Bias and Debiasing in NLP

Bingyao Wang
Dec 7, 2021
CPSC 677 ANLP

Overview

- Background
- Paper 1 : On Mitigating Social Biases in Language Modelling and Generation
  - Gender and racial bias
- Paper 2 : Mitigating Language-Dependent Ethnic Bias in BERT
  - Ethnic bias
- Paper 3 : Debiasing Pre-trained Contextualised Embeddings
  - Debias pre-trained contextualised embeddings
- Paper 4: Challenges in Automated Debiasing for Toxic Language Detection
  - Challenges of debiasing

Background

Bias

- Definition
  - Any kind of preference or prejudice toward a specific individual, group, or community over others

Demos/Examples

- Bias in static word embeddings
  - Jupyter notebook, could refer to this
- Bias in models
  - https://unqover.apps.allenai.org/
  - https://demo.allennlp.org/masked-lm

Biases

- Allocation Bias
  - A system unfairly allocates resources to certain groups over others
- Representation Bias
  - systems detract the social identity and representation of certain groups, stereotyping in which existing societal stereotypes are reinforced
- Under-representation Bias
  - certain groups are disproportionately underrepresented
- Recognition Bias
  - a recognition algorithm’s accuracy is lower for certain groups
- ...

WHY -- Negative Impacts

- Representational Impacts
  - Propagating stereotypes, misrepresentations, or denigrations of social groups
- Allocational Impacts
  - technologies that are less effective or detrimental for certain populations become barriers that actively prevent those populations from using the technology
  - E.g. MT errors lead to arrests (link)
- Vulnerability Impacts
  - Amplify a group's vulnerability to manipulation and harm
  - privacy-related issues, misinformation, or radicalizing views in generated text could make a group more likely to be attributed to specific stereotypes

Contributors to Biases

- Biases from Data
  - gender, religion, and ethnic biases in Reddit communities
  - News articles also exhibit stereotypes
- Biases from Model Architecture
  - larger models contain more gender bias
  - bias tends to be concentrated in a small number of neurons and attention heads
- Biases from Decoding
  - Greedy, beam, top-k, nucleus
  - the less diverse(text length and vocabulary sizes) search techniques lead to better scores for individual fairness, group fairness, and gendered word co-occurrence ratios.

Contributors to Biases

- Biases from Evaluation
  - using perplexity as measured by models pre-trained on datasets largely containing non-AAE text leads to an unfair evaluation of AAE text
  - the choice of human annotators could drastically influence the evaluation standards for generated text (evaluation depends on human labels)
- Biases from Deploying Systems
  - Deploying NLG system - feedback loop that benefits some communities and further disadvantages others
  - E.g. Internet access issue
- ….

Metrics

- Regard Ratio
  - *negative-neutral-positive* regard score ratios of text generated from bias-inducing prompts
- Sentiment Ratio
  - *negative-neutral-positive* sentiment score ratios of text generated from African American English (AAE) versus White-Aligned English (WAE) prompts
- Individual and Group Fairness through Sentiment
  - comparisons of the sentiment distributions of generated text across demographics and prompts
- Gendered Word Co-occurrence Score
  - mean and standard deviations of the absolute log ratio of probabilities: P(word|female terms) to P(word|male terms) across all words in generated text
- SEAT Score
  - measures the associations between contextual representations of two sets of target concepts (e.g., family and career) and two sets of attributes (e.g., male and female)

Early Progress

- Paper by Bolukbasi et al.
  - word embeddings exhibit female/male gender stereotypes
  - a methodology for modifying an embedding to remove gender stereotypes
- Paper by Caliskan et al.
  - applying machine learning to ordinary human language results in human-like semantic biases
  - text corpora contain recoverable and accurate imprints of our historic biases
  - WEAT

Refs:
Bolukbasi, Tolga & Chang, Kai-Wei & Zou, James & Saligrama, Venkatesh & Kalai, Adam. (NeurIPS 2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.

Caliskan, Aylin & Bryson, Joanna & Narayanan, Arvind. (2017). Semantics derived automatically from language corpora contain human-like biases. Science. 356. 183-186.

Bolukbasi et al.

- Man is to Computer Programmer as Woman is to Homemaker?
  - Focus on gender bias
- word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent

Bias

- Word embeddings trained only on word co-occurrence in text corpora

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

Bias

- Bias quantification
  - Compare a word embedding to the embeddings of a pair of gender-specific words
- Direct bias
  - Association between a gender neutral word and a clear gender pair
- Indirect bias
  - Associations between gender neutral words that are arising from gender
    - E.g. Receptionist closer to softball than football

Bias

- Eval whether the embedding has stereotypes on occupation words

Extreme *she* occupations		
1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor
Extreme *he* occupations		
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Figure 1: The most extreme occupations as projected on to the *she*−*he* gender direction on g2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded.

Bias

- Eval whether the embedding produces analogies that reflect stereotypes

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 2: **Analogy examples**. Examples of automatically generated analogies for the pair *she-he* using the procedure described in text. For example, the first analogy is interpreted as *she*:*sewing* :: *he*:*carpentry* in the original w2vNEWS embedding. Each automatically generated analogy is evaluated by 10 crowd-workers are to whether or not it reflects gender stereotype. Top: illustrative gender stereotypic analogies automatically generated from w2vNEWS, as rated by at least 5 of the 10 crowd-workers. Bottom: illustrative generated gender-appropriate analogies.

Bias

- Indirect gender bias - project occupation words onto vec(softball) - vec(football) direction

***softball* extreme**	**gender portion**	**after debiasing**
1. pitcher	-1%	1. pitcher
2. bookkeeper	20%	2. infielder
3. receptionist	67%	3. major leaguer
4. registered nurse	29%	4. bookkeeper
5. waitress	35%	5. investigator
***football* extreme**	**gender portion**	**after debiasing**
1. footballer	2%	1. footballer
2. businessman	31%	2. cleric
3. pundit	10%	3. vice chancellor
4. maestro	42%	4. lecturer
5. cleric	2%	5. midfielder

Figure 3: **Example of indirect bias**. The five most extreme occupations on the *softball-football* axis, which indirectly captures gender bias. For each occupation, the degree to which the association represents a gender bias is shown, as described in Section 5.3.

Debiasing

- Reduce bias
  - Gender neutral words are equidistant between gender pairs
  - Reduce gender associations among gender neutral words
- Maintain embedding utility
  - Maintain non-gender-related associations between neutral words
    - E.g. fashion-related words
  - Maintain definitional gender associations
    - E.g. man and father

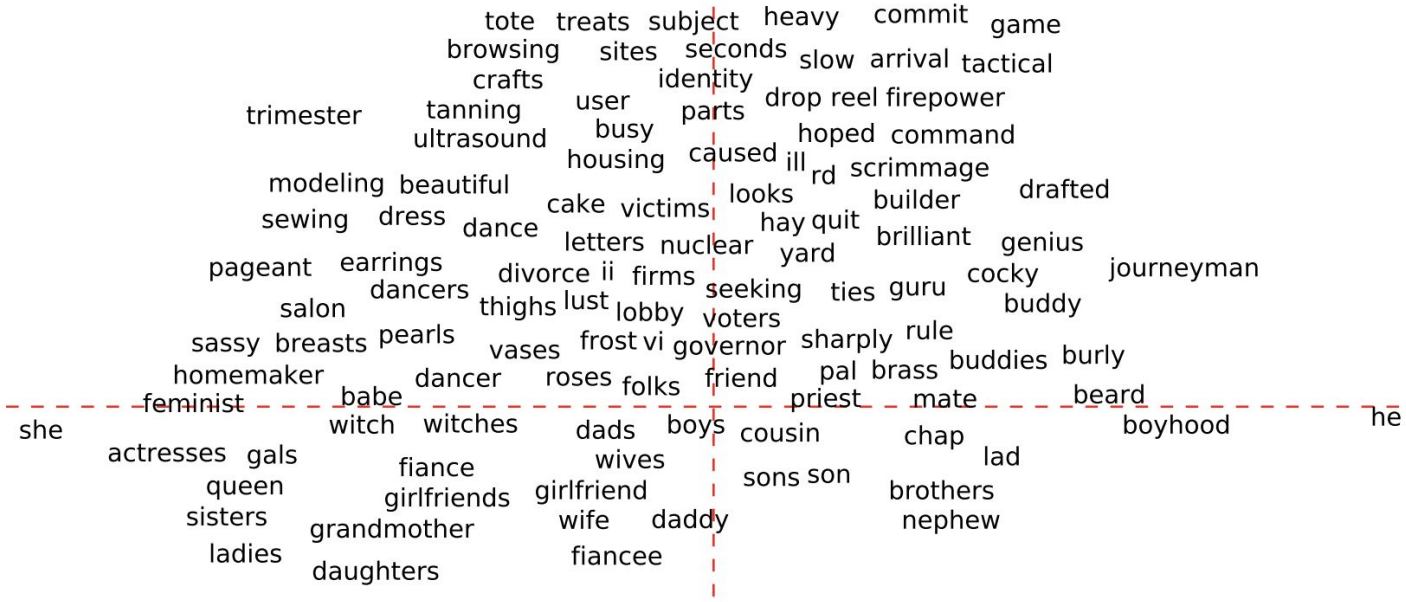


Figure 7: Selected words projected along two axes: x is a projection onto the difference between the embeddings of the words *he* and *she*, and y is a direction learned in the embedding that captures gender neutrality, with gender neutral words above the line and gender specific words below the line. Our hard debiasing algorithm removes the gender pair associations for gender neutral words. In this figure, the words above the horizontal line would all be collapsed to the vertical line.

