Measuring and Mitigating Unintended Bias in Text Classification

Lucas Dixon
ldixon@google.com

John Li*
jetpack@google.com

Jeffrey Sorensen
sorenj@google.com

Nithum Thain
nthain@google.com

Lucy Vasserman*
lucyvasserman@google.com

** AIES Presenters*

Conversation-AI

ML to improve online conversations at scale

Research Collaboration

Jigsaw, CAT, several Google-interna teams, and external partners (NYTimes, Wikimedia, etc)

Perspective API

Unintended Bias

Model falsely associates frequently attacked identities with toxicity: *False Positive Bias*

Sentence	model score
"i'm a proud **tall** person"	0.18
"i'm a proud **lesbian** person"	0.51
"i'm a proud **gay** person"	0.69

Bias Source and Mitigation

Bias caused by dataset imbalance

- Frequently attacked identities are overrepresented in toxic comments
- Length matters

Add *assumed non-toxic data* from Wikipedia articles to fix the imbalance.

- Original dataset had 127,820 examples
- 4,620 non-toxic examples added

	Comment Length				
Term	**20-59**	**60-179**	**180-539**	**540-1619**	**1620-4859**
ALL	17%	12%	7%	5%	5%
gay	88%	77%	51%	30%	19%
queer	75%	83%	45%	56%	0%
homosexual	78%	72%	43%	16%	15%
black	50%	30%	12%	8%	4%
white	20%	24%	16%	12%	2%
wikipedia	39%	20%	14%	11%	7%
atheist	0%	20%	9%	6%	0%
lesbian	33%	50%	42%	21%	0%
feminist	0%	20%	25%	0%	0%
islam	50%	43%	12%	12%	0%
muslim	0%	25%	21%	12%	17%
race	20%	25%	12%	10%	6%
news	0%	1%	4%	3%	3%
daughter	0%	7%	0%	7%	0%

Measuring Unintended Bias - Synthetic Datasets

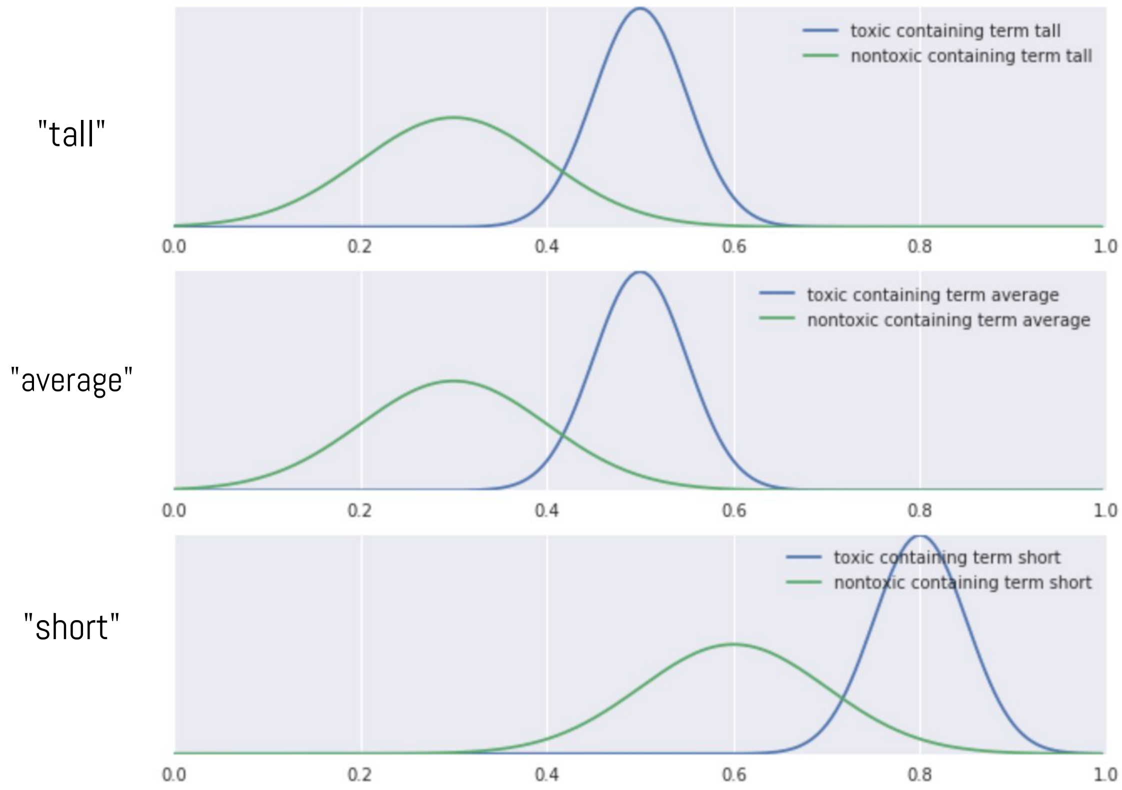
Challenges with real data:

- Existing datasets are small and/or have false correlations
- Each example is completely unique: not easy to compare for bias

Approach: "bias madlibs": a synthetically generated 'templated' dataset for evaluation

Sentence	model score
"i'm a proud **tall** person"	0.18
"i'm a proud **lesbian** person"	0.51
"i'm a proud **gay** person"	0.69
"audre is a **brazilian** computer programmer"	0.02
"audre is a **muslim** computer programmer"	0.08
"audre is a **transgender** computer programmer"	0.56

Measuring Unintended Bias - Metrics Challenges



Equality of Odds

- Requires choosing a threshold, not aligned with real-world usage
- Choice of threshold can drastically change results!

