The Impact of Debiasing on the Performance of Language Models in Downstream Tasks is Underestimated

Masahiro Kaneko[1,2] **Danushka Bollegala**[3,4*] **Naoaki Okazaki**[2]
[1]MBZUAI [2]Tokyo Institute of Technology [3]University of Liverpool [4]Amazon
Masahiro.Kaneko@mbzuai.ac.ae
danushka@liverpool.ac.uk okazaki@c.titech.ac.jp

Abstract

Pre-trained language models trained on large-scale data have learned serious levels of social biases. Consequently, various methods have been proposed to debias pre-trained models. Debiasing methods need to mitigate only discriminatory bias information from the pre-trained models, while retaining information that is useful for the downstream tasks. In previous research, whether useful information is retained has been confirmed by the performance of downstream tasks in debiased pre-trained models. On the other hand, it is not clear whether these benchmarks consist of data pertaining to social biases and are appropriate for investigating the impact of debiasing. For example in gender-related social biases, data containing female words (e.g. *"she, female, woman"*), male words (e.g. *"he, male, man"*), and stereotypical words (e.g. *"nurse, doctor, professor"*) are considered to be the most affected by debiasing. If there is not much data containing these words in a benchmark dataset for a target task, there is the possibility of erroneously evaluating the effects of debiasing. In this study, we compare the impact of debiasing on performance across multiple downstream tasks using a wide-range of benchmark datasets that containing female, male, and stereotypical words. Experiments show that the effects of debiasing are consistently *underestimated* across all tasks. Moreover, the effects of debiasing could be reliably evaluated by separately considering instances containing female, male, and stereotypical words than all of the instances in a benchmark dataset.

1 Introduction

Unfortunately, Pre-trained Language Models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) easily learn discriminatory social biases expressed in human-written texts in massive datasets (Kurita et al., 2019; Zhou et al., 2022; Kaneko et al., 2022). For example, if a model is given "*[MASK] is a nurse.*" as the input, a gender biased PLM would predict "*She*" with a higer likelihood score than for "*He*" when filling the [MASK]. Various debiasing methods have been proposed to mitigate social biases in PLMs. Zhao et al. (2019); Webster et al. (2020) proposed a debiasing method by swapping the gender of female and male words in the training data. Kaneko and Bollegala (2021) proposed a method for debiasing by orthogonalising the vectors representing gender information with the hidden layer of a language model given a sentence containing a stereotypical word. Webster et al. (2020) showed that dropout regularization can reduce overfitting to gender information, thereby can be used for debiasing PLMs.

	All	Female	Male	Occ.
CoLA	1,043	174	722	96
MNLI	9,832	3,467	8,875	1,415
MRPC	408	101	391	96
QNLI	5,463	2,149	5,371	1,066
QQP	40,430	7,415	29,638	3,331
RTE	277	113	269	94
SST-2	872	187	691	75
STS-B	1,500	513	1,277	151
WNLI	71	27	71	6

Table 1: The total number of instances containing female, male, and occupational (Occ.) words in the GLUE development data.

The debiasing method should mitigate only discriminatory information, while pre-trained useful information should be retained in the model. Evaluations in downstream tasks often employ the GLEU benchmark (Wang et al., 2018), which measures the ability to understand language (Kaneko and Bollegala, 2021; Guo et al., 2022; Meade et al., 2022). The data for downstream tasks are not selected in terms of whether they reflect the impact of

[*] Danushka Bollegala holds concurrent appointments as a Professor at University of Liverpool and as an Amazon Scholar. This paper describes work performed at the University of Liverpool and is not associated with Amazon.

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 29–36
November 1–4, 2023. ©2023 Association for Computational Linguistics

debiasing. To mitigate gender bias, data containing female words such as *"she"* and *"woman"*, male words such as *"he"* and *"man"*, and stereotypical words such as *"doctor"* and *"nurse"* would be most affected by debiasing.

