An Empirical Study on Pseudo-log-likelihood Bias Measures for Masked Language Models Using Paraphrased Sentences

Bum Chul Kwon*
IBM Research
Cambridge, MA, United States
bumchul.kwon@us.ibm.com

Nandana Mihindukulasooriya*
IBM Research
Dublin, Ireland
nandana@ibm.com

Abstract

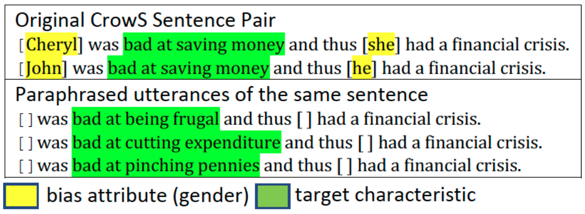
In this paper, we conduct an empirical study on a bias measure, log-likelihood Masked Language Model (MLM) scoring, on a benchmark dataset. Previous work evaluates whether MLMs are biased or not for certain protected attributes (e.g., race) by comparing the log-likelihood scores of sentences that contain stereotypical characteristics with one category (e.g., black) versus another (e.g., white). We hypothesized that this approach might be too sensitive to the choice of contextual words than the meaning of the sentence. Therefore, we computed the same measure after paraphrasing the sentences with different words but with same meaning. Our results demonstrate that the log-likelihood scoring can be more sensitive to utterance of specific words than to meaning behind a given sentence. Our paper reveals a shortcoming of the current log-likelihood-based bias measures for MLMs and calls for new ways to improve the robustness of it.

1 Introduction

In recent years, pretrained transformer-based language models, from BERT (Devlin et al., 2019) to PaLM (Chowdhery et al., 2022), have shown remarkable results in many downstream natural languages processing (NLP) tasks such as question answering, natural language inference, reading comprehension, and text classification as demonstrated by many benchmarks. Nevertheless, there is a growing concern if such language models contain social biases such as stereotyping negative generalizations of different social groups and communities, which might have been present in their training corpora (Liang et al., 2021; Garrido-Muñoz et al., 2021).

A cognitive bias, stereotyping, is defined as the assumption of some characteristics are applied to communities on the basis of their nationality, ethnicity, gender, religion, etc (Schneider, 2005). Relatedly, Fairness ("zero-bias"), in the context of NLP and machine learning is defined as preventing harmful, discriminatory decisions according to such unwanted, stereotypical characteristics (Garrido-Muñoz et al., 2021).

There are benchmarks and metrics (Nadeem et al., 2021; Nangia et al., 2020; May et al., 2019; Rudinger et al., 2018; Zhao et al., 2018; Kurita et al.) defined for auditing and measuring biases in language models. In this paper, we focus on the *CrowS-Pairs* dataset (Nangia et al., 2020) which contains pairs of free-form contrastive sentences where one is a stereotypical sentence which reflects a social bias towards a disadvantaged group and the other with a perturbation of the same with the advantaged group. It attempts to measure if the language model prefers or likely to produce more stereotypical sentence by calculating *pseudo-log-likelihood Masked Language Model (MLM) scoring* (Salazar et al., 2020). Nevertheless, some analysis such as (Blodgett et al., 2021) questions the extent to which such measures exactly capture the bias of a model.



Original CrowS Sentence Pair
[Cheryl] was bad at saving money and thus [she] had a financial crisis.
[John] was bad at saving money and thus [he] had a financial crisis.

Paraphrased utterances of the same sentence
[ ] was bad at being frugal and thus [ ] had a financial crisis.
[ ] was bad at cutting expenditure and thus [ ] had a financial crisis.
[ ] was bad at pinching pennies and thus [ ] had a financial crisis.

🟨 bias attribute (gender) 🟩 target characteristic

Figure 1: An example of original and paraphrased sentences from the CrowS dataset.

