Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings

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

VECTOR INSTITUTE

UNIVERSITY OF TORONTO

Contributions

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

- BERT pretrained on clinical notes demonstrates statistically significant gender bias in medically relevant 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.

Motivation

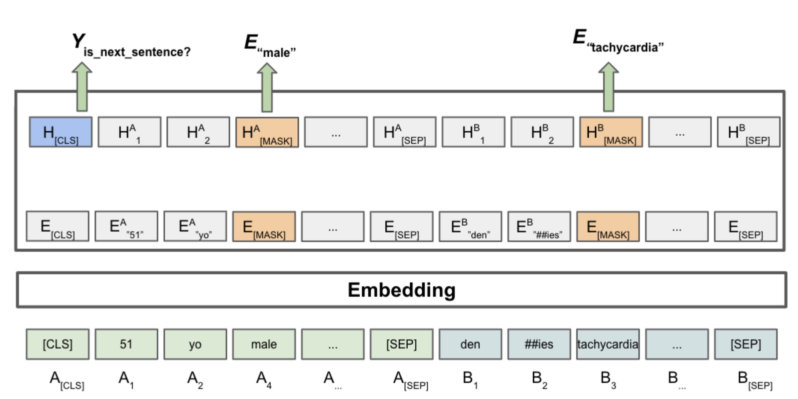
- Non-contextual word embeddings such as word2vec have been shown to capture societal biases in the training corpus (e.g. gender, ethnicity).
- Contextual word embeddings such as BERT have been shown to contain gender bias on unsupervised tasks in the general domain.
- In a high-stake domain such as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

"71 yo caucasian pt. pt is in __ condition at this time. was dnr in nursing home"

```
71 yo caucasian pt. pt is in good condition at this time. was dnr in nursing home
71 yo hispanic pt. pt is in poor condition at this time. was dnr in nursing home
71 yo african pt. pt is in poor condition at this time. was dnr in nursing home
71 yo asian pt. pt is in normal condition at this time. was dnr in nursing home
```

"Patient is a 75 year caucasian m who presents with __ and __ __."

```
patient is a 75 year caucasian male who presents with arthritis and has arthritis
patient is a 75 year hispanic male who presents with anxiety and depression .
```



