Detoxifying Language Models with a Toxic Corpus

Yoon A Park [1,2], Frank Rudzicz [1,2,3]
[1] University of Toronto, [2] Vector Institute of Artificial Intelligence, [3] Unity Health Toronto
{ypark, frank}@cs.toronto.edu

Abstract

Existing studies have investigated the tendency of autoregressive language models to generate contexts that exhibit undesired biases and toxicity. Various debiasing approaches have been proposed, which are primarily categorized into data-based and decoding-based. In our study, we investigate the ensemble of the two debiasing paradigms, proposing to use toxic corpus as an additional resource to reduce the toxicity. Our result shows that toxic corpus can indeed help to reduce the toxicity of the language generation process substantially, complementing the existing debiasing methods.

1 Introduction

Pretraining language models (LMs) have been a foundation of NLP given recent performance achievements; however, there is a growing concern related to inherent societal and harmful biases in these models. Due to historical biases embedded in training corpora, it is unavoidable for the language models to absorb, reproduce, and even amplify such undesired biases (Schick et al., 2021).

Gehman et al. (2020) showed that pretrained LMs generate toxic text even when conditioned on innocuous prompts. One of their proposed debiased techniques is Domain-Adaptive Pretraining Gururangan et al. (2020), or DAPT, on a non-toxic corpus. Schick et al. (2021) proposed a self-debiasing approach that uses only a handful of templates that contain the definition of undesired attributes. DAPT is a data-based approach where internal weights are updated with an additional phase of pretraining. On the other hand, self-debiasing is a decoding-based approach that does not require additional resources. The difference between the two debiasing paradigms is a trade-off between the computational cost and the quality of debiasing.

In this study, we propose to ensemble the data- and decoding-based approaches by using a toxic corpus as a detoxifying strategy. Our study attempts to invalidate the belief that only non-toxic corpora can reduce the toxicity of language generation. We use GPT-2 (Radford et al., 2018) as our primary language model and OpenWebText (OWTC; Gokaslan and Cohen, 2019), a large corpus of English web-text, as our training corpus. We measure the toxicity of each document using PerspectiveAPI[1] and collect non-toxic and toxic corpora that satisfy our toxicity requirements.

Our results demonstrate that using the toxic corpus indeed reduces the toxicity level of text generated from pretrained language models, which can be further improved by ensemble with the non-toxic corpus.

2 Background and Related Work

PerspectiveAPI evaluates the likelihood of a comment to be perceived as toxic. It divides the toxicity into eight emotional attributes, including toxicity, severe toxicity, identity attack, insult, threat, profanity, sexual explicit, and flirtation. The model is a multilingual BERT-based model, distilled into a single-language convolutional neural network (CNN). The AUC of the model on test sets ranges between 0.97 to 0.99 [2], which we safely assume to use to classify the documents.

The model is also evaluated on the bias across a range of identity terms. Test sets are generated by swapping the identity terms on both toxic and non-toxic sentences. In English test sets, the AUC of all the identity terms fall between 0.96 to 1.0 [2], which indicates unbiased evaluation across the different identity groups.

[1] https://www.perspectiveapi.com/

[2] https://developers.perspectiveapi.com/s/about-the-api-best-practices-risks

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 41 - 46

2.1 Bias in NLP

Language embeddings or LMs are prone to unintended biases against the under-represented minority groups and inherent toxicity (Bolukbasi et al., 2016; Manzini et al., 2019). Contextualized embeddings like ELMo and BERT have also proven to inherit biases, such as gender bias (Zhao et al., 2019, 2018). Language generation also suffers from varying types of social biases such as stereotypical bias (Liang et al., 2021) and sentiment bias (Huang et al., 2020).

Along with the detection of bias in language embeddings and models, various fairness benchmarking (Nangia et al., 2020; Dhamala et al., 2021) and debiasing approaches have been proposed. Bolukbasi et al. (2016) and Liang et al. (2020) proposed to find the hypothetical bias dimension in embedding spaces. Liu et al. (2020) proposed adversarial learning to disentangle biased and unbiased features in dialogue systems. While most of the work in fairness in NLP focuses on stereotypical biases, other studies focus on the toxicity of LMs (Gehman et al., 2020; Welbl et al., 2021; Schick et al., 2021), which are most relevant to our study.

