Gender Bias Amplification in Natural Language Processing (NLP)

Lindsey Ehrlich

Agenda

- Motivation
- Guiding question
- Choosing the task
- Choosing the dataset
- Procedure
- Results
- Future steps

Motivation

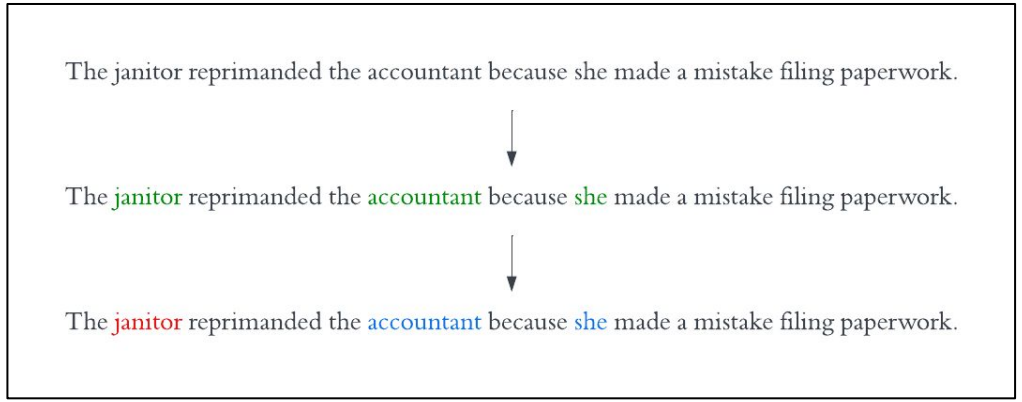
- Directional Bias Amplification (Wang & Russakovsky)
  - Rectifies three shortcomings in the bias amplification metric introduced in “Men Also Like Shopping” (Zhao et al)
- A Systematic Study of Bias Amplification (Hall et al)
  - Tackles six research questions to better understand bias amplification trends in computer vision tasks
    - RQ1: how does bias amplification vary as the bias in the data varies?
  - Study limited to “binary classification tasks in the image-recognition domain”
- Bias amplification study lacking in the NLP space
- Why NLP?

https://arxiv.org/pdf/2102.12594.pdf
https://arxiv.org/pdf/2201.11706.pdf

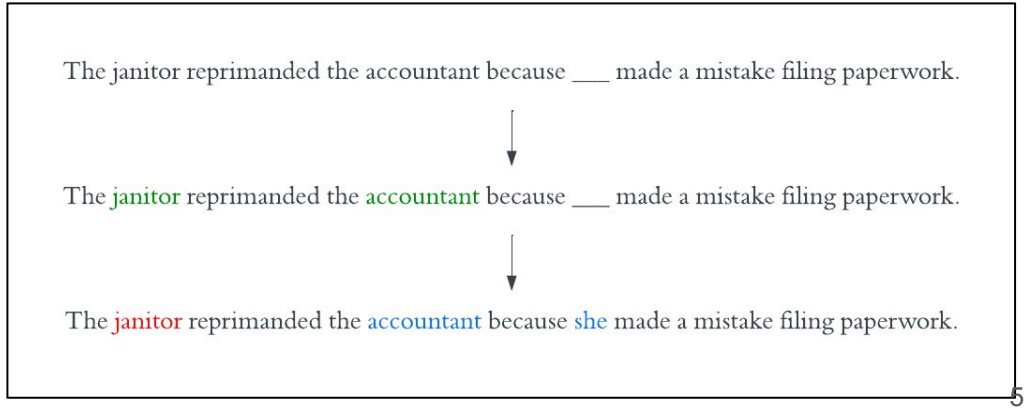
Guiding question: how does bias amplification vary as bias in the data varies?

