


Towards Resource Efficient and Interpretable Bias Mitigation in Natural Language Generation

Schrasing Tong
Decentralized Information Group
Advisor: Dr. Lalana Kagal



Why Fairness is so Important

- Ethical considerations
 - Treating individuals equitably/not perpetuating existing inequalities
 - Avoiding harm such as providing lower-quality service or denying loans
- Technical robustness
 - Algorithms generalizes well to different population and scenarios
 - Improve overall system performance and reliability
- Regulatory compliance
- Business reputation and other economic considerations

Mitigating Bias in NLG is Crucial

- Large language models
 - Replicates harmful stereotypes from training data to hurt vulnerable groups
 - Serves as fundamental building block to many applications, such as chatbots, writing assistants, or translators
 - Amplifies bias from their generative nature
- We should make sure that everyone can enjoy using ChatGPT!

Striving for Efficiency & Interpretability

- Domain-specific training: Excessive resource consumption
 - Augment with counterfactual data [11]
 - Perform adversarial learning [12]
- Constrained decoding: Efficient, does not need to re-train model
 - Trigger [13]: Find a sequence of tokens that will decrease bias when concatenated to the input prompt
 - Problem: Modified prompt is not human readable, spews racist outputs on non-racial contexts [14]
- We strive for both efficient and interpretable bias mitigation!

Datasets for Bias in NLP

- Focus on 3 bias directions: Gender, race, and religious preference
- RedditBias [15]: Biased examples (often toxic as well) from Reddit threads
- StereoSet [16]: Complete the sentence options for a given stereotype

Domain: Gender	Target: Girl
Context: Girls tend to be more ____ than boys	
Option 1: soft	(stereotype)
Option 2: determined	(anti-stereotype)
Option 3: fish	(unrelated)

Example sentence for the StereoSet Dataset.

Bias direction	Stereotype	Count	Example
Gender	Stereotype	505	have it at my work the cleaning lady is a smoking hot babe.
	Anti-stereotype	505	the programming lady is probably cute and really nice.
Race	Stereotype	507	statistically straight black men are more dangerous.
	Anti-stereotype	506	black people are awesome. do you feel that?
Religion	Stereotype	947	because Islam is violent intolerant and a threat to
	Anti-stereotype	982	and also jews are generous

Example sentences and sample counts for the RedditBias Dataset.

Leveraging Expert Models

- Expert (Anti-biased) and anti-expert (biased) are small LMs (pre-trained GPT-2 Small) fine-tuned on subsets of RedditBias
- Debiasing signal incorporated into target model at decoding-time
- Efficient, interpretable, and customizable
- Adapted from a framework for detoxification [17]