Table 1 shows the total number of instances containing female, male, and occupational (Occ.) words in the development data in the GLUE benchmark suite (Wang et al., 2019), which is widely recognised as a standard evaluation benchmark for LLMs. Occupational words have been used for probing LLMs for stereotypical social biases (Bolukbasi et al., 2016). From Table 1, we see that the GLEU benchmark has little data related to females and occupations. Therefore, the impact of debiasing on data related to females and occupations may be potentially underestimated when LLMs are evaluated on GLUE.

We first extract instances containing female words, data containing male words, and data containing stereotypical words from the benchmarks. We then calculated the performance difference between the original model and the debiased model for each category and compared it to the performance difference using the entire benchmark. The results showed that the debiased model performed worse than the original model on data related to females and occupations compared to the original model when evaluated on the entire dataset. Therefore, existing evaluations underestimate the impact of debiasing on the performance of the downstream task.

It is important to be able to compare how well the effects of debiasing are captured in the datasets related to females, males, and occupations. We propose a method to control the degree of debiasing of PLMs and investigate whether the performance difference between original and debiased models widens as the degree of debiasing increases. Experimental results showed that the proportion of female, male and occupational words in the dataset is related to the susceptibility of the dataset to debiasing.

2 Experiments

2.1 Debiasing Methods

We use the following three commonly used debiasing methods in our experiments. We apply these debiasing methods during fine-tuning in downstream tasks.

Counterfactual Data Augmentation (CDA) debiasing: CDA debiasing (Webster et al., 2020) swaps the gender of gender words in the training data. For example, *"She is a nurse"* is swapped to *"He is a nurse"*, and the swapped version is appended to the training dataset. This enables to learn a less biased model because the frequency of female and male words will be the same in the augmented dataset.

Dropout debiasing: Webster et al. (2020) introduced dropout regularisation as a method to mitigate biases. They enhanced the dropout parameters for the attention weights and hidden activations of PLMs. Their research demonstrated that intensified dropout regularisation diminishes gender bias in these PLMs. They showed that dropout interfers with the attention mechanism in PLMs, and prevents undesirable associations between words. However, it is also possible that the model may no longer be able to learn desirable associations.

Context debiasing: Kaneko and Bollegala (2021) proposed a method to debias MLMs through fine-tuning. It preserves semantic information while removing gender-related biases using orthogonal projections at token- or sentence-level. This method targets male and female words and occupational words in the text for debiasing. This method can be applied various MLMs, independent of the model architectures and pre-training methods. Token-level debiasing across all layers produces the best performance.

2.2 Settings

Although we use BERT (bert-base-cased[1]) (Devlin et al., 2019) as our PML here as it has been the focus of much prior work on bias evaluations (Kaneko and Bollegala, 2021; Guo et al., 2022; Meade et al., 2022), the evaluation protocol we use can be applied to any PLM. We used the word lists[2] proposed by Kaneko and Bollegala (2021) as female words, male words, and occupational words for extracting data instances and debiasing.

We use the following nine downstream tasks from the GLEU benchmark: CoLA (Warstadt et al., 2019), MNLI (Williams et al., 2018), MRPC (Dolan and Brockett, 2005), QNLI (Ra-

[1] `https://huggingface.co/bert-base-cased`

[2] `https://github.com/kanekomasahiro/context-debias`

	CDA				Dropout				Context			
	All	Female	Male	Occ.	All	Female	Male	Occ.	All	Female	Male	Occ.
CoLA	-1.36	**-3.42**	-2.01	-1.45	0.42	-0.14	**-0.21**	-0.07	-0.32	**-0.86**	-0.71	-0.55
MNLI	-0.55	**-0.90**	-0.71	-0.63	0.23	0.13	**0.01**	0.05	-0.05	**-0.47**	-0.43	-0.32
MRPC	-0.96	-1.28	**-1.31**	-1.03	-0.82	**-1.12**	-1.02	-1.04	-0.88	-1.01	**-1.06**	-0.92
QNLI	-1.13	**-1.42**	-1.19	-1.27	-1.01	-1.11	-1.07	**-1.21**	0.25	**-0.19**	-0.06	-0.04
QQP	-0.21	**-0.69**	-0.32	-0.25	0.53	**0.13**	0.47	0.30	0.14	**-0.12**	0.03	-0.05
RTE	-1.16	**-1.21**	-1.02	-1.13	-1.01	**-1.24**	-0.96	-1.13	-0.43	-0.65	-0.51	**-0.73**
SST-2	-0.11	**-0.81**	-0.34	-0.25	0.45	0.20	**0.12**	0.23	0.22	**-0.15**	-0.02	-0.12
STS-B	-1.01	**-1.95**	-1.34	-1.10	0.21	0.09	-0.03	**-0.11**	-0.08	-0.31	**-0.38**	-0.34
WNLI	-2.82	**-3.07**	-2.82	-2.71	-2.01	-2.21	-2.01	**-2.33**	-1.52	**-1.88**	-1.52	-1.61

