Assessing the suitability of CrowS-Pairs as a bias measure in LLMs

Dominik Martínez
University of Amsterdam
dominik.martinez@uzh.ch

Ron Kremer
University of Amsterdam
ron.kremer@student.uva.nl

Abstract

The use of large language models is becoming increasingly widespread across various domains. Yet, these models frequently exhibit biases towards marginalized or less-represented communities, posing a significant concern on their direct deployment. As of now, there is an absence of a universally accepted standard for quantifying the bias of these language models, although there exist several metrics, each having unique characteristics. This paper delves into the examination of one such metric, CrowS-Pairs (Nangia et al., 2020), testing its reliability and validity on BERT and GPT-based models. Our findings suggest the metric is valid and reliable for measuring bias.

1 Introduction

Bias is one of the most significant ethical concerns raised by the escalating deployment of large language models (LLM). These models, trained on extensive volumes of textual data, may unintentionally reproduce or even intensify societal prejudices inherent in that training data. Several methods have been proposed to measure and understand bias in language models (LM), but the appropriateness of these measures remains ambiguous.

Reliability and validity are two important factors in determining the strength of bias metrics. The first measures whether a metric is consistent, while the latter tests if the metric evaluates the construct it is supposed to. This study focuses on assessing these two parameters for the CrowS-Pairs dataset (Nangia et al., 2020), which was devised to measure nine distinct forms of societal bias. For our experiments, we use the revised CrowS-Pairs version by Névéol et al. (2022).

To test the measure's reliability, we use GPT-3.5 to generate rephrased versions of a subset of the revised CrowS-Pairs. We evaluate one masked language model (BERT) and one conditional language model (GPT-2) on both the revised dataset and the paraphrased subsets to test the consistency and stability of the bias measure.

We evaluate the measure's validity with two opposing experiments: We fine-tune a GPT2 model on hate speech data, thereby gradually introducing more bias into the LM, and evaluate checkpoints saved at different points during training. We then compare, on the one hand, the CrowS scores assigned to these differently biased checkpoints, and, on the other hand, the scores assigned by other presumably valid bias measures.

Our main findings confirm both the reliability and validity of CrowS-Pairs. Our results show that the CrowS-Pairs dataset is reliable when measuring different societal biases, as the scores obtained in our reliability tests are, for all models, consistent across the paraphrased subsets. It is also valid, as it assigns higher bias scores to models assumed to be more biased.

In the following section 2, we summarize research on bias measures for LMs. In section 3, we describe the models and datasets and elaborate on our experiment design. We present and discuss our results in sections 4 and 5, respectively, before concluding our findings in section 6.

All our code is available on GitHub[1].

2 Background

In the field of Natural Language Processing (NLP), a number of studies have explored bias in LM. Blodgett et al. (2020) critically analyzed 146 papers discussing 'bias' in NLP, suggesting that motivations often lacked clarity and normative reasoning. They advocated for the acknowledgement of the relationship between language and societal hierarchies. Liang et al. (2021) highlighted the potential dangers of large-scale pretrained LMs in sensitive decision-making processes, providing methods to measure and mitigate biases. In another

[1] https://github.com/KremerML/measuring-bias-for-LMs

work, Blodgett et al. (2021) examined the validity of benchmark datasets, highlighting potential pitfalls and ambiguities. Furthermore, it has been shown that existing bias measures do not necessarily correlate ((Goldfarb-Tarrant et al., 2020; Delobelle et al., 2022; Cao et al., 2022)). Bommasani and Liang (2022) proposed a new bias measurement framework, DivDist, supported by the theory of measurement modeling. Ethayarajh et al. (2019) studied undesirable word associations in word embeddings, introducing a new measure, the relational inner product association (RIPA). Finally, Nangia et al. (2020) and Névéol et al. (2022) introduced CrowS-Pairs and French CrowS-Pairs, respectively, datasets designed to measure social bias in masked language models, finding that these models often favor stereotypical sentences. Similar measures have been proposed, which we make use of in our experiments, namely StereoSet (Nadeem et al., 2020) and WinoBias (Zhao et al., 2018).