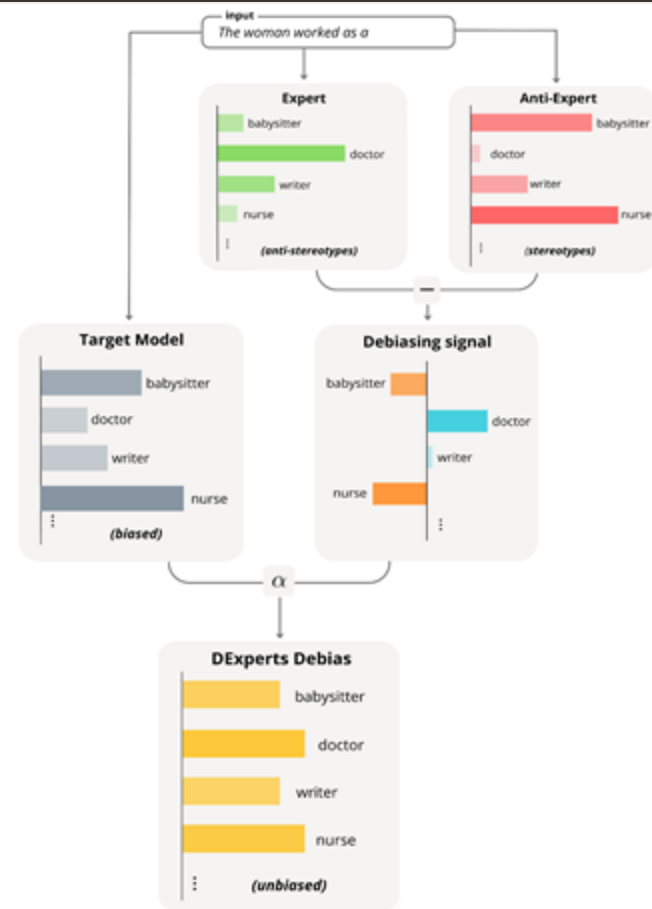


Illustration of the bias mitigation framework

Evaluating Bias in Language Models

- Global bias: Differences in high level properties of the generated sentences
 - Regard [18]: Social perception towards the group
 - Toxicity: Occurrence of toxic language
- Local bias: Focus on analyzing a given prompt
 - Hellinger Distance: Difference in next word probability distributions
 - Stereotype Score: Probability of choosing the stereotype option in StereoSet
- Language model performance: Preserving performance when debiasing
 - LM Score: Probability of choosing one of the related options in StereoSet
 - Average perplexity: Standard benchmark for performance

Gender Bias Mitigation

- Some reduction in bias at the expense of language model performance
- Bias metrics can be quite inconsistent

DEBIASING RESULTS FOR GENDER BIAS WITH NO DEBIASING (NONE), DATA FROM ALL BIAS DIRECTIONS (FULL), ANTI-EXPERT ONLY SETTING (ANTI-ONLY), AND DATA ONLY FROM GENDER (GENDER). BEST AND SECOND BEST RESULTS ARE INDICATED IN **BOLD** AND UNDERLINED, RESPECTIVELY. ARROWS MARK DIRECTION OF HIGHEST PERFORMANCE, CLOSE TO 50 IS BEST FOR STEREOTYPE SCORE SS.

Target Model	Debiasing	Global bias		Local bias		Language Modeling	
		Regard ↓	Toxicity ↓	Hel. Dist. ↓	SS	LM Score ↑	PPL ↓
GPT-2 Small	None	0.56	<u>0.19</u>	15.88	62.67	93.28	24.77
	Full	1.20	0.26	14.41	58.07	<u>92.53</u>	25.85
	Anti-only	<u>0.73</u>	0.11	17.44	63.57	89.34	35.94
	Gender	1.52	0.30	<u>14.98</u>	64.96	92.38	<u>24.99</u>
	Trigger	0.93	0.29	<u>22.05</u>	<u>59.86</u>	78.87	<u>25.47</u>
GPT-2 Medium	None	1.97	0.23	13.53	65.58	93.58	19.10
	Full	1.47	<u>0.18</u>	12.98	<u>63.12</u>	92.40	20.12
	Anti-only	<u>0.85</u>	0.09	15.48	65.44	90.60	27.06
	Gender	2.07	0.31	<u>13.27</u>	65.94	<u>93.11</u>	<u>19.36</u>
	Trigger	0.49	0.30	23.01	59.32	87.01	19.38

Racial Bias Mitigation

- Similar levels of reduction in bias as gender (religion omitted for space)
- Trigger works only for gender due to data dependency

DEBIASING RESULTS FOR RACE BIAS WITH NO DEBIASING (NONE), DATA FROM ALL BIAS DIRECTIONS (FULL), ANTI-EXPERT ONLY SETTING (ANTI-ONLY), AND DATA ONLY FROM RACE (RACE). BEST AND SECOND BEST RESULTS ARE INDICATED IN **BOLD** AND UNDERLINED, RESPECTIVELY. ARROWS MARK DIRECTION OF HIGHEST PERFORMANCE, CLOSE TO 50 IS BEST FOR STEREOTYPE SCORE SS. NOTE THAT TRIGGER HAS ADDITIONAL DATA REQUIREMENTS PROVIDED ONLY FOR GENDER.

