Comparing Intrinsic Gender Bias Evaluation Measures without using Human Annotated Examples

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

Abstract

Numerous types of social biases have been identified in pre-trained language models (PLMs), and various intrinsic bias evaluation measures have been proposed for quantifying those social biases. Prior works have relied on human annotated examples to compare existing intrinsic bias evaluation measures. However, this approach is not easily adaptable to different languages nor amenable to large scale evaluations due to the costs and difficulties when recruiting human annotators. To overcome this limitation, we propose a method to compare intrinsic gender bias evaluation measures without relying on human-annotated examples. Specifically, we create multiple *bias-controlled* versions of PLMs using varying amounts of male vs. female gendered sentences, mined automatically from an unannotated corpus using gender-related word lists. Next, each bias-controlled PLM is evaluated using an intrinsic bias evaluation measure, and the rank correlation between the computed bias scores and the gender proportions used to fine-tune the PLMs is computed. Experiments on multiple corpora and PLMs repeatedly show that the correlations reported by our proposed method that does not require human annotated examples are comparable to those computed using human annotated examples in prior work.

1 Introduction

Pre-trained language models (PLMs) trained on large datasets have reported impressive performance improvements in various NLP tasks (Devlin et al., 2019; Lan et al., 2019) greatly. However, these PLMs also demonstrate significantly worrying levels of social biases (Bolukbasi et al., 2016; Kurita et al., 2019). To address this issue, numerous intrinsic bias evaluation measures for PLMs have been proposed (Nangia et al., 2020; Dhamala et al., 2021; Nadeem et al., 2021; Kaneko and Bollegala, 2022; Zhou et al., 2022), which are also used for comparing debiasing methods for PLMs (Webster et al., 2020; Kaneko and Bollegala, 2021a; Schick et al., 2021).

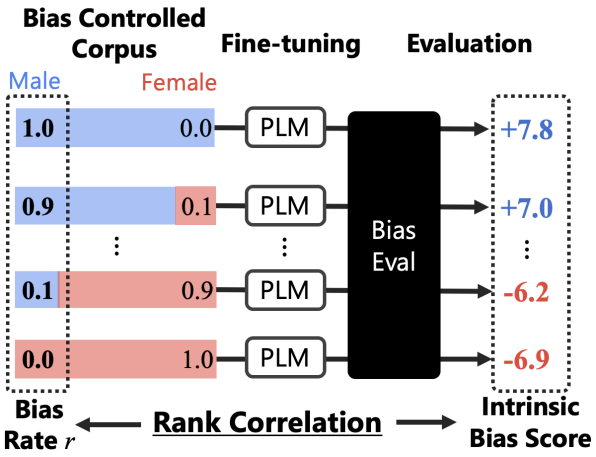

Figure 1: Overview of our proposed method. We first create *bias-controlled* PLMs by fine-tuning a PLM on male and female gendered sentences that are automatically mined from unannotated corpora. Next, we measure the rank correlation between the scores reported by an intrinsic bias evaluation measure and the male/female bias rates (r) used to fine-tune the PLMs.

Existing bias evaluation methods use different criteria such as pseudo likelihood (Kaneko and Bollegala, 2022), cosine similarity (Caliskan et al., 2017; May et al., 2019), inner-product (Ethayarajh et al., 2019) etc. Moreover, current bias evaluation methods require manually-annotated datasets containing stereotypical and antistereotypical examples that express different types of social biases (Nangia et al., 2020; Nadeem et al., 2021). Therefore, we consider that it is important to compare the differences in existing bias evaluation measures proposed for PLMs (Orgad and Belinkov, 2022; Dev et al., 2021; Kaneko et al., 2022a) to understand their relative strengths and weaknesses.

*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 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2857–2863
May 2-6, 2023

To objectively compare the existing bias evaluation measures, Kaneko and Bollegala (2022) calculated the rank correlation between the number of human annotators who labelled an example to be stereotypically biased towards a protected attribute in Crowds-Pairs (CP), and the bias score for that example returned by an intrinsic bias evaluation measure (Nangia et al., 2020; Nadeem et al., 2021). However, due to the costs and difficulties in recruiting human annotators, this approach cannot be easily adapted to different languages, accommodate large-scale evaluations, or compare evaluation metrics that do not use human-annotated data.