Tasks

- Sentiment analysis
- Coreference resolution



- Pronoun prediction

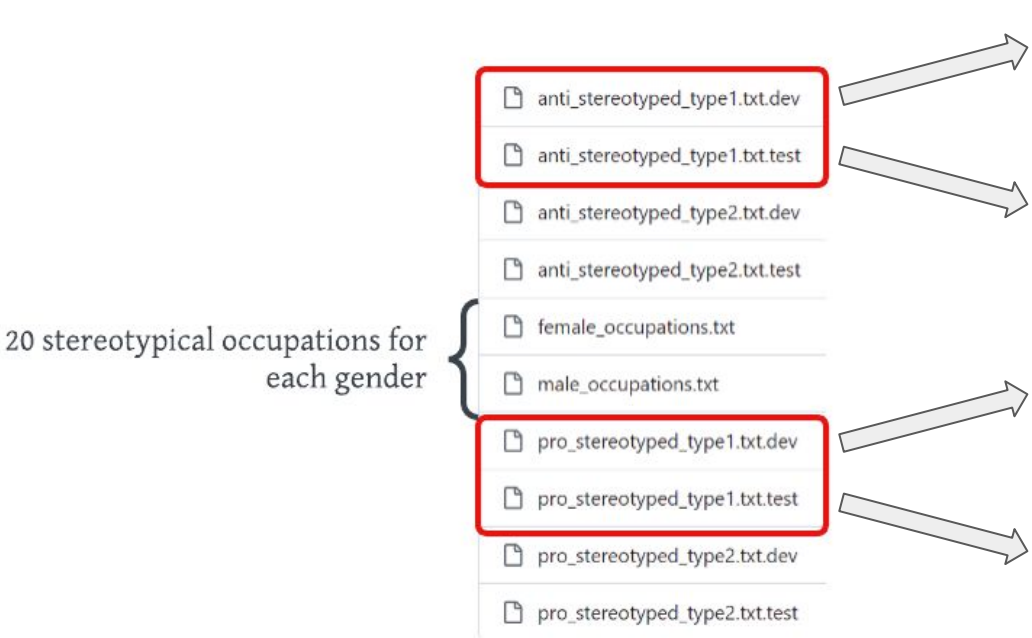


WinoBias Dataset

- Sentence type 1: [entity1] [interacts with] [entity2] [conjunction] [pronoun] [circumstances]
  - E.g. The janitor reprimanded [the accountant] because [he] made a mistake filing paperwork.
- Sentence type 2: [entity1] [interacts with] [entity2] and then [interacts with] [pronoun] for [circumstances]
  - E.g. The janitor met [the accountant] and wished [him] well.