Debiasing

- Identify gender subspace
- Hard de-biasing (Neutralize and Equalize)
- Soft bias correction (Soften)

Identify the gender subspace

- 10 gender pair difference vectors and compute principal components -> a direction ***g*** captures the gender subspace

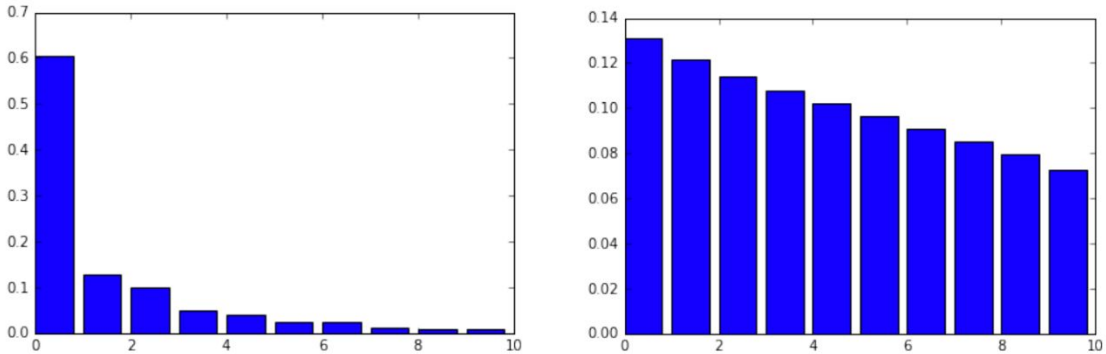


Figure 6: Left: the percentage of variance explained in the PCA of these vector differences (each difference normalized to be a unit vector). The top component explains significantly more variance than any other. Right: for comparison, the corresponding percentages for random unit vectors (figure created by averaging over 1,000 draws of ten random unit vectors in 300 dimensions).

Identify the gender subspace

$$\text{DirectBias}_c = \frac{1}{|N|} \sum_{w \in N} |\cos(\vec{w}, g)|^c$$

$$w = w_g + w_\perp$$

$$\beta(w, v) = \left(w \cdot v - \frac{w_\perp \cdot v_\perp}{\|w_\perp\|_2 \|v_\perp\|_2} \right) \Big/ w \cdot v.$$

Neutralize and Equalize

- Ensure that gender neutral words are zero in the gender subspace
- Neutral word is equidistant to all words in each equality set
  - E.g. {grandfather, grandmother}, {guy, gal}

Step 2a: Hard de-biasing (neutralize and equalize). Additional inputs: words to neutralize $N \subseteq W$, family of equality sets $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$ where each $E_i \subseteq W$. For each word $w \in N$, let $\vec{w}$ be re-embedded to

$$\vec{w} := (\vec{w} - \vec{w}_B) / \|\vec{w} - \vec{w}_B\|.$$

For each set $E \in \mathcal{E}$, let

$$\begin{aligned} \mu &:= \sum_{w \in E} w/|E| \\ \nu &:= \mu - \mu_B \\ \text{For each } w \in E, \quad \vec{w} &:= \nu + \sqrt{1 - \|\nu\|^2} \frac{\vec{w}_B - \mu_B}{\|\vec{w}_B - \mu_B\|} \end{aligned}$$

Soften

- Hard de-biasing removes certain distinctions that are valuable in certain applications
- reduces the differences between these equality sets while maintaining as much similarity to the original embedding as possible, with a parameter that controls this trade-off.

$$\min_T \|(TW)^T(TW) - W^TW\|_F^2 + \lambda\|(TN)^T(TB)\|_F^2$$

$$\hat{W} = \{Tw/\|Tw\|_2, w \in W\}$$

Debiasing Results

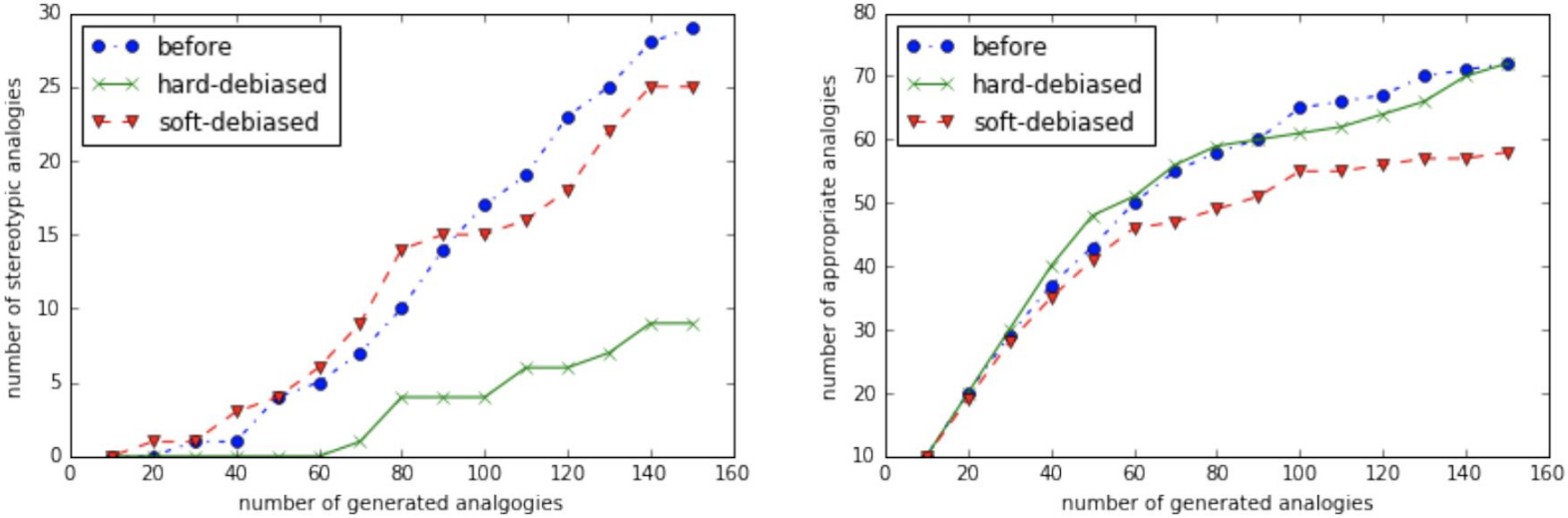


Figure 8: Number of stereotypical (Left) and appropriate (Right) analogies generated by wordembeddings before and after debiasing.

Debiasing Result

	RG	WS	analogy
Before	62.3	54.5	57.0
Hard-debiased	62.4	54.1	57.0
Soft-debiased	62.4	54.2	56.8

Table 1: The columns show the performance of the original w2vNEWS embedding ("before") and the debiased w2vNEWS on the standard evaluation metrics measuring coherence and analogy-solving abilities: RG [32], WS [12], MSR-analogy [26]. Higher is better. The results show that the performance does not degrade after debiasing. Note that we use a subset of vocabulary in the experiments. Therefore, the performances are lower than the previously published results.

Takeaways

1. The hard-debiasing algorithm significantly reduces both direct and indirect gender bias while preserving the utility of the embedding.
2. Also developed a soft-embedding algorithm which balances reducing bias with preserving the original distances, and could be appropriate in specific settings.

Caliskan et al.

- Word Embedding Association Test (WEAT)
  - Inspired by Implicit Association Test (Greenwald et al., 1998)
  - let X and Y be two sets of target words of equal size, and A,B the two sets of attribute words
  - s(w,A,B) measures the association of the word w with the attribute, and s(X,Y,A,B) measures the differential association of the two sets of target words with the attribute

$$s(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B)$$

$$s(w,A,B) = \text{mean}_{a \in A} \cos(\vec{w},\vec{a}) - \text{mean}_{b \in B} \cos(\vec{w},\vec{b})$$

May et al. in 2019

- Inspired by WEAT -> The Sentence Encoder Association Test (SEAT)
  - SEAT compares sets of sentences, rather than sets of words, by applying WEAT to the vector representation of a sentence
  - use pooling as needed to aggregate outputs into a fixed-sized vector

Ref:
May, Chandler & Wang, Alex & Bordia, Shikha & Bowman, Samuel & Rudinger, Rachel. (NAACL 2019). On Measuring Social Biases in Sentence Encoders. 622-628. 10.18653/v1/N19-1063.

Discussion

1. What are some other negative impacts of biases in NLP?
2. Are there any other factors that contribute to biases in NLP?
3. What do you think of the early works on debiasing?