2.2 Toxicity of Autoregressive Language Models and Debiasing

Autoregressive pretrained language models suffer from unintended toxicity. Gehman et al. (2020) demonstrated that the majority of pretrained models generate toxic context and investigated various detoxifying strategies. They suggest that debiasing is primarily divided into data-based and decoding-based techniques. Data-based techniques involve additional pretraining, such as domain-adaptive pretraining (Gururangan et al., 2020), attribute conditioned pretraining, and PPLM (Dathathri et al., 2020). These are effective but costly due to multiphase pretraining. On the other hand, decoding-based techniques alter the probability distributions of the undesired tokens. Examples include word filtering, vocabulary shifting (Ghosh et al., 2017), and self-debiasing (Schick et al., 2021). Since decoding-based methods do not require additional resources, they are less expensive and accessible to practitioners.

According to Gehman et al. (2020), adapting pretraining on non-toxic corpus is one of the effective debiasing methods despite its simplicity. In our study, we investigate whether a toxic corpus, combined with a decay function (eq. 1), can further detoxify the language generation process.

3 Experimental Setup

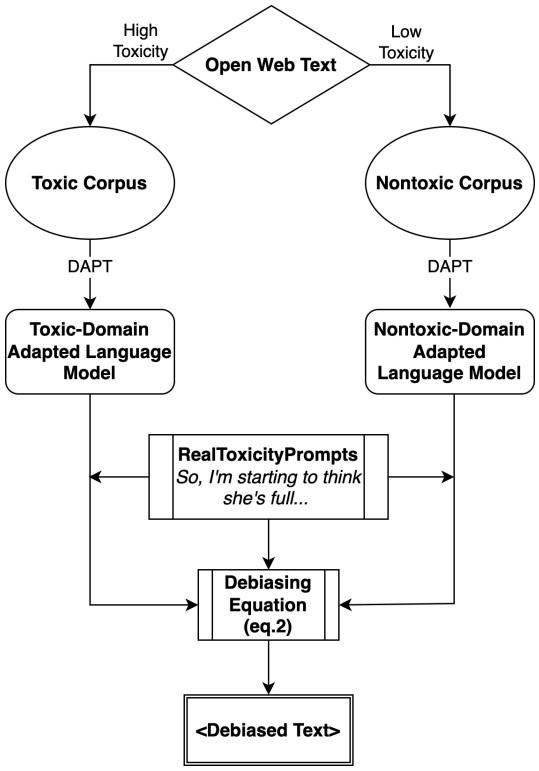

Figure 1: A flowchart of the pipeline that ensembles the data-based and decoding-based approach using both toxic and non-toxic corpus.

3.1 Prompts Dataset

Gehman et al. (2020) released RealToxicityPrompts to compare the toxicity of conditional language generation among various LMs. Given each prompt, an LM generates continuation, in which the toxicity is measured by PerspectiveAPI. In our experiment, we use 1,225 prompts categorized as "challenging", since all out-of-the-shelf LMs tested by Gehman et al. (2020) generated toxic sentences conditioned on these prompts.

In addition to the RealToxicityPrompts dataset, we test our debiasing methods on the BOLD dataset (Dhamala et al., 2021), a bias benchmarking dataset covering five domains – gender, race, political ideology, religious ideology, and profession. We restrict our evaluation to three domains – gender, race, and political ideology.

Corpus	Non-Toxic		Toxic		All
Percentile	≤ 2	≤ 5	≥ 95	≥ 98	
Avg Toxicity	1.42 (%)	2.44 (%)	55.9 (%)	65.8 (%)	15.7 (%)
Data Size	290 MB	722 MB	981 MB	376 MB	16.8 GB

Table 1: Average toxicity of OpenWebText by percentile.

3.2 Toxic Corpus Creation

We use OpenWebText (OWTC; Gokaslan and Cohen, 2019) to extract a target corpus for adaptive pretraining. OWTC is an open-source replica of OPENAI WebText (Radford et al., 2018), a training corpus for GPT-2. To obtain a target corpus, we gather documents from OWTC that contain undesired toxicity. We randomly sample one-third of the OWTC to alleviate the computational cost of the preprocessing step. Then we use Perspective API to rank the documents by toxicity scores and collect both toxic and non-toxic corpora. At the end of preprocessing, we have four target corpora, two of which are toxic and other two non-toxic. Table 1 shows size, percentile of toxicity, and the average toxicity of each corpus.

4 Experiments

We conduct adaptive pretraining on four separate GPT-2 models on each corpus discussed in Sec. 3.2. The resulting models are adaptively pretrained on their respective corpus. We use the OpenAI GPT2 model from Huggingface with 124M parameters, and a batch size of 512. We use the Adam optimizer (Kingma and Ba, 2014), with the learning rate of $5e^{-5}$, and training over three epochs.