Table 2: Performance difference between the original model and debiased model for each dataset. **Bolded** values indicate the largest drop in performance of the debiased model.

jpurkar et al., 2016), QQP[3], RTE (Dagan et al., 2006; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), SST-2 (Socher et al., 2013), STS-B (Cer et al., 2017), and WNLI (Levesque et al., 2012). Hyperparameters for debiasing follow previous studies (Kaneko and Bollegala, 2021; Webster et al., 2020), and we used the default values of huggingface for downstream task hyperparameters.[4] For fine-tuning we use the entire training dataset for each corresponding task, without splitting into male, female and occupational instances. We evaluate the performance on all tasks using the official development data.

2.3 Performance of Original vs. Debiased Models

We extract instances containing female words, male words, and stereotypical words from each of the datasets. We then calculate the performance difference between the original model and the debiased model for each dataset, and compare against the performance differences obtained when using all instances. If the performance difference for all instances is smaller than that when evaluated for the female, male, and occupational instances, it would indicate that the effect of debiasing is underestimated when evaluated on the entire dataset.

Table 2 shows the performance differences between the original model and the debiased model for each dataset/task in the GLEU benchmark. All, Female, Male, and Occ. are the performance differences when evaluated on the entire task dataset, instances containing female words, instances containing male words, and instances containing occupational words, respectively.

From the results in Table 2, it can be seen that the performance difference between the original model and the debiased model is larger for the Female, Male, and Occ. instances compared to that when using all instances. In particular, instances related to females exhibit a significant decrease in performance after debiasing.

It can be seen that different word lists used for debiasing have different effects on the performance degradation in downstream tasks due to debiasing. Context debiasing uses occupational words for debiasing, while CDA debiasing does not. Therefore, in CDA debiasing, Occ. does not have a large performance difference compared to female- and male-related instances. Therefore, in CDA debiasing, the performance difference for occupation-related instances is smaller than that for the female and male-related instances. On the other hand, in Context debiasing, occupation-related instances has the largest performance difference as well as female- and male-related instances. Dropout debiasing does not use word lists for debiasing. Therefore, unlike CDA and context debiasing, we see large drops in performance for female, male and Occ. across tasks with Dropout debiasing.

2.4 Debias Controlled Method

To understand how debiasing of an PLM affects the performance of individual downstream benchmark datasets, following the probing technique proposed by Kaneko et al. (2023), we apply different levels of debiasing to bert-base-cased PLM and measure the difference in performance with respect to its original (non-debiased) version. For this purpose we use CDA as the debiasing method, where we swap the gender-related pronouns in $r \in [0, 1]$ fraction of the total N instances of a dataset (i.e.the total number of gender swapped instances in a dataset will be

[3] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

[4] https://github.com/huggingface/transformers/tree/main/examples/pytorch/text-classification