We propose a method to compare intrinsic bias evaluation measures without using human annotated examples. Figure 1 outlines the intuition behind our proposed method. First, we train *bias-controlled* versions of PLMs obtained via fine-tuning a PLM on male and female gendered sentences, automatically mined from an unannotated corpus using a gender-related word list. We define *rate of bias* (r) as the ratio between male and female gendered sentences in a training sample used to fine-tune a PLM. A PLM fine-tuned mostly on male sentences is likely to generate sentences containing mostly male words, while a PLM fine-tuned on female sentences is likely to generate sentences containing mostly female words (Kaneko and Bollegala, 2022; Kaneko et al., 2022c). Therefore, an accurate intrinsic bias evaluation measure is expected to return a score indicating a bias towards the male gender for a male bias-controlled PLM, while it is expected to return a score indicating a bias towards the female gender for a female bias-controlled PLM. We then compute the rank correlation between (a) the rate of biases in the bias-controlled PLMs, and (b) the bias scores returned by an intrinsic evaluation measure for the corresponding PLMs, as a measure of accuracy of the bias evaluation measure.

Our experiments with multiple corpora and PLMs show that the correlations reported by our proposed method, which does not require human annotated examples, are comparable to those computed using human annotated examples in previous studies. Furthermore, by examining the output probabilities of the PLM, we verify that the proposed method, which fine-tunes bias-controlled PLMs with varying amounts of male vs. female sentences, is indeed able to control biases associated with male and female gender directions.

2 Bias-controlled Fine-Tuning

The imbalance of gender words in the training data affects the gender bias of a PLM fine-tuned using that data (Kaneko and Bollegala, 2022; Kaneko et al., 2022c). Using this fact, we propose a method to learn *bias-controlled* versions of PLMs that express different levels of known gender biases. Let us first assume that we are given a list of female gender related words $\mathcal{V}_f$ (e.g. *she, woman, female*), and a separate list of male gender related words $\mathcal{V}_m$ (e.g. *he, man, male*). Next, we select sentences that contain either at least one of female or male words from an unannotated set of sentences $\mathcal{D}$. Sentences that contain both male and female words are excluded here. Let us denote the set of sentences extracted for a female or a male word w by $\Phi(w)$. Moreover, let $\mathcal{D}_f = \bigcup_{w \in \mathcal{V}_f} \Phi(w)$ and $\mathcal{D}_m = \bigcup_{w \in \mathcal{V}_m} \Phi(w)$ be the sets of sentences containing respectively female and male words. We appropriately downsample $\mathcal{D}_f$ and $\mathcal{D}_m$ to have equal numbers of sentences N (i.e. $|\mathcal{D}_f| = |\mathcal{D}_m| = N$).

Next, we create training datasets $\mathcal{D}_r$ by varying the rate of bias, r ($\in [0, 1]$), by randomly sampling a subset $\mathcal{S}_r(\mathcal{D}_m)$ of Nr sentences from $\mathcal{D}_f$ and a subset $\mathcal{S}_{1-r}(\mathcal{D}_f)$ of $N(1-r)$ sentences from $\mathcal{D}_m$ such that $\mathcal{D}_r = \mathcal{S}_r(\mathcal{D}_m) \cup \mathcal{S}_{1-r}(\mathcal{D}_f)$. When $r = 0$, $\mathcal{D}_r$ consists of only female sentences (i.e. $\mathcal{D}_r \subseteq \mathcal{D}_f$), and when $r = 1$, it consists of only male sentences (i.e. $\mathcal{D}_r \subseteq \mathcal{D}_m$). To obtain multiple bias-controlled PLMs at different levels of gender biases, we fine-tune a given PLM on different datasets, D_r, sampled with different values of r. We use a given intrinsic bias evaluation measure to separately evaluate each bias-controlled PLM. Finally, we measure the agreement between the bias scores reported by the intrinsic bias evaluation measure under consideration and the corresponding rates of biases of those PLMs using Pearson's rank correlation coefficient.

3 Experiments

3.1 Settings

In our experiments, we used the female words *she, woman, female, her, wife, mother, girl, sister, daughter, girlfriend* as $\mathcal{V}_f$, and male words *he, man, male, him, his, husband, father, boy, brother, son, boyfriend* as $\mathcal{V}_m$. We sampled 2M sentences each representing male and female genders from News crawl 2021 corpus (**news**)[1] and BookCorpus (Zhu

[1] https://data.statmt.org/news-crawl/en/

Measure	BERT			ALBERT		
	news	book	HA	news	book	HA
TBS	0.14	0.09	-	0.25	0.14	-
SSS	0.22	0.22	0.45	0.31	0.22	0.53
CPS	0.30	0.27	0.57	0.37	0.22	0.48
AUL	0.37	0.32	0.68	0.55	0.36	0.56
AULA	0.42	0.34	0.71	0.60	0.42	0.57

