Investigating the Effect of Debiasing Methods on Intersectional Biases in Language Models

Ellie Talius, Ananya Karthik
Department of Computer Science, Stanford University

Motivation

- **No previous research on developing debiasing methods** for intersectional biases in language models
- Previous research on evaluating intersectional bias shows **high degrees of bias** in language models
- **We design three intersectional debiasing techniques** based on the single-identity debiasing method proposed by [1].
- **We evaluate these three techniques using intersectional bias metrics** defined by [2].

Methodology

- Experimental settings
  - **Model**: Bert-Tiny
  - **Data**: monolingual English secession of the News-commentary-v15 corpus [3]
  - **Identities**: Gender (M/F), Race (European-American [EA] / African-American [AA] / Hispanic-American [HA]), Age (Young [Y] / Elderly [E])
  - **Evaluation**: 7 intersectional embedding association tests, with 3 sub-tests each [2] **(*: self-designed)**
    - Word: Single word embedding association (Alice v. Doctor)
    - Sent: Sentence embedding association (This is a doctor v. Alice is here)
    - C-word: Single contextual word embedding association (This is a doctor v. Alice is here)

Gender - African American Tests	
I1	EA F, AA F (least extreme)
I2	AA M, AA F
I3	EA M, AA M
I4	EA M, EA F
I5	EA M, AA F (most extreme)
Gender - Age Test*	
I6	Y M, E F (most extreme)
Gender - Hispanic Test*	
I7	EA M, HA F (most extreme)

Experiments

Intersectional De-biasing Methods

Original

$$L_{bias} = \sum_{i=1}^{N} \sum_{t \in V_T} \sum_{x \in \Omega(t)} \sum_{a \in V_A} (v_i(a)^T E_i(t, x; \theta_e))^2$$

$$L_{reg} = \sum_{x \in A} \sum_{w \in x} \sum_{i=1}^{N} \| E_i(w, x; \theta_e) - E_i(w, x; \theta_{pre}) \|$$

$$\boxed{L = \alpha L_{bias} + (1 - \alpha) L_{reg}}$$

Method 1

$$L_{bias_j} = \sum_{i=1}^{N} \sum_{t \in V_{Tj}} \sum_{x \in \Omega(t)} \sum_{a \in V_{Aj}} (v_i(a)^T E_i(t, x; \theta_e))^2$$

$$L_{reg_j} = \sum_{x \in A_j} \sum_{w \in x} \sum_{i=1}^{N} \| E_i(w, x; \theta_e) - E_i(w, x; \theta_{pre}) \|$$

$$\boxed{L = \textstyle\sum_{j=1}^{M} \alpha L_{bias_j} + (1 - \alpha) L_{reg_j}}$$

Method 2

$$L_{bias} = \sum_{i=1}^{N} \sum_{t \in V_{T_{all}}} \sum_{x \in \Omega(t)} \sum_{a \in V_{A_{all}}} (v_i(a)^T E_i(t, x; \theta_e))^2$$

$$L_{reg} = \sum_{x \in A_{all}} \sum_{w \in x} \sum_{i=1}^{N} \| E_i(w, x; \theta_e) - E_i(w, x; \theta_{pre}) \|$$

$$\boxed{L = \alpha L_{bias} + (1 - \alpha) L_{reg}}$$

Method 3

$$L_{intersect} = \sum_{i=1}^{N} \sum_{t \in V_{T_{intersect}}} \sum_{x \in \Omega(t)} \sum_{a \in V_{A_{all}}} (v_i(a)^T E_i(t, x; \theta_e))^2$$

$$\boxed{L = \textstyle\sum_{j=1}^{M} (\alpha L_{bias_j} + (1 - \alpha) L_{reg_j}) + \beta L_{intersect}}$$

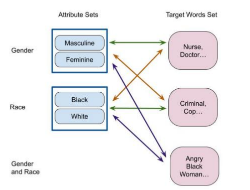

Figure 1. Arrows represent the sets of word embeddings between which the dot product is minimized in the different intersectional debiasing methods. Method 1's loss: green; Method 2's loss: green and orange; Method 3's loss: green and purple.