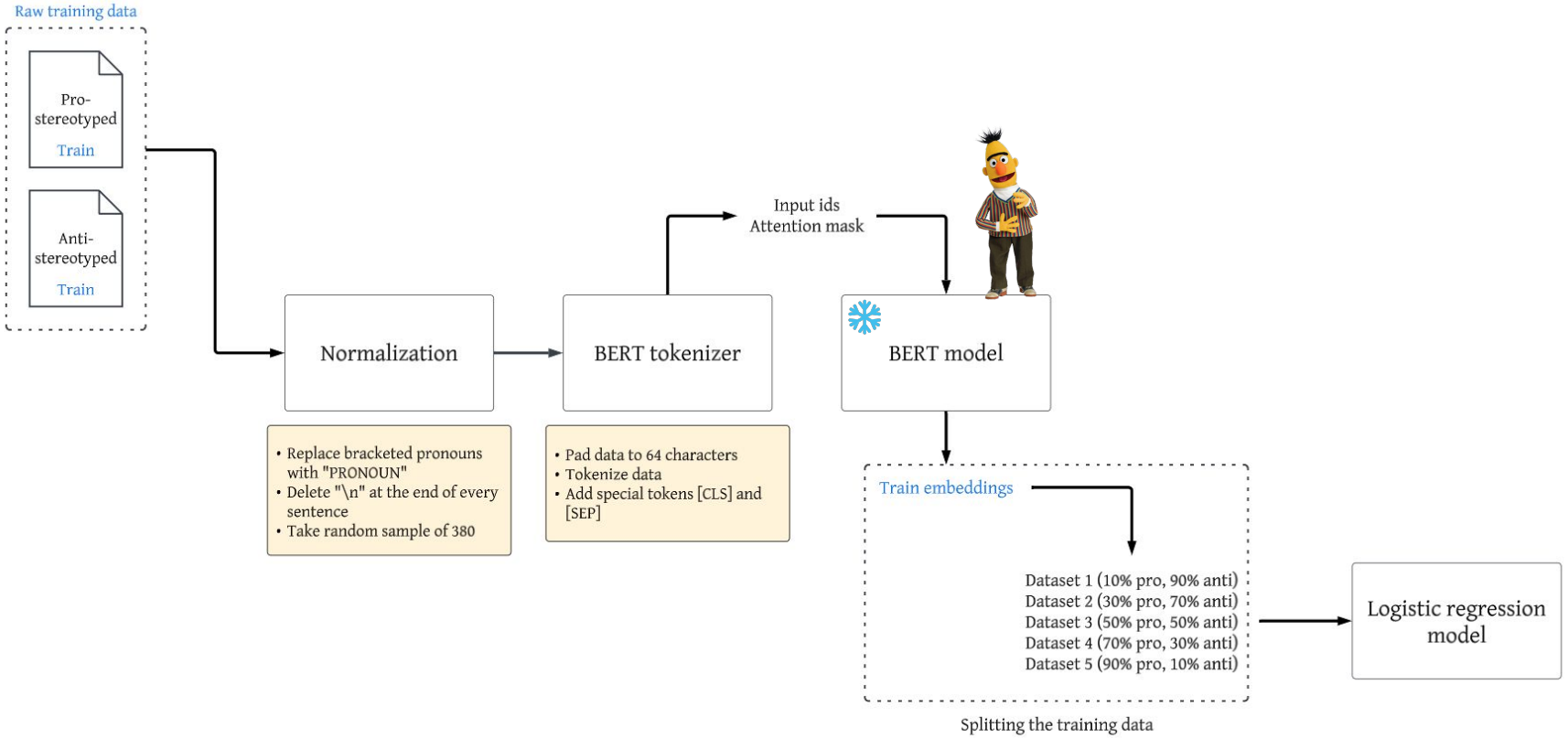
https://arxiv.org/pdf/1804.06876.pdf

WinoBias Dataset



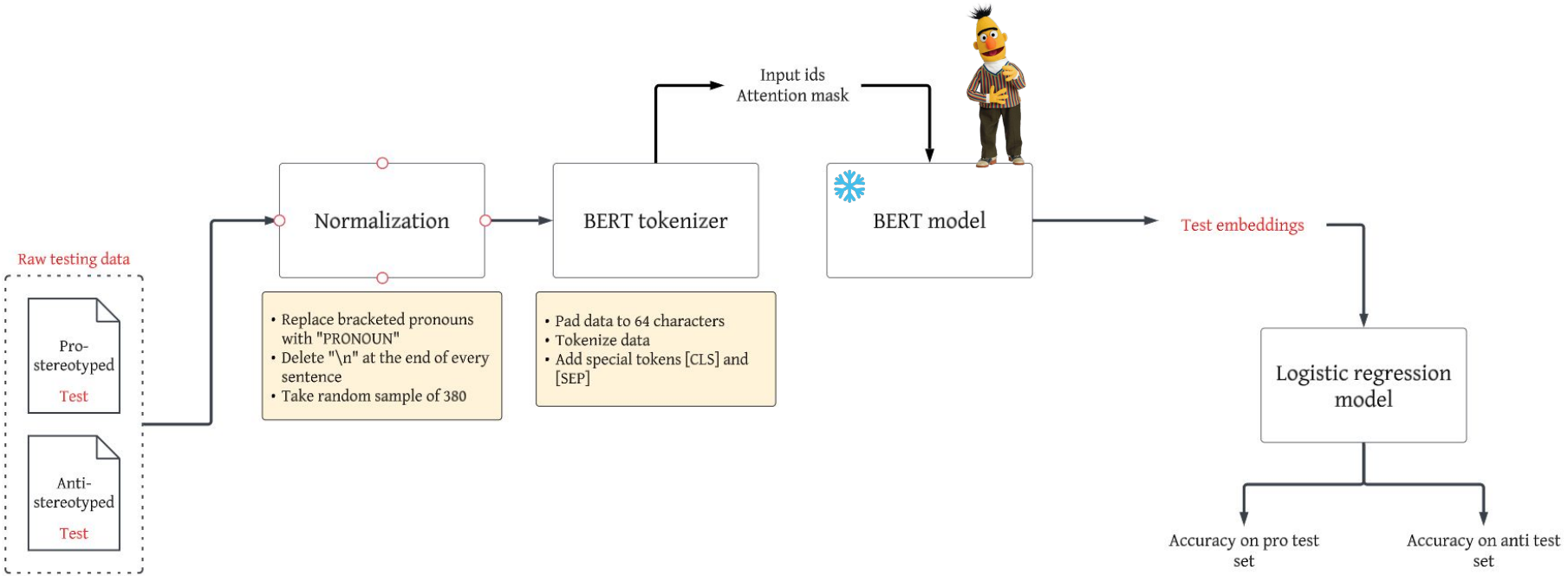
https://github.com/uclanlp/corefBias/tree/master/WinoBias/wino/data