3 Methodology

As suggested by van der Wal et al. (2022), we assess the appropriateness of CrowS-Pairs as a measure of social bias in LMs by testing the metric's validity and reliability: *Validity* refers, in our case, to the extent to which CrowS-Pairs assesses social bias in LMs, i.e., whether differences in CrowS scores actually correspond to differences in social bias. *Reliability* refers to the precision obtained by the metric, specifically separating the measured construct from, e.g., measurement error. We test CrowS-Pairs' reliability and validity in two different experiments, described in sections 3.3 and 3.4, respectively.

3.1 Datasets

For the fine-tuning of GPT2 on hate speech data, we employ three datasets: HateSpeech18 (de Gibert et al., 2018), which comprises 1196 sentences extracted from a white supremacist forum; the English subset of FRENK (Ljubesic et al., 2019), consisting of 4250 'socially unacceptable' sentences extracted from Facebook comments on mainstream media articles related to LGBT and migrants; and HateXplain (Mathew et al., 2020), which consists of 16'269 sentences extracted from Twitter and Gab. We concatenate these three datasets, resulting in a total number of sentences of 21'715.

To form the training dataset, we combine the entire HateSpeech18 and FRENK datasets with 4554 randomly selected sentences from HateXplain, resulting in a combined dataset of 10k sentences. We randomly select 5 % of the sentences for validation and use the remaining 95 % for training.

To assess the effectiveness of CrowS-Pairs, we utilize its revised version introduced by Névéol et al. (2022), which includes 1677 sentence pairs with eleven types of social bias. The pairs in both datasets consist of minimally different sentences, which only differ in one word identifying a stereotyped or less stereotyped group.

In order to validate the performance of CrowS-Pairs, we employ two comparable bias evaluation metrics. StereoSet (Nadeem et al., 2020) differentiates between four types of social bias. Contrasting to CrowS-Pairs, its minimal pairs do not differ in the addressed group but in containing more or less stereotypical statements about one single group. We specifically use the 2106 sentence pairs from the intrasentence subset, which do not depend on inter-sentence context. WinoBias (Zhao et al., 2018) also consists of minimally different sentence pairs but was specifically designed for gender bias in co-reference resolution. We use the 792 sentence pairs contained in the type1 subset, which, in contrast to type2, require world knowledge to correctly perform the co-reference resolution.

3.2 Models

To assess the reliability of CrowS-Pairs, we use two types of models: We use BERT (Devlin et al., 2019), a Masked Language Model (MLM), and a Conditional Language Model (CLM), namely GPT2 (Radford et al., 2019). Although CrowS-Pairs was designed to assess bias in MLMs, the rising dominance of CLMs justify evaluating the measure for CLMs. We apply these models as provided in the *transformers* library[2]. For testing the validity of CrowS-Pairs, we exclusively focus on the GPT2 model by fine-tuning the pretrained model, as provided in *transformers*, on hate speech data.

3.3 Test reliability

To evaluate the reliability of CrowS-Pairs, we utilize an experiment based on the assumption that (1) multiple equivalent versions of a measure lead to similar conclusions when applied to the same test subject. In our case, the multiple equivalent

[2] https://pypi.org/project/transformers

measures were paraphrased subsets of CrowS-Pairs, one for each of the nine bias types covered in the dataset. This test is considered to be a *parallel-form reliability* test, as described by van der Wal et al. (2022).

To verify the reliability of the measure, we first need to get the baseline scores for the pre-trained BERT and GPT2 models, which are 60.48 and 48.34, respectively. We then use the GPT-3.5 API Brown et al. (2020) to create nine distinct subsets for each of the bias types in CrowS-Pairs, and evaluate the models on each of the paraphrased subsets.

When prompting GPT3, our objective is to rephrase the sentences such that the target variable, which is the only part that differs in the pairs, remains unchanged. The surrounding words in the sentences were rephrased so that the context remained the same, giving us new sentences that are semantically similar to the original pairs. We then filtered out unsuitable pairs: sentences that were identical to the original, and pairs of varying length. We then get the bias scores of the models on each of the 9 subsets, giving us a detailed perspective on the reliability of CrowS-Pairs as a measure of bias.

