Language Generation, Modeling and Bias Evaluation on BOLD Dataset

Jingyu Wu
jw6761@nyu.edu

Cora Mao
ym1596@nyu.edu

Tharangini Sankarnarayanan
ts4180@nyu.edu

Abstract

Large-scale pre-trained language models (LMs) capture not only linguistic knowledge but also general knowledge that is implicit in their training data. LMs have the potential to generate potentially dangerous biases resulting from stereotypes that spread hostile generalizations. Text is generated using prompts of the Bias in the Open-Ended Language Generation Dataset (BOLD). Various evaluation metrics are implemented on the generated texts to evaluate the implicit bias in pre-trained language models. Our experiments research the different degrees of biases across social groups in society between GPT-2, CTRL, XL-Net, GPT-Neo and OPT.

1 Introduction

Language models (LMs) are deployed in an increasing number of downstream applications to generate open-ended text. Understanding LMs' behaviour in generating potentially biased texts is crucial to avoid harm to the end-users. Related work points out that bias in NLP research needs a more concrete conceptualization of what bias is and matching quantitative analysis (Blodgett et al., 2020). On this direction, Nadeem et al. used a crowd-sourced dataset containing different levels of stereotypical sentences for specific targets to test if LMs would prefer stereotypes (Nangia et al., 2020). However, it does not directly study the LMs' generated texts. This project aims to use The Bias in Open-Ended Language Generation Dataset (BOLD) (Dhamala et al., 2021) to systematically review social biases in LMs' generated texts by triggering LMs with curated prompts that match the distribution of the human-written text. We replicated most of the experiments in the paper and extended on evaluating the biases in three more LMs.

2 Data

BOLD contains 23,678 text generation prompts in English for assessing social bias in a specific profession, gender, race, religion, or political ideology (Dhamala et al., 2021). The prompts are from English Wikipedia pages related to the identified domains. Because Wikipedia is an online content encyclopedia with authors and contributors from all over the world, and the data is parsed through quality checks, it becomes a representative text for language generation. The authors contended that BOLD reflects the diversity and structure of sentence beginnings for text generation models compared to existing bias benchmarking datasets compiled by experts or crowd-workers.

In our replication and extension experiments, we used a subset of the BOLD data containing prompts in four domains and subgroups with the counts recorded in Table1.

Domain	# Groups	# Prompts
Gender	2	3,204
Race	4	7,657
Religious & spiritual beliefs	7	639
Political ideology	12	1,984

Table 1: BOLD Statistics

3 Experiments and Methods

Our experiments were in two steps:

1. Use BOLD prompts to generate texts with replication models: GPT-2 and CTRL (3 submodels with specific control words), and extension models: XL-Net, GPT-Neo, and OPT.

2. Evaluate biases in generated texts with three metrics: Regard, Sentiment, and Toxicity.

3.1 Text generation

3.1.1 GPT-2

GPT-2 is pre-trained on the WebText dataset, and is trained with a causal language modelling (CLM) objective and predicts the next token in a sequence in an auto-regressive manner, allowing the generation of syntactically coherent text.

3.1.2 CTRL

CTRL is an unidirectional transformer language model that learned the conditional probability of tokens given the control word by being trained on related subsets of training texts prepended with control codes (Keskar et al., 2019). Replicating the base paper, we generated text using CTRL with three control words: Wikipedia (C-WIKI), Opinion (C-OPN), and Thoughts (C-THT), each associated with subsets of the training data from English Wikipedia, Reddit r/changemyview, and Reddit r/showerthoughts respectively. CTRL text generation was done with HuggingFace's transformer package setting repetition penalty to 1.2, top-p to 0.9, and other parameters as default.

3.1.3 XL-Net

The probability of any sequence is modelled using any permutation in an auto-regressive manner in XL-Net (Yang et al., 2019). We begin with a prompt phrase and text generation from right to left on its left side for beams of a certain length. The following tokens are selected with top-K sampling at each stage of beam search. The resulting new phrase generates beams on the right side of the new start phrase, which serves as the starting point for the next iteration. The Hugging Face tokenizer with top-K to 20 and num-beams to 5 was applied.

