Identifying and Reducing Gender Bias in Word-Level Language Models

Shikha Bordia Yu Wang Jason Cramer Samuel R. Bowman
New York University

Problem Formulation

- Most text corpora exhibit implicit gender bias.
- The machine learning models that are trained using such text data exhibit similar and amplified bias in their predictions.
- Towards this pursuit, we introduce a regularization term to reduce bias in word-level language models trained on biased text.

Quantifying Bias and de-biasing the Language Model

Bias Regularization

- We propose a bias regularization term that penalizes the projection of embeddings learned by the model onto the gender subspace.

$$\mathcal{L}_B = \lambda \|NB\|_F^2$$

- λ controls the importance of minimizing bias in the embedding matrix.

Bias Measure

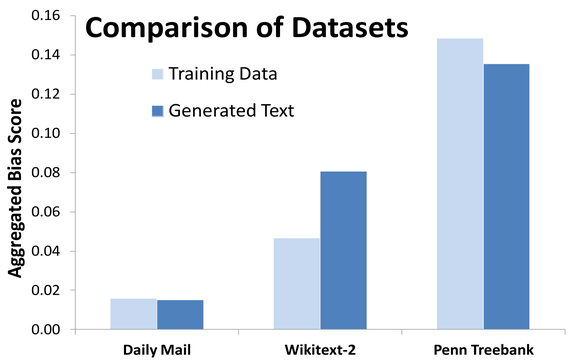
- The conditional probability of a word given a specific gendered word ***g*** is defined as below.

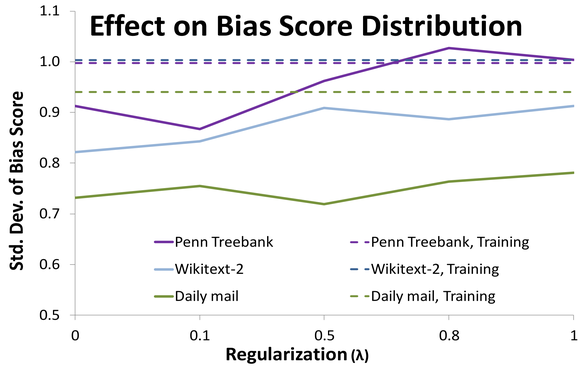
$$P(w|g) = \frac{P(wg)}{P(g)} = \frac{C(wg)/\sum_i C(w_i g)}{C(g)/\sum_i C(w_i)}$$

- We define bias measure *as*

$$b = \log \frac{P(w|f)}{P(w|m)}$$

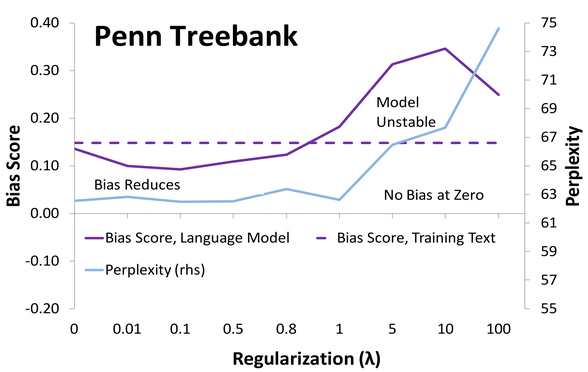
Experiments and Observations

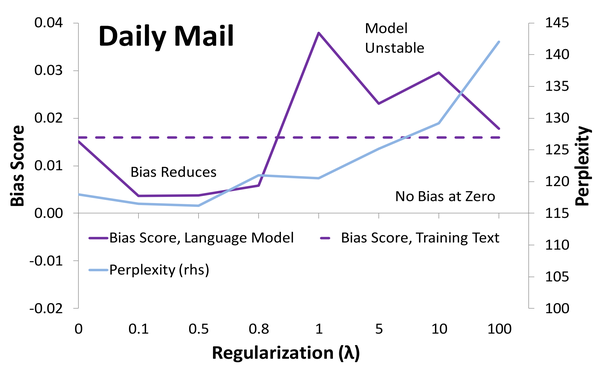


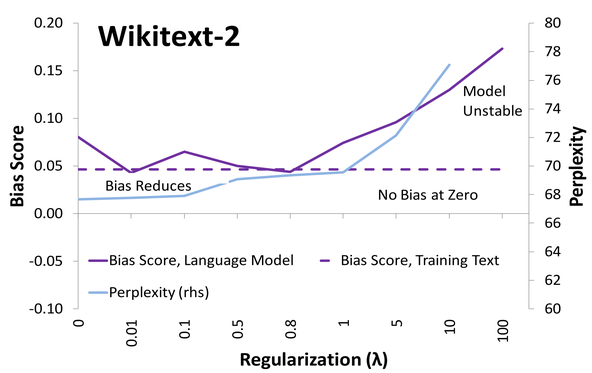


Conclusions

- Perplexity bias trade-off
- Effect of de-biasing encoder versus decoder embeddings
- Relative Bias Scores of Datasets: Daily Mail < Wikitext-2 < Penn Treebank







Future Work

- Data Augmentation : Men and women at work
- Debiasing multi-faceted biases e.g. racial biases
- Improving Bias Metric
- Modifying conditional probabilitities: Token level smoothing

References

Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *In Advances in Neural Information Processing Systems.*

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models. *CoRR abs/1708.02182.*