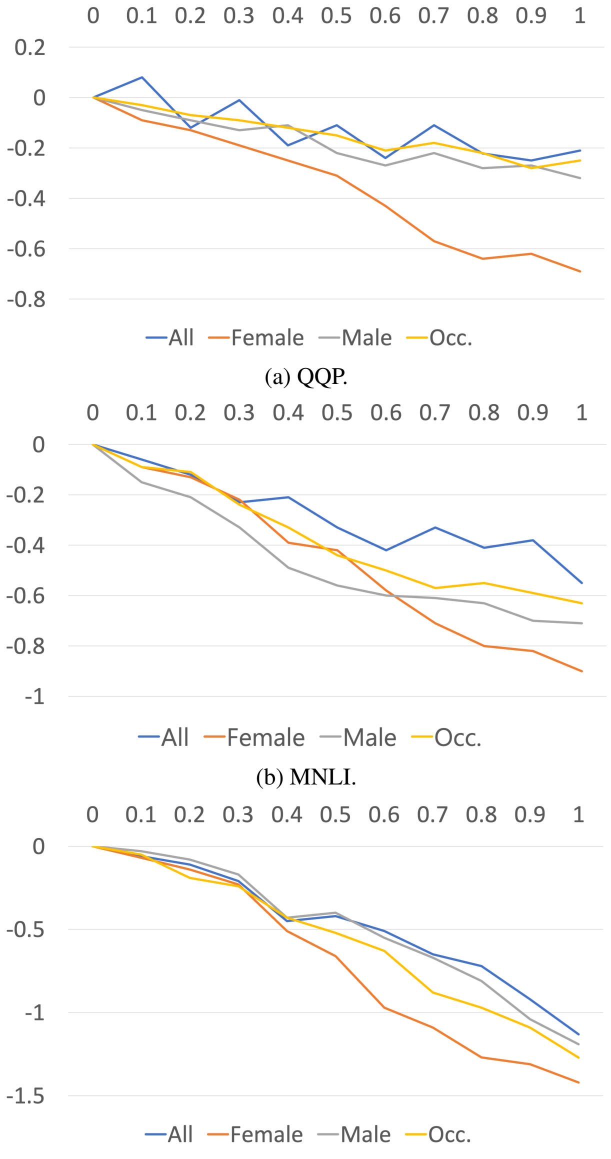
$r \times N$). $r = 0$ corresponds to not swapping gender in any training instances of the dataset, whereas $r = 1$ swaps the gender in all instances. We increment r in step size 0.1 to obtain increasingly debiased version of the PLM, which is then fine-tuned for the downstream task[5]. Figure 1 shows the difference in performance between the original vs. debiased versions of the PLM for QQP, MNLI, and QNLI, which have the largest numbers of instances in the GLEU benchmark.

Note that CDA debiasing reverses gender without considering the context, as in *"He gets pregnant"* for *"She gets pregnant"*. This is problemantic because it eliminates even useful gender-related information learnt by the PLM via co-occurring contexts. Therefore, CDA debiasing has a negative impact on the performance of downstream tasks (Zmigrod et al., 2019) as shown by all three subplots in Figure 1. In fact, Table 2 shows that the performance difference of CDA debiasing is larger than that of dropout debiasing and context debiasing. Therefore, the larger r is for CDA, the more gender instances is balanced and debiased, but the performance is unfortunately degraded.

If the dataset of the downstream task is sensitive to the effect of debiasing, the performance difference between the original model and the debiased model widens as r increases. On the other hand, if the data set is insensitive to the effect of debiasing, the performance difference between the original model and the debiased model is unlikely to increase with the value of r.

We find that the performance differences for the female, male, and occupational instances in the QQP, MNLI, and QNLI datasets increase with the value of r. On the other hand, for QQP and MNLI, there is a rise and fall in the performance difference when all data are used. These results indicate that All, which includes instances related to non-gender, are less sensitive to the effect of debiasing compared to Female, Male, and Occupational instances.

On the other hand, for QNLI, All has a small rise and fall in the performance difference. As seen from Table 1, QNLI contains more gender-related instances than QQP and MNLI. Therefore, it is likely that the performance decreases with r even for All. All and Male instances have a similar trend in performance difference with r.



(a) QQP.

(b) MNLI.

(c) QNLI.

Figure 1: Performance difference between original and debiased models by debias rate r. The vertical axis shows the performance difference between the original and debiased models, and the horizontal axis shows the debias rate.

