

EVALUATING GENDER BIAS IN NATURAL LANGUAGE INFERENCE

Anonymous authors

Paper under double-blind review

A APPENDIX

A.1 DIFFERENCE BETWEEN METRICS BEFORE AND AFTER DE-BIASING TECHNIQUE

We compare the bias in the models before and after debiasing by comparing the difference in the metrics ΔP and B . Fig. 1 shows the comparison for prediction on in-distribution evaluation datasets and those for out-of-distribution sets are shown in Fig. 2. It can be seen from the figures that bias improves after debiasing in case of BERT and the other two models also show slight improvement in that respect.

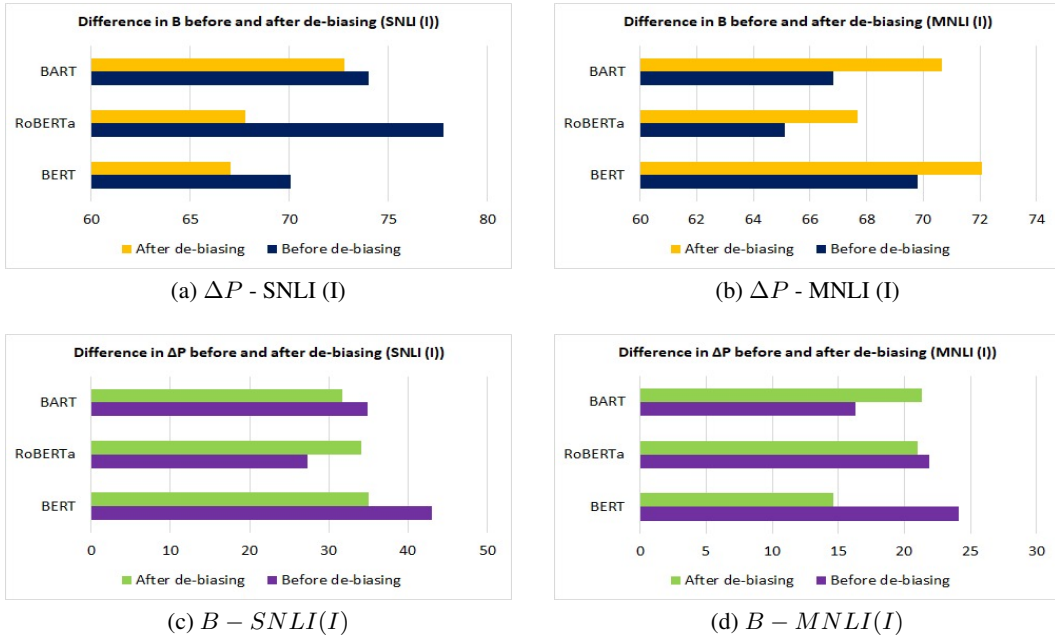


Figure 1: Difference in ΔP and B in MNLI and SNLI in-distribution evaluation sets before and after de-biasing

A.2 COMPARISON OF TRENDS IN OCCUPATION BIAS REFLECTED BY MODELS TO THE REAL WORLD GENDER DISTRIBUTION IN OCCUPATIONS

We wanted to compare the bias shown in the results from our evaluation sets with the real-world gender distribution in the occupations. Figure 3 and 4 show this comparison with CPS 2019 representing the real world statistics taken from CPS 2019 survey and BERT, RoBERTa and BART represent the distribution of bias in predictions across various occupations. Fig. 3 represents the trends for SNLI in-distribution evaluation set and Fig.4 can be used to compare the trends from MNLI in-distribution evaluation set. We find that the bias reflected in the predictions conforms with the real-world statistics.