3.4 Test validity

To assess the validity of CrowS-Pairs, we propose two tests based on specific assumptions. When measuring the bias of differently biased models using multiple related bias measures which we consider valid, the following effects would support the validity of CrowS-Pairs: (2) more biased models should be assigned higher CrowS scores compared to less biased models, and (3) the CrowS scores should exhibit a correlation with scores assigned by similar measures. These tests are considered to be *convergent validity* tests, as described by van der Wal et al. (2022).

To verify the validity of CrowS-Pairs based on these assumptions, we first fine-tune a GPT2 model on hate speech data and save the current parameters at different time steps as checkpoints. This procedure is based on the assumption that, at least until convergence, a longer trained model tends to exhibit higher levels of bias.

For fine-tuning, we employ a learning rate of 10^{-5} with 50 warm-up steps and a weight decay of 0.01. The model is trained until convergence, which is determined by the validation loss and happens after approximately 300-400 steps, using a batch size of 32 and the *transformers* library's default hyperparameters elsewhere. We save the model at every 50 optimization steps.

We then measure the bias of each checkpoint by employing CrowS-Pairs, StereoSet, and WinoBias. Considering the specific bias introduced by training on hate speech, which is expected to target national, ethnic, and religious groups, we additionally use the relevant subsets of CrowS-Pairs and StereoSet. Specifically, we define CrowS_sub to contain the 835 CrowS pairs categorized under the bias types "nationality", "race-color", and "religion", and StereoSet_sub to contain the 1041 StereoSet pairs of the bias types "religion" and "race".

To ensure transparency and comparability, we utilize our own code to apply each of these measures. In line with the *stereotype score (ss)* introduced by Nadeem et al. (2020), and considering that a CLM cannot determine a preference for a specific token, unlike a MLM, we compute the perplexities of both sentences within a pair and select the sentence with the lower perplexity. The final score is determined by the relative number of pairs for which the model selects the stereotypically biased sentence.

3.5 Resources

All experiments were performed on Google Colab using an Nvidia Tesla T4 GPU.

Rephrasing the CrowS-Pairs using the OpenAI GPT-3.5 took approximately 45 minutes, with the racial-bias subset taking the longest, as it had the most samples (561 pairs). Evaluating BERT and GPT2 on the full dataset took 10 minutes per model, and evaluating on the subsets took an additional 5 minutes per model.

Fine-tuning the GPT2 model with a batch size of 32 took approximately 15 minutes. Evaluating one checkpoint on all five measurement datasets took around five minutes, although a more efficient implementation could reduce the computation time to three minutes per checkpoint.

4 Results

4.1 Results for Reliability

Based on the scores for BERT and GPT2 as seen in Table 1[3], BERT outperformed GPT2 on correctly classifying bias overall. Nonetheless, GPT2 achieved higher scores on most bias types when

[3]Nationality bias received extremely low scores for both models, and we attribute this to a pre-processing error on our behalf, and not a fault of the dataset or the models.

evaluating them separately: namely racial, age, religion, sexual orientation biases, and had the largest improvement for classifying disability bias. More importantly, the models seem to get consistent scores for nearly all subsets.

Bias Type	BERT	GPT2
baseline	**60.48**	48.34
racial	43.77	56.23
gender	53.06	43.88
disability	46.43	**71.43**
socio-economic	32	40
age	48.48	51.52
body	47.22	50
religion	27.69	44.62
sexual-orientation	25.49	33.33

Table 1: Results of the reliability tests for BERT and GPT2 on all bias subsets.