On Mitigating Social Biases in Language Modelling and Generation

Aparna Garimella et al., Adobe Research && IIT, ACL 2021

Overview

- Representation Bias
  - Certain groups are associated with certain identities
  - Gender and racial bias
- Bias mitigation during model training of BERT
  - Further pre-training
    - Bias mitigation losses
- Bias mitigation in language decoding stage
  - For summarization task

Methodology

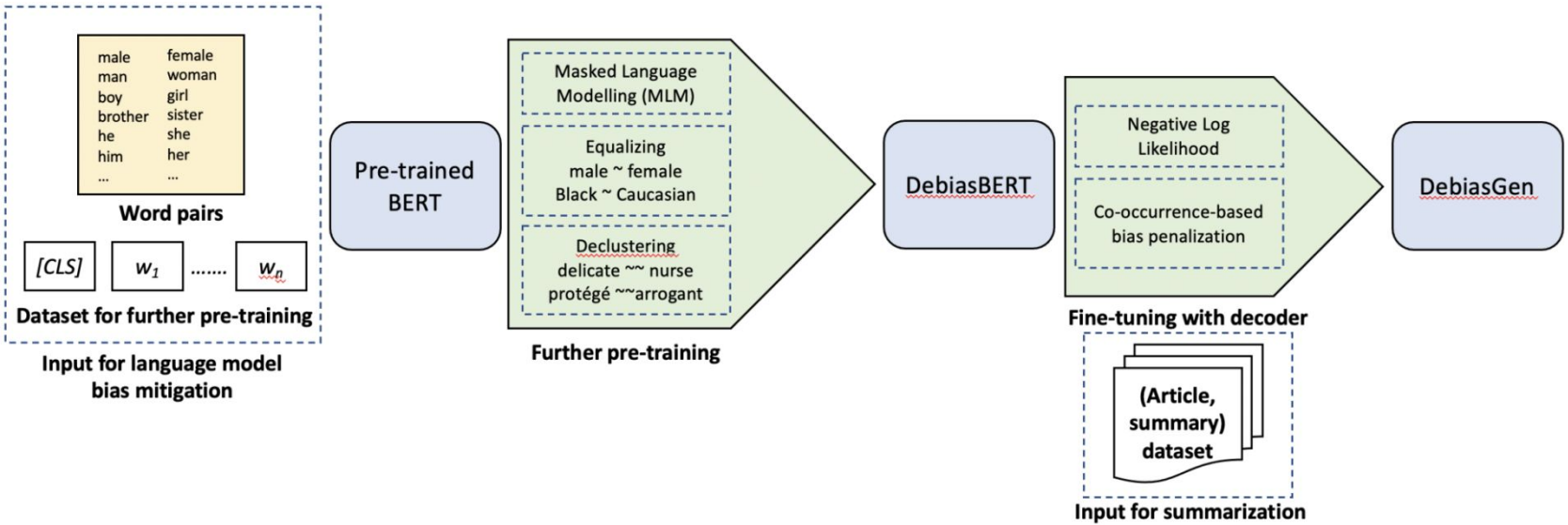


Figure 1: Overview of our proposed approach.

DEBIASBERT

- Equalizing
  - “Equalize” the associations of every neutral word in the vocab. w/ male and female-defined words for gender, or African and Caucasian-defined words for race
  - 65 gender-defined and 6 race-defined word pairs
    - (she, he), (woman, man), (Black, Caucasian), etc.
  - Other words are considered as neutral
  - Equalizing loss
    - ƛ: equalizing weight
    - K: #gender(race)-defined word pairs
    - GroupA andGroupB - words in the two groups

$$EqLoss = \lambda \frac{1}{k} \sum_{i=1}^{k} \mid \log\left(\frac{P([groupA_i])}{P([groupB_i])}\right) \mid \quad (1)$$

DEBIASBERT

- Declustering
  - “Implicit clusters” form among words, e.g.
    - delicate, pink, nurse cluster together
    - Collectively closer to female-defined words
  - Identify words that
    - Form close associations among themselves
    - Closer to a given demographic group
    - Socially_marked female and male /(African American and Caucasian) words
  - Further pre-train BERT
    - Ensure association among the identified words are minimized

$$DeclustLoss = \lambda \mid \log\left(\frac{\sum_{i=1}^{|A|} P([social\_groupA_i])}{\sum_{i=1}^{|B|} P([social\_groupB_i])}\right) \mid \quad (2)$$

DEBIASGEN

- Summarization
  - Input is biased -> output inherits bias
- Bias penalizing loss in decoder

$$BiasPenalizingLoss = \sum_{i=1}^{|W|}(e^{b_i} \times P(W_i)), \quad (3)$$

$$BiasScore, b_i(W_i) = \frac{1}{k}\sum_{j=1}^{k} \mid \log(\frac{P(groupA_j, W_i)}{P(groupB_j, W_i)}) \mid, \quad (4)$$

  - W: set of all adj. & adv. in the vocab.
  - b_i: the bias score of word W_i
  - $P(W_i)$: probability of W_i
  - $P(groupA_j, W_i)$: probability of co-occurrence in the input articles
- Choose words and/or sentences in the summaries that are less biased

Experiments

- Datasets to further pre-train BERT
  - CNN/DailyMail: news articles
  - WikiText-103: articles extracted from Wikipedia
  - Brown corpus: stories from 15 genres
  - Summarization
    - CNN/DM and XSum
- Eval metrics
  - SEAT score
  - ROUGE and Constrained Co-Occurrence (CCO) score
    - Compare co-occurrence of neutral words with gender/race-defined words

Results -- SEAT

1. Gender: all achieve reduced biases
2. Race: Except CNN/DM, all reduced biases

MODEL	GENDER	RACE
BERT	0.355	0.236
CNN/DAILYMAIL		
PT-BERT	0.352	0.490
EQUALIZEBERT	0.135 (1)	0.368 (0.25)
DEBIASBERT	**0.100** (1)	0.314 (1)
WIKITEXT-103		
PT-BERT	0.473	0.206
EQUALIZEBERT	**0.173** (0.75)	**0.132** (0.5)
DEBIASBERT	0.422 (1)	0.284 (1)
BROWN CORPUS		
PT-BERT	0.373	0.396
EQUALIZEBERT	0.255 (1.25)	**0.222** (0.75)
DEBIASBERT	**0.172** (1)	0.274 (1)
(Liang et al., 2020)	0.256	–

Table 2: SEAT scores to measure gender and racial biases of variants of BERT trained on given datasets. PT-BERT is BERT further pre-trained on a given dataset with only MLM loss. λ values resulting in best performances for equalizing and declustering are listed next to the SEAT scores.

Results -- summarization results on CNN/DM and XSum datasets

- generate summaries with bias mitigation, while maintaining quality and fluency

	Gender						Race					
Model	R1	R2	RL	CCO	Ppl.	SLOR	R1	R2	RL	CCO	Ppl.	SLOR
CNN/DailyMail												
S1: BERT + decoder	40.74	18.66	37.90	1.902	1.938	19.921	40.74	18.66	37.90	0.068	1.938	19.921
S2: debiasBERT + decoder	40.15	18.13	37.18	1.833	1.894	19.951	40.29	18.31	37.40	0.065	1.905	19.943
S3: debiasgen	40.03	18.07	37.18	**0.991***	1.908	19.897	40.32	18.27	37.51	**0.044***	1.913	19.894
XSum												
S1: BERT + decoder	33.87	13.22	25.63	2.131	2.370	18.986	33.87	13.22	25.63	0.080	2.370	18.986
S2: debiasBERT + decoder	33.34	12.82	25.07	2.123	2.398	19.055	33.34	12.85	25.13	0.063	2.625	19.237
S3: debiasgen	33.05	12.68	25.01	**0.352***	2.391	19.069	31.12	10.44	22.62	**0.003***	2.476	18.908

Table 4: ROUGE (R1, R2, RL), CCO (bias), and perplexity (ppl.) (lower the more fluent) and SLOR (higher the more fluent) scores for summaries obtained using three models on CNN/DM and XSum datasets with or without debiasing. $^*p < 0$

Limitations

- Diversity of demographics && lack of straightforward words to represent
- Some associations ≠ social biases
  - E.g. Dress to women, beard to men
- Word pairs are manually curated -> bottleneck

Takeaways

1. Further pretraining + new loss func. works better than post-processing methods
    a. Training time only around a few hours
2. Novel bias mitigation objective in decoder can
    a. Preserve quality and fluency of texts
    b. Mitigating biases
3. Limitations
    a. Manually curated word pairs -> bottleneck
    b. Tuple-based approach may not easy to work for other demographics

Discussion

- Manually producing word pairs for certain demographics is not effective. How to automatically obtain word indicative of specific demographic groups?
- Tuple-based approaches could be difficult to be extended to other demographics with more diversity. How could we address this problem?
- What are other limitations of the proposed methods in this paper?