Experimental Procedure



https://carboncostume.com/bert-from-sesame-street/

Experimental Procedure



https://carboncostume.com/bert-from-sesame-street/

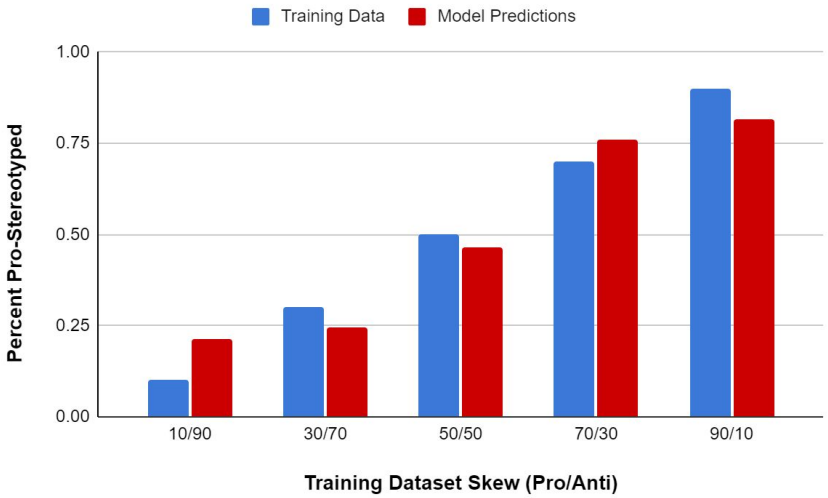
Accuracy Results

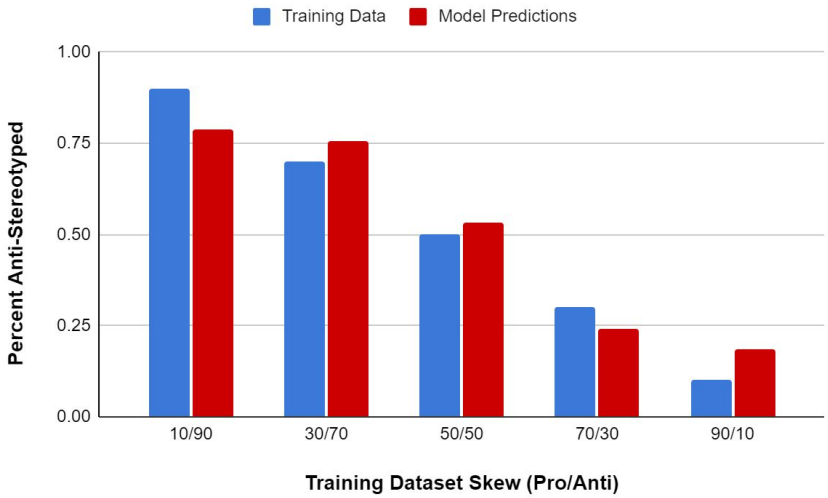
Pro/Anti Split	Unbiased accuracy	Pro Accuracy	Anti Accuracy
10/90	0.5157894736842106	0.2	0.8131578947368421
30/70	0.5210526315789473	0.25	0.7657894736842106
50/50	0.5210526315789473	0.45789473684210524	0.55
70/30	0.48947368421052634	0.7736842105263158	0.22105263157894736
90/10	0.48157894736842105	0.8184210526315789	0.16842105263157894

Accuracy Results

Pro/Anti Split (Training Data)	Unbiased accuracy	Pro Accuracy	Anti Accuracy
10/90	0.5158	0.2	0.81312
30/70	0.5211	0.25	0.7658
50/50	0.5211	0.4579	0.55
70/30	0.4895	0.7737	0.2211
90/10	0.4816	0.8184	0.1684