ROC-AUC

- Doesn't capture bad orderings between groups

	AUC
Tall	0.93
Average	0.93
Short	0.93
Combined	0.79

Measuring Unintended Bias - Pinned AUC

Pinned AUC metric measures unintended bias on *real-valued scores* directly

$$\boldsymbol{D} = \text{full dataset}$$

$$\boldsymbol{D_t} = \text{subset of } \boldsymbol{D} \text{ containing term } \boldsymbol{t}$$

$$PinnedAUC(t) = AUC(\underbrace{D_t + sample(D)}_{\text{"Pinned" Dataset for term } t}), \text{ where } |D_t| = |sample(D)|$$

Pinned AUC

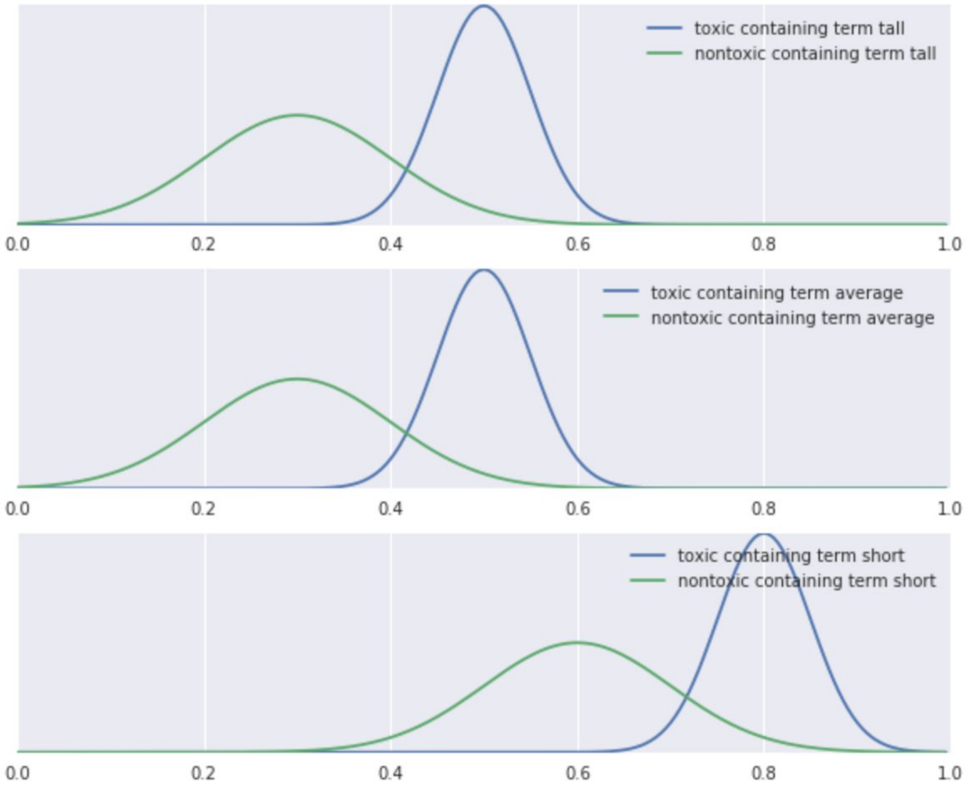
$$PinnedAUC(t) = AUC(D_t + sample(D))$$

for identity term ***t*** and full dataset ***D***

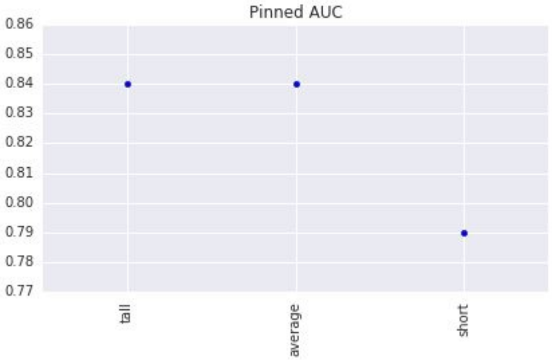
"tall"

"average"

"short"



	AUC	Pinned AUC
Tall	0.93	0.84
Average	0.93	0.84
Short	0.93	0.79
Combined	0.79	N/A



Pinned AUC Equality Difference

For identity terms, **t**, in a balanced test set **D**:

$$PinnedAUC\Delta(t) = |AUC(D) - PinnedAUC(t)|$$

$$\boxed{\text{Pinned AUC Equality Difference} = \sum PinnedAUC\Delta(t), \text{ for all terms } \boldsymbol{t}}$$

- A single number that measures how much a model treats different identity terms differently.
- Generalizes to identity groups if data exists.
- Questions to consider:
  - What is the set of identities?
  - What is the appropriate test set?
  - Squared error?