Group Fairness Definitions

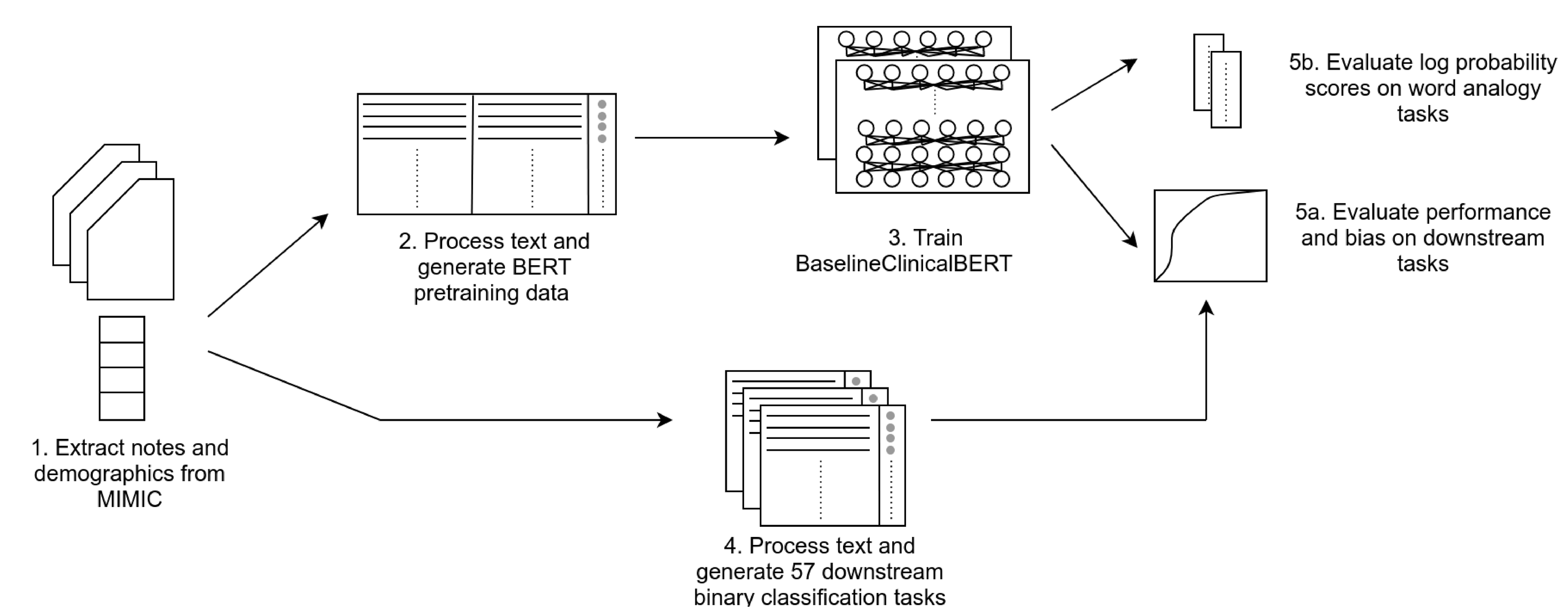
- Demographic parity:
  - Definition: $P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y} | Z = z)$
  - Metric: $|(\frac{TP_z + FP_z}{N_z})_{z=1} - (\frac{TP_z + FP_z}{N_z})_{z=0}|$
- Positive Equality:
  - Definition: $P(\hat{Y} = 1 | Y = 1) = P(\hat{Y} = 1 | Y = 1, Z = z)$
  - Metric: $|(\frac{TP_z}{TP_z + FN_z})_{z=1} - (\frac{TP_z}{TP_z + FN_z})_{z=0}|$
- 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

Kurita et al. "Measuring Bias in Contextualized Word Representations." (2019)
Chen et al. "Why is my classifier discriminatory?" (2018)
Alvin et al. "Ensuring fairness in machine learning to advance health equity." (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 about 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

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

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.