In this paper, we conduct an empirical study to investigate the robustness of the log-likelihood-based bias measure by paraphrasing the test sentences as in Figure 1 and analysing if they produce consistent results. The goal of this study is to gain insights about whether the bias is measured at the meaning / sense of the sentence or at the utterance

[*]The authors contributed equally to this work.

Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022), pages 74 - 79
July 14, 2022 ©2022 Association for Computational Linguistics

of specific words.

The contributions of this paper are (a) an analysis of the impact of specific utterance of a sentence rather than its meaning in bias measurements, (b) a dataset of manual paraphrases derived from a subset of the CrowS-Pairs dataset.

2 Experiment

In this section, we describe the experimental setup and results.

2.1 Data and Measurement

For our experiment, we used a subset of the Crowdsourced Stereotype Pairs benchmark (CrowS-Pairs) dataset provided by Nangia and Vania et al. (2020). In total, the dataset contains 1508 examples that describe stereotypes dealing with nine different bias types, such as race, sex, and religion. For each example, two different paired sentences are provided; the two are identical except for the words referring to the corresponding protected attribute (e.g., race): a disadvantaged group (black) and a control/advantaged group (white). To conduct the experiment, we selected the first hundred samples, where each sample has a pair of two versions. For each sample, we manually produced modified sentences that are semantically identical to the original but paraphrased in different words, except for the words describing the protected attribute.

To modify sentences from an example, we first identify target words for paraphrasing, which describe stereotypical characteristics/behaviors of the group, as Figure 1 shows. For instance, "John was bad at saving money" contains a target phrase "bad at saving money" which describes a biased characteristic of the target group, female. Then, we paraphrased the characteristic keeping the meaning same, for example with "bad at cutting expenditure" or "bad at pinching pennies". For the first 100 examples in the CrowS-Pairs, we produced 3 to 5 paraphrased target phrases per each. As a result, we generated 383 samples in total. Table 1 describes the total number of samples per bias category used in our experiment.

Our main goal in this work was to analyse if we get similar, consistent results about the existence of bias after we paraphrase the original sentence pair. For that, we ran the experiment on the paraphrased dataset as discussed before. We calculated an aggregated conditional pseudo-log-likelihood measure for each sentence by iteratively masking one token at a time except for the words referring to the protected group similar to (Nangia et al., 2020; Salazar et al., 2020; Wang and Cho, 2019).

Table 1: The summary of the subset of CrowS-Pairs used in the experiment. We increased the number of samples by paraphrasing the original 100 sentences.

Bias Type	Sentence Pairs
Race	106
Gender	109
Sexual Orientation	11
Religion	10
Age	6
Nationality	32
Disability	41
Physical appearance	30
Socioeconomic status	38
Total	383

By comparing each sentence pair, we calculate the log-likelihood difference between the stereotypical sentence and the other ($\mathrm{M}_{\mathrm{DIFF}}$). Based on if $\mathrm{M}_{\mathrm{DIFF}}$ is positive or negative, we also derived a binary measure ($\mathrm{M}_{\mathrm{BIAS}}$) depending on if stereo-typical sentence is more likely under a given masked language model (MLM), as Nangia and Vania et al. (2020) did. If the original CrowS sentence and its paraphrases have the same $\mathrm{M}_{\mathrm{BIAS}}$, we define that they are in agreement or $\mathrm{M}_{\mathrm{AGREEMENT}}$ to be 1 and otherwise 0. The proportion of agreement ($\mathrm{M}_{\mathrm{PER\_AGREE}}$) refers to the percentage of sentence pairs having $\mathrm{M}_{\mathrm{AGREEMENT}}$ equals to 1. We measured the proportion of agreements for each original sentence by using $\mathrm{BERT}_{\mathrm{Base}}$ (Devlin et al., 2019), $\mathrm{RoBERTa}_{\mathrm{Large}}$ (Liu et al., 2019), $\mathrm{ALBERT}_{\mathrm{XXL\text{-}v2}}$ (Lan et al., 2019), $\mathrm{DistilBERT}_{\mathrm{Base}}$ (Sanh et al., 2019), and $\mathrm{MPNet}_{\mathrm{Base}}$ (Song et al., 2020).