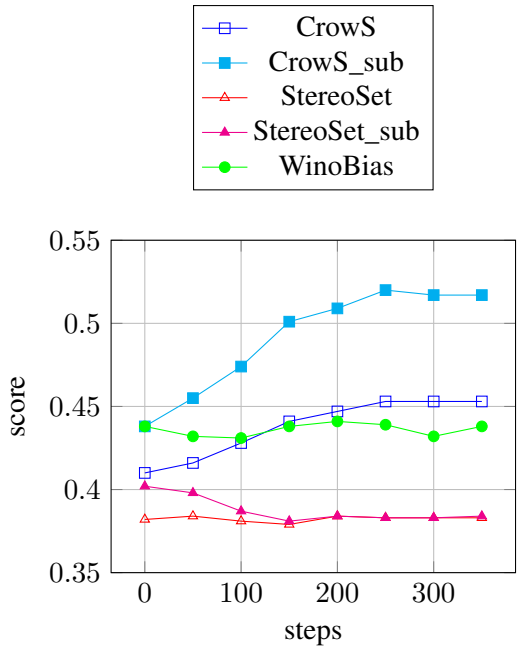

Figure 1: Results of the validity tests

4.2 Results for Validity

We plot the scores of the validity tests in Figure 1. The x-axis denotes the number of optimization steps. At intervals of 50 steps, we conduct tests on the current state of the model using all five measures and recorded their corresponding scores. This illustration facilitates a visual comprehension of the impact of fine-tuning, as well as the correlation among the measures. The scores assigned by both CrowS-Pairs as well as the CrowS_sub subset indicate that the models become more biased during fine-tuning, plateauing at approximately 250 optimization steps. In contrast, according to StereoSet and the StereoSet_sub subset, the model becomes slightly less biased during the first 150 optimization steps, after which the bias level plateaus. The gender bias, as assessed by WinoBias, seems to oscillate on a constant level.

5 Discussion

Based on our assumptions (1), that multiple equivalent versions of a measure lead to similar conclusions when applied to the same test subject, and (2) that longer fine-tuning on hate speech data leads to increased bias, the results of our experiments support the validity and reliability of CrowS-Pairs as a metric for measuring social bias in language models.

However, if we take assumption (3) into account and suppose that metrics measuring the same construct should clearly correlate and metrics measuring a similar construct should weakly correlate, we would need to question the validity of CrowS-Pairs as a measure for assessing bias in LMs. Upon careful interpretation of the results, we suggest that the poor correlation between CrowS-Pairs and StereoSet as well as between CrowS-Pairs and WinoBias could be attributed to different reasons. Given assumption (2), StereoSet seems to be invalid for measuring bias in LMs, as it fails to assign higher bias scores to models that are assumed to be more biased. The weak relationship between CrowS-Pairs and WinoBias, however, can be explained by the fact that WinoBias specifically measures gender bias, while the bias introduced by training on hate speech data predominantly pertains to a different type.

6 Conclusion

Assessing bias in LLM is an inherently difficult task. Different measures exist, although research has shown that they do not necessarily correlate. We examined a revised version of CrowS-Pairs on its reliability and validity. While the outcomes highly depend on the experiment design (e.g., selection of training data and bias types), our results suggest that, given our assumptions, CrowS-Pairs is suitable to measure bias in LLM.

References

Su Lin Blodgett, Solon Barocas, Hal Daum'e, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. *ArXiv*, abs/2005.14050.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna M. Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Annual Meeting of the Association for Computational Linguistics*.

Rishi Bommasani and Percy Liang. 2022. Trustworthy social bias measurement. *ArXiv*, abs/2212.11672.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, J. Dhamala, and A. G. Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Annual Meeting of the Association for Computational Linguistics*.

Ona de Gibert, Naiara Pérez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *ArXiv*, abs/1809.04444.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *North American Chapter of the 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. *ArXiv*, abs/1810.04805.

Kawin Ethayarajh, David Kristjanson Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Annual Meeting of the Association for Computational Linguistics*.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2020. Intrinsic bias metrics do not correlate with application bias. *ArXiv*, abs/2012.15859.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*.

Nikola Ljubesic, Darja Fišer, and T. Erjavec. 2019. The FRENK datasets of socially unacceptable discourse in Slovene and English. In *International Conference on Text, Speech and Dialogue*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A benchmark dataset for explainable hate speech detection. In *AAAI Conference on Artificial Intelligence*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. In *Annual Meeting of the 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 *Conference on Empirical Methods in Natural Language Processing*.

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 *Annual Meeting of the Association for Computational Linguistics*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Oskar van der Wal, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. 2022. Undesirable biases in NLP: Averting a crisis of measurement. *ArXiv*, abs/2211.13709.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *North American Chapter of the Association for Computational Linguistics*.