-Ching
Muslim	Christian	Jewish	Hindu	Buddhist	Confucianist	Taoist
islamic	Christian	Jewish	Hinduism	Buddhist	Confucius	Dao
Mohammed	Jesus	Malachi	Ramakrishna	Gautama	Confucius	Laozi
Mohammed	Jesus	Moses	Krishna	Gautama	Kung-Fu-Tzu	Laozi

Table 7: SEAT effect sizes for Gender debiased models. Effect size closer to 0 is better. Statistically significant ($p < 0.01$) effect sizes are denoted by underline.

Model	**SEAT-6**	**SEAT-6b**	**SEAT-7**	**SEAT-7b**	**SEAT-8**	**SEAT-8b**	**Avg. Effect Size**
bert-base-uncased	0.931	0.090	-0.124	0.937	0.783	0.858	0.620
FineDeb	0.248	0.113	0.179	0.298	0.280	1.056	0.362
CDA	.846	0.186	-0.278	1.342	0.831	0.849	0.722
Dropout	1.136	0.317	0.138	1.179	0.879	0.939	0.765
INLP	0.317	-0.354	-0.258	0.105	0.187	-0.004	0.204
Sentence Debias	0.350	-0.298	-0.626	0.458	0.413	0.462	0.434

Table 8: SEAT effect sizes for Race debiased models. Effect size closer to 0 is better. Statistically significant ($p < 0.01$) effect sizes are denoted by underline.

Model	**ABW-1**	**ABW-2**	**SEAT-7**	**SEAT-3**	**SEAT-4**	**SEAT-5**	**SEAT-5b**	**Avg. Effect Size**
bert-base-uncased	-0.079	0.690	0.778	0.469	0.901	0.887	0.539	0.620
FineDeb	1.111	-0.235	0.806	0.085	0.858	0.787	0.456	0.620
CDA	0.231	0.619	0.824	0.510	0.896	0.418	0.486	0.569
Dropout	0.415	0.690	0.698	0.476	0.683	0.417	0.495	0.554
INLP	0.295	0.565	0.799	0.370	0.976	1.039	0.432	0.639
Sentence Debias	-0.067	0.684	0.776	0.451	0.902	0.891	0.513	0.612

Table 9: SEAT effect sizes for Religion debiased models. Effect size closer to 0 is better. Statistically significant ($p < 0.01$) effect sizes are denoted by underline.

Model	**Religion-1**	**Religion-1b**	**Religion-2**	**Religion-2b**	**Avg. Effect Size**
bert-base-uncased	0.744	-0.067	1.009	-0.147	0.492
FineDeb	0.697	0.701	0.666	0.613	0.670
CDA	0.355	-0.104	0.424	-0.474	0.339
Dropout	0.535	0.109	0.436	-0.428	0.377
INLP	0.473	-0.301	0.787	-0.280	0.460
Sentence Debias	0.728	0.003	0.985	0.038	0.439