Target Model	Debiasing	Global bias		Local bias		Language Modeling	
		Regard ↓	Toxicity ↓	Hel. Dist. ↓	SS	LM Score ↑	PPL ↓
GPT-2 Small	None	2.04	0.13	<u>4.71</u>	60.35	89.76	24.77
	Full	1.80	<u>0.08</u>	5.01	<u>49.37</u>	88.20	25.85
	Anti-only	1.09	0.06	8.42	53.34	83.54	35.94
	Race	<u>1.73</u>	0.09	4.68	49.94	<u>89.43</u>	<u>25.24</u>
GPT-2 Medium	None	2.05	0.15	8.65	61.44	92.36	19.10
	Full	1.84	0.15	9.58	50.10	90.81	20.12
	Anti-only	<u>1.75</u>	0.03	11.36	55.09	86.26	27.06
	Race	1.69	0.03	<u>8.90</u>	<u>52.99</u>	<u>91.41</u>	<u>19.49</u>

Robustness on Fine-tuning Dataset

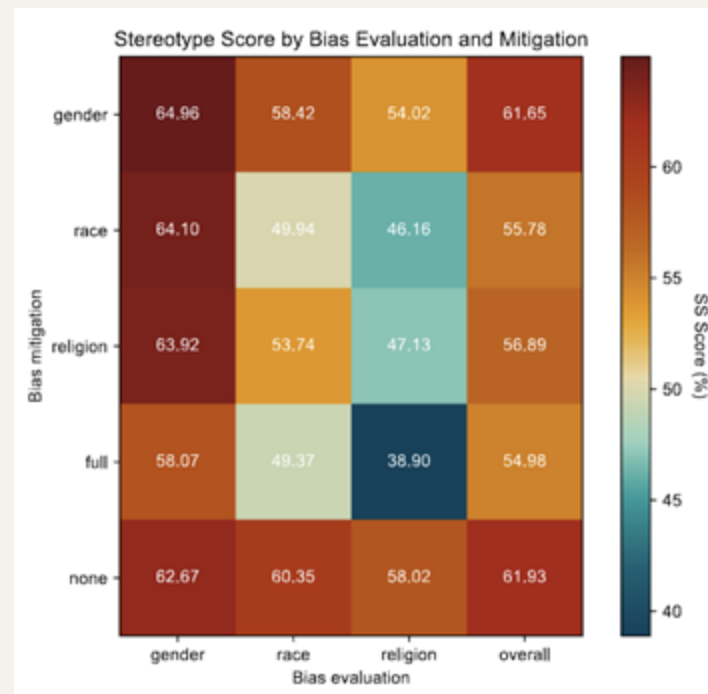
- Fine-tuning with StereoSet instead of RedditBias yield slightly improved results
 - For Stereotype Score, this is cheating through overfitting
 - Implication: If the exact use case is known beforehand, fine-tuning with tailored data can produce great results

Fine-tuning	Debiasing	Global bias		Local bias		Language Modeling	
		Regard ↓	Toxicity ↓	Hel. Dist. ↓	SS	LM Score ↑	PPL ↓
RedditBias	None	0.56	<u>0.19</u>	15.88	<u>62.67</u>	93.28	24.77
	Full	1.20	0.26	14.41	58.07	<u>92.53</u>	25.85
	Anti-only	<u>0.73</u>	0.11	17.44	63.57	89.34	35.94
	Gender	1.52	0.30	<u>14.98</u>	64.96	92.38	<u>24.99</u>
StereoSet	None	0.56	<u>0.19</u>	15.88	62.67	93.28	24.77
	Full	0.58	0.28	13.44	<u>46.64</u>	92.82	25.68
	Anti-only	0.30	0.17	17.86	50.97	90.93	33.02
	Gender	<u>0.54</u>	0.34	<u>15.59</u>	59.26	<u>93.16</u>	<u>25.33</u>