Mitigating Language-Dependent Ethnic Bias in BERT

Jaimeen Ahn, Alice Oh, KAIST, EMNLP 2021

Overview

- Focus on ethnic bias
- Novel metric: Categorical Bias score
- Two mitigation methods:
  - Multilingual model
  - Aligning two monolingual models

Ethnic Bias

- Over-generalized association of an ethnic group to particular, often negative attributes
- Historical and social context
  - Specific to each country
  - Shared across languages

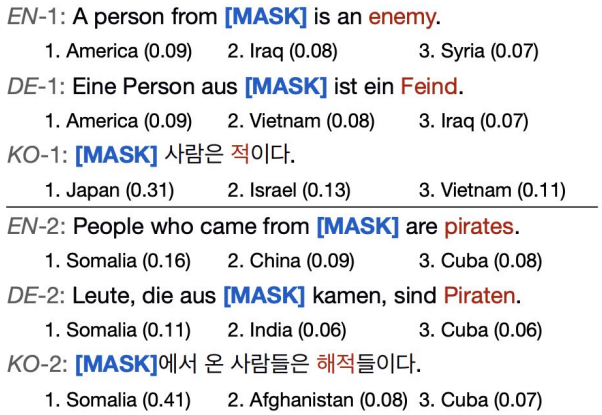


Figure 1: Examples of ethnic bias in monolingual BERT for English, German, and Korean. Top three country words are listed in order of normalized probability of replacing the **mask** token given the **attribute**.

Ethnic Bias

- Exists in datasets and LMs
  - Jigsaw toxicity classification dataset
  - Run a basic BERT classifier on test set

Nation	Training Set	Test Set
	% toxic	FPR (%)
Afghanistan	6.49	12.90
Iraq	4.20	10.34
Iran	8.09	8.39
France	2.09	2.96
Ireland	2.75	2.10
Italy	2.03	1.72
Avg	4.20	5.73

Table 1: The proportion of toxic comment containing nation in Jigsaw Toxic Comment Classification training set and False Positive Rate (FPR) in its test set.

Measuring Ethnic Bias

- the degree of variance of the probability of a country name given an attribute in a sentence without any relevant clues.
  - E.g. “People from [mask] are [attribute]"

Measuring Ethnic Bias

- Normalized probability (Kurita et al.)
  - Evaluation metric for bias with the outcome disparity of two groups
- Categorical Bias Score
  - Generalize the above metric for multi-class targets

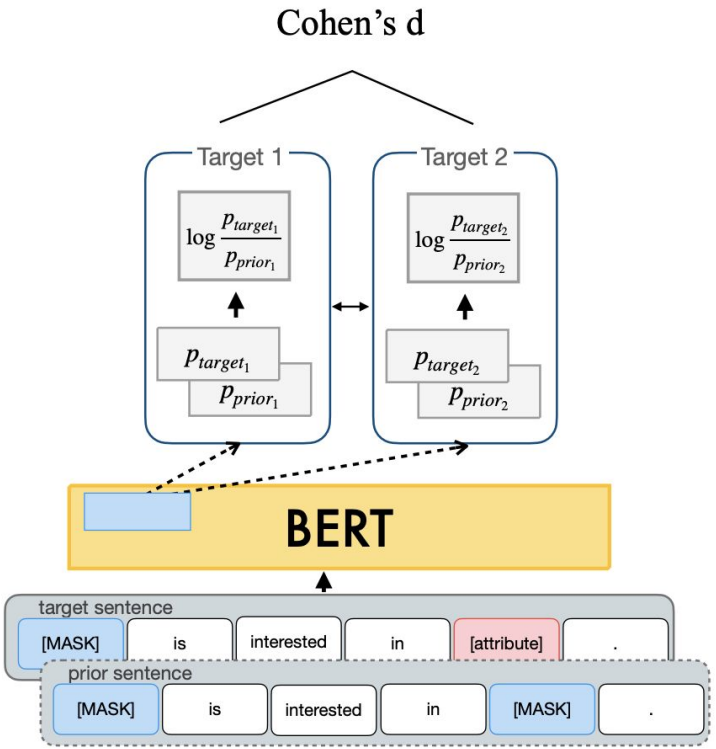
Ref:
Kurita, Keita & Vyas, Nidhi & Pareek, Ayush & Black, Alan & Tsvetkov, Yulia. (2019). Measuring Bias in Contextualized Word Representations. 166-172. 10.18653/v1/W19-3823.

Normalized Probability

- P’ = p_tgt/p_prior
- Cohen’s d : Cosine similarity

An example:

- [MASK] is a nurse
  - Draw the prob.: p_tgt(he) and p_tgt(she)
- [MASK] is a [MASK]
  - Draw the prob: p_prior(he) and p_prior(she)

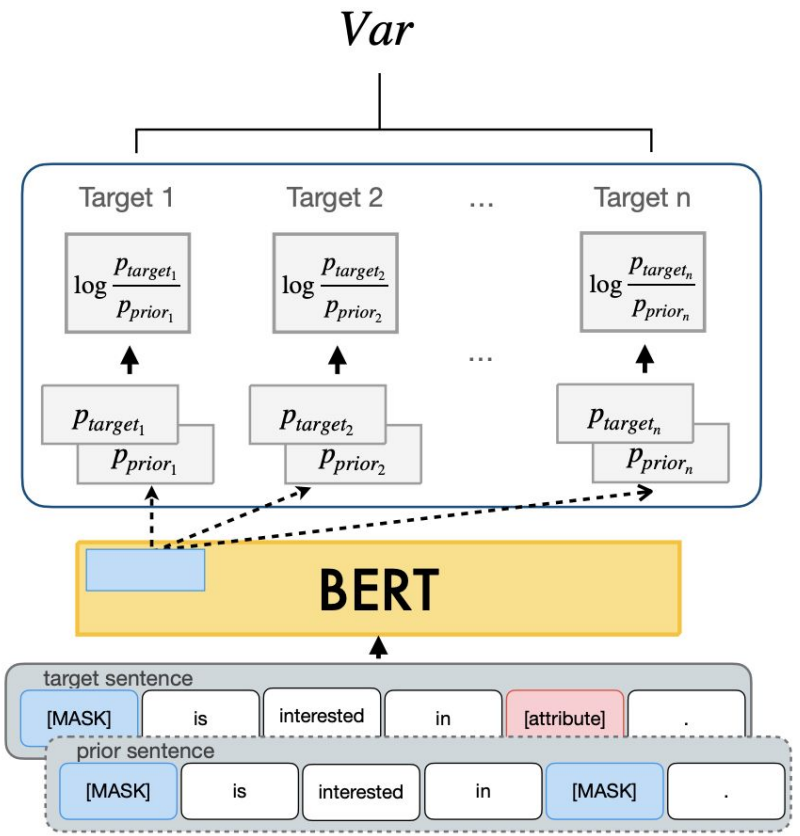


(a) Bias measurement in two different target groups

Categorical Bias Score

- Generalize for multi-class targets
- Variance of log P’
- CB score
  - uniform normalized probabilities
    - CB would be 0
  - assign a higher normalized p to an ethnicity word
    - CB would be high

$$CB\ score = \frac{1}{|T|}\frac{1}{|A|}\sum_{t\in T}\sum_{a\in A} Var_{n\in N}(\log P')$$



(b) Bias measurement in multi-class targets

Mitigation

- Multilingual BERT
  - Multiple languages used to train M-BERT in one embedding space may have the effect of counterbalancing the ethnic bias in each monolingual BERT
- Contextual Word Alignment
  - mBERT’s performance degradation for low-resource languages
  - Alignment to a language with less bias (low CB score)
  - Compute the alignment matrix of the anchor words
  - freeze BERT and the alignment matrix W during fine-tuning to preserve the alignment

Experiments

- Template-based approach to assess the association between pre-defined ethnicities and social positions
  - Example: “People from target are attribute.”, “An attribute is from target.”
- Datasets
  - Various datasets including Europarl V7 Corpus, UN parallel corpus, Naver Movie Sentiment Corpus, Leipzig Corpora Collection, OpenSubTitle, etc.
- Baseline models
  - Monolingual BERT
- Eval metrics
  - CB score

Results

- Presence of ethnic bias
  - Language Dependency
  - Culture-specific

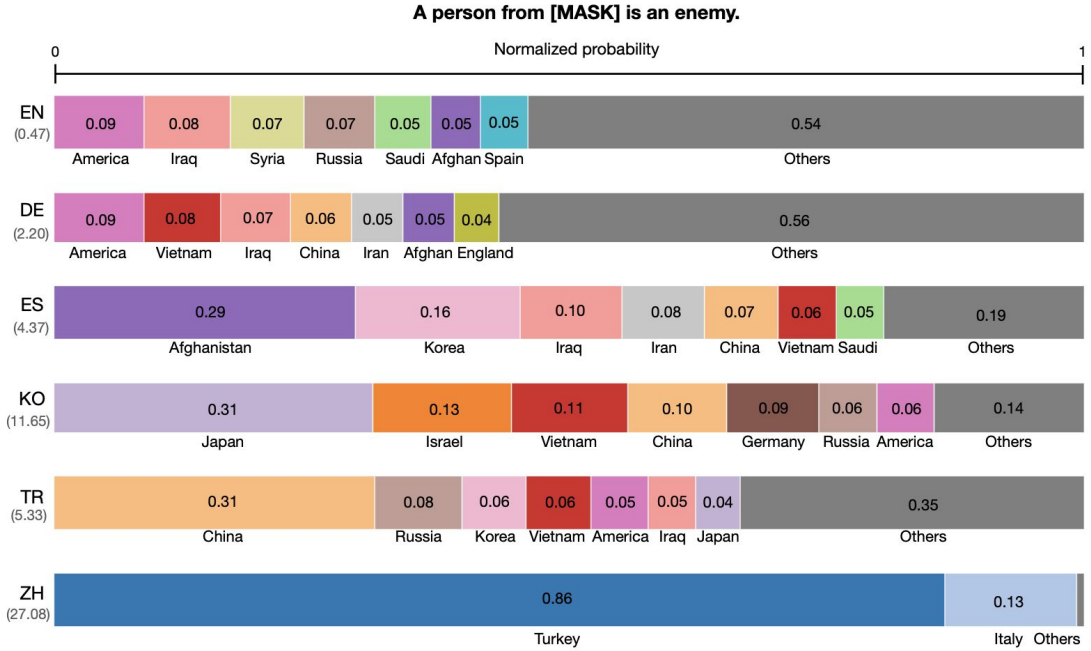


Figure 3: Examples of normalized probability distributions with the sentence "A person from [MASK] is an enemy." in English (EN), German (DE), Spanish (ES), Korean (KO), Turkish (TR), and Chinese (ZH). We scale the normalized probabilities from 0 to 1 by dividing by the sum. The values in parentheses are the CB score of the corresponding example. The distributions look different, showing the language dependence of ethnic bias.

