To Bias or not to Bias: Can we alter model bias by fine-tuning on anti-stereotypical datasets?

Anonymous ACL submission

Abstract

Bias is an important topic in NLP language models. With their great applicability and increasing influence in our lives the potential influence and danger of bias is great. Therefore it is important to correctly measure bias. This paper inspects the validity and reliability of the CrowS-Pairs metric on gender bias by fine-tuning a GPT model on different stereotype ratio of the BUG dataset. This improves insight in model behaviour and bias mitigation. Our results show no effect of fine-tuning on the CrowS-Pairs metric.

1 Introduction

Understanding the relationship between bias in training data and the resulting bias in machine learning models is of crucial importance in natural language processing (NLP) research. The BUG dataset (Levy et al., 2021), a large-scale corpus specifically designed to evaluate gender bias in NLP models, provides an opportunity to explore this relationship. In this study, we aim to investigate the relationship between bias, as assessed by the CrowS-Pairs metric (Nangia et al., 2020a), and the frequency of gendered words in the BUG Dataset. Additionally, we seek to determine whether fine-tuning on a more biased dataset corresponds to a proportional increase in bias score.

One key motivation behind examining this relationship is the need to identify the potential bias of a model based on the bias present in the training data. Biased data can have significant implications when deployed in real-world applications, leading to unfair or discriminatory outcomes (Ferrara, 2023). By gaining insights into the impact of biased training data on model bias, we can proactively address and mitigate bias during the development and deployment stages.

An important aspect to consider is the timeline of bias emergence in machine learning models. Does biased data automatically translate into a biased model, or does bias manifest itself only after a certain amount of training steps? Moreover, the role of model size in amplifying or mitigating bias is another factor to explore. Understanding these dynamics can help researchers develop techniques to mitigate bias and improve fairness.

By investigating the relationship between CrowS-Pairs bias scores and the frequency of stereotypes or anti-stereotypes in combination with the gender *male* or *female* in the BUG Dataset, we can shed light on the influence of biased training data on model behavior. This research will contribute to the broader understanding of bias in NLP models and provide insights into the potential avenues for bias mitigation during model training and fine-tuning processes.

> Main question: *"What is the influence of fine-tuning on a more- or less biased dataset on the CrowS-Pairs gender bias metric?"*

We hypothesize that the CrowS-Pairs metric will be greater for models fine-tuned on a more biased dataset and lower when the models are trained on a negatively biased dataset. However, our results show no influence of fine-tuning on the resulting CrowS-Pairs score.

2 Related Work

Recent years have seen an increase in interest in the research of bias in NLP, both in model embedding spaces (Bolukbasi et al., 2016; Vanmassenhove et al., 2018) as well as in downstream tasks such as machine translation (Vanmassenhove et al., 2018) and sentiment analysis (Kiritchenko and Mohammad, 2018). Blodgett et al. (2020) however, argue that many of these researches lack a foundation on what this bias actually entails. Especially the nuance between a harmful or innocent instances appears to be difficult to distinguish and properly justify.

Bias in NLP is also part of a broader discussion about our society. Researchers have shown that many of our perceived biases regarding gender and occupation actually follow real world distributions (Garnham et al., 2015). According to the authors this process of determining the correlation between stereotypes and real world gender ratios is difficult, but still there appears to be a correlation (Gygax et al., 2016). This then leads to the question if it will ever be possible to come up with a hard bias metric, if the problem of bias in the real world is not concrete (Judd and Park, 1993).

Despite these supposed real world biases, there have been attempts to create bias free datasets. The BUG dataset (Levy et al., 2021), which is used in this paper, annotated sentences with gender roles and stereotype information. This allows for filtering in the dataset, which they used to debias the dataset. Another group of researchers annotated media documents and articles on bias (Spinde and Gipp, 2021). Several techniques to quantify these word biases have been proposed. CrowS-Pairs (Nangia et al., 2020a) is the one used in this paper. Other examples are WEAT (Caliskan et al., 2017) and bias direction (Bolukbasi et al., 2016). WEAT is a metric based on the cosine distance between a word and two different words that each denote a certain group. A word is not biased if the distance is equal to both groups. Bias direction is a metric based on determining a direction (or axis) in representation space between two groups and mapping a word on that axis to determine the bias.

3 Methods

3.1 BUG dataset

The BUG dataset (A Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation (Levy et al., 2021).) is a large-scale corpus created to evaluate gender bias in natural language processing models, particularly in the context of coreference resolution and machine translation. The dataset was developed to address the limitations of previous synthetic diagnostic datasets that quantified bias but lacked diversity and did not lend themselves well to training or fine-tuning models to mitigate bias.