3.1.4 GPT-Neo

We used the pre-trained GPT-Neo 1.3B model to generate text through HuggingFace's transformer package. GPT-Neo is a GPT-2 like causal language model trained on The Pile dataset. Its architecture is similar to GPT-2's, except that the model uses local attention in every other layer with a window size of 256 tokens (Gao et al., 2020).

3.1.5 OPT

Meta-developed OPT is a decoder-only pre-trained transformer model. Using a self-supervised causal language modeling (CLM) objective, the model is predominantly trained with English text. OPT is used for prompting for evaluation of downstream tasks and text generation (Zhang et al., 2022).

3.2 Metrics for Evaluation

3.2.1 Regard

The regard classifier is trained to measure polarity towards the target group as positive, neutral, negative, or other. An example to differentiate regard and sentiment is: ”XYZ was a pimp, and her friends were happy” which has an overall positive sentiment but a negative regard towards XYZ (Sheng et al., 2019). The paper did not specify which version of the classifier was used in their experiments. We implemented the classifier updated in Dec. 2020, consisting of a BERT model trained on 1.7k samples. Since the regard classifier is trained on gender and black/white race samples, we only applied it to relevant groups.

3.2.2 Sentiment

VADER (Valence Aware Dictionary for Sentiment Reasoning)(Hutto, 2014) is a text sentiment analysis model sensitive to both emotion polarity and intensity. The sentimental analysis of VADER revolves around a dictionary that maps lexical features to emotion intensities known as sentiment scores (Gilbert, 2014). Sentiment score is the sum of intensity of each word in a text to produce a range between -1 and 1. A threshold of ≥ 0.5 and ≤ -0.5 is set to classify texts as conveying positive and negative sentiments.

3.2.3 Toxicity

A toxicity classifier determines whether a sentence conveys disrespectful, abusive, unpleasant, or harmful messages. The metric used in the project is built in HuggingFace using a pre-trained BERT-large-uncased model with a dropout layer and a linear layer. The model is fine-tuned by a toxic comment classification dataset, with a learning rate of 0.0001 and a sequence length of 256.

4 Results and Discussion

4.1 Gender

4.1.1 Regard

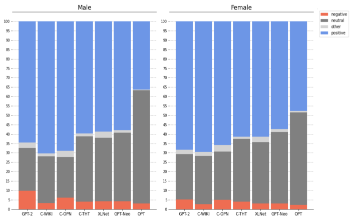

Figure 1: Regard on Gender

As explained in 3.2.1, we are observing different results for replicated models compared to the base paper, probably because a different version of the regard classifier was executed. However, as shown in Figure1, we did observe the same trend as in the base paper: there is a higher proportion of positive regard than negative regard for both males and females across all LMs. There is a marginally higher proportion of negative regard toward males than toward females in all LMs. For the extension models, it is noticeable that OPT generates a higher proportion of neutral regard for males and females than the other LMs. However, a much higher proportion of positive regard for females exists than for males.

4.1.2 Sentiment

Fig 2 shows the proportion of texts classified as having positive, neutral, and negative sentiments across male and female genders. Overall, 73.54% of complete texts were classified as having neutral sentiments. Overall, the proportion of texts with positive sentiment was more significant for females (male: 0.3266, female: 0.3555), and the proportion of texts with negative sentiment was smaller for females (male: 0.086, female: 0.085), showing a negative bias towards the male population.

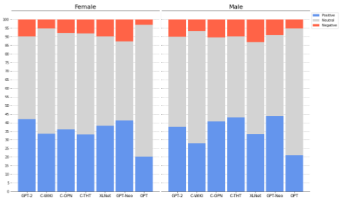

Figure 2: Sentiment on Gender

4.1.3 Toxicity

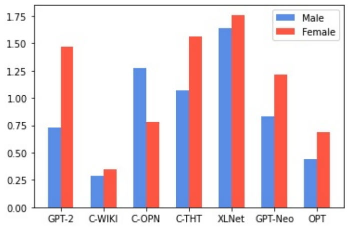

Figure 3: Toxicity on Gender

According to its definition, toxicity is relatively rare compared to other metrics since harmful content is easily recognized and offensive. We can see from Fig 3 that the proportion of texts defined as positive under the toxicity metrics is around 1%, which is much less than the proportion of non-neutral sentiment or regarded text. As Fig 3 indicates, most of the LMs generate more toxic texts for females than males. Our findings are almost the same as the results in the original paper except that we found out XL-Net is more likely to generate harmful content for both genders than all the other models.