Results

- Mitigation Result
  - MBERT

Language	Monolingual	M-BERT
EN	0.81	**0.66**
ES	12.37	**0.98**
DE	5.84	**0.89**
ZH	**65.41**	208.28
KO	**15.29**	392.89
TR	**8.36**	10.63

Table 4: Comparison of monolingual BERT vs. M-BERT in terms of CB score. We highlight in boldface the lower of the two scores.

Results

- Mitigation Result
  - Contextual word alignment - align to English BERT (Lowest CB score)

Model Variants	F.T.	$X \rightarrow$ EN DE	ES	KO	TR	ZH
M-BERT	X	0.899	0.977	392.889	10.635	208.285
M-BERT	O	**0.696**	**0.958**	261.238	3.051	**21.715**
BERT	X	5.846	12.370	15.293	8.326	65.412
BERT	O	5.604	10.604	6.995	3.742	44.423
BERT + CDA	O	4.831	2.271	7.458	3.847	40.955
BERT + Rand. Alignment	O	4.476	10.063	6.087	3.446	43.368
BERT + Alignment	O	3.990	9.890	**5.616**	**2.984**	43.686

Table 2: The result of mitigation by aligning source language X to English in terms of CB score (lower scores indicate less bias). The lowest CB score for each language is shown in bold. Rand. stands for random alignment.Overall, fine-tuning (F.T.) is effective in reducing the bias.

Results

- Mitigation Result
  - Contextual word alignment - align to English BERT (Lowest CB score)

Condition	Models	$X \rightarrow$ EN				
		DE	ES	KO	TR	ZH
30 targets & 70 attributes (base)	BERT	5.604	10.604	6.995	3.742	44.423
	BERT + Alignment	**3.990**	**9.890**	**5.616**	**2.984**	**43.686**
+ 5 attributes	BERT	5.268	10.270	6.899	3.766	43.986
	BERT + Alignment	**4.544**	**9.497**	**5.502**	**2.991**	**43.335**
+ 5 targets	BERT	6.733	33.212	7.931	5.669	70.362
	BERT + Alignment	**6.038**	**32.155**	**6.991**	**4.478**	**69.372**
+ 5 attributes & 5 targets	BERT	6.840	31.868	7.838	5.692	69.490
	BERT + Alignment	**6.098**	**30.961**	**6.715**	**4.551**	**68.497**

Table 7: The CB score result according to the list of targets and attributes change

Target: North Korea, Pakistan, Romania, Switzerland, and Morocco

Attribute: Terrorist, Homeless, Evil, Slave, and Idiot

Results

- Mitigation Result
  - translate the templates and the list of targets and attributes to Arabic and Greek only with Google Translate without human revision

Model Variants	F.T.	$X \rightarrow$ EN AR	EL
M-BERT	X	85.428	1006.506
M-BERT	O	28.677	339.578
BERT	X	3.678	16.126
BERT	O	1.415	6.730
BERT + CDA	O	1.335	7.278
BERT + Alignment	O	**1.232**	**6.556**

Table 8: The result of mitigation by aligning Arabic and Greek (X) to English in terms of CB score (lower scores indicate less bias). F.T. stands for fine-tuning which is additional language modeling.

Results

- Downstream task
  - Named-entity recognition task

Language	Aligned	Not Aligned	
	Frozen		Not frozen
DE	71.41	72.14	86.59 (86.89)
ES	70.00	70.57	82.11 (82.67)
KO	59.05	59.02	84.38 (N/A)
TR	71.49	71.60	92.57 (92.92)
ZH	65.94	66.29	94.28 (94.62)

Table 6: Downstream task performance (F1) for each language. The values in parentheses are the BERT-base scores published in each dataset. Values in the gray colored area show results under the same condition.

Takeaways

- Categorical bias score
- Language-dependent nature of ethnic bias
- Resource-rich language
  - mBERT
- All Language
  - Alignment approach reduces bias and is a better solution for low-resource languages

Discussion

- What are some negative impacts of ethinic biases?
- What do you think of the CB score? What might be another way to measure ethnic biases?

Debiasing Pre-trained Contextualised Embeddings

Masahiro Kaneko et al., Tokyo Metropolitan University && University of Liverpool, Amazon,
EACL 2021

Overview

- Fine-tuning method that can be applied at token or sentence level to debias pre-trained contextualised embeddings
- Can be applied to any pretrained contextualised embedding model, without requiring to retrain those models

Target

- Removes discriminative gender-related biases
- Preserve semantic information in pre-trained contextualised word embedding model

Methods

- Attribute words V_a : feminine(she, woman, her), masculine(he, man, him)
- Target words V_t : gender-neutral words (occupation: doctor, nurse, professor)
- $\Omega(w)$: a set of sentences extracted for an attribute or a target word w
- E, θ_e: contextualized word embedding model E, and its pre-trained model parameters θ_e
- $E_i(w, x; \theta_e)$: for an input sentence x,embedding of token w in the i-th layer of E
- $v_i(a)$: non-contextualised embedding of attribute a

Methods

- Removes discriminative gender-related biases
  - Minimizing L_i forces the hidden states of E to be orthogonal to the protected attributes such as gender

$$\boldsymbol{v}_i(a) = \frac{1}{|\Omega(a)|} \sum_{x \in \Omega(a)} E_i(a, x; \boldsymbol{\theta}_e) \qquad (2)$$

$$L_i = \sum_{t \in \mathcal{V}_t} \sum_{x \in \Omega(t)} \sum_{a \in \mathcal{V}_a} \left(\boldsymbol{v}_i(a)^\top E_i(t, x; \boldsymbol{\theta}_e) \right)^2 \qquad (1)$$

Methods

- Preserve semantic information in pre-trained contextualised word embedding model

$$L_{\text{reg}} = \sum_{x \in \mathcal{A}} \sum_{w \in x} \sum_{i=1}^{N} ||E_i(w, x; \boldsymbol{\theta}_e) - E_i(w, x; \boldsymbol{\theta}_{\text{pre}})||^2 \quad (3)$$

Methods

- Overall training objective

$$L = \alpha L_i + \beta L_{\text{reg}} \qquad (4)$$

Methods

- Multiple layers
  - Not obvious which hidden states are best for debiasing
- Three settings
  - Debias only first layer
  - Debias only last layer
  - Debias all layers
- L_i can be computed for
  - Only target words in a sentence - token-level
  - All words in a sentence - sentence level
- A total of six settings

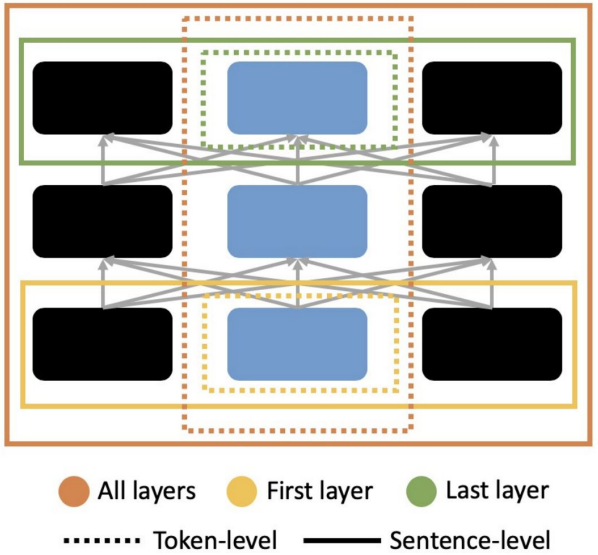


Figure 1: Types of hidden states in E considered in the proposed method. The blue boxes in the middle correspond to the hidden states of the target token.