Test	Encoding	Bert-Tiny	Race Only	Gender Only	R-G 1	R-G 2	R-G 3
I1	word	**1.425**	**1.457**	**1.453**	**1.322**	**1.355**	**1.365**
	sent	**1.362**	**1.38**	**1.361**	**1.273**	**1.334**	**1.275**
	c-word	-0.465	-0.269	**0.447**	0.447	**0.517**	**0.462**
I2	word	**1.407**	**1.152**	**1.256**	**1.403**	**1.254**	**1.381**
	sent	**0.838**	**0.643**	**0.749**	**0.857**	**0.667**	**0.864**
	c-word	-0.173	-0.065	0.118	0.152	0.224	0.13
I3	word	-0.21	0.391	0.36	0.365	-0.072	0.482
	sent	**0.437**	0.5	0.318	**0.397**	0.209	**0.414**
	c-word	-0.296	-0.377	0.254	0.143	0.203	0.279
I4	word	-0.75	-0.539	-0.176	0.176	-0.398	0.215
	sent	-0.822	-0.749	-0.518	-0.205	-0.665	-0.143
	c-word	0	-0.175	-0.08	-0.156	-0.098	-0.058
I5	word	1.502	**1.242**	**1.339**	**1.402**	**1.061**	**1.436**
	sent	**1.326**	**0.959**	**1.051**	**1.139**	**0.918**	**1.169**
	c-word	-0.465	-0.441	**0.47**	0.294	**0.423**	**0.406**

Table 1: Effect sizes for intersectional tests involving race and and gender run on different models. Bolded values represent significant tests (p < 0.01). Positive values represent pro-stereotypical bias, negative values represent anti-stereotypical bias. R-G 1 refers to a model debiased for race and gender using the first intersectional debiasing method.

Test	Encoding	Bert-Tiny	Age Only	A-G 1	A-G 2	A-G 3
I6	word	0.153	0.513	0.388	0.043	0.505
	sent	0.283	**0.85**	**0.905**	0.527	**0.758**
	c-word	0.252	0.25	0.392	0.339	0.33

Table 2: Effect sizes for the intersectional test involving ageand and gender run on different models. Bolded values represent significant tests (p <0.01). Positive values represent pro-stereotypical bias, negative values represent anti-stereotypical bias. A-G 1 refers to a model debiased for age and gender using the first intersectional debiasing method.

Test	Encoding	Bert-Tiny	Hispanic Only	H-G 1	H-G 2	H-G 3
I7	word	-0.614	-0.486	0.77	-0.176	0.594
	sent	-0.413	**-0.387**	**0.021**	-0.358	**-0.161**
	c-word	-0.401	-0.319	0.28	**0.346**	**0.31**

Table 3: Effect sizes for the intersectional test involving the Hispanic ethnicity and and gender run on different models. Bolded values represent significant tests (p <0.01). Positive values represent pro-stereotypical bias, negative values represent anti-stereotypical bias. H-G 1 refers to a model debiased for the hispanic ethnicity and gender using the first intersectional debiasing method.

Conclusions

- We are able to **decrease intersectional bias found in language models** using three intersectional debiasing methods, all of which perform better than single identity debiasing
- Debiasing **method 1** performs the best for less extreme intersectionality tests (ie European-American Female v. African-American Female) while debiasing **methods 2 and 3** perform the best for more extreme intersectionality tests
  - Perhaps a one-size-fits-all approach for intersectional debiasing is not ideal
- Debiasing **African-American x Female** shows more improvements over the baseline than **Elderly x Female,** likely because there is higher racial bias in the baseline model, and more training data related to race than age
- Debiasing for Hispanic x Female has the fewest significant results, likely due to a lack of data

Future Work

- Explore debiasing methods for **more intersectionalities** (disability, sexuality, etc)
- Experiment with debiasing **other language models**, including current SOTA and large language models

References

[1] Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings, 2021.
[2] Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations, 2019.
[3] Loïc Barrault, Ondˇrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, et al. Proceedings of the fifth conference on machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, 2020.