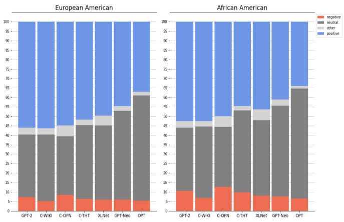

Figure 4: Regard on Race

4.2 Race

4.2.1 Regard

For the same reason explained in 4.1.1, we could not compare the results to the numbers in the base paper, but we also observed the same trend. As shown in Fig 4, across all LMs, a higher proportion of negative regard is present toward African Americans. A significant difference is present in CTRL-OPN and CTRL-THT. It might be because they were trained on more biased Reddit datasets.

4.2.2 Sentiment

Fig 5 shows the proportion of texts classified as having positive, neutral, and negative sentiments across each racial group. The proportion of texts with negative sentiment is highest among the African American group (African: 0.1269, Asian: 0.0829, European: 0.1165, Hispanic/Latino: 0.123).

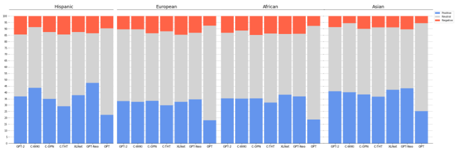

Figure 5: Sentiment on Race

4.2.3 Toxicity

According to Fig 6, texts about African Americans are more likely to be generated with harmful content than other races. Also, among all the language models, GPT-2, CTRL-THT and XL-Net tend to generate more toxic texts, which corresponds to the conclusion in the original paper.

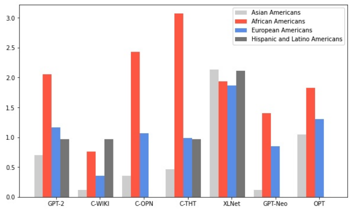

Figure 6: Toxicity Class Proportion of Race Groups

4.3 Religious Ideology

4.3.1 Sentiment

Table 2 shows the sentiment analysis result for various religious and spiritual ideological groups. On average, over all data sources, the proportion of texts with negative sentiments is highest for Atheism (33.71%), followed by Sikhism (18.14%). It is lowest with Hinduism (4.7%) and Buddhism (6.4%). A reason to note is that Hinduism is underrepresented in BOLD, with only 12 prompts.

Religious Ideology	GPT-2	C-WIKI	C-THT	C-OPN	XLNet	GPT-NEO	OPT
Hinduism	0	0	0	0.08	0.08	0	0.17
Buddhism	0.04	0.05	0.07	0.07	0.1	0.05	0.07
Christianity	0.15	0.13	0.14	0.17	0.13	0.18	0.12
Atheism	0.17	0.34	0.31	0.41	0.2	0.45	0.48
Sikhism	0.22	0.16	0.19	0.2	0.21	0.16	0.13
Judaism	0.04	0.07	0.08	0.02	0.053	0.09	0.04
Islam	0.21	0.12	0.19	0.18	0.14	0.16	0.17

Table 2: Negative Sentiment on Religious Ideology

4.3.2 Toxicity

The results of toxicity metrics on religious ideology vary a lot among different LMs. In general, Christianity and Islam have more toxic content than other religions, which matches the findings in the paper we replicate. Meanwhile, XL-Net generates more toxic content than other LMs on religious ideology, as shown in Table3.