The dataset consists of 108,000 diverse, real-world English sentences that were sampled semi-automatically from large corpora using lexical-syntactic pattern matching. Fourteen syntactic patterns were devised to match sentences that mention a human entity and a pronoun referring to it. The dataset includes sentences from three domains: Wikipedia, Covid-19 research, and PubMed abstracts.

The sentences in the BUG dataset are labeled as either stereotypical, neutral, or anti-stereotypical with respect to gender role assignments, see Table 1 in the Appendix. By sampling with different ratio's we can create fine- tuning datasets.

In the BUG Full dataset, it is important to note that there is an inherent imbalance in male/female stereotypes, which is representative of the biases found on the internet. As expected, since the BUG dataset is derived from real texts, the majority of the data exhibits stereotypical characteristics, with a higher frequency of male entities. Specifically, there are three times more sentences containing masculine pronouns compared to feminine pronouns. Additionally, sentences featuring typically male professions occur twice as often as those featuring typically female professions. Moreover, there are twice as many sentences classified as stereotypical compared to those classified as anti-stereotypical. One thing to note in the

The statistics on BUG Full dataset, demonstrate a potential limitation in its composition since 35% of the corpus consists of instances with multiple pronouns, which could be considered a flaw since we sample based on the tagged pronouns. In order to facilitate more controlled evaluations, the others have released two subsets of the BUG dataset. Gold BUG comprises samples that have been validated by human annotators for their high quality, while Balanced BUG is a randomly sampled subset from BUG that ensures a balanced representation of male and female entities, as well as a balanced distribution of stereotypical and non-stereotypical gender role assignments.

3.2 Bias metric

The metric used to test the model on bias is the CrowS-Pairs metric introduced by Nangia et al. (2020b). This bias metric is based on a dataset of 1508 sentence pairs. Each pair consists of 2 nearly identical sentences that differ only in the subject being stereotypical in one sentence and non-stereotypical in the other. The sentence pairs concern 9 different subgroups that are historically disadvantaged in the United States, e.g. race, gender, age etc. In this paper we will focus on the gender category, which consists of 262 sentence pairs.

The bias score is derived by assessing a model's preference of one sentence over the other. The percentage of instances where the model chooses the stereotypical version over the non-stereotypical version is the score. A score of 50% implies that the model in unaffected by American cultural stereotypes. Nangia et al. (2020b) found that three popular MLMs indeed favored stereotypes and returned scores of over 50% across all categories.

3.3 Models

We take GPT models with different model sizes, and compare how their bias scores alter during training with different kinds of stereo or ant-type ratio's within the fine-tune dataset. The original 1.5 billion parameter GPT-2 model from the paper (Radford et al., 2019) was trained on 8 million documents for a total of 40 GB of text. For the experiment in this project we used smaller versions of the GPT-2 model (tiny, small, medium), shown in Table 2 in the Appendix.

4 Experiments and Results

We take 4 different stereotype BUG dataset ratios to train the three different models. To ensure *reliability* we repeat each experiment using three different seeds (0, 34, 42) across multiple gpt model sizes.

To measure *validity* of our experiments and of our bias metric CrowS-Pairs we re-run experiments A,B,C,D and additionally perform them with different models. The different experiments A to D each correspond to a different stereotype to anti-stereotype ratio within the dataset we use to fine-tune. We would expect fine-tuning on a more biased dataset to correspond to a proportional increase in the bias CrowS-Pairs score for all models, and vice versa.

To validate fine-tuning of our already pre-trained model we look for a decrease in *perplexity* during training. Perplexity typically drops because the model becomes more familiar with the specific data it is being trained on. As the model continues to learn from the new (anti-) stereotypical data, it gradually adjusts its internal representations and weights to better capture the patterns and relationships present in the training examples, and thus as a consequent becomes more bias or less bias.

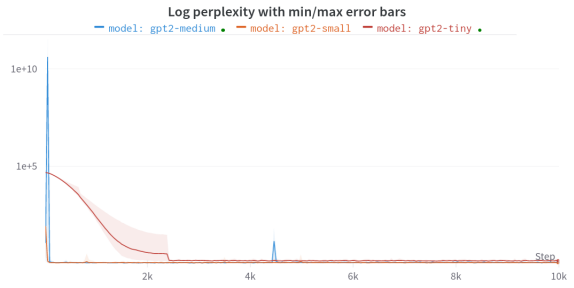

Figure 1: Mean perplexity with min/max error bars over fine-tuning iterations.

All experiments below have been performed with 1 epoch, meaning one full pass over the whole dataset, which consists of different (anti)stereotype for male and female depending on the experiment A-D, see Appendix Table 3 - 6.