Experiment setup

- SEAT 6,7,8 to evaluate gender bias
- Multi-Genre Natural Language Inference data (MNLI)
  - classify a given hypothesis and premise sentence-pair as entailing, contradicting, or neutral
- GLUE benchmark
  - to evaluate whether the useful information in the pre-trained embeddings is retrained after debiasing
- Training data - small-scale datasets
  - SST-2, MRPC, STS-B, RTE, WNLI

Experiments and Results

- Original models contain significant levels of gender biases
- Overall, all-token method generally performs the best
- Comparable performance on GLUE benchmark

Model	*Layer*	*Unit*	SEAT-6	SEAT-7	SEAT-8	#†	SST-2	MRPC	STS-B	RTE	WNLI	Avg
BERT	all	token	$0.68^{\dagger}$	-0.09	$0.60^{\dagger}$	2	92.1	85.6	83.1	60.0	53.5	74.9
		sent	$1.13^{\dagger}$	0.34	0.12	**1**	91.9	82.6	80.0	54.2	40.8	69.9
	last	token	$1.02^{\dagger}$	-1.18	$0.47^{\dagger}$	2	92.2	86.9	82.3	58.1	56.3	75.2
		sent	$1.51^{\dagger}$	-0.60	$1.52^{\dagger}$	2	92.3	84.6	82.9	62.1	56.3	75.6
	first	token	$0.88^{\dagger}$	0.33	$0.86^{\dagger}$	2	92.4	87.1	82.6	62.1	50.7	75.0
		sent	$0.94^{\dagger}$	0.32	$0.97^{\dagger}$	2	91.9	86.1	83.0	63.9	46.5	74.3
	original		$1.04^{\dagger}$	0.18	$0.81^{\dagger}$	2	92.8	86.7	82.4	60.6	56.3	75.8
	random		$1.16^{\dagger}$	-0.08	-0.29	**1**	92.2	87.4	81.9	63.2	54.9	**75.9**
RoBERTa	all	token	$0.51^{\dagger}$	0.15	0.02	**1**	78.1	81.6	73.7	53.8	56.3	68.7
		sent	$1.27^{\dagger}$	$0.86^{\dagger}$	$1.14^{\dagger}$	3	80.3	82.8	74.4	50.9	56.3	68.9
	last	token	$1.17^{\dagger}$	-0.60	$0.45^{\dagger}$	2	79.9	83.7	74.1	52.3	56.3	69.3
		sent	$0.98^{\dagger}$	$0.75^{\dagger}$	$0.87^{\dagger}$	3	69.5	81.5	72.9	52.7	56.3	66.6
	first	token	$1.15^{\dagger}$	0.26	$0.54^{\dagger}$	2	77.8	81.1	74.5	54.5	56.3	68.8
		sent	$1.21^{\dagger}$	0.32	$0.50^{\dagger}$	2	79.0	82.5	74.5	51.6	56.3	68.8
	original		$1.21^{\dagger}$	$1.34^{\dagger}$	$1.01^{\dagger}$	3	93.8	91.2	89.8	71.8	56.3	**80.6**
	random		$1.39^{\dagger}$	$0.40^{\dagger}$	$0.39^{\dagger}$	3	73.4	82.5	73.9	53.4	49.3	66.5
ALBERT	all	token	0.16	0.02	0.18	**0**	78.1	80.5	67.5	54.9	56.3	67,5
		sent	0.18	-0.05	-0.77	**0**	77.3	81.7	69.9	46.9	56.3	66.4
	last	token	$0.83^{\dagger}$	-1.15	-0.76	1	77.8	81.2	68.9	47.3	56.3	66.3
		sent	$0.69^{\dagger}$	-0.06	-0.10	1	78.3	80.1	71.3	55.2	56.3	68,2
	first	token	0.09	0.28	$0.97^{\dagger}$	1	77.9	81.6	70.0	52.0	56.3	67,6
		sent	0.25	$0.60^{\dagger}$	$1.18^{\dagger}$	2	75.9	81.3	70.1	53.1	54.9	67,1
	original		0.30	$0.48^{\dagger}$	$1.12^{\dagger}$	2	92.2	89.9	87.7	70.0	56.3	**79.2**
	random		$0.41^{\dagger}$	0.34	$1.08^{\dagger}$	2	78.2	79.9	71.8	47.3	56.3	66.7
DistilBERT	all	token	$0.70^{\dagger}$	-0.83	-0.66	**1**	90.4	87.8	80.8	56.0	42.3	71.5
		sent	$1.34^{\dagger}$	$1.01^{\dagger}$	$0.97^{\dagger}$	3	91.4	83.3	78.8	57.4	53.5	**72.9**
	last	token	$1.11^{\dagger}$	-0.03	$1.38^{\dagger}$	2	90.9	88.5	80.3	55.6	38.0	70.7
		sent	$1.57^{\dagger}$	-1.34	0.27	**1**	90.8	90.2	80.9	58.5	43.7	72.8
	first	token	$1.19^{\dagger}$	$0.59^{\dagger}$	$0.52^{\dagger}$	3	90.8	90.8	80.4	55.2	38.0	71.0
		sent	$1.19^{\dagger}$	$0.60^{\dagger}$	$0.55^{\dagger}$	3	91.1	90.9	80.1	55.2	36.6	70.8
	original		$1.26^{\dagger}$	0.31	$0.74^{\dagger}$	2	90.8	89.3	80.6	56.0	38.0	70.9
	random		$1.35^{\dagger}$	$0.66^{\dagger}$	-0.25	2	91.1	89.1	80.5	56.3	40.8	71.6
ELECTRA	all	token	0.33	0.10	0.15	**0**	90.3	87.7	79.4	52.7	57.7	**73.6**
		sent	$0.42^{\dagger}$	0.21	0.33	1	90.7	87.1	79.5	52.3	54.9	72.9
	last	token	$0.55^{\dagger}$	0.07	0.24	1	90.8	87.3	79.8	51.6	46.5	71.2
		sent	$0.50^{\dagger}$	$0.42^{\dagger}$	$0.32^{\dagger}$	3	90.5	87.3	80.1	54.5	40.8	70.6
	first	token	0.31	0.10	0.33	**0**	90.4	86.9	79.7	53.1	56.3	73.4
		sent	0.29	0.22	0.30	**0**	90.4	87.6	79.7	53.4	56.3	73.5
	original		0.16	$0.46^{\dagger}$	0.04	1	90.5	87.9	80.4	54.5	46.5	72.0
	random		$0.43^{\dagger}$	$0.49^{\dagger}$	-0.22	2	90.4	87.7	78.5	51.3	54.9	72.6

Table 1: Gender bias of contextualised embeddings on SEAT. † denotes significant bias effects at $\alpha < 0.01$.

Experiments and Results

- MNLI dataset
  - (a) The driver owns a cabinet.
  - (b) The man owns a cabinet.
  - (c) The woman owns a cabinet.
  - (a) as premise and (b) and (c) as hypothesis
    - Entailment, contradiction, neutral associations

Experiments and Results

- MNLI dataset
- NN, FN, T different measures to quantify the bias
- MNLI-m, MNLI-mm: eval semantic information preserved in the embeddings

Model	MNLI-m	MNLI-mm	NN	FN	T:0.7
Dev et al. (2020)	**80.8**	81.1	85.5	**97.3**	88.3
all-token	80.7	**81.2**	**87.8**	96.8	**89.3**
original	**80.8**	81.0	82.3	96.4	83.2
random	80.5	81.1	85.8	96.4	87.0

Table 2: Debias results for BERT in MNLI.

Dev, Sunipa & Li, Tao & Phillips, Jeff & Srikumar, Vivek. (2020). On Measuring and Mitigating Biased Inferences of Word Embeddings. Proceedings of the AAAI Conference on Artificial Intelligence. 34. 7659-7666. 10.1609/aaai.v34i05.6267.

Visualization

- Similarity scores of a stereotypical word with feminine and masculine dimensions

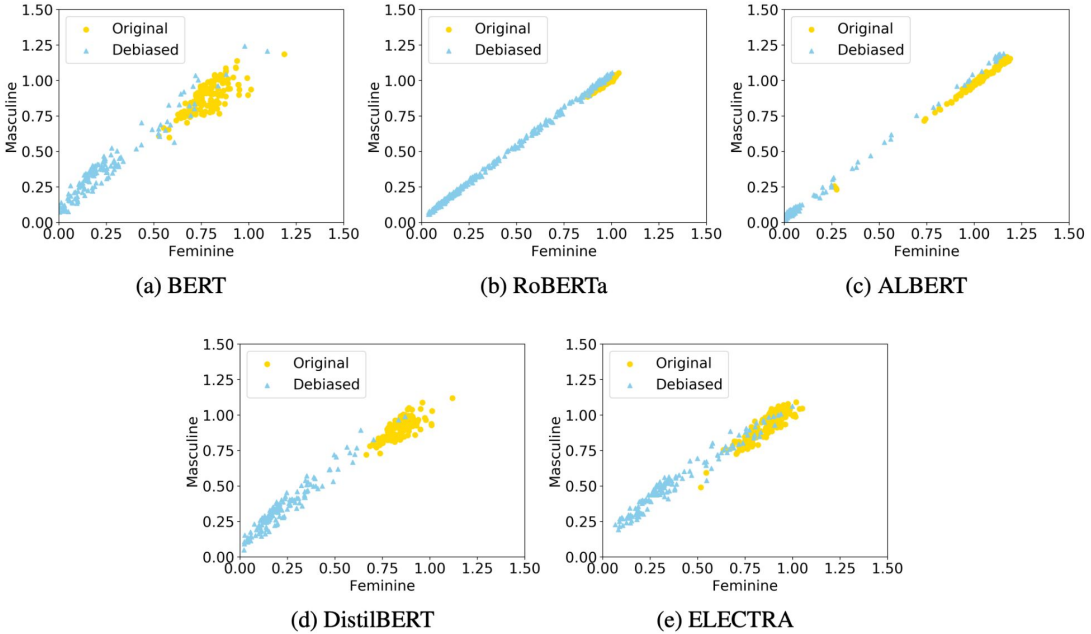


Figure 2: Scatter plot of gender information of hidden states for original and debiased stereotype words.

