# Quantifying and Removing Biases in Clinical Contextual Word Embeddings

Haoran Zhang* Amy Lu* Mohamed Abdalla
Marzyeh Ghassemi

Submitted to ACM FAT* 2020

Vector NLP Meeting - August 29, 2019

# Motivation

Machine learning models can contain biases against specific groups.

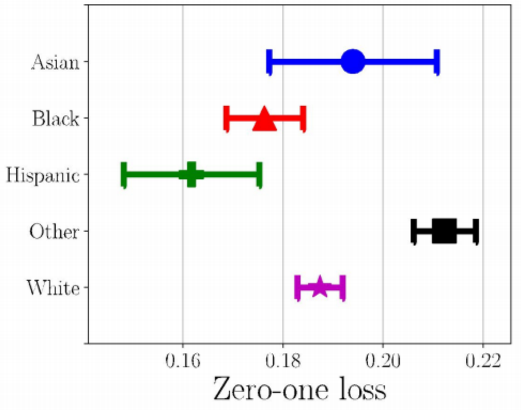
- Gender bias in predicting occupation from biographies (De-Arteaga et al., 2019)
- Facial analysis algorithm performs worse on dark-skinned women (Buolamwini and Gebru, 2018)
- Recidivism predictor has higher false positive rate for blacks than whites (Chouldechova, 2017)

# Motivation



In healthcare:

- Model to predict ICU mortality from clinical notes performs better on certain ethnicities than others (Chen, Johansson, and Sontag, 2018)



# Motivation



In word embeddings:

- "Man is to Computer Programmer as Woman is to Homemaker?" (Bolukbasi et al., 2016)
- "Word embeddings quantify 100 years of gender and ethnic stereotypes" (Garg et al., 2018)
- Contextual word embeddings show statistically significant gender and race bias on word analogy tasks (Kurita et al., 2019)

# Contributions



- We show that contextual word embeddings trained on clinical notes (MIMIC-III) exhibit undesired biases and performance gaps between protected groups.
- We propose an adversarial debiasing scheme for contextual word embeddings, and show that its effect is limited, which is in line with previous work (Elazar and Goldberg, 2018).

# Fairness in ML

Consider a task to predict binary output label $Y$ given $X$, while remaining unbiased with respect to some categorical variable $Z$. The predictor is $\hat{Y} = f(X)$.

- **Demographic parity**:

$$P(\hat{Y} = y) = P(\hat{Y} = y | Z = z)$$

- Define

$$ProbTrue_z = P(\hat{Y} = 1 | Z = z) = \frac{TP_z + FP_z}{N_z}$$

- Demographic gap ($|Z| = 2$):

$$dgap = ProbTrue_1 - ProbTrue_0$$

# Fairness in ML



- **Equality of odds**:

$$P(\hat{Y} = y | Y = y) = P(\hat{Y} = \hat{y} | Y = y, Z = z)$$

- Define

$$ProbCorrect_{1,z} = P(\hat{Y} = 1 | Z = z, Y = 1) = \frac{TP_z}{TP_z + FN_z}$$

$$ProbCorrect_{0,z} = P(\hat{Y} = 0 | Z = z, Y = 0) = \frac{TN_z}{TN_z + FP_z}$$

- Equality Gap ($|Z| = 2$):

$$egap_{y=1} = ProbCorrect_{y=1,z=1} - ProbCorrect_{y=1,z=0}$$

$$egap_{y=0} = ProbCorrect_{y=0,z=1} - ProbCorrect_{y=0,z=0}$$

# Fairness in ML



- **Equality of opportunity**:

$$P(\hat{Y} = y | Y = y) = P(\hat{Y} = \hat{y} | Y = y, Z = z)$$

for one specific value of $Y$

- Equal TPR **or** equal TNR

# What if $|Z| > 2$?



- Binary case: $egap_{y=1} = TPR_{z=1} - TPR_{z=0}$
- Option: Combining all other classes

$$\text{egap}_{j,y=1} = P(\hat{Y}=1|Y=1, Z=z_j) - P(\hat{Y}=1|Y=1, Z \neq z_j)$$

$$= TPR(z_j) - TPR(\tilde{z}_j)$$

- Proposed method: max absolute gap

$$ind = arg \max_{i \in Z} |TPR(z_j) - TPR(z_i)|$$

$$egap_{j,y=1} = TPR(z_j) - TPR(z_{ind})$$

- $egap_{j,y=1} = 0$ iff $TPR(z_i) = TPR(z_j)\ \forall z_i \in Z$

# SciBERT



- All of the existing "ClinicalBERT" models are initialized from BioBERT or $\text{BERT}_{\text{BASE}}$ (Alsentzer et al., 2019; Huang, Altosaar, and Ranganath, 2019)
- SciBERT (Beltagy, Cohan, and Lo, 2019) is better for several reasons:
  - Better benchmarking performance
  - Trained from scratch; not initialized from $\text{BERT}_{\text{BASE}}$
  - Uses an improved vocabulary (`SciVOCAB`)