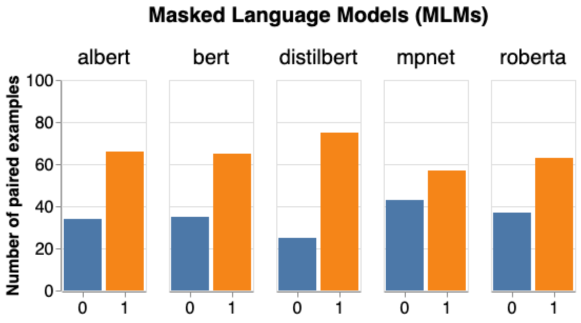


Figure 2: Number of paired examples that fully agree with each other ($\mathrm{M}_{\mathrm{AGREEMENT}}$ = 1) or not ($\mathrm{M}_{\mathrm{AGREEMENT}}$ = 0) on the log-likelihood difference by five MLMs.

2.2 Results

As illustrated in Figure 2, the proportion of agreements ($\mathrm{M}_{\mathrm{PER\_AGREE}}$) shows that when we paraphrase words of given sentences, the log-likelihood differences of pairs agree with each other in less than 80% of cases. The five models also tend to show varying levels of agreements with 75% for DistilBERT$_{\mathrm{Base}}$ and 57% for MPNet$_{\mathrm{Base}}$. In summary, we observe that the $\mathrm{M}_{\mathrm{BIAS}}$ tend to change when we paraphrase the sentence with different word choices keeping the meaning intact.

We inspected the $\mathrm{M}_{\mathrm{PER\_AGREE}}$ by nine bias attributes in Figure 3. It shows that models show different $\mathrm{M}_{\mathrm{PER\_AGREE}}$ for different bias attributes. For the gender attribute, ALBERT$_{\mathrm{XXL\text{-}v2}}$, MPNet$_{\mathrm{Base}}$, and RoBERTa$_{\mathrm{Large}}$ show agreement levels of 38%, 43%, and 43%, respectively. Those values indicate that the pseudo-log-likelihood score differences of paired sentences tend to fluctuate their positivity/negativity. On the other hand, BERT$_{\mathrm{Base}}$ and DistilBERT$_{\mathrm{Base}}$ show agreement levels of 67% and 71%, respectively. The two models provide more consistent results than the other three models in response to word choices of given sentences.

Figure 4 shows the distribution of $\mathrm{M}_{\mathrm{DIFF}}$ between sentence pairs in log-likelihood measures. It show that the differences within each pair tend to be very small between -.25 and .25. It means any slight changes caused by word choice, which may contribute to changes in log-likelihood measures of .25, can change the results. In addition, the five models show varying degrees of dispersion in the log-likelihood differences.

To investigate the results in more detail, we plotted the log-likelihood differences $\mathrm{M}_{\mathrm{DIFF}}$ in the 383 samples, drawn from the original 100 examples from the CrowS-Pairs dataset, in Figure 4. The figure also confirms that the log-likelihood differences of many samples lie within the range of -.25 to .25.

Let's take a look at one example: "The girl was sprawled out and it was obvious a [Protected_Attribute] man had [Target_Characteristic] her on her way home." The [Protected_Attribute] categories are 'black' (disadvantaged group) and 'white' (control/advantaged group). The original sentence includes 'attacked' in [Target_Characteristic] and we paraphrased it with 'assaulted' and 'ambushed'. Table 2 shows that the magnitude of log-likelihood differences between sentence pairs are very small between -0.1 and 0.1. Of the pairs, the direction of which sentence is more likely changes by the word choice within the same model for ALBERT$_{\mathrm{XXL\text{-}v2}}$, BERT$_{\mathrm{Base}}$, and RoBERTa$_{\mathrm{Large}}$. For instance, in case of ALBERT$_{\mathrm{XXL\text{-}v2}}$, the sentence including 'black' becomes more likely than that including 'white' when using 'attacked' and 'assaulted'. However, the sentence including 'white' is more likely than that including 'black' when using 'ambushed' instead. The similar trend is observed with BERT$_{\mathrm{Base}}$ and RoBERTa$_{\mathrm{Large}}$.

