A Dataset for Tackling Gender Bias in Text

Yasmeen Hitti[1], Eunbee Jang[1], Ines Moreno[1], Carolyne Pelletier[1] and Jasleen Ashta[2]

[1] Mila [2] Independent (Equal contributions [1] and [2])
biasly4good@gmail.com

Gender bias is found in personal conversations, in the media, in historical writings, popular culture, in the labor force, household responsibilities and now in machines (Fiebert and Meyer, 1997; Kingdon, 2005). Gender bias occurs in machine learning models when they are trained with data that contains human-like biases (Haussler, 1988). Current research is focused on detecting and correcting for gender bias in existing machine learning models, such as word embeddings (Zhao et al., 2018: Bolukbasi et al., 2016), coreference resolution (Zhao et al., 2018) and visual recognition tasks involving language (captioning) (Zhao et al., 2017). Rather than removing gender bias in current machine learning models, we are tackling the issue at its root and creating a gender bias dataset with which to train a machine learning model. Enabling a model to learn gender bias would allow for gender bias detection and possibly correction in text.

To the best of our knowledge, there is no existing gender bias dataset to encompass all fields where bias is present. Therefore it is necessary to use a linguistic approach and analyze the construction of sentences to build a clean dataset. As a proof of concept, biaslyAI.com was created as a platform to crowdsource labels; the sentences were labeled as gender biased or non gender biased. The sentences presented to the human labelers were scraped from news articles and online magazines. In the span of a single week we reached 365 participants with each labeling 10 random sentences. A sentence received a final label after it met the confidence interval set to 80% (cf. Table 1). The goal of this preliminary experiment was to collect a small set of data and to analyze the accuracy of the labeling.

Sentences	**Labeled Gender Bias**	**Labeled Non Gender Bias**	**Final Label**
Studies have also shown that women do not enjoy the same level of professional status as male physicians.	60%	40%	More labels needed
Men feel secure in knowing that their partners approve of them and where they are in their career.	57%	43%	More labels needed
Existing research shows gender roles can harm both sexes.	13%	87%	Non Gender Bias
In general, a nanny is concerned about her reputation amongst parents.	91%	9%	Gender Bias

Table 1: Shows a sample of the results of our prototype experiment conducted at biaslyAI.com 'More labels needed' indicates that the sentence has not met the confidence interval of 80%

To further this proof of concept, we are investigating techniques to refine the data collection and labeling process. The first step is finding a baseline definition of gender bias by outreaching to sociologists, linguists, psychologists, gender-related studies and any other relevant field. This baseline definition will be used to better guide our labelers, since gender bias is not fully understood by the vast majority of people, as seen in Table 1 (Alvesson and Billing, 2009). Data collection can be achieved in different ways; for example, through web scraping or data augmentation with existing text data. Data augmentation techniques on existing text data can be applied using a linguistic construct to breakdown the sentences and replace words or groups of words with gender related content (Van Dyk and Meng, 2001). A first linguistic construct approach to explore is semantic role labeling; this technique allows for a better understanding of the arguments of the predicates in the sentences (Bjrkelund al., 2009). Once semantic roles have been established within the data, the relationship between the structure of the sentence and its content can be analyzed. To improve our labeling, we propose a best-worst scaling technique rather than having a binary choice. Best-worst scaling is a known measurement method for rating or paired comparisons, and will help identify the strongest bias sentences amongst the set shown to the labeler (Louviere and Flynn, 2010). More precisely, best-worst scaling with 4-tuples is known for its efficiency in annotating and can reveal 5 different relationships amongst sentences presented. In our case, 4 sentences such as A, B, C and D, if A is the most gender bias and D is the least gender bias then A > B, A > C, A > D, B > D, and C > D (Mohammad, 2017).

A clear definition of gender bias from experts is crucial. With a linguistic approach (i.e. semantic role labeling), data collection can be enhanced and augmented. Our labeling technique can be improved by educating our labelers with the obtained definition and by using best-worst scaling. A model trained on a robust gender bias dataset could directly address the negative preconceived ideas people have about gender and guide human judgments by recognizing their gender biases. This could then initiate reflections on gender-sensitive topics and empower the movement of fairness and equity for all genders.

Acknowledgments

We would like to acknowledge the AI for Social Good Summer Lab of 2018 (aiforsocialgood.ca) and their mentors for their support and guidance throughout the design of the biaslyAI.com prototype.

References

Alvesson, M., & Billing, Y. D. (2009). Understanding gender and organizations. Sage.

Bjrkelund, A., Hafdell, L., & Nugues, P. (2009, June). Multilingual semantic role labeling. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task (pp. 43-48). Association for Computational Linguistics.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems (pp. 4349-4357).

Fiebert, M. S., & Meyer, M. W. (1997). Gender stereotypes: A bias against men. The Journal of psychology, 131(4), 407-410.

Haussler, D. (1988). Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. Artificial intelligence, 36(2), 177-221.

Kingdon, G. G. (2005). Where has all the bias gone? Detecting gender bias in the intrahousehold allocation of educational expenditure. Economic Development and Cultural Change, 53(2), 409-451.

Louviere, J. J., & Flynn, T. N. (2010). Using best-worst scaling choice experiments to measure public perceptions and preferences for healthcare reform in Australia. The Patient: Patient-Centered Outcomes Research, 3(4), 275-283.

Mohammad, S. M. (2017). Word affect intensities. arXiv preprint arXiv:1704.08798.

Van Dyk, D. A., & Meng, X. L. (2001). The art of data augmentation. Journal of Computational and Graphical Statistics, 10(1), 1-50

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:1804.06876.

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. W. (2018). Learning Gender-Neutral Word Embeddings. arXiv preprint arXiv:1809.01496.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457.