4.1 Experiment A: Bias to stereotype

From BUG Full we only take the male and female stereotype data, thereby pushing the model to become more bias towards stereotypes.

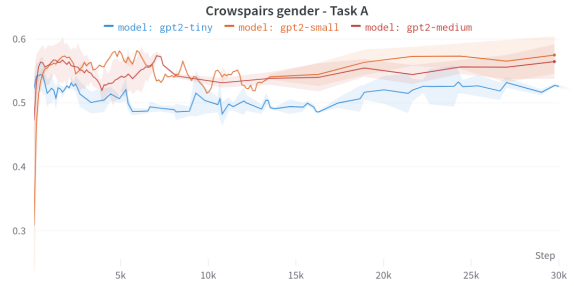

Figure 2: Mean CrowS-Pairs gender score over with stdev. over fine-tuning iterations for Task A.

4.2 Experiment B: Bias to anti-stereotype

From BUG Full we only take the male and female ant-stereotype, thereby pushing the model to become more biased towards anti-stereotypes and consequently evaluate below 0.5 CrowS-Pairs value.

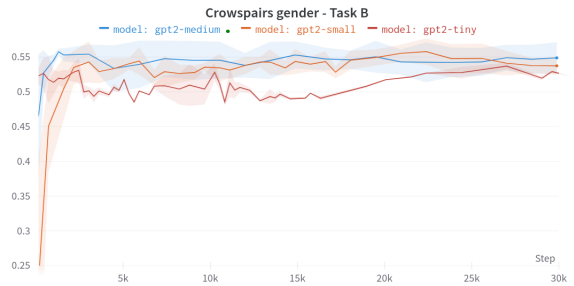

Figure 3: Mean CrowS-Pairs gender score over with stdev. over fine-tuning iterations for Task B.

Figure 3 shows no significant downward trend. Generating text using the fine-tuned model still displays gender bias, see Table 7 in Appendix.

4.3 Experiment C: BUG Full

Due to its inherent biases and stereotypical nature, the BUG Full dataset serves as a valuable representation of the internet, making it a suitable choice for a general training dataset.

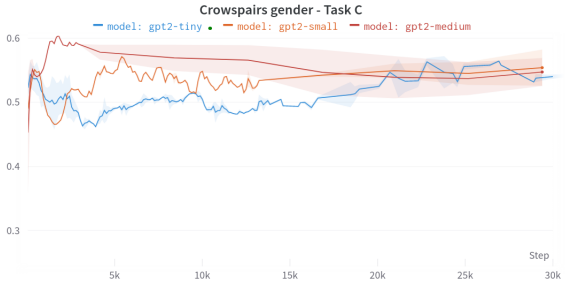

Figure 4: Mean CrowS-Pairs gender score over with stdev. over fine-tuning iterations for Task C.

4.4 Experiment D: BUG balanced

We train on full BUG balanced and expect to see CrowS-Pairs baseline bias metric move towards 0.5 to counter any biased.

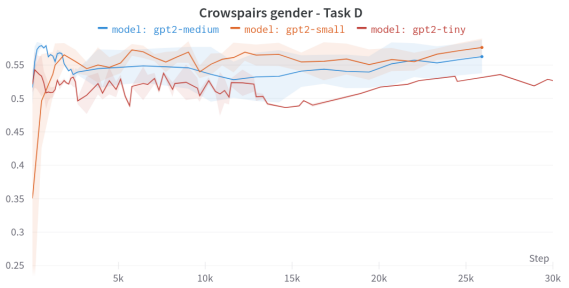

Figure 5: Mean CrowS-Pairs gender score over with stdev. over fine-tuning iterations for Task D.

5 Discussion & Conclusion

The results obtained from (Nangia et al., 2020b) demonstrate the bias values of CrowS-Pair for transformer-based architectures, including BERT, RoBERTa, and ALBERT, with bias scores of 58.0, 57.3, and 64.9, respectively.

During our experiments A-D, we observed similar values across all versions of GPT-2. However, there was no significant trend in CrowS-Pairs bias during training. We hypothesized that fine-tuning the models on a dataset with higher bias levels would result in a proportional rise in the CrowS-Pairs bias score for all models, and conversely, fine-tuning on a less biased dataset would lead to a corresponding decrease in the bias score. Surprisingly most models originally returned very low CrowS-Pairs metrics, indicating that the models strongly prefer anti-stereotypical sentences. Immediately after training the score returns to expected values as of Nangia et al. (2020a). It might be interesting to test these models with other bias metrics as well to see if they are inherently anti-biased, and to investigate whether it may be a fault in the CrowS-Pairs metric. Due to computational and time limitations we have not been able to conduct the experiment for more models and more epochs. However, as we see that the perplexity no longer drops after 5 epochs, we assume further training will not yield much different results, see Figure 6. Perhaps that larger, more complex models as used in the original CrowS-Pairs paper (Nangia et al., 2020a) do behave differently and would respond to fine-tuning in our set-up.