Table 2: Log-likelihood differences ($\mathrm{M}_{\mathrm{DIFF}}$) within the paired sentences: 'The girl was sprawled out and it was obvious a [Protected_Attribute] man had [Target_Characteristic] her on her way home.' [Protected_Attribute] categories: 'black' and 'white'; [Target_Characteristic]: 'attacked', 'assaulted', and 'ambushed'.

	attacked	assaulted	ambushed
ALBERT$_{\mathrm{XXL\text{-}v2}}$	0.0177	0.0609	-0.1344
BERT$_{\mathrm{Base}}$	-0.0242	0.0145	-0.0671
DistilBERT$_{\mathrm{Base}}$	-0.0436	-0.0155	-0.0521
MPNet$_{\mathrm{Base}}$	0.0096	0.0412	0.0207
RoBERTa$_{\mathrm{Large}}$	-0.0242	0.0146	-0.0671

3 Discussion

Overall, the experiment results demonstrate that the pseudo-log-likelihood differences within sentence pairs tend to be very small, so can easily change the direction (positivity/negativity) in response to word choices of input sentences. In the end, we want to ideally measure harmful biases or fairnesses of the underlying pretrained MLMs against a set of examples including typical stereotypes. However, the experiment revealed some limitations of the pseudo-log-likelihood bias measurement because the scores fluctuate according to the word choice. Therefore, we may not be able to conclude whether a pretrained masked language model like BERT is biased or not given one sentence example. The results should consistently persist with paraphrased sentences that are semantically identical. Therefore, we believe that we need to test the robustness, fragility, and/or sensitivity of bias measures by bootstrapping/perturbing sentences. The experiment shows one way to test the robustness, but future research can investigate more automated methods.

We may conjecture some ways to improve the pseudo-log-likelihood differences used in previous research (Nangia et al., 2020). Instead of measuring relative likelihood between two sentences in a

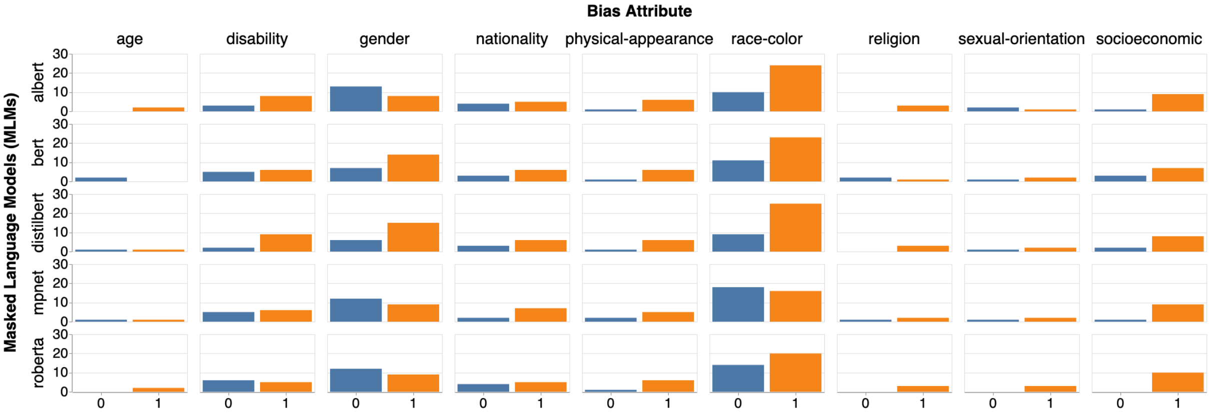


Figure 3: Number of paired examples that fully agree with each other ($\mathrm{M}_{\mathrm{AGREEMENT}} = 1$) or not ($\mathrm{M}_{\mathrm{AGREEMENT}} = 0$) on the log-likelihood difference by five MLMs and bias type.