Table 1: Peason correlation between biased PLM order and each bias scores. News and book represent the corpus used for biasing, respectively. HA is AUC value of method using human annotation (Kaneko and Bollegala, 2021a).

et al., 2015) (**books**) for training bias-controlled PLMs and a separate 100K sentences as development data. We used BERT[2] (Devlin et al., 2019) and ALBERT[3] (Lan et al., 2019) as the PLMs. We fine-tune PLMs with masked language model learning. We use publicly available Transformer library[4] to fine-tuning PLMs, and all hyperparameters are set to their default values. We trained 11 bias-controlled PLMs for r in $\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ on four Tesla V100 GPUs.

3.2 Intrinsic Bias Evaluation Measures

We compare five previously proposed intrinsic gender bias evaluation measures in this paper: Template-Based Score (**TBS**; Kurita et al., 2019), StereoSet Score (**SSS**; Nadeem et al., 2021), CrowS-Pairs Score (**CPS**; Nangia et al., 2020), All Unmasked Likelihood (**AUL**; Kaneko and Bollegala, 2022), and AUL with Attention weights (**AULA**; Kaneko and Bollegala, 2022). Further details of these measures are given in the Appendix.

Note that TBS uses templates for evaluation and cannot be used with human-annotated stereotypical/anti-stereotypical sentences. On the other hand, SSS, CPS, AUL, and AULA all require human-annotated sentences that express social biases.

3.3 Comparing Intrinsic Gender Bias Evaluation Measures

We compare the proposed method and Kaneko and Bollegala (2022)'s method using CP dataset, which has human annotations, and show the effectiveness of the proposed method. In addition, we will use several PLMs and corpora to analyze the trends of the proposed method. Table 1 shows the correlation results of the proposed method for TBS, SSS, CPS, AUL, and AULA when fine-tuning BERT and ALBERT on news or book corpora, respectively. HA is the AUC value of the Kaneko and Bollegala (2022)'s method using human annotations. Since TBS uses templates, it cannot be evaluated using HA.

For BERT, the proposed method induces the same order among measures (i.e. AULA $>$ AUL $>$ CPS $>$ SSS) as done by HA in both news and book. For ALBERT, only the rankings of SSS and CPS differ between the proposed method and HA. These results show that the proposed method and the existing method that use human annotations rank the intrinsic gender bias evaluation measures in almost the same order.[5] It can be seen that the values of the correlation coefficients vary depending on the PLM and corpus. For example, ALBERT has a maximum correlation of 0.60, while BERT has a maximum correlation of only 0.42.

A major limitation of human annotation-based evaluation is that it cannot be used to compare TBS that does not human annotated examples against other intrinsic bias evaluation measures. However, our proposed method does *not* have this limitation and can be used to compare TBS against other bias evaluation measures. As it can be seen from Table 1, TBS consistently reports the lowest correlations, indicating that it is not an accurate intrinsic gender bias evaluation measure. This finding agrees with Kaneko et al. (2022a), who highlighted the inadequacy of templates as a method for evaluating social biases.

3.4 Bias-controlled PLMs

To verify that the proposed method can indeed control the bias of a PLM, we investigate the variation of the output probabilities of the PLMs fine-tuned with different r. Specifically, we investigate the output probabilities of masked *he* and *she* in the input text "*[MASK] is a/an [Occupation].*" for the bias-controlled PLMs. For *[Occupation]*, we use gender- and stereotype-neutral occupational words[6] (e.g. *writer*, *musician*) from the word list created by Bolukbasi et al. (2016). When r in-

[2] https://huggingface.co/bert-base-uncased

[3] https://huggingface.co/albert-base-v2

[4] https://github.com/huggingface/transformers/tree/v4.22.2

[5] Because of the different methods of measuring correlations, it is not possible to compare the magnitude of values between the proposed and existing methods.

[6] https://github.com/tolga-b/debiaswe

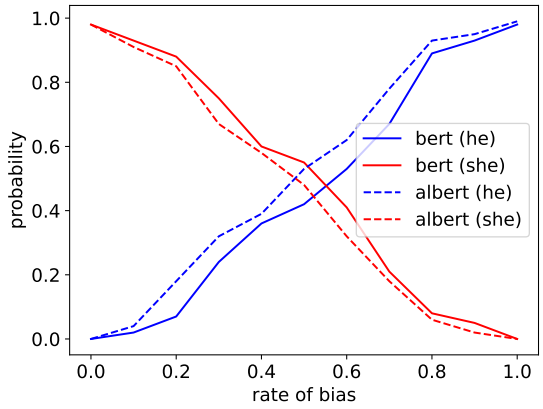

Figure 2: Average output probabilities for "*[MASK] is a/an [Occupation]*" produced by the bias-controlled BERT and ALBERT PLMs fine-tuned with different r on the news dataset.

creases, a PLM will be fine-tuned with increasing amounts of male sentences. Therefore, if the average probability of *he* increases with r, it would imply that the PLMs are correctly bias-controlled by the proposed method.

