Word embeddings quantify 100 years of gender and ethnic stereotypes

Anonymous Author(s)
email

Abstract

1 Summary

Several researchers demonstrated in the past that word embeddings can be biased [10, 4] and offered solutions to debias them [3]. Garg et al. [6] embrace this shortcoming and use it as a quantative measure of the temporal changes in stereotypes towards women and ethnic minorities. They demonstrate that the embedding bias (the average distance of one group to some words from a fixed word list, e.g. jobs, minus the average distance of another group) correlates with demographic and societal measures over time.

2 Critique

Overall, I really like the novel idea of exploiting the embedding bias to quantify stereotypes. Although their methodology has some shortcomings (that they mostly address and justify), they show the validity of this measure and therefore answer the main research question. The Corpus of Historical American English and Google Books dataset seem to be well-suited for the historical temporal analysis as they are very large. Furthermore, special care has been taken to avoid selection bias issues according to the authors. However, they do not explain or point to literature that goes into detail which methods the creators used (and how they validated them) to avoid selection bias, so it is quite hard to judge this statement. It is not clear to me how representative the dataset is for the social attitude of the time (because it contains for instance mainly texts by professional authors which may have different attitudes), but this issue is hard to avoid in general. For more recent years, it would be interesting to also incorporate other text sources besides the Google News dataset (e.g., sources like social media, which may be more representative of the general population) and analyze if the results change. I am not convinced that the Amazon Mechanical Turk experiment measures gender stereotypes among the population well. Because these experiments are usually poorly paid [9], I would expect that low-income individuals (with potentially different stereotypes) are severely overrepresented. A similar argument holds for the Princeton trilogy. College students (more likely to have a high socieconomic status background [2] and an above-average IQ [8]) may not be representative for the whole population. However, they pointed out most of these limitations (although I would have liked a more elaborate discussion of them) and used different datasets, which may average out these effects. The plots are very helpful to support their claims visually and I really liked that they included confidence intervals as I wondered while reading how strong the reliance on the chosen word groups is. I would have still liked it if they compiled the word lists more systematically and documented this procedure (like they did with the surnames) or even externally validated them. Although they did not do this, they at least suggested it for further work ("These were hand-coded from the overall list of occupations, follow-on work should study this more systematically" and "A more complete analysis would first collect externally validated lists of words that describe each such dimension and then measure the embedding association with respect to these lists over time.").

Submitted as response essay for the *Sequencing Legal DNA: NLP for Law and Political Economy* course at ETH Zurich. Do not distribute.

In some parts of the paper, I had the impression of slight confirmation bias on their side. In the discussion of Figure 4, they point out that transition in the gender embeddings from 1960 to 1970 is statistically significant and attribute this to the strong influence of the women's movement. However, when we look at the results in Table B.23, we can see that the results in two other decades (1940-1950 and 1970-1980) are statistically significant as well. They do not address these decades in the discussion, whereas they explicitly mention for Figure 5 / Table B.24 that all significant shifts correspond to historical events. A discussion and analysis of the two other significant shifts for the gender embeddings would be a valuable addition, in my opinion.

It would have been very interesting to compare their framework with other automated approaches to quantify stereotypes such as simpler, lexicon-based ones. However, there has been further research in this direction which came to the conclusion that end-to-end deep learning models outperform lexicon-based approaches [5]. Furthermore, similar to their analysis of the reliance on the relative norm bias metric (compared with cosine similarity), an interesting investigation would be the reliance on the chosen metric for measuring bias (i.e., the difference of the average distances). Alternatives to examine would be the projection on the gender direction or the values on reserved gender dimensions, as described by Cryan et al [5].

3 Further Work

In my opinion, they have convincingly demonstrated that their framework is useful to quantify stereotypes and it would be very interesting to apply it to different datasets and answer different research questions. For instance, one could investigate if gender stereotypes are stronger on social networking sites (as suggested by other research [1]) or if they differ among voters with different political attitudes, which other researchers also pointed out [7]. Consistent results in these investigations would further support the validity of the framework.

References

[1] Jane Bailey et al. "Negotiating With Gender Stereotypes on Social Networking Sites: From "Bicycle Face" to Facebook". en. In: *Journal of Communication Inquiry* 37.2 (Apr. 2013), pp. 91–112. ISSN: 0196-8599. DOI: 10.1177/0196859912473777.

[2] Alanna Bjorklund-Young. "Family Income and the College Completion Gap". In: *Johns Hopkins Institute for Education Policy* (2016).

[3] Tolga Bolukbasi et al. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., Dec. 2016, pp. 4356–4364. ISBN: 978-1-5108-3881-9.

[4] Marc-Etienne Brunet et al. "Understanding the Origins of Bias in Word Embeddings". en. In: *International Conference on Machine Learning*. PMLR, May 2019, pp. 803–811.

[5] Jenna Cryan et al. "Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–11. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376488.

[6] Nikhil Garg et al. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes". en. In: *Proceedings of the National Academy of Sciences* 115.16 (Apr. 2018), E3635–E3644. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1720347115.

[7] Ted G. Jelen. "The Effects of Gender Role Stereotypes on Political Attitudes". en. In: *The Social Science Journal* 25.3 (Jan. 1988), pp. 353–365. ISSN: 0362-3319. DOI: 10.1016/0362-3319(88)90036-5.

[8] Walter T Plant and Harold Richardson. "The IQ of the Average College Student." In: *Journal of Counseling Psychology* 5.3 (1958), p. 229.

[9] Alana Semuels. *The Internet Is Enabling a New Kind of Poorly Paid Hell*. en. https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/. Jan. 2018.

[10] Jieyu Zhao et al. “Gender Bias in Contextualized Word Embeddings”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 629–634. DOI: 10.18653/v1/N19-1064.