# MIMIC-III



- EHR for 38,597 adults admitted to ICU of the Beth Israel Deconess Medical Center between 2001 and 2012
- Contains about 2 million clinical notes of varying types (discharge summaries, nursing notes, radiology reports, etc)
- Self-reported patient demographic information such as ethnicity, language spoken, insurance status
- Also contains labs, vitals, medications, etc

# Baseline Clinical BERT



- Initialized from SciBERT (Beltagy, Cohan, and Lo, 2019)
- Additional pre-training on MIMIC notes
  - one epoch (8 million examples) at combined sequence length 128
  - one epoch (4 million examples) at combined sequence length 512
- Using whole-word masking

# Downstream Tasks

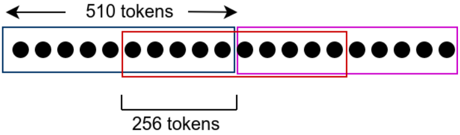


1. **In-hospital mortality**. Predict whether a patient will die in hospital given notes from the first 48 hours of their ICU stay.
2. **Phenotyping using all notes**. Predict whether a patient will have ICD-9 codes belonging to one of 25 HCUP CCS code groups (+any chronic, any acute, any disease), using all notes available.
3. **Phenotyping using the first note**. Same target as above, but using only the first note within the first 48 hours of a patient's stay.

# Document-level Predictions



- BERT has a fixed maximum input sequence length of 512
- Many notes are longer than 512 tokens
- Can create subsequences using sliding window approach:




- Assign label for each sequence to be the label from the source document
- Merge predicted probabilities with a function (Huang, Altosaar, and Ranganath, 2019):

$$P(Y=1) = \frac{P^n_{max} + P^n_{mean} n/c}{1 + n/c}$$

# Finetuning Procedure



Will use feature-based approach, keep BERT weights frozen

1. Representations for subsequences are extracted from BERT (concatenating last 4 hidden layers of `[CLS]` token)
2. Acuity scores and age are appended
3. Train fully-connected NN for binary classification problem
   - Grid search over # layers, # neurons, dropout probability
4. Predictions are aggregated to the patient level

# Qualitative Evaluation: Log Probability Bias Scores



- Uses the "log probability bias score" for assessing bias in BERT, as proposed in Kurita et al., 2019
- Procedure:
  1. Prepare template sentences with both a **medical context** and **gendered keyword** (e.g. "55 yo caucasian [MASK] with a hx of hiv"). Pass this sentence into the BERT model.
  2. For the softmax-normalized vector indicating probabilities for the [MASK] position, select the likelihood of predicting each gendered key word as $p_{target}$
  3. To control for the natural propensity for BERT to favour a certain demographic token, calculate $p_{prior}$ using the same procedure as above, using a template sentence without the medical contexts (e.g. "55 yo caucasian [MASK] with a hx of [MASK]"
  4. Calculate the log-probability bias score as $\frac{p_{target}}{p_{prior}}$

# Qualitative Evaluation: Log Probability Bias Score



- To increase the sample size of tested sentences, we permute through lists of:
  1. **Medical contexts** for each outcome category (e.g. "Heart disease": ["cvd", "heart disease", "heart failure", ...]
  2. **Gendered keywords** (e.g. "male": ["male", "m", "he"])
- Test for a statistical significant difference between the two gender groups using Wilcoxon Signed-Ranks test

# Baselines: Log Probability Bias Score



| | **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 |

Table: Log probability scores for different groups of words denoting various medical attributes. **7 out of 8 categories** have a statistically significant difference in bias scores between the two genders.

# Qualitative Evaluation: Sentence Completion

- Uses templates adapted from real medical notes
- SciBERT can generate radically different symptoms and outcomes for different racial populations

***"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 .

*Note on abbreviations: yo – years old; pt – patient; dnr – do not resuscitate*

# Qualitative Comparison of Baseline Clinical BERT with Previous Models



## Baseline Clinical BERT (Ours)

***40 year old caucasian man with hiv and new dx of nhl now with fever***
{14: ['fever', 'hypotension', 'neutrop', 'diarrhea', 'thrombocytop', 'nausea', 'gib', 'new', 'vomiting', 'br', 'dyspnea', 'lymphoma', 'a', 'abdominal', 'hem', 'neutropenia', 'hypoxia', 'ams', 'hiv', 'sob']}