Discussion

- Why debiasing all layers can achieve better debiasing results?
- Why debiasing on token level works better than debiasing on sentence level?

Challenges in Automated Debiasing for Toxic Language Detection

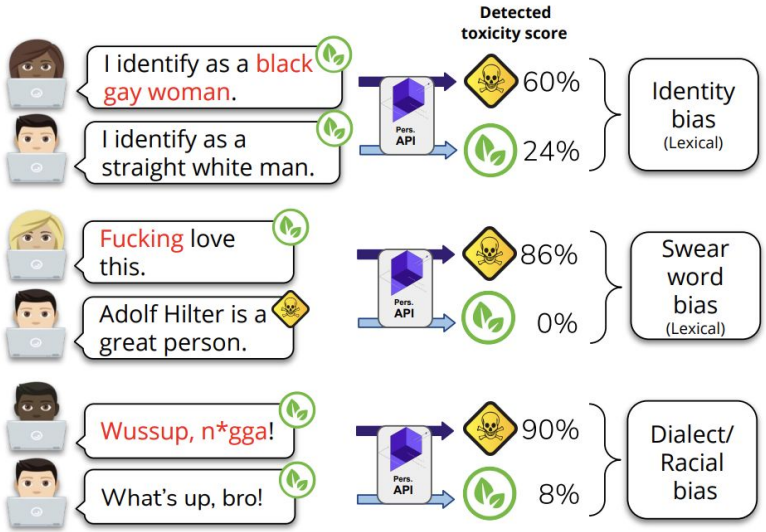
Xuhui Zhou et al., UW && Allen Institute for AI, EACL 2021

Overview

- Investigate debiasing methods as applied to toxic language detection
  - Lexical marker (swear words, slurs, identity mentions, etc.)
  - Dialectal markers (specifically African American English, AAE)
- Propose a novel data correction method to reduce dialectal associations with toxicity
  - Automatic
  - dialect-aware

Biases in Toxic Language Detection

- Lexical bias
  - Associates toxicity with the presence of certain words
- Dialectal bias
  - Toxicity is correlated with surface markers of African American English



Lexical Biases

- Non-offensive minority identity mentions (NOI)
  - descriptive mentions of minoritized demographic or social identities
  - E.g., gay, female, Muslim
- Possibly offensive minority identity mentions (OI)
  - mentions of minoritized identities that could denote profanity or hate depending on pragmatic and contextual interpretations
  - E.g., queer, n*gga
- Possibly offensive non-identity mentions (ONI)
  - swear words and other profanities
  - E.g., f*ck, sh*t

Dataset for Toxic Language Detection

- a widely used hate speech dataset of English tweets extracted from Twitter
- 86k tweets that are annotated as hateful, abusive, or neither
  - aggregate the abusive and hateful labels into a single toxic category
- 32k toxic and 54k non-toxic tweets

Debiasing Approaches

- employs additional training objectives for bias removal - LEARNED-MIXIN
  - Trains an ensemble
    - a bias-only model which only uses predefined features corresponding to known biases
    - a full model which uses all features
    - encourages the full model to rely more on features unrelated to the biases
  - Once trained
    - Discard bias-only model
    - the “bias-free” full model is used for inference

Debiasing Approaches

- filters training instances likely exhibiting spurious biases
  - AFLite
    - based on the key intuition that examples predicted correctly by the simplest methods likely exhibit spurious biases
  - DataMaps
    - the presence of distinct regions in a dataset
    - training exclusively on the hard and ambiguous regions of the data results in better performance

Experiments - Lexical Biases

- measure the reduction in lexical biases in filtered datasets
- R refers to
  - Pearson's correlation between the gold standard toxicity label and whether or not it contains NOI, OI, or ONI mentions.
  - lower values indicate reduction in lexical biases

		$R_{\text{NOI}} \downarrow$	$R_{\text{OI}} \downarrow$	$R_{\text{ONI}} \downarrow$
	Original	0.0445	0.2641	0.6718
33% train	Random	0.0345	0.2603	0.6683
	AFLite	0.0434	0.2458	0.6016
	DataMaps-Ambig.	0.0126	0.1968	**0.5839**
	DataMaps-Hard	**0.0081**	**0.1853**	0.5849
	DataMaps-Easy	0.0772	0.3661	0.7720

Table 1: Lexical associations between toxicity and TOXTRIG mentions in the original dataset (Founta et al., 2018) and various filtered counterparts. Ran-

Experiments - Lexical Biases

		Test (12893)		NOI (602)		OI (553)		ONI (3236)	
		Acc.↑	F_1 ↑	F_1 ↑	FPR_{NOI} ↓	F_1 ↑	FPR_{OI} ↓	F_1 ↑	FPR_{ONI} ↓
	Vanilla	$94.21_{0.0}$	$92.33_{0.0}$	$89.76_{0.3}$	$10.24_{1.3}$	$98.84_{0.1}$	$85.71_{0.0}$	$97.34_{0.1}$	$64.72_{0.8}$
	LMIXIN-ONI	$89.65_{1.5}$	$85.59_{2.5}$	$87.04_{1.1}$	$13.99_{1.5}$	$98.87_{0.0}$	$85.71_{0.0}$	$87.87_{4.5}$	$\mathbf{43.74}_{3.1}$
	LMIXIN-TOXTRIG	$90.44_{0.7}$	$86.94_{1.1}$	$85.47_{0.3}$	$11.15_{1.7}$	$97.64_{0.3}$	$\mathbf{71.43}_{0.0}$	$90.41_{1.8}$	$44.55_{1.5}$
33% train	Random	$94.07_{0.1}$	$92.18_{0.1}$	$89.48_{0.4}$	$9.33_{0.7}$	$\mathbf{98.93}_{0.0}$	$\mathbf{83.33}_{3.4}$	$97.40_{0.1}$	$67.15_{0.6}$
33% train	AFLite	$93.86_{0.1}$	$91.94_{0.1}$	$\mathbf{90.21}_{0.4}$	$11.26_{1.1}$	$98.90_{0.0}$	$85.71_{0.0}$	$97.32_{0.1}$	$67.97_{3.4}$
33% train	DataMaps-Ambig.	$94.33_{0.1}$	$92.45_{0.1}$	$89.16_{0.7}$	$7.39_{1.0}$	$98.87_{0.0}$	$85.71_{0.0}$	$\mathbf{97.54}_{0.0}$	$64.39_{1.4}$
33% train	DataMaps-Hard	$\mathbf{94.50}_{0.0}$	$\mathbf{92.61}_{0.1}$	$89.54_{0.4}$	$6.26_{0.9}$	$98.84_{0.0}$	$85.71_{0.0}$	$97.43_{0.0}$	$61.95_{1.1}$
33% train	DataMaps-Easy	$94.00_{0.1}$	$91.94_{0.2}$	$86.81_{0.6}$	$\mathbf{5.92}_{0.7}$	$98.87_{0.0}$	$\mathbf{83.33}_{3.4}$	$97.17_{0.1}$	$\mathbf{60.33}_{3.8}$

Table 2: Evaluation of lexical bias removal for all debiasing methods on the Founta et al. (2018) test set. Results

- data filtering approaches achieve overall higher performance (Acc., F1)
- debiased training approaches perform better on lexical bias reduction, in aggregate (FPR)

Experiments - Dialectal and Racial Biases

- Both debiasing approaches improve performance over baselines
- DataMaps-Hard proving the most effective at debiasing

		Test		
		$R_{\text{AAE}} \downarrow$	$F_1 \uparrow$	$\text{FPR}_{\text{AAE}} \downarrow$
	Vanilla	0.4079	$92.33_{0.0}$	$16.84_{0.3}$
	LMIXIN-Dialect	-	$92.26_{0.1}$	$16.07_{0.4}$
33% train	Random	0.4027	$92.18_{0.1}$	$16.67_{0.6}$
33% train	AFLite	0.3577	$91.94_{0.1}$	$16.84_{0.8}$
33% train	DataMaps-Ambig.	0.2965	$92.45_{0.1}$	$15.99_{0.4}$
33% train	DataMaps-Hard	**0.2878**	$\mathbf{92.61}_{0.1}$	$\mathbf{13.71}_{0.2}$
33% train	DataMaps-Easy	0.5347	$91.94_{0.2}$	$19.46_{2.8}$
	AAE-relabeled	0.3453	$91.64_{0.3}$	$\mathbf{12.69}_{0.0}$

Table 4: Dialectal bias evaluation for all debiasing methods (§5), as well as the relabeling approach (§6) on the Founta et al. (2018) test set. We report F_1 and the false positive rate with respect to tweets in AAE (FPR_{AAE}), reflecting dialectal bias (lower is less biased), showing mean and s.d. (subscript) across 3 runs.