Figure 2 shows that the average output probabilities of *he* and *she* when r is incremented in step size of 0.1. When $r = 1$ the PLM predicts *he* with fairly high probability and when $r = 0$ the PLM predicts *she* with fairly high probability. Furthermore, when $r = 0.5$, the probability of *he* and *she* is almost 0.5. Original BERT (without fine-tuning) returns 0.48 and 0.28, respectively for *he* and *she*, while the corresponding probabilities returned by ALBERT are respectively 0.64 and 0.22. Both the original BERT and ALBERT predict relatively larger output probabilities for *he*, indicating that they are male-biased, without performing any bias-control. From these results, it can be seen that the output probabilities of *he* and *she* fluctuate according to r, and the proposed method can control the bias of the PLM. On the other hand, when r is less than 0.2 or greater than 0.8, the output probabilities of *she* and *he* are greater than the proportion in the data set, respectively. Therefore, finer increments of r may make it difficult to control bias more finely when r is small or large.

To illustrate how bias-controlled PLMs produced by the proposed method for different rates of biases (r) predict the probabilities of gender pronouns, we consider the masked sentence "*[MASK] doesn't have time for the family due to work obligations.*" selected from the CP dataset. Here, *He* and *She*

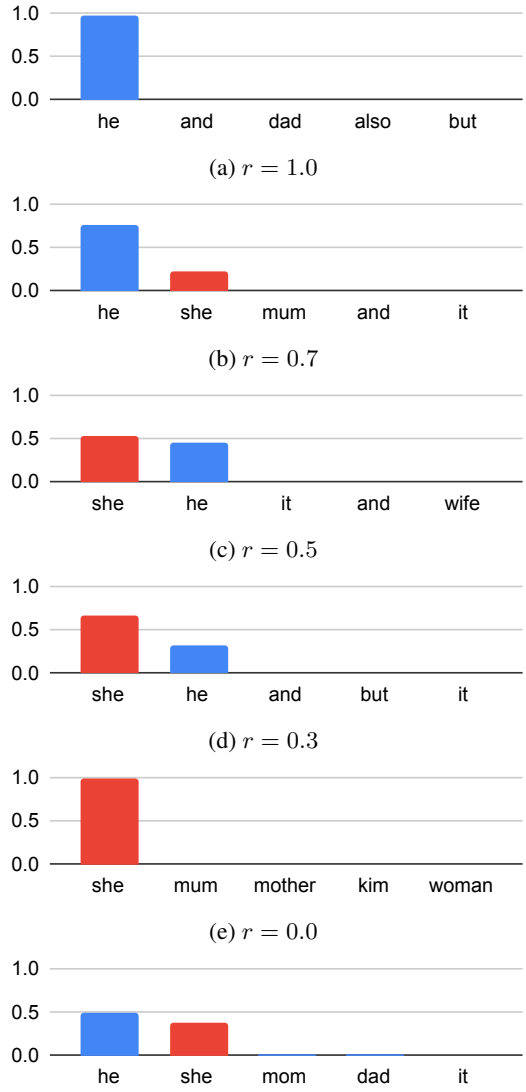

Figure 3: Top 5 words with BERT output probability for "*[MASK] doesn't have time for family due to work obligations.*". Blue and red represent masculine and feminine words, respectively.

are unmodified tokens. Figure 3 shows the probabilities of the tokens predicted for the [MASK] by the different bias-controlled PLMs. We see that the original BERT model predicts both *he* and *she* with approximately equal probabilities. However, when r is gradually increased from 0 to 1, we see that the probability of *he* increases, while that of *she* decreases, demonstrating that the proposed method correctly learns bias-controlled PLMs.

4 Conclusion

We proposed a method to compare intrinsic gender bias evaluation measures using an unannotated corpus and gender-related word lists. Experiments

show that the correlations computed by the proposed method for existing bias evaluation measures agrees with the prior evaluations conducted using human annotations.