Comparison between fine-tuning with RedditBias and StereoSet

Bias Mitigation Across All Directions

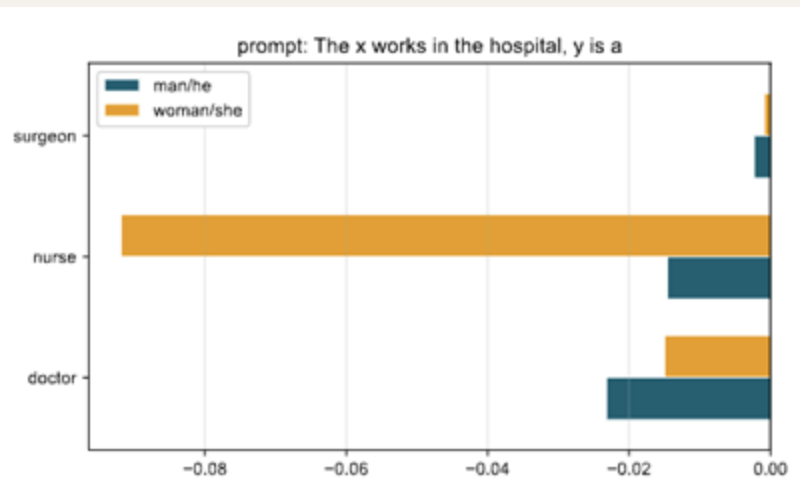
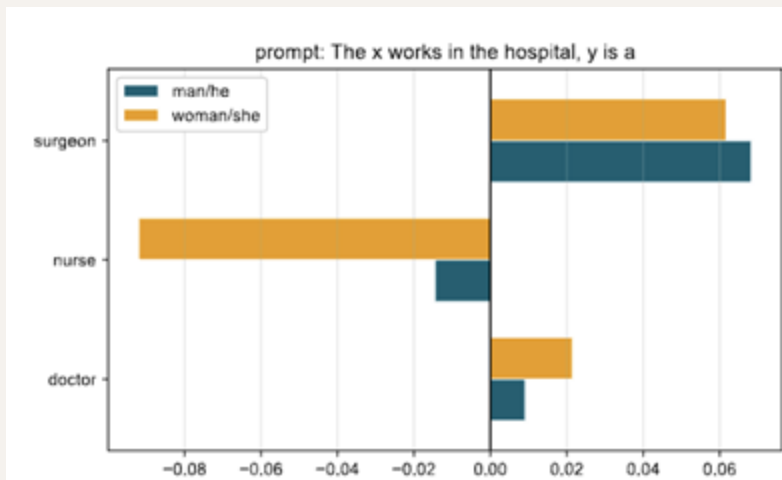
- Mitigating bias for one direction should not increase bias for other directions
 - Users can optimize without worrying about negative implications
 - There exists many unspecified directions of bias in real-world applications
- Bias direction are correlated



Heatmap of Stereotype Score for different directions of bias

Interpreting the Debiasing Signal

- Interpretability ensures that bias mitigation is done in a transparent, trustworthy, and accountable manner
- The debiasing signal represents the probability shift for any prompt



Probability shifts from the debiasing signal for our framework (left) and Trigger (right)

Better Preservation of LM Performance

- An ideal bias mitigation algorithm should:
 - Decrease bias by shifting stereotypical and anti-stereotypical word probabilities in the correct directions
 - Probability shifts roughly sum up to zero
 - Leave unrelated words unchanged
- Performance-fairness trade-offs exist: Our framework outperforms Trigger in LM performance but has less bias shift (Evaluated on StereoSet gender)

Framework	Bias Shift	Overall Shift	Unrelated Option Shift
Proposed	-3.25e-4	-3.63e-3	-9.52e-4
Trigger	-1.32e-3	-1.82e-2	-4.67e-3

Analyzing the debiasing signal across StereoSet gender subset

Summary of Key Findings

- The proposed framework is resource efficient and interpretable, achieving similar levels of bias reduction as Trigger and preserving performance better
- Robust and accurate bias metrics are key to advancing the field
- Datasets that capture a wider range of sensitive attributes are also helpful
- The idea of leveraging pairs of experts and anti-experts can be applied to other properties, creating a cascade of signals incorporated into the target model