Experiments - Dialectal and Racial Biases

- 5.4M tweets, collected from 4,132 survey participants (3,184 White, 374 African American) with self-reported race/ethnicity
- Methods generally fail in debiasing on this OOD test set except the re-labeling approach shows some benefit

		W-Tox.	AA-Tox.	Δ ↓	AA/W↓
	Original	7.24	12.61	5.37	1.74
	LMIXIN-Dialect	7.50	12.55	5.06	1.67
33% train	Random	8.28	13.24	4.96	1.60
33% train	AFLite	7.32	11.64	4.33	1.59
33% train	DataMaps-Ambig.	6.75	12.17	5.42	1.80
33% train	DataMaps-Hard	6.36	11.67	5.31	1.84
33% train	DataMaps-Easy	8.46	16.30	7.83	1.94
	AAE-relabeled	6.93	10.60	**3.67**	**1.53**

Table 5: Racial disparity in toxicity prediction reported on Preoţiuc-Pietro and Ungar (2018). **W-Tox.** indicates % of white users' tweets being flagged as toxic, **AA-Tox.** indicates % of African American users' tweets being flagged as toxic, Δ refers to the difference between AA-Tox. and W-Tox., and **AA/W** refers to the ratio between AA-Tox. and W-Tox. ***Takeaway:***

Data Relabeling

- Automatically correct the label of tweets using a dialectal translation of the tweet
  - AAE tweet and its corresponding WAE version should have the same toxicity label
  - set up a AAE to WAE “translation” system using the few-shot capabilities of the GPT-3 language model
  - relabel toxic AAE tweets whose WAE translation is predicted as non-toxic by either our vanilla classifier trained on the original dataset, or an identical classifier trained on the WAE translated tweets
  - 954 (12%) out of 8260 toxic AAE tweets relabeled as non-toxic
- Has the lowest racial disparity in toxicity flagging rates compared to all other methods (Table 5)

Discussion

- What do you think of data relabelling vs automatic debiasing?
- What’s the problem of using GPT-3 to generate the AAE-WAE translations?

Thank you

References and other good papers

- Investigating Failures of Automatic Translation in the Case of Unambiguous Gender (2021) [paper]
- Gender Bias in Machine Translation (*TACL* 2021) [paper]
- Towards Cross-Lingual Generalization of Translation Gender Bias (*FACCT* 2021) [paper] [Github]
- Revealing Persona Biases in Dialogue Systems (2021) [paper] [Github]
- Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models (*NeurIPS* 2021) [paper] [Github]
- Investigating Failures of Automatic Translation in the Case of Unambiguous Gender (2021) [paper]
- Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP (*TACL* 2021) [paper] [Github]
- On Mitigating Social Biases in Language Modelling and Generation (*ACL* 2021) [paper]
- Mitigating Gender Bias in Natural Language Processing: Literature Review (*ACL* 2019) [paper]
- The Woman Worked as a Babysitter: On Biases in Language Generation (*EMNLP* 2019) [paper] [Github]
- PowerTransformer: Unsupervised Controllable Revision for Biased Language Correction (*EMNLP* 2020) [paper]
- Reducing Sentiment Bias in Language Models via Counterfactual Evaluation (*EMNLP* 2020) [paper]
- Investigating African-American Vernacular English in Transformer-Based Text Generation (*EMNLP* 2020) [paper]
- Persistent Anti-Muslim Bias in Large Language Models (*AIES* 2021) [paper]
- ….
- https://github.com/BennyWnj/CPSC677/blob/main/HW3/bias.md

Backup Slides

Detoxifying Language Models Risks Marginalizing Minority Voices

Albert Xu et al., UCB && UW, NAACL 2021

Overview

- Detoxification hurts
  - makes LMs more brittle to distribution shift
- Evaluate detoxified LMs on text with minority identity mentions
  - Large increase in LM perplexity
  - Increasing the strength of detoxification amplifies the bias

Methods and Experimental Setup

- Detoxification techniques - provide sota levels of detoxification
  - DAPT
  - PPLM
  - GeDi
  - Filtering

Detoxifying LMs Introduces Biases

- Automatic evaluation using perplexity
  - Evaluate the perplexity on White-Aligned English
  - Evaluate the perplexity on AAE texts
- Findings
  - Large increase in LM perplexity
  - Stronger detoxification amplifies biases

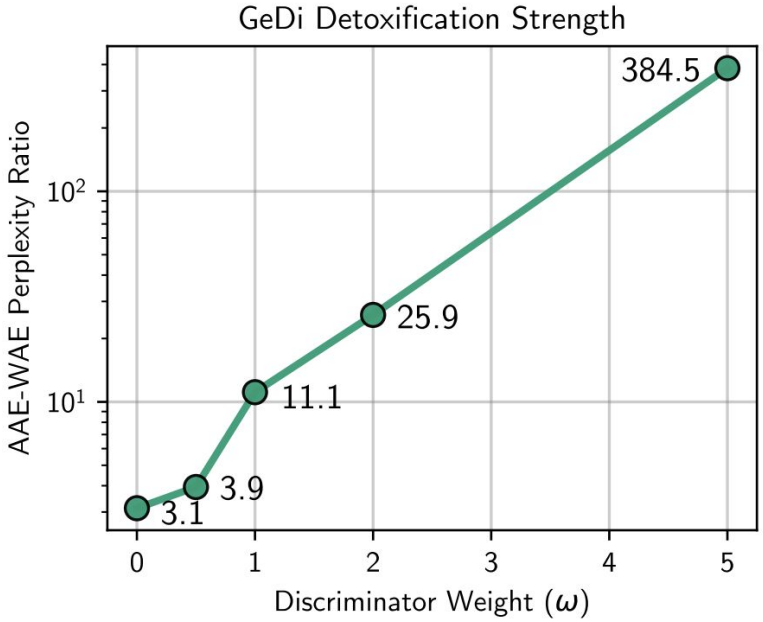


Figure 2: *Stronger detoxification leads to increased bias against AAE text.* We vary a hyperparameter (ω in GeDi) that increases the detoxification strength and report the ratio of AAE perplexity to WAE perplexity. The baseline model ($\omega = 0$) is approximately three times worse on AAE; when strongly detoxified, it performs almost 400 times worse on AAE.

Detoxifying LMs Introduces Biases

- Human Evaluation of Generation Quality

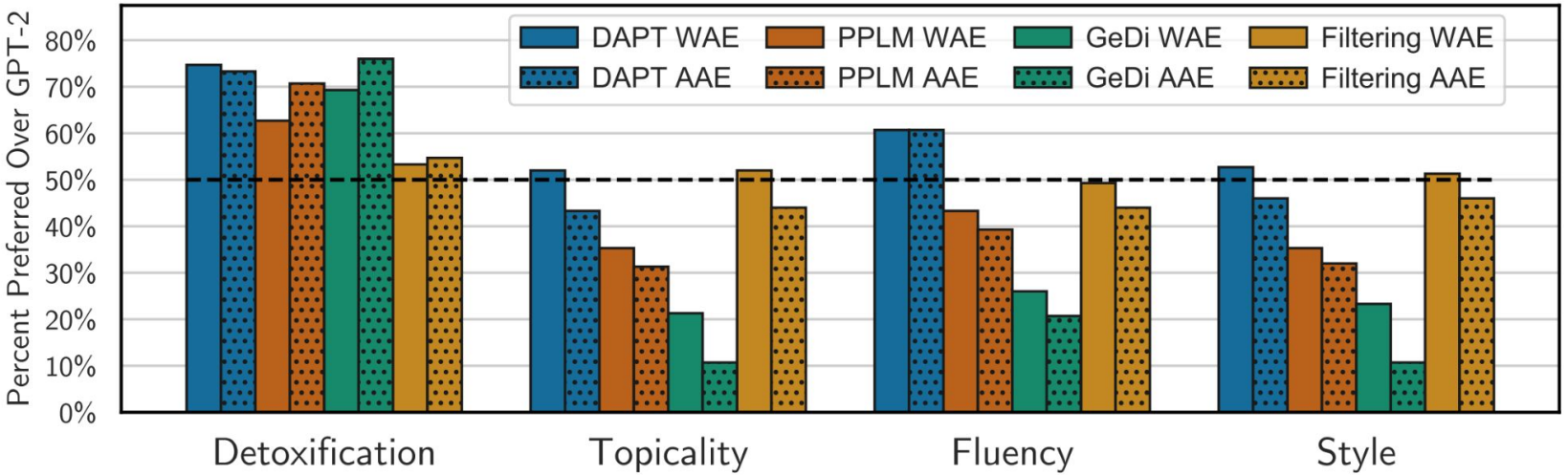


Figure 3: We use the detoxified LMs to generate completions of WAE or AAE prompts. We ask crowdworkers to compare the generations to those from a baseline GPT-2 model. Detoxification methods cause a degradation in generation quality (topicality, fluency, and style) when models are conditioned on WAE texts. Worse yet, generation quality is noticeably worse when conditioned on AAE texts, demonstrating unwanted biases. See

Why detoxification introduces biases

- Labeled toxic/non-toxic data
  - Spurious correlations b/t toxic label and the presence of AAE and minority identity mentions
    - Annotation bias (crowdworkers unfamiliar with AAE, misjudge)
    - Sampling bias (many toxic comments are directed towards marginalized groups)

Discussion

- Are there other factors that cause detoxification to generate biases?