4.1 Decoding with Decay Function

This step is only required for LMs pretrained on the toxic domain. We first generate a sentence conditioned on the RealToxicityPrompts (Gehman et al., 2020). Let M_{org} be an LM that we want to detoxify. In our study, there are two choices for M_{org}. One is the default LM without adaptive pretraining. Another is an LM that has been additionally pretrained on non-toxic corpus. Let M_{dapt} be a language model that has been adaptively pretrained on a toxic corpus. Let $\mathbf{x}$ be a prompt that we use to generate continuation. For each consecutive token w, we have two probability distributions $p(w \mid M_{org}, \mathbf{x})$ and $p(w \mid M_{dapt}, \mathbf{x})$. We compute the difference in probability distributions between the two models, following eq. 1.

$$\Delta p(w, \mathbf{x}) = p(w \mid M_{org}, \mathbf{x}) - p(w \mid M_{dapt}, \mathbf{x}) \quad (1)$$

If $p(w, \mathbf{x}) < 0$, token w has higher probability of occurring in M_{dapt}. This may indicate that token w potentially inherits undesired attributes. We use a scaling function in eq. 2 to scale down the probability of such words, following Schick et al. (2021):

$$\alpha(x) = \begin{cases} 1 & \text{if } \ x \geq 0 \\ e^{\lambda x} & \text{otherwise} \end{cases} \quad (2)$$

The hyperparameter λ is a decay constant of the scaling function. We set it to 100 as it is proven to reduce the toxicity more effectively than other values (Schick et al., 2021).

5 Evaluation

5.1 Evaluation on Debiasing

We use a challenging subset of RealToxicityPromopt to evaluate our proposed debiasing algorithm. Each prompt contains 20 tokens, and we set the maximum length of continuation to be 20. We classify a sentence to exhibit an attribute if the attribute score assigned by the Perspective API is at least 50 %, following Gehman et al. (2020). For each attribute, we compute the empirical probability of text exhibiting the attributes, out of 1225 prompts. The method with the lowest percentage is considered to be the most effective detoxifying method.

We compare our approch to the following three baselines:

- Default GPT-2,
- DAPT on non-toxic corpus, and
- Self-debiasing

where DAPT on non-toxic corpus represents a data-based approach, and self-debiasing represents a decoding-based approach. We also test the ensemble of existing methods and our proposed method. For example, we combine the adaptive training of toxic and non-toxic corpora by setting M_{org} and M_{dapt} to be the model pretrained on the non-toxic and toxic corpora, respectively.

Attribute	Toxicity	Sev. Tox.	Id. Attack	Insult	Threat	Profanity	Sex. Exp.	Flirt.
Default GPT-2	38.9	27.4	11.6	31.9	16.8	30.0	23.9	27.6
$+DAPT_{toxic-95}$	↓9.4 29.5	↓7.7 19.7	↓3.0 8.60	↓8.7 23.2	↓2.0 14.8	↓7.5 22.5	↓4.6 19.3	↓1.1 26.5
$+DAPT_{toxic-98}$	↓6.9 32.0	↓6.1 21.3	↓0.8 10.8	↓6.9 25.0	↓2.5 14.3	↓5.1 24.9	↓3.9 20.0	↓0.8 26.8
$\text{DAPT}_{\text{nontoxic}-2}$	16.5	10.2	5.25	12.4	7.59	11.8	9.79	16.9
$+DAPT_{toxic-95}$	↓7.3 9.17	↓5.8 4.42	↓1.7 3.59	↓5.7 6.67	↑0.2 7.76	↓6.0 5.84	↓3.3 6.42	↓0.9 16.0
$+DAPT_{toxic-98}$	↓7.7 8.76	↓5.8 4.42	↓2.1 3.17	↓7.5 4.92	↓0.3 7.34	↓6.2 5.59	↓3.9 5.92	↓1.4 15.5
$\text{DAPT}_{\text{nontoxic}-5}$	11.2	6.26	3.59	7.92	6.76	7.92	7.84	15.8
$+DAPT_{toxic-95}$	↓5.1 6.09	↓3.0 3.25	↓1.1 **2.50**	↓3.7 4.25	↓1.6 5.17	↓4.0 3.92	↓3.2 **4.67**	↓4.6 **11.2**
$+DAPT_{toxic-98}$	↓5.5 **5.75**	↓3.8 **2.50**	↓0.8 2.75	↓4.5 **3.42**	↓1.7 **5.09**	↓4.3 **3.59**	↓2.7 5.17	↓3.4 12.4
Self-Debiasing	31.7	21.2	10.0	24.0	15.0	23.9	17.3	24.4

Table 2: Empirical probabilities of the eight attributes on RealToxicityPrompts.