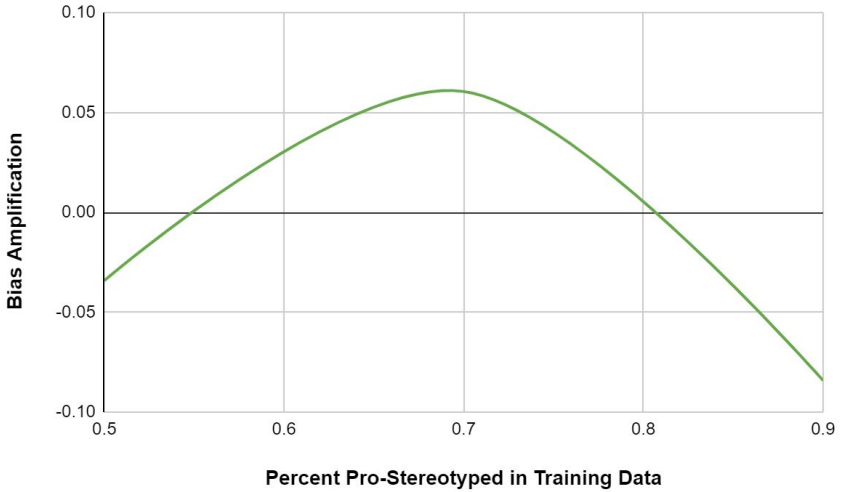
Results

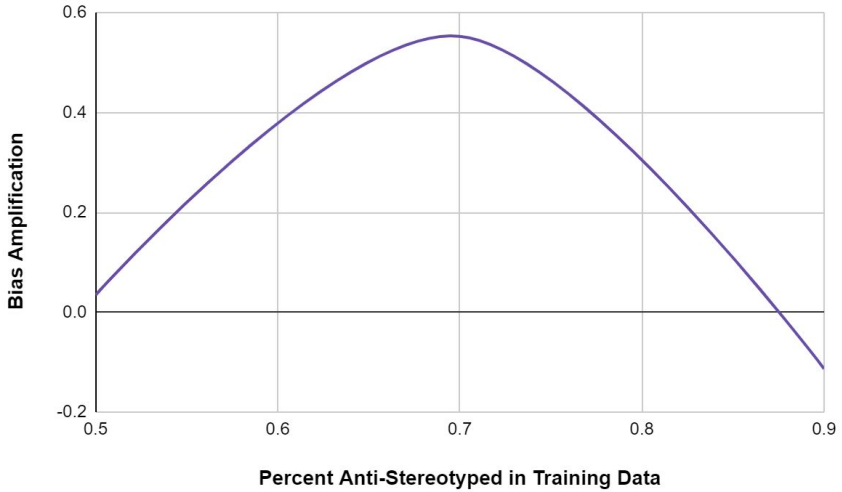




As we move toward either extrema (away from the 50/50 split), bias amplification increases, then sharply decreases.

Results





As we move toward either extrema (away from the 50/50 split), bias amplification increases, then sharply decreases.

Future Steps

- Examine bias amplification with respect to each occupation in the dataset
- Examine bias amplification when training on a soft label dataset vs. a hard label dataset
- Examine how bias amplification changes as the model trains

Thank You!

- Dr. Russakovsky
- Allison Chen
- Angelina Wang
- Austin Hanjie
- Princeton University

Preprocessing Example Sentence

original sentence: "[The developer] argued with the designer because [she] did not like the design.\n"

↓

normalized sentence: "[The developer] argued with the designer because PRONOUN did not like the design."

↓

input_id: tensor([101, 1996, 5160, 2246, 2046, 6206, 13519, 2114, 1031, 1996, 5356, 3771, 1033, 1010, 2021, 4013, 3630, 4609, 2089, 2031, 2042, 23123, 5496, 1012, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])

[CLS] = 101
[SEP] = 102

attention_mask: tensor([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])

↓

embedding: tensor([-0.7561, -0.5056, -0.9446, 0.5786, 0.6703, -0.1555, 0.7214, 0.3716, -0.8180, -1.0000, -0.4903, 0.8223, 0.9648, 0.7889, 0.8260, -0.6572 . . . -0.5132, -0.1393, 0.7039, -0.0365, 0.9298, 0.7405, -0.6300, -0.1625, 0.7226, -0.6779, -0.6149, 0.7686])