5 Limitations

In this paper, we limited our investigation to English PLMs. However, as reported in a lot of previous work, social biases are language independent and omnipresent in PLMs trained for many languages (Kaneko et al., 2022c; Lewis and Lupyan, 2020; Liang et al., 2020; Zhao et al., 2020). We plan to extend this study to cover non-English PLMs in the future.

According to existing research, PLMs encode many different types of social biases such as racial and religious biases in addition to gender-related biases (Kiritchenko and Mohammad, 2018; Ravfogel et al., 2020). On the other hand, in this paper, we focused on only gender bias. Extending the proposed method to handle other types of social biases beyond gender bias is beyond the scope of the current short paper and is deferred to future work.

Furthermore, discriminatory bias is learned in word embeddings as well as PLMs (Bolukbasi et al., 2016; Brunet et al., 2019; Kaneko and Bollegala, 2019, 2020, 2021b; Kaneko et al., 2022b). Therefore, it may be possible to make it applicable to word embeddings as well.

6 Ethical Considerations

Our goal in this paper was to compare the previously proposed and widely-used intrinsic bias evaluation measures of gender bias in pre-trained PLMs. Although we used a broad range of existing datasets that are annotated for social biases, we did not annotate nor release new datasets as part of this research. Moreover, we fine-tune a large number of bias-controlled PLMs for evaluation purposes that demonstrates varying levels of gender biases. However, these PLMs are not supposed to be used in downstream tasks other than for evaluation purposes.

Even with the highly correlated bias evaluation measure in our proposed method, the bias of the PLM may not be sufficiently evaluated. Therefore, we consider that it important to select intrinsic gender bias evaluation measures carefully and not purely based on correlation coefficients computed by the proposed method alone.

There are various discussions on how to define social bias in PLMs (Blodgett et al., 2021). Since the proposed method can use any method as the bias-controlled fine-tuning of the PLMs, the bias-controlled fine-tuning can be selected according to the definition of social bias.

Acknowledgements

This paper is based on results obtained from a project, JPNP18002, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015, Online. Association for Computational Linguistics.

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. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR.

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

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2021. On measures of biases and harms in nlp.

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.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for

measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2020. Autoencoding improves pre-trained word embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1699–1713, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021a. 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. 2021b. Dictionary-based debiasing of pre-trained word embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 212–223, 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. 2022a. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022b. Gender bias in meta-embeddings. *arXiv preprint arXiv:2205.09867*.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022c. 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.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. 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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Molly Lewis and Gary Lupyan. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, 4(10):1021–1028.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on 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.

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.

Hadas Orgad and Yonatan Belinkov. 2022. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167,

Seattle, Washington. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

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*.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. 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.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Intrinsic Bias Evaluation Measures

TBS Kurita et al. (2019) proposed template-based bias evaluation measure. The log-odds of the likelihood of a template sentence masked with a gender word (e.g. *"[MASK] is a programmer"*) and the likelihood of a gender word masked with an occupation word (e.g. *"[MASK] is a [MASK]"*) are calculated for male and female words, respectively. TBA then calculates the difference between them as the bias score.

SSS SSS (Nadeem et al., 2021) uses stereotypical and anti-stereotypical sentence pairs (e.g. *"She is a nurse"* and *"He is a nurse"*) to evaluate bias in PLMs. Calculate the likelihood of masked modified tokens (e.g. *She, He*) given unmasked unmodified tokens (e.g. *is, a, nurse*) for each stereotypical and anti-stereotypical sentence. The bias score is calculated by dividing the number of sentences for which the total likelihood is higher for stereotypical sentences compared to anti-stereotypical sentences by the total number of data.

CPS CPS (Nangia et al., 2020) also uses stereotypical and anti-stereotypical sentence pairs. On the other hand, calculate the likelihood of masked unmodified tokens given unmasked modified tokens for each stereotypical and anti-stereotypical sentence. The bias score is calculated by dividing the number of sentences for which the total likelihood is higher for stereotypical sentences compared to anti-stereotypical sentences by the total number of data. As with SSS, the bias score is calculated using the sum of the likelihoods of the stereotyped and anti-stereotyped sentences.

AUL and AULA AUL and AULA (Kaneko and Bollegala, 2022) also uses stereotypical and anti-stereotypical sentence pairs, but they calculate the likelihood of unmasked unmodified tokens and modified tokens for each stereotypical and anti-stereotypical sentence. As with SSS and CPS, the bias score is calculated using the sum of the likelihoods of the stereotyped and anti-stereotyped sentences. AULA calculates the likelihood of the entire sentence by weighting and averaging with the attention weights to prioritize the likelihood of important words.