***45 year old caucasian man with hiv and new dx of hiv***
{11: ['hiv', 'lymphoma', 'renal', 'diagnosis', 'phinamephi', 'malignant', 'aml', 'leukemia', 'diagnosed', 'cell', 'hod', 'aids', 'dl', 'pcp', 'liver', 'newly', 'of', 'derm', 'malignancy', 'thym']}

***50 year old caucasian man with hiv hep c acute renal failure : reason for this examination : hiv***
{18: ['hiv', 'aids', 'renal', 'interval', 'cirrhosis', 'liver', 'pre', 'please', 'indication', 'fever', '?', 'evaluate',

## Clinical BERT, Alsentzer et al.

***40 year old caucasian man with hiv and new dx of nhl now with p***
{19: ['p', 'f', 'u', 'pm', 'c', 'a', 'd', 'r', 'new', 'h', 'ch', 'g', 'n', 't', 'di', 'am', 'fever', 're', 'si', 's']}

***45 year old caucasian man with hiv and new dx of p***
{15: ['p', 'c', 'f', 'd', 'a', 'g', 'r', 'u', 's', '1', 'n', 't', '2', 'b', 'di', 'y', 'te', 'l', 'pm', 'j']}

***50 year old caucasian man with hiv hep c acute renal failure : reason for this examination : 1***
{23: ['1', 'reason', '.', ':', 'Reason', 'for', 'patient', 'n', 'no', 'to', 'r', 'data', 'order', 'cause', 'pro', 'do', 'a', '-',

# Baseline: Downstream Task Results - Gender



TPR gap in gender: 12 significant, 8 favoring male

| Task Type | Task | Recall Gap (F-M) | AUROC | AUPRC | Prevalence in Males | Prevalence in Females |
|---|---|---|---|---|---|---|
| All | Gastrointestinal hemorrhage | -20.52% | 87.32% | 50.31% | 6.81% | 7.71% |
| All | Other liver diseases | -13.39% | 86.91% | 53.40% | 6.84% | 8.48% |
| All | Shock | -11.33% | 87.74% | 44.34% | 7.63% | 6.89% |
| All | Disorders of lipid metabolism | -9.84% | 76.65% | 51.55% | 23.72% | 27.56% |
| First | Coronary atherosclerosis | -9.06% | 82.73% | 71.84% | 25.41% | 39.48% |
| All | Comp. of surgical procedures | -7.25% | 73.77% | 41.83% | 19.90% | 20.79% |
| All | Coronary atherosclerosis | -4.68% | 86.70% | 80.73% | 25.80% | 38.03% |
| First | Comp. of surgical procedures | -2.51% | 66.21% | 27.27% | 15.76% | 16.87% |
| All | Any disease | 0.40% | 91.62% | 99.25% | 93.16% | 92.29% |
| All | Respiratory failure | 5.94% | 89.96% | 67.98% | 19.23% | 16.20% |
| All | Cardiac dysrhythmias | 6.80% | 77.90% | 64.06% | 30.76% | 32.18% |
| All | Pneumonia | 11.12% | 81.61% | 43.31% | 8.73% | 8.32% |

# Baseline: Downstream Task Results - Ethnicity & Insurance

Ethnicity:

| | # Significant | # Favoring |
|---|---|---|
| **White** | 3 | 3 |
| **Black** | 11 | 1 |
| **Hispanic** | 6 | 0 |
| **Asian** | 10 | 3 |
| **Other** | 17 | 2 |

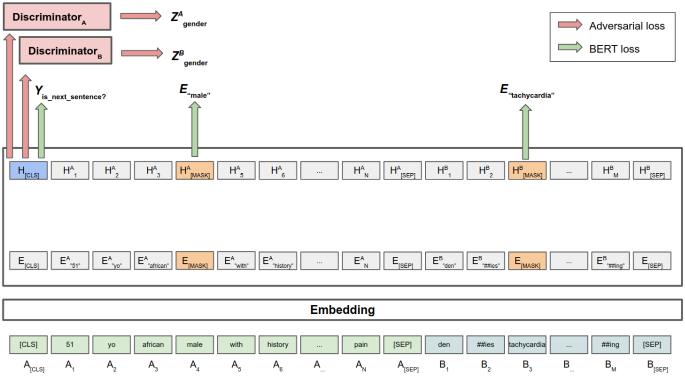
Insurance:

| | # Significant | # Favoring |
|---|---|---|
| **Medicare** | 25 | 20 |
| **Private** | 13 | 2 |
| **Medicaid** | 20 | 6 |

Language: 14 significantly different performances (12 favoring English speakers)