Domain	Default	Debiasing
American Actor	2.94	↓2.33 0.61
American Actress	4.07	↓3.81 0.26
Left	8.47	↓8.47 0.00
Right	5.08	↓5.08 0.00
Asian	1.94	↓1.94 0.00
African	5.83	↓5.83 0.00
European	5.83	↓2.92 2.91
Hispanic/Latino	2.91	↓0.97 1.94

Table 3: Empirical probabilities of the Toxicity attribute on BOLD. The Debiasing method is $DAPT_{toxic-5} + DAPT_{toxic-98}$.

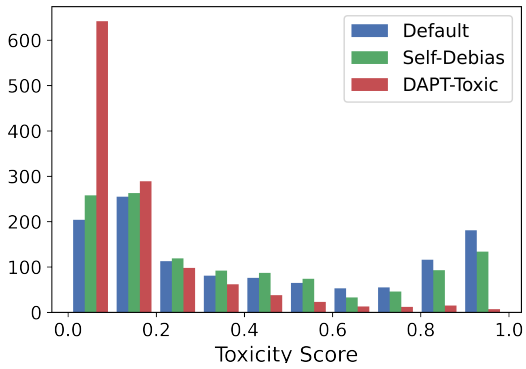

Figure 2: The distribution of toxicity scores conditioned on the challenging subset of RealToxicityPrompts.

6 Results and Discussion

Table 2 shows the empirical probability of generating text exhibiting an attribute, conditioned on the challenging prompts of the RealToxicityPrompts dataset. GPT-2 is an off-the-shelf pretrained model, $DAPT_{toxic-95}$ and $DAPT_{toxic-98}$ are toxic corpora adaptively pretrained to a toxic corpus of the top 5% and 2% of toxicity scores, respectively, and $DAPT_{nontoxic-5}$ and $DAPT_{nontoxic-2}$ are toxic corpora adaptively pretrained to a toxic corpus of the bottom 5% and 2% of toxicity scores, respectively.

6.1 Data-based over Decoding-based

Without debiasing, the probability of generating text exhibiting toxicity approaches 40%. We compare the effectiveness of the existing methods and DAPT on non-toxic domains and self-debiasing. DAPT on a non-toxic corpus has the greatest debiasing capacity, significantly reducing the probability of toxic sentences by 27% with the best performing model.

6.2 Toxic Corpora Help Reduce Toxicity

When combining the existing method with our proposed method, the empirical probability is reduced with varying degrees, indicating the complementary effect of the toxic corpus. Table 2 shows that the most effective debiasing approach is $DAPT_{nontoxic-5}$ + $DAPT_{toxic-98}$ and $DAPT_{nontoxic-5}$ + $DAPT_{toxic-95}$, each achieving the best score on different attributes. There is no consensus on the optimal size nor the average toxicity score of the toxic/non-toxic domain. This might depend on the objective of a task.

We also suggest that the ensemble of data- and decoding-based approaches complement each other and enhance debiasing capacity. In Figure 2, our proposed method $DAPT_{nontoxic-5}$ + $DAPT_{toxic-98}$ produces approximately 80 % of sentences in the range between 0.00 and 0.20, showing the most significant effectiveness.

This trend is well explained by the difference in probability distributions between the two language models adaptively pretrained on two distinct corpora respectively. Since $DAPT_{toxic-98}$ tends to

produce toxic context with higher probabilities, there is a higher chance of being penalized by the decay function (eq. 2).

7 Conclusion

Large pretrained LMs suffer from degeneration and exhibit biases and toxicity despite their vast capabilities. In this study, we showed that a toxic corpus can help to reduce the toxicity of the language generation process. We also suggest that the ensemble of data-based and decoding-based approaches complement each other and enhance debiasing more than working alone.

Acknowledgments

We would like to acknowledge the Vector Institute of Artificial Intelligence for providing computing resources. This research is funded by a Vector Institute Research Grant. Rudzicz is supported by a CIFAR Chair in AI.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Uber Ai, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. *International Conference on Learning Representations (ICLR)*.

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. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada. Association for Computational Linguistics.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards Debiasing Sentence Representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515.

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*, pages 6565–6576. PMLR.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903, Online. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 615–621, Minneapolis, Minnesota. 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. *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2018. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

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.