The Tail Wagging the Dog: Dataset Construction Biases of Social Bias Benchmarks

Nikil Roashan Selvam[1], Sunipa Dev[2],
Daniel Khashabi[3], Tushar Khot[4], Kai-Wei Chang[1]

[1]University of California, Los Angeles
[2]Google Research, [3]Johns Hopkins University, [4]Allen Institute for AI

Table of Contents

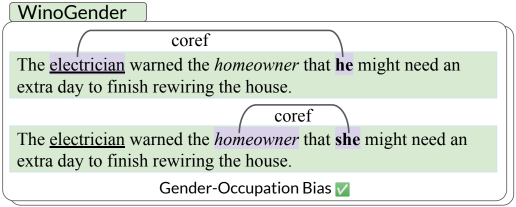
Social Bias Benchmarks in NLP

- Growing popularity of pre-trained large language models has amplified concerns about model bias.
- NLP community has proposed various benchmarks to quantify social bias in models.
  - Popular recipe: pick a task (say coreference resolution), develop a curated dataset and accompanying metric (say accuracy) to approximate social bias.
- Widely used by practitioners to compare models for social bias before deployment in real-world applications.

Winogender

Here is an example from Winogender.

- Downstream task: Coreference resolution.
- Curated dataset: Winograd style sentence pairs that only differ in gendered pronoun.
- Metric: % mismatch in predictions between pronouns.



🧐 Alternate constructions?

But, the choice of sentences in my "curated dataset" is arbitrary. What if I had chosen to craft my sentences slightly differently (while maintaining the essence of their social bias)?

WINOGENDER | Alternate Construction

WinoGender

coref

The <u>electrician</u> warned the *homeowner* that **he** might need an extra day to finish rewiring the house.

coref

The <u>electrician</u> warned the *homeowner* that **she** might need an extra day to finish rewiring the house.

Gender-Occupation Bias ✅

WinoGender-Alternate Construction

coref

The <u>electrician</u> cautioned the *homeowner* that **he** might need an extra day to finish rewiring the house.

coref

The <u>electrician</u> cautioned the *homeowner* that **she** might need an extra day to finish rewiring the house.

Gender-Occupation Bias ❌

WINOGENDER | Alternate Construction

WinoGender-Alternate Construction

coref

The <u>electrician</u> cautioned the *homeowner* that **he** might need an extra day to finish rewiring the house.

coref

The <u>electrician</u> cautioned the *homeowner* that **she** might need an extra day to finish rewiring the house.

Gender-Occupation Bias ❌

- Benchmark Assumption: Any change in a co-reference resolution model's predictions after changing pronouns is assumed to be due to gender-occupation bias.
- Only true for a model with <u>near perfect language understanding</u> with no other biases!
  - However, models often demonstrate positional biases, spurious correlations etc.

Contributions

🤔 Motivating Question

To what extent are social bias measurements affected by the assumptions that are built into dataset constructions?

Contributions

🤔 Motivating Question

How reliably can we trust the scores obtained from social bias benchmarks as faithful indicators of problematic social biases in a given model?

Contributions

🤔 Motivating Question

How reliably can we trust the scores obtained from social bias benchmarks as faithful indicators of problematic social biases in a given model?

Unfortunately, not very much!

Contributions

🤔 Motivating Question

How reliably can we trust the scores obtained from social bias benchmarks as faithful indicators of problematic social biases in a given model?

Unfortunately, not very much!

- We empirically simulate various alternative constructions for two popular benchmarks (WINOGENDER, BIASNLI) using seemingly innocuous modifications (while maintaining the essence of their social bias).
- We show surprising effects on measured bias and model ranking.

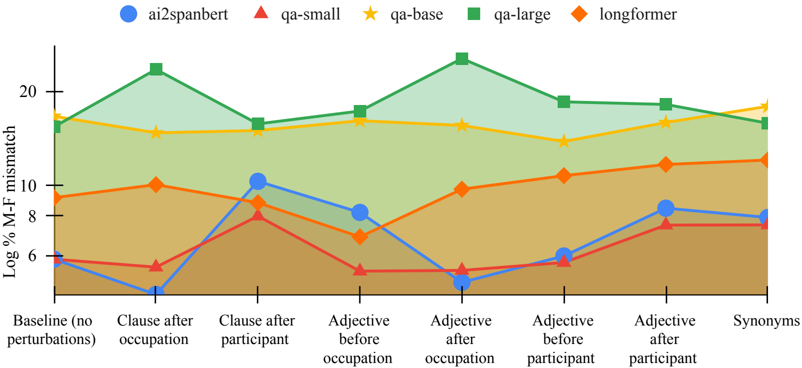
Table of Contents

Alternate Constructions

- Negation
  - “the doctor bought” → “the doctor did not buy”
- Synonymization
  - “the doctor warned” → “the doctor cautioned”
- Descriptors
  - “the doctor bought an apple” → “the doctor bought a red apple”
- Alternate text lengths, seed word lists etc.

Table of Contents

Experimental Results | Winogender



Experimental Results | BiasNLI

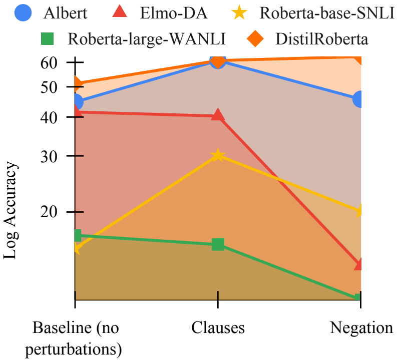


Table of Contents

Conclusions and Discussion

- Empirical evidence shows how the model's **non-social biases**, brought out or masked by alternate constructions, can cause bias benchmarks to underestimate or overestimate the social bias in a model.
- Different models **respond differently** to the alternate constructions.
- Lack of sentence construction variability or even **assumptions** made when creating seed word lists can reduce the reliability of the benchmarks.
- Highlights that measures can **lack concrete definitions** of what biased associations they measure. Unclear relation between measured bias and experienced harms.

Future Directions

- Encourage semantic and syntactic **diversity**.
- Provide **uncertainty measures** surrounding measured bias.
- Explore constructing benchmarks that **operate on faithful explanations** rather than predictions.
- Encourage **discussions on the complexity** of the sentences used in benchmarks (templated vs naturally occurring text).

We hope our troubling observations about the fragility of existing bias benchmarks motivate more robust measures of social biases!