# Debiasing Method





# Debiasing Method



- Sequence representations $h = f(x_1, x_2)$ are extracted from BERT
- Classifiers $a_1$ and $a_2$ try to predict $\hat{z}_1 = a_1(h)$ and $\hat{z}_2 = a_2(h)$

$$L = \sum_{(x_1,x_2)\in X} L_{adv}(a_1(J(h)), z_1) + L_{adv}(a_2(J(h)), z_2) + L_{LM} + L_{NS}$$

- Gradient reversal: $J(h) = h, \frac{dJ}{dx_1} = -\lambda \frac{dh}{dx_1}$
- Note: $\lambda$ is a hyperparameter tuning for the strength of fairness

# Drawbacks

1. Large difference in model capacity between the generator and discriminator $\rightarrow$ likely weak debiasing effect
2. Only debiases the [CLS] token; not useful for sequence output tasks
3. Can only give demographic parity

# Debiased: Log Probability Bias Score



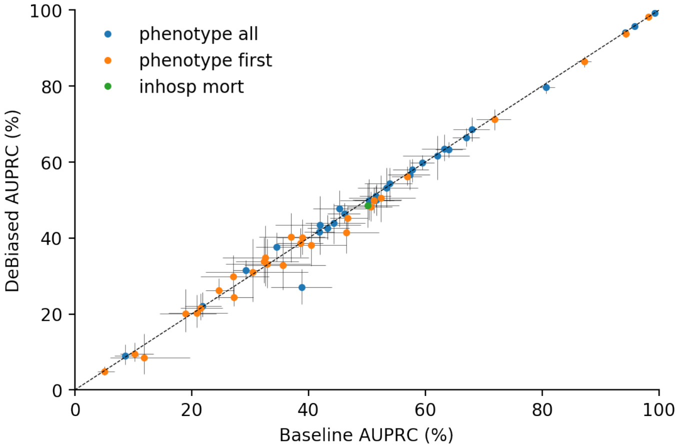
| | **Male** | **Female** | **P-value** | $n$ |
|---|---|---|---|---|
| Addiction | -0.306 | -0.438 | $p = 0.02$ | 2048 |
| Heart Disease | 0.262 | -0.236 | $p = 0.141$ | 18000 |
| Diabetes | -0.247 | -0.520 | $p < 0.01$ | 3600 |
| "Do Not Resuscitate" | -1.693 | -1.544 | $p = 0.838$ | 256 |
| Analgesics | -0.566 | 0.036 | $p = 0.024$ | 480 |
| HIV | 0.388 | -0.832 | $p < 0.01$ | 3600 |
| Hypertension | -0.255 | -0.410 | $p < 0.01$ | 10800 |
| Mental Illness | 0.123 | -0.190 | $p < 0.01$ | 9000 |

Table: Log probability scores for different groups using the **debiased model.** Notice that **4 out of 8 categories** have a statistically significant difference between the genders, **compared to 7 out of 8 categories** for the baseline.

# Debiased: Performance on Downstream Tasks



Debiasing does not lead to a significant change in performance.



# Debiased: Downstream Task Results - Gender



TPR gaps for tasks where baseline was biased:

| **Task Type** | **Task** | **Recall Gap Baseline (F-M)** | **Recall Gap Debiased (F-M)** |
|---|---|---|---|
| All | Gastrointestinal hemorrhage | -20.52% | -14.40% |
| All | Other liver diseases | -13.39% | -6.14% |
| All | Shock | -11.33% | -13.98% |
| All | Disorders of lipid metabolism | -9.84% | -11.98% |
| First | Coronary atherosclerosis | -9.06% | -11.75% |
| All | Complications of surgical procedures | -7.25% | -0.94% |
| All | Coronary atherosclerosis | -4.68% | 0.09% |
| First | Complications of surgical procedures | -2.51% | 0.37% |
| All | Any disease | 0.40% | 0.50% |
| All | Respiratory failure | 5.94% | 4.99% |
| All | Cardiac dysrhythmias | 6.80% | 6.27% |
| All | Pneumonia | 11.12% | 9.00% |

But number of significant gaps increased to 13!

# Debiased: Downstream Task Results - Gender



| | | **# Significant Gaps** | **# Favoring Male** | **# Favoring Female** |
|---|---|---|---|---|
| **Parity** | Baseline | 20 | 8 | 12 |
| | Debiased | 17 | 6 | 11 |
| **Recall** | Baseline | 12 | 8 | 4 |
| | Debiased | 13 | 8 | 5 |
| **Specificity** | Baseline | 16 | 12 | 4 |
| | Debiased | 12 | 10 | 2 |

# Conclusions



- BERT word embeddings trained on MIMIC-III exhibit undesired biases between protected groups on downstream clinical tasks
- We propose an adversarial debiasing scheme, and show that its ability to remove bias from downstream tasks is poor.