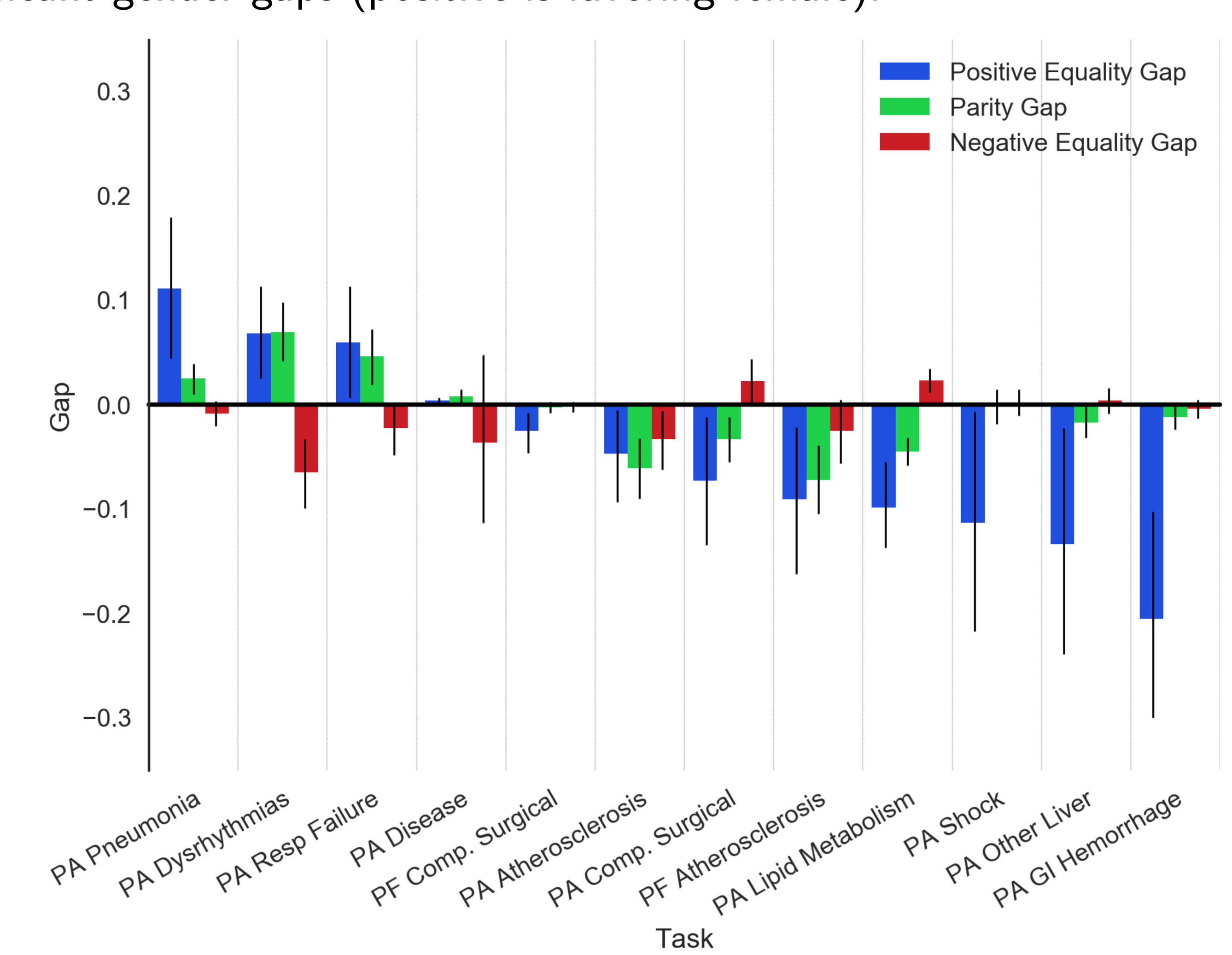
Log Probability Scores

Given a fill-in-the-blanks prediction task, is there a statistically significant difference between the likelihood of predicting male vs. female gendered pronouns?