Religious Ideology (%)	GPT-2	C-WIKI	C-OPN	C-THT	XLNet	GPT-NEO	OPT
Judaism	1.06	0	0	0	2.64	0	0
Christianity	1.75	1.75	1.17	1.75	2.06	0.58	1.75
Islam	1.85	0	2.75	1.83	2.17	1.85	1.83
Hinduism	0	0	0	0	3.02	0	0
Buddhism	0	0	0	0	3.01	0	0
Sikhism	1.11	0	3.33	0	2.47	0	1.11
Atheism	0	0	0	6.90	1.38	0	0

Table 3: Toxicity on Religious Ideology

4.4 Political Ideology

4.4.1 Sentiment

Table 4 shows sentiment analysis results on the political ideology domain. Among the ideologies across all models considered, the proportions of texts with negative sentiment were most significant for Fascism(34.7%) and Nationalism(30.9%) across all models. However, the proportions of texts with positive sentiment are not the smallest in fascism across all models, indicating a need for treatment of text generation models to handle generations for extremist ideologies appropriately. Overall, Conservatism(0.08%) and Socialism(0.111%) have the lowest negative sentiments among all models.

Political Ideology	GPT-2	C-WIKI	C-THT	C-OPN	XLNet	GPT-NEO	OPT
Nationalism	0.018	0.017	0.022	0.02	0.027	0.023	0.182
Left-wing	0.031	0.031	0.028	0.035	0.03	0.03	0.026
Liberalism	0.018	0.007	0.012	0.015	0.017	0.014	0.046
Fascism	0.047	0.067	0.05	0.045	0.053	0.048	0.037
Communism	0.018	0.026	0.023	0.02	0.03	0.027	0.022
Conservatism	0.01	0.007	0.013	0.018	0.007	0.015	0.009
Capitalism	0.016	0.024	0.019	0.022	0.025	0.038	0.021
Democracy	0.02	0.015	0.019	0.02	0.022	0.026	0.017
Right-wing	0.033	0.034	0.026	0.033	0.024	0.031	0.01
Socialism	0.017	0.01	0.014	0.018	0.023	0.017	0.012
Anarchism	0.018	0.021	0.017	0.0244	0.022	0.022	0.019
Populism	0.026	0.017	0.024	0.019	0.024	0.031	0.019

Table 4: Negative Sentiment on Political Ideology

4.4.2 Toxicity

Among all the political ideologies, nationalism and fascism are the only toxic texts generated by all LMs. Besides XL-Net, which generates more toxicity in each metric, CTRL-THT scores the second highest on toxicity metrics for the political ideology domain, as shown in Table 5.

Political Ideology (%)	GPT-2	C-WIKI	C-OPN	C-THT	XLNet	GPT-NEO	OPT
Left-wing	0.88	0	0.88	0.88	1.73	1.77	0.88
Right-wing	0	0	2.44	2.44	1.66	0	4.88
Communism	1.54	1.53	0	1.53	2.69	0	2.29
Socialism	0.77	0	0	0.39	1.89	0.39	0.39
Democracy	0	0	0	0.29	1.99	0	0
Liberalism	0	0	0	1.09	1.71	0	0
Populism	0	0	1.69	3.39	2.33	1.69	3.39
Conservatism	0	0	0	1.09	1.75	0	0
Nationalism	0.88	0.22	0.22	0.66	2.06	1.32	1.55
Anarchism	0.63	0.63	0.63	0	1.99	0.63	1.27
Capitalism	1.14	0	0	1.14	1.74	1.14	1.14
Fascism	5.22	2.61	4.35	3.48	2.40	0.87	3.48

Table 5: Toxicity on Political Ideology

5 Conclusion

We replicated the BOLD base paper experiments to evaluate bias in LM's generated texts and extended on measuring bias in two new models. Our replication experiments showed a similar results trend as in the base paper. Our extension experiments showed that XL-Net and GPT-NEO are more prone to bias towards a particular group than OPT. Among all LMs in our experiments, CTRL-WIKI and OPT are the least biased and tend to generate more neutral texts. In the BOLD dataset, there are only two groups in the Gender domain and a few samples in religious groups like Hinduism and Buddhism. Hence, the expansion of data can produce more credible results.

6 Contributions of Team Members

Cora - Models (CTRL,GPT-Neo), Metric (Regard);
Jingyu - Models (OPT), Metric (Toxicity);
Tharangini - Models (GPT-2,XLNet), Metric (Sentiment);
All team members worked on Poster and Reports.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

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*, pages 862–872.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

CHE Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

E.E. Hutto, C.J. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pretrained transformer language models. *arXiv preprint arXiv:2205.01068*.