Experiments and Results

Three Models

Baseline

- 127,820 Wikipedia comments

Control

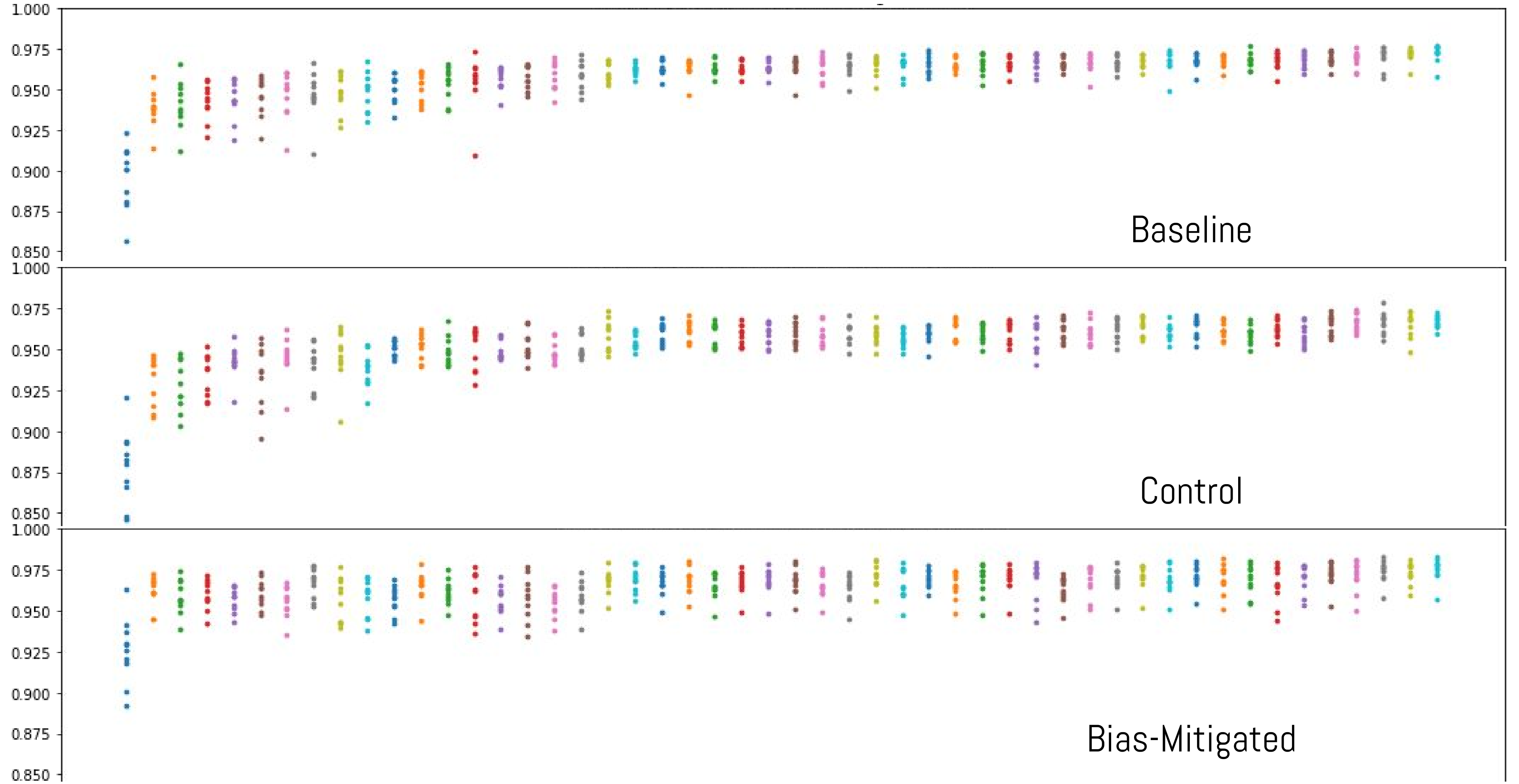
- 4,620 Wikipedia article snippets, *randomly selected*

Bias-Mitigated

- 4,620 Wikipedia article snippets, *selected to balance toxicity distribution for specific terms*

Model	**Pinned AUC Equality Difference**
Baseline	6.37
Control	6.84
Bias-Mitigated	4.07

Per-term Pinned AUC



Summary

- Unintended bias can be mitigated by strategically adding data
- Synthetic datasets enable bias measurement
- Pinned AUC metric measures bias on real-valued scores

Future work

- Beyond synthetic datasets
- Additional mitigation techniques