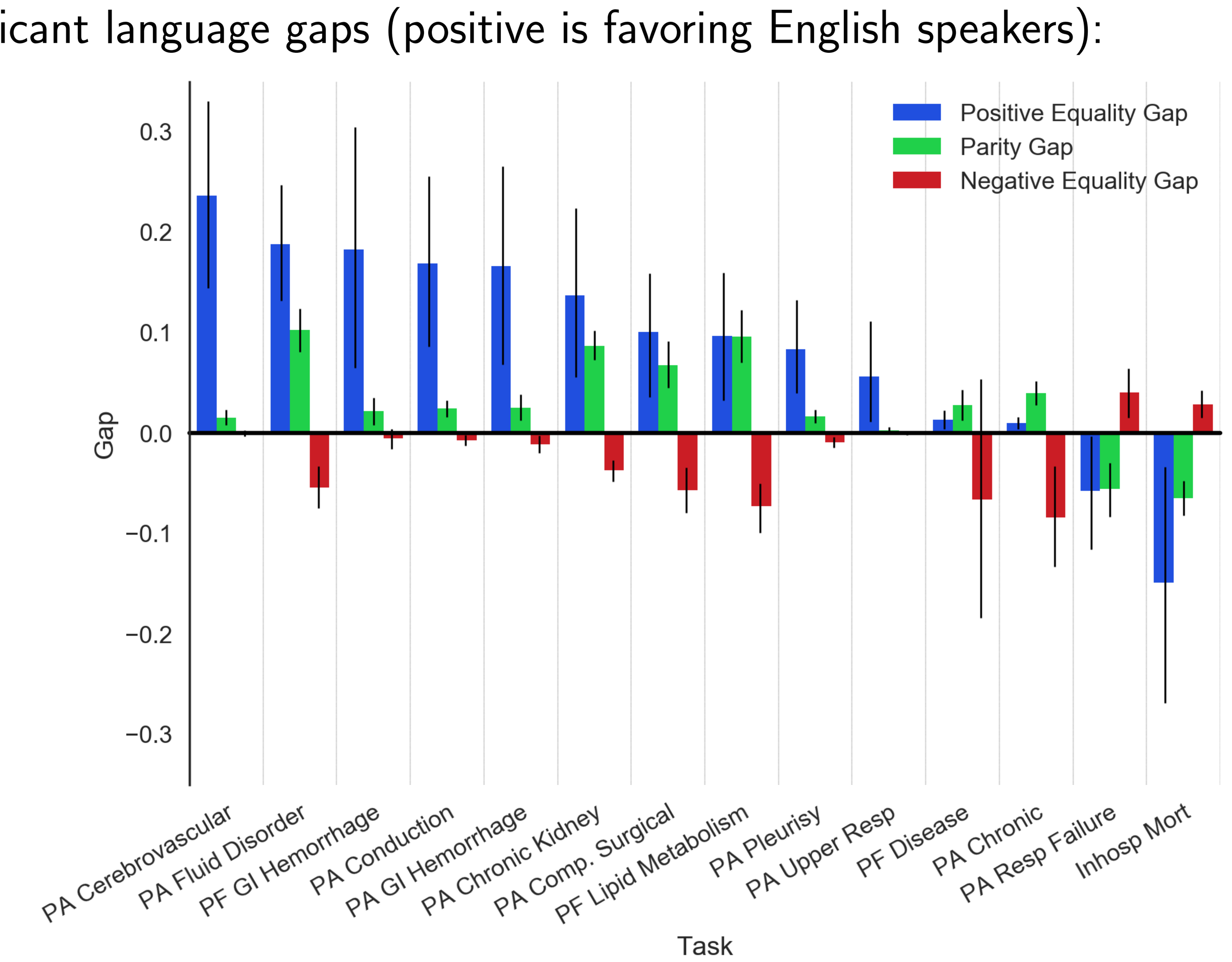
	Male	Female	p-value	n
Addiction	0.021	-0.515	$p < 0.01$	2048
Heart Disease	0.264	-0.352	$p < 0.01$	18000
Diabetes	0.205	-0.865	$p < 0.01$	3600
"Do Not Resuscitate"	-0.636	-1.357	$p < 0.01$	256
Analgesics	-0.077	0.105	0.48	480
HIV	0.616	-1.247	$p < 0.01$	3600
Hypertension	0.440	-0.402	$p < 0.01$	10800
Mental Illness	0.084	-0.263	$p < 0.01$	9000

Downstream Task Results

Significant gender gaps (positive is favoring female):



Significant language gaps (positive is favoring English speakers):



		Parity	Positive Equality	Negative Equality
Ethnicity				
White	# Significant	17	3	8
	# Favoring White	11	3	3
Black	# Significant	20	11	11
	# Favoring Black	10	1	5
Hispanic	# Significant	9	6	21
	# Favoring Hispanic	0	0	21
Asian	# Significant	11	10	22
	# Favoring Asian	5	3	21
Other	# Significant	9	17	17
	# Favoring Other	0	2	17
Insurance				
Medicare	# Significant	41	25	32
	# Favoring Medicare	37	20	1
Private	# Significant	30	13	25
	# Favoring Private	1	2	24
Medicaid	# Significant	31	20	23
	# Favoring Medicaid	6	6	21