Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings

VECTOR INSTITUTE

UNIVERSITY OF TORONTO

Haoran Zhang*[1,2], Amy X. Lu*[1,2], Mohamed Abdalla[1,2], Matthew McDermott[3], Marzyeh Ghassemi[1,2]

*Equal Contribution [1]University of Toronto [2]Vector Institute [3]MIT

Published in ACM CHIL 2020

Contributions

Contextual word embeddings can perpetuate statistically significant biases when applied to clinical notes in downstream tasks.

- BERT pretrained on clinical notes demonstrates statistically significant gender differences in unsupervised sentence completion tasks.
- BERT pretrained on clinical notes results in statistically significant performance gaps when applied to downstream clinical tasks.
- These biases often favor the majority group with regards to gender, language, ethnicity, and insurance status.
- Our paper is available at https://arxiv.org/abs/2003.11515

Motivation

- Non-contextual word embeddings such as word2vec have been shown to capture societal biases in the training corpus (e.g. gender, ethnicity).
- Contextual word embeddings such as BERT have been shown to contain gender bias on unsupervised tasks in the general domain.
- In a high-stake domain such as clinical notes, do BERT embeddings exhibit bias when qualitatively and quantitatively examined?

Prompt: **[**RACE**] pt became belligerent and violent . sent to [**TOKEN**] [**TOKEN**]**

SciBERT:
caucasian pt became belligerent and violent . sent to **hospital .**
white pt became belligerent and violent . sent to **hospital .**
african pt became belligerent and violent . sent to **prison .**
african american pt became belligerent and violent . sent to **prison .**
black pt became belligerent and violent . sent to **prison .**

Group Fairness Definitions

- Demographic parity:
  - Definition: $P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y} | Z = z)$
  - Metric: $|(\frac{TP_z + FP_z}{N_z})_{z=1} - (\frac{TP_z + FP_z}{N_z})_{z=0}|$
- Positive Equality:
  - Definition: $P(\hat{Y} = 1 | Y = 1) = P(\hat{Y} = 1 | Y = 1, Z = z)$
  - Metric: $|(\frac{TP_z}{TP_z + FN_z})_{z=1} - (\frac{TP_z}{TP_z + FN_z})_{z=0}|$
- Negative Equality:
  - Definition: $P(\hat{Y} = 0 | Y = 0) = P(\hat{Y} = 0 | Y = 0, Z = z)$
  - Metric: $|(\frac{TN_z}{TN_z + FP_z})_{z=1} - (\frac{TN_z}{TN_z + FP_z})_{z=0}|$
- Multi-group Fairness Expansion:
  - $i_j^* = \text{argmax}_{i \in z} |m_j - m_i|$
  - $gap_j = m_j - m_i$

Relevant Prior Work

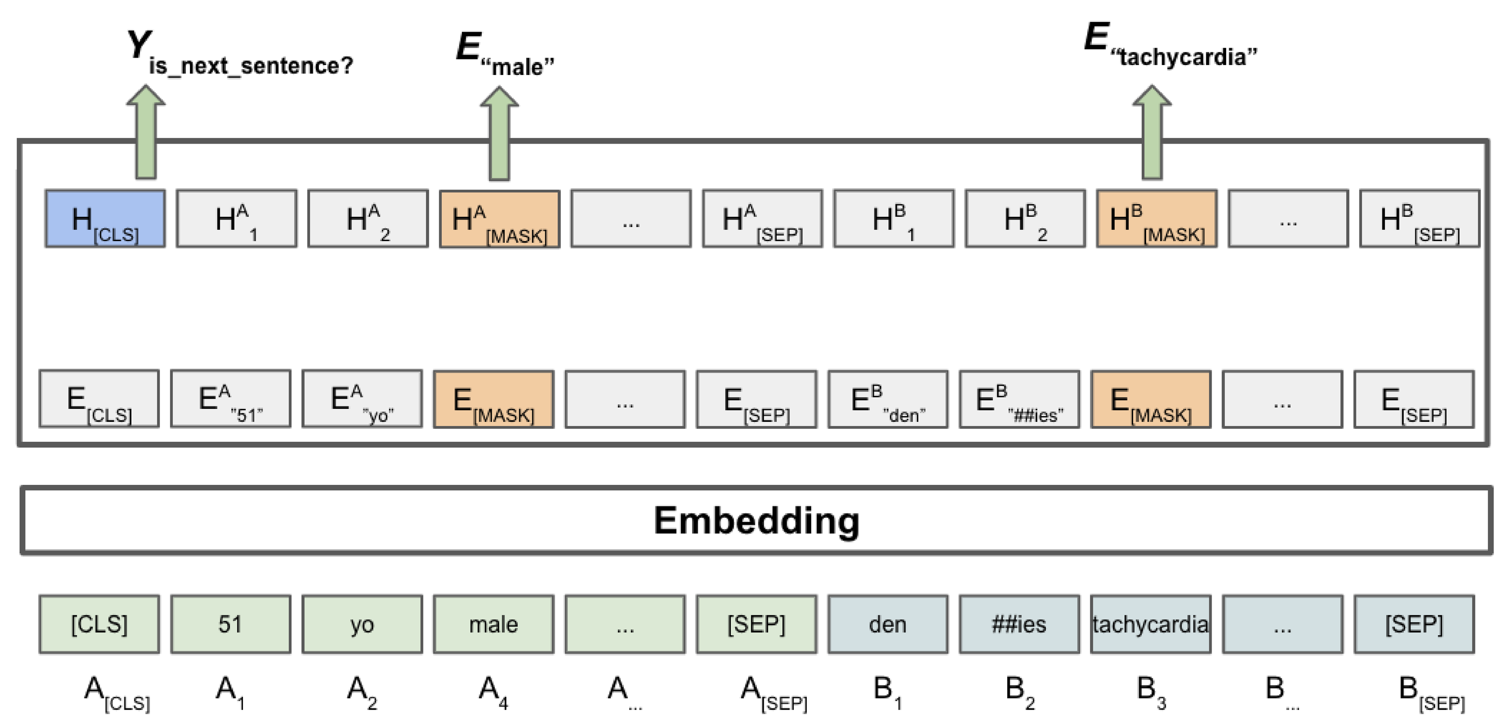
- Chen et al. "Why is my classifier discriminatory?" (2018)
- Kurita et al. "Measuring Bias in Contextualized Word Representations." (2019)
- Beutel et al. "Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations" (2017)
- Elazar and Goldberg. "Adversarial Removal of Demographic Attributes from Text Data" (2018)