3 Conclusion

This study focused on gender-related social biases and the presence of female, male, and stereotypical words in benchmark datasets. Prior work had used the performance on downstream tasks to prove the usefulness of debiasing methods, overlooking the fact that only a small fraction of those downstream benchmark datasets contain gender-related instances. On the contrary, we found that the effects of debiasing an PLM were consistently underestimated across all tasks. We recommend that the

[5] In Appendix A, we show that the debias controlled method is able to debias the model according to r.

evaluation of debiasing effects must be more reliably conducted by considering instances containing specific gender-related words separately rather than evaluating all instances in a benchmark dataset.

4 Ethical Considerations

This study uses existing methods and datasets for experiments and does not propose a debiasing method or create a new dataset for social bias. This study evaluates the impact of debiasing on the performance of the downstream task, and it is not possible to evaluate how much bias is mitigated in the PLMs. Therefore, when evaluating the bias of PLMs, it is necessary to use evaluation methods such as StereoSet (Nadeem et al., 2021), Crowds-Pairs (Nangia et al., 2020), and All Unmasked Likelihood (Kaneko and Bollegala, 2022).

In this study, we only included binary gender as a gender bias. However, gender bias regarding non-binary gender has also been reported (Cao and Daumé III, 2020; Dev et al., 2021). It is necessary to verify whether there is a similar trend in debiasing for non-binary genders.

5 Limitations

Many previous studies have shown that various social biases other than gender bias are learned in PLMs. This study targets only gender bias. While existing studies (Webster et al., 2020; Zhao et al., 2019) have debiasing various PLMs, we have experimented only with bert-base-cased. Furthermore, although this study targets only English, which is a morphologically limited language. On the other hand, various types of social biases are also learned in the PLMs across many languages (Kaneko et al., 2022; Névéol et al., 2022). Therefore, if the proposed method is to be used with other social biases and PLMs, it is necessary to properly verify its effectiveness in languages other than English. Moreover, we have not verified the use of debias controlled methods in languages such as Spanish and Russian, where gender swapping is not easy from a grammatical point of view (Zmigrod et al., 2019).

Acknowledgements

These research results were obtained from the commissioned research (No.225) by National Institute of Information and Communications Technology (NICT), Japan.

References

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*. Citeseer.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190. Springer.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7659–7666.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask–evaluating social biases in masked language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11954–11962.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2023. Comparing intrinsic gender bias evaluation measures without using human annotated examples. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2857–2863, Dubrovnik, Croatia. Association for Computational Linguistics.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

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.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019.

GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yi Zhou, Masahiro Kaneko, and Danushka Bollegala. 2022. Sense embeddings are also biased – evaluating social biases in static and contextualised sense embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1924–1935, Dublin, Ireland. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

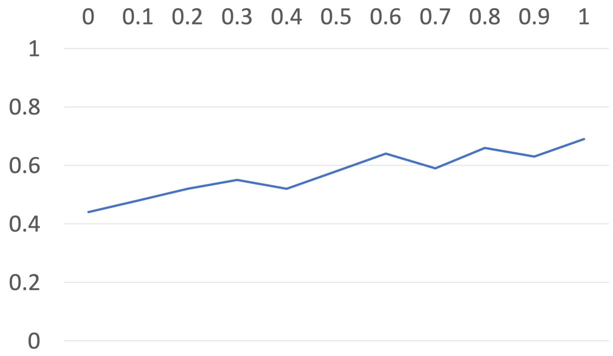


Figure 2: Evaluation of debias controlled models using FN evaluation method. The vertical axis shows the bias score, and the horizontal axis shows r.

A Bias Evaluation in NLI task

We show that the debias controlled method is appropriately debiasing according to r. We use Fraction Neutra (**FN**; Dev et al., 2020) as the bias evaluation method. The FN method evaluates bias in the NLI by considering the percentage of neutral labels predicted by the model for the premise sentence (e.g. *The driver owns a cabinet.*) and the hypothesis sentence (e.g. *The man owns a cabinet.*) generated with the template. The FN method indicates that the lower the score, the more bias there is in the model. We evaluate PLMs trained on MNLI with FN method.

Figure 2 shows the bias scores of FN method for each debias controlled model. It can be seen that the bias of the model is decreasing with r. Therefore, the debias controlled method is able to debias the models according to r.