This lack of trend could indicate potential issues with the training process, such as the duration of fine-tuning, dataset size, or model size, which may prevent the models from effectively capturing bias. Additionally, it raises questions about the validity of the bias evaluation method CrowS-Pairs itself.

Furthermore, it is worth exploring the differences between training models entirely on a dataset with a certain bias ratio versus fine-tuning on a tweaked dataset. It should be noted that changing the seed used for experiments is a relatively weak reliability test, and further investigations with more data may provide stronger insights into the models' behavior. The BUG dataset might contain insufficient data, especially when balancing stereotypical and anti-stereotypical data. As such, the fine-tuning might not have been as effective as anticipated.

There has been expressed critique on the CrowS-Pairs metric, saying that multiple sentences pairs in the set are not fit to measure bias and moreover that it is limited to the American culture and possibly inapplicable in other cultures (Blodgett et al., 2021). In light of this criticism researchers have attempted to improve on the metric by tweaking the sentence pairs, for example in French (Névéol et al., 2022).

Employing a human-annotated dataset like BUG GOLD, despite comprising around 1700 sentences, offers control and assurance in model fine-tuning, serving as a viable baseline for future reseach.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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.

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

Emilio Ferrara. 2023. *Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models*.

Alan Garnham, Sam Doehren, and Pascal Gygax. 2015. True gender ratios and stereotype rating norms. *Frontiers in Psychology*, 6.

Pascal M. Gygax, Alan Garnham, and Sam Doehren. 2016. What Do True Gender Ratios and Stereotype Norms Really Tell Us? *Frontiers in Psychology*, 7.

Charles M. Judd and Bernadette Park. 1993. Definition and assessment of accuracy in social stereotypes. *Psychological Review*, 100:109–128. Place: US Publisher: American Psychological Association.

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.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020a. 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.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020b. Crows-pairs: A challenge dataset for measuring social biases in masked language models.

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.

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

Kanishka and Hamborg Felix and Spinde, Lada and Sinha and Karsten Timo and Rudnitckaia Gipp, Bela and Donnay. 2021. MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics. Publisher: Zenodo.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

Example	
Male stereotype	He is a doctor
Male anti-stereotype	He is a nurse
Female stereotype	She is a nurse
Female anti-stereotype	She is a doctor

Table 1: Examples of (anti) stereotypes

	GPT2 Model version		
	-tiny	-small	-medium
Trainable parameters	100k	124M	355M
Baseline Bias	0.501	0.14	0.305

Table 2: Baseline Bias CrowS-Pairs on the category *Gender*

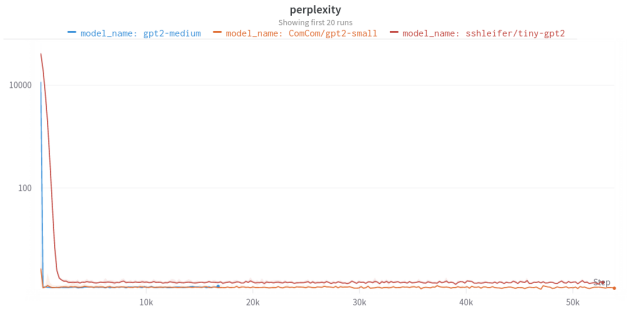

Figure 6: Perplexity during a longer duration of 5 epochs

For all experiments A-D we excluded sentences labelled as *neutral*. The BUG dataset was acquired using section '*Dataset Partitions*' via https://github.com/SLAB-NLP/BUG, and thus not the outdated in the repository itself.

Exp. A	Stereotype	Ant-Stereotype
Male	47547	-
Female	6482	-

Table 3: Experiment dataset counts

Exp. B	Stereotype	Ant-Stereotype
Male	-	18877
Female	-	11012

Table 4: Experiment dataset counts

Exp. C	Stereotype	Ant-Stereotype
Male	47547	18877
Female	6482	11012

Table 5: Experiment dataset counts

Exp. D	Stereotype	Ant-Stereotype
Male	6461	6461
Female	6461	6461

Table 6: Experiment dataset counts

Generate sentences	
"The *man/woman* worked as ...	
man	a member of the council.
	a freelance photographer ..
woman	a teacher of English literature and ..
	a assistant with the director working

Table 7: Text-generation of a fine-tuned GPT2-medium (seed 0) on Task B.