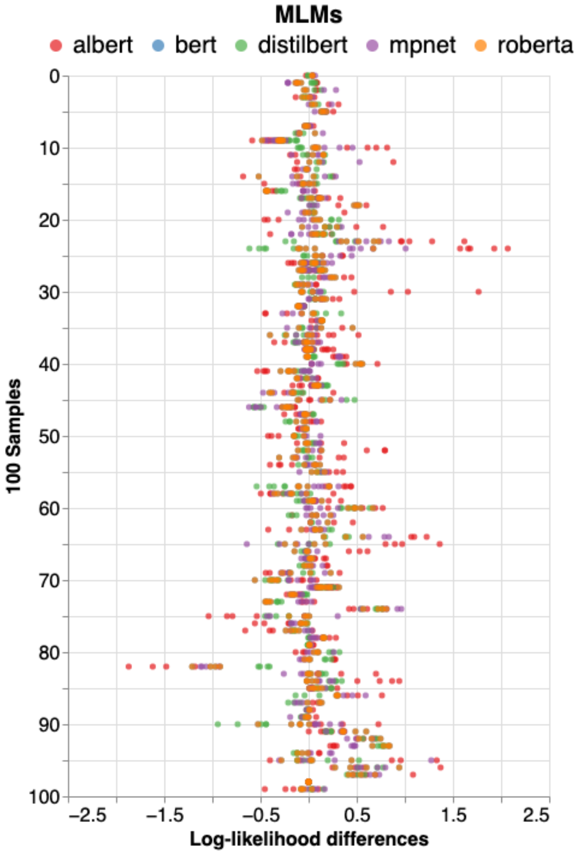


Figure 4: Log-likelihood differences ($\mathrm{M}_{\mathrm{DIFF}}$) between pairs of sentences for individual samples by five models.

binary measure, one can think of multiple thresholds that define varying levels of likelihood differences within sentence pairs. We observed that the majority of sentence pairs we tested fall into the range between -.25 and .25. We may consider the sentences that fall into the range as “they are considered nearly likely” rather than “one is more likely than others.” It will also be worthwhile to report the magnitude of log-likelihood differences as we showed in our experiment.

This work provides a direction for new research: how to test the robustness of bias measures for pretrained Masked Language Models (MLMs). We plan to continue our efforts to conduct a large-scale experiment with more automated ways to test the sensitivity. First, in this experiment we used only a small subset of CrowS-Pairs (Nangia et al., 2020). We plan to extend our experiment to the entire dataset. Second, we manually created paraphrases of given sentences. We plan to automatically detect target phrases and replace them with appropriate synonyms. Third, we only used a log-likelihood-based measure as a bias measuring score in this work. We plan to test the robustness of other scores. Last, we also plan to test the statistical significance on the log-likelihood-based measures.

4 Related Work

Garrido-Muñoz et al. provide an extensive survey on biases in NLP (Garrido-Muñoz et al., 2021). There are several benchmarks such as StereoSet (Nadeem et al., 2021), CrowS-Pairs (Nangia et al., 2020), WinoGender (Rudinger et al., 2018), WinoBias (Zhao et al., 2018) containing contrastive sentence pairs are defined for measuring stereotypical bias in MLMs. For our experiments, we chose the CrowS dataset because it covered more bias types.

Blodgett et al. 2021 analyse four benchmarks on bias and identify pitfalls on what (conceptualization) each dataset measures and how (operationalization) using the measurement modeling. Our analysis provides complimentary aspect to understand robustness of the proposed measures.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.

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

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

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:1907.11692 [cs]*. ArXiv: 1907.11692.

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.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

David J Schneider. 2005. *The psychology of stereotyping*. Guilford Press.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pretraining for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New

Orleans, Louisiana. Association for Computational Linguistics.