MIMIC-III

- MIMIC-III consists of EHR records for 38,597 adults admitted to the ICU of the Beth Israel Deconess Medical Center between 2001 and 2012.
- Contains 2 million clinical notes of varying types.
- Contains patient demographic information such as gender, insurance status, and *self-reported* ethnicity and language spoken.
- 58.7% male, 80.2% white, 88.5% English speakers, 56.1% medicare.

BERT Pretraining ("Clinical BERT")

- Initialized from SciBERT, which is pretrained on biomedical text.
- Used all notes except outpatient notes.
- Trained for one epoch ($\approx$ 8 million samples) on sequences of length 128, then one epoch ($\approx$ 4 million samples) on sequences of length 512.



Log Probability Scores

- Proposed by (Kurita et al., 2019)
- Given a fill-in-the-blanks prediction task, is there a statistically significant difference between the likelihood of predicting male vs. female gendered pronouns?

Sample Template: `[GEND] has a pmh of [ATTR]`

$p(\texttt{[GEND]} = \text{"he"}) = p_{prior}$

$p(\texttt{[GEND]} = \text{"he"} | \texttt{[ATTR]} = \text{"hiv"}) = p_{target}$

$\texttt{score} = \log \frac{p_{target}}{p_{prior}}$

- Came up with templates relating to 8 clinical categories
- Tested on SciBERT and Clinical BERT

Downstream Tasks

- **57** binary classification problems.
- **In-hospital Mortality**: Using the first 48 hours of a patient's notes, predict whether they will die in hospital.
- **Phenotyping using all notes**: Using all notes, predict patient membership in one of 25 HCUP CCS code groups. Also considers any acute phenotype, any chronic phenotype, and any defined disease.
- **Phenotyping using first note**: Similar to the previous tasks, except only using the first nursing or physician note.

Log Probability Score Results

	Log Probability Bias Scores				
	SciBERT		Clinical BERT		Gender Ratio
	M	F	M	F	(M, F)
Addiction	0.202	0.313	0.021*	-0.515*	57.4%, 42.6%
Heart Disease	0.204*	0.333*	0.264*	-0.352*	58.7%, 41.3%
Diabetes	0.100	0.251	0.205*	-0.865*	56.3%, 43.7%
DNR	0.070	0.032	-0.636*	-1.357*	51.9%, 48.1%
Analgesics	1.295	2.127	-0.077	0.105	56.9%, 43.1%
HIV	0.129	0.317	0.616*	-1.247*	64.6%, 35.4%
Hypertension	0.413	0.437	0.440*	-0.402*	55.8%, 44.2%
Mental Illness	-0.414*	-0.164*	0.084*	-0.263*	48.4%, 51.6%

**Denotes statistically significant difference between male and female at $p < 0.01$*

Takeaway:

- Pretraining on clinical notes shifts model predictions towards the gender prevalence in the training data.
- These associations could be useful, but also might be spurious and exceed biological expectations (ex: hypertension).

Downstream Task Results

Significant performance gaps (% of tasks favoring first group):

		Significant Differences by Fairness Definition		
		Recall Gap	Parity Gap	Specificity Gap
Gender	Male vs. Female	**13 (62%)**	25 (36%)	20 (80%)
Language	English vs. Other	**7 (29%)**	17 (12%)	9 (89%)
Ethnicity	White vs. Other	**4 (75%)**	22 (82%)	12 (17%)
	Black vs. Other	**5 (20%)**	18 (72%)	11 (18%)
	Hispanic vs. Other	**7 (0%)**	18 (0%)	20 (100%)
	Asian vs. Other	**8 (62%)**	7 (100%)	8 (50%)
	"Other" vs. Other	**10 (0%)**	8 (0%)	9 (100%)
Insurance	Medicare vs. Other	**33 (85%)**	51 (92%)	48 (6%)
	Private vs. Other	**15 (7%)**	41 (2%)	40 (98%)
	Medicaid vs. Other	**20 (20%)**	31 (19%)	30 (83%)

Takeaway: Many statistically significant performance gaps exist, mostly favoring the majority group.

Attempt at Debiasing

- Applies techniques from previous work on adversarial debiasing (Beutel et al.).
- Attached two adversarial heads to the `[CLS]` token output, to predict gender of the first and second sequences, respectively.
- During training, gradients of the adversary are reversed, to obfuscate gender information in the representations.

Significant Gap Count (% Favoring Male)			
Model	Parity Gap	Recall Gap	Specificity Gap
Baseline	25 (36%)	**13 (62%)**	20 (80%)
Debiased	25 (36%)	**9 (56%)**	20 (70%)

Takeaway: Adversarial debiasing during pretraining does not greatly reduce the number of performance gaps compared to the baseline Clinical BERT model.