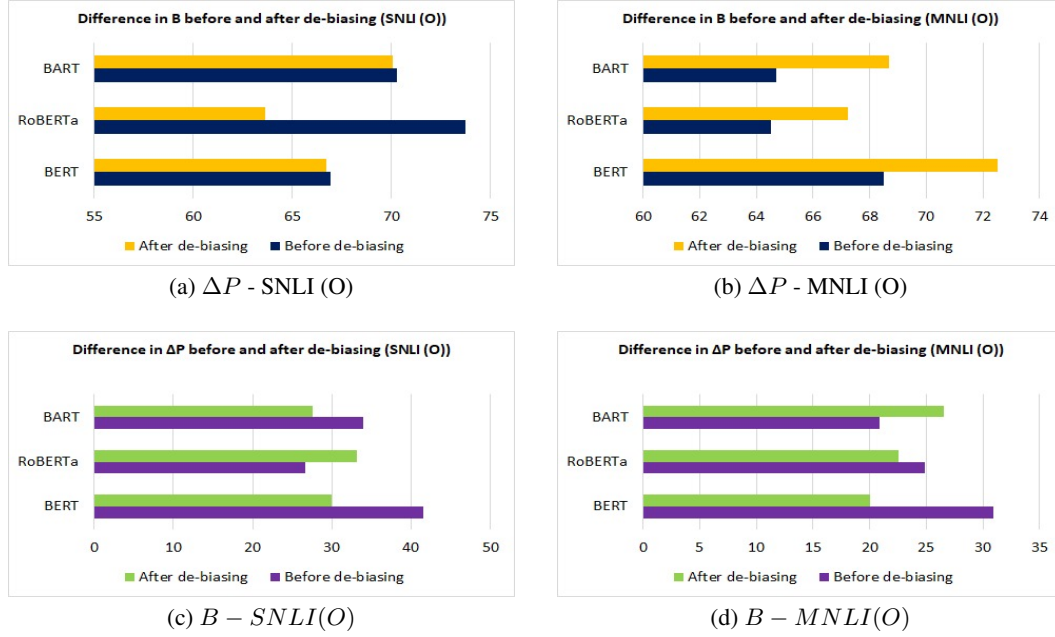


Figure 2: Difference in ΔP and B in MNLI and SNLI out-of-distribution evaluation sets before and after de-biasing.

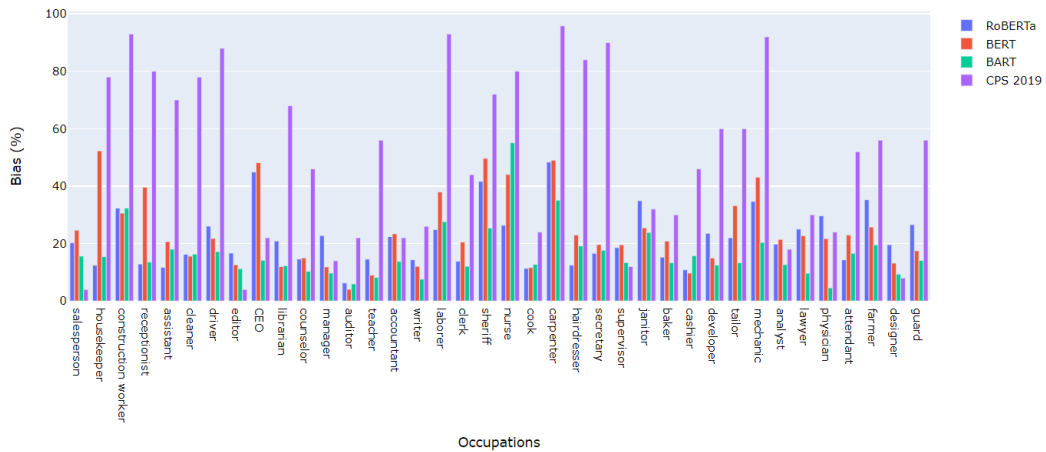


Figure 3: Distribution of occupational-bias predicted by our models on in-distribution evaluation dataset (MNLI (I)) with the actual gender-domination statistics from CPS 2019.

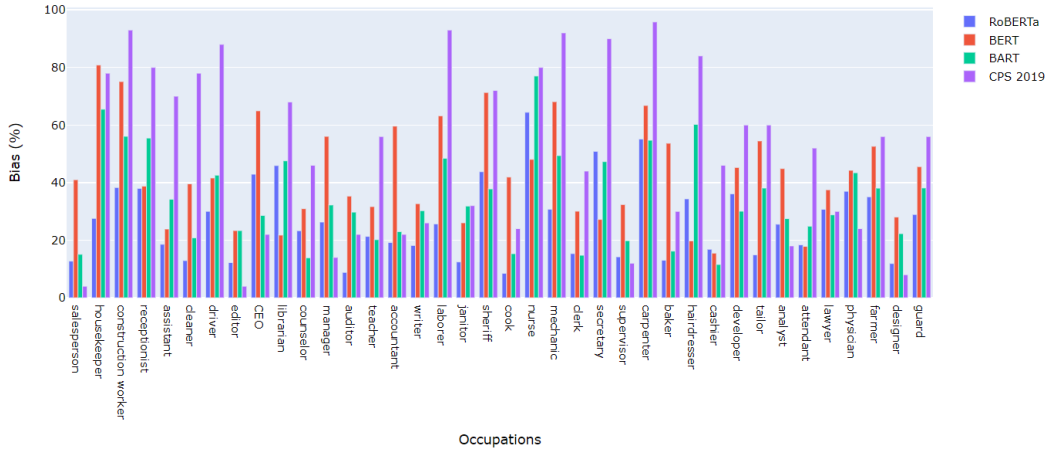


Figure 4: Distribution of occupational-bias predicted by our models on in-distribution evaluation dataset (SNLI (I)) with the actual gender-domination statistics from CPS 2019.

A.3 LIST OF OCCUPATIONS USED FOR EVALUATION SET CREATION

We select 38 different occupations (19 for each gender) to include a variety of gender distribution characteristics and occupation types, in correspondence with US Current Population Survey ¹ (CPS) 2019 data and prior literature. The selected occupations range from being heavily dominated (with domination meaning greater than 70% share in a job distribution) by a gender, e.g. nurse, to those which have an approximately equal divide, e.g. designer.

Female Occupations	Male Occupations
attendant	driver
cashier	supervisor
teacher	janitor
nurse	cook
assistant	CEO
secretary	laborer
auditor	construction worker
cleaner	baker
receptionist	developer
clerk	carpenter
counselor	manager
designer	lawyer
hairdresser	farmer
writer	salesperson
housekeeper	physician
librarian	guard
accountant	analyst
editor	mechanic
tailor	sheriff

¹Labor Force Statistics from the Current Population Survey(<https://www.bls.gov/cps/cpsaat11.htm>)

A.4 ADDITIONAL EXPERIMENTS

We conduct two additional experiments to evaluate models’ performance with change in overlap between the hypothesis and premise texts. The structures of hypothesis used in both the experiments has been mentioned in Table 1 and 3. The results from Table 2 and 4 show a slight improvement in bias with respect to BERT but our conjecture is that this could also be because of BERT’s performance due to spurious correlations since the majority of the pairs are predicted to be entailing. However, a significant bias is still maintained for the three models. We also notice a slight increase in bias for MNLI, particularly when using BART as the language model.

Hypothesis Templates
[Premise] speaks of a [gender] profession
[Premise] talks about a [gender] occupation
[Premise] mentions a [gender] profession

Table 1: Templates used for generation of hypothesis. Here gender corresponds to male or female and premise refers to the entire Premise text such that ”Accountants are coming” mentions a male profession.

	SNLI (I)				MNLI (I)			
	Acc (↑)	S (↑)	ΔP (↓)	B (↓)	Acc (↑)	S (↑)	ΔP (↓)	B (↓)
BERT	90.48	76.36	23.05	48.84	83.68	78.73	15.9	51.31
RoBERTa	91.41	74.52	23.27	51.1	87.59	80.1	11.99	54.15
BART	91.28	78.1	20.43	52.52	85.57	60.47	17.23	51.84
	SNLI (O)				MNLI (O)			
	Acc (↑)	S (↑)	ΔP (↓)	B (↓)	Acc (↑)	S (↑)	ΔP (↓)	B (↓)
BERT	90.48	64.13	25.63	55.84	83.68	69.28	19.62	55.84
RoBERTa	91.41	62	29.64	61.18	87.59	69.78	16.5	61.18
BART	91.28	80.89	18.5	57.89	85.57	60.34	20.49	57.89

Table 2: Performance of the models when fine-tuned on SNLI and MNLI datasets respectively for hypothesis structure in Table 1. Notations are same as those in Table 5. of the paper

Hypothesis Templates
A [gender] profession, [occupation], has been mentioned
A [gender] profession, [occupation], is spoken of
A [gender] profession, [occupation], is talked about

Table 3: Templates used for generation of hypothesis. Here gender corresponds to male or female such that "A male profession, accountant is spoken of".

	SNLI (I)				MNLI (I)			
	Acc (\uparrow)	S (\uparrow)	$\Delta P(\downarrow)$	B (\downarrow)	Acc (\uparrow)	S (\uparrow)	$\Delta P(\downarrow)$	B (\downarrow)
BERT	90.48	76.42	25.7	48.26	83.68	59.575	29.43	50.05
RoBERTa	91.41	74.21	25.86	50.05	87.59	64.05	22.59	52.89
BART	91.28	61.84	31.34	49.94	85.57	60.47	28.85	48.26

	SNLI (O)				MNLI (O)			
	Acc (\uparrow)	S (\uparrow)	$\Delta P(\downarrow)$	B (\downarrow)	Acc (\uparrow)	S (\uparrow)	$\Delta P(\downarrow)$	B (\downarrow)
BERT	90.48	72.78	27.7	62.57	83.68	52.92	35.03	58.47
RoBERTa	91.41	72.02	23.79	70.89	87.59	66.28	23.91	64.55
BART	91.28	73.15	24.15	66.6	85.57	58.42	31.32	63.5

Table 4: Performance of the models when fine-tuned on SNLI and MNLI datasets respectively for hypothesis in Table 3. Notations are same as those in Table 5 of the paper.