Did You Just Assume My Vector?
Detecting Gender Stereotypes in Word Embeddings

Amir Bakarov
National Research University Higher School of Economics, Moscow, Russia
amirbakarov@gmail.com

Abstract. Recent studies found out that supervised machine learning models can capture prejudices and stereotypes from training data. Our study focuses on the detection of gender stereotypes in relation to word embeddings. We review prior work on the topic and propose a comparative study of existing methods of gender stereotype detection. We evaluate various word embeddings models with these methods and conclude that the amount of bias does not depend on the corpora size and training algorithm, and does not correlate with embeddings performance on the standard evaluation benchmarks.

Keywords: Word embeddings · Gender stereotypes

1 Introduction

Word embeddings (real-valued word representations produced by neural distributional semantic models) are ubiquitous tools in contemporary NLP. They are mostly treated as black-boxes, and it is unclear how to evaluate them. Recent studies propose various approaches to their evaluation [1,2]. We suppose, though, that certain properties of word embeddings were not considered in recent evaluation studies. One of them is the amount of bias in word vector spaces.

The concept of bias in machine learning commonly refers to prior information [3], but in this work, we will use the notion of bias from ethics studies, which is regularities in training data relating to prejudices or stereotypes about a person’s physiological features (race, gender, orientation, etc.). Training data historically contains implicit stereotypes and discrimination, so the supervised machine learning models unintentionally capture this bias [4]. It can cause unfair decisions in the model for certain sets of a person’s characteristics – for instance, while deciding whether to approve a personal loan [5].

It is usually hard to track such stereotypes in training data for supervised models. For instance, corpus statistics can contain bias that even does not exist in the cognition of the language speakers. The collocation black sheep can be more frequent than the collocation white sheep in the corpora, but in the real world, black sheep are rarer than white ones [6].
Word embeddings tend to frequently capture these unobvious statistics in the data. Such correlations inevitably appear in the resulting models, so the word vectors produce unfair distances between words characterizing men and women. For example, the situation when the word *researcher* is closer to the word *man* than *woman* can be considered as unfair, since there is no reason why women should be less related to research than men. Such situations make the work of systems based on word embeddings “biased”: for instance, in the search engine system the query “researchers” can give results more related to men than to women, and then it could be even harder for women to become recognized as researchers. This problem can be solved from the corpus side as well as from the model side, but the main issue is how to evaluate whether the model contains such bias (and how much of it).

The current paper aims to overview recent advances in the detection of gender stereotypes and to propose a comparative analysis of existing gender bias detection methods, finding out whether the amount of bias depends on the diverting factors (corpus size and training algorithm), or embeddings performance on the traditional evaluation benchmarks (word similarity and analogy reasoning). The contribution of our work is to structurize recent studies related to gender bias in word embeddings, to overview existing metrics for gender bias detection, and to compare them with each other. We are the first to propose a comprehensive comparison of gender bias detection metrics for word embeddings.

This work is organized as follows. In Sect. 2 we survey the recent works related to the problem of gender bias in word embeddings and describe existing methods for detecting gender bias. In Sect. 3 we describe our evaluation experiments and make a discussion of the obtained results, while Sect. 4 concludes the paper and reveals our plans on future work on the subject.

# Gender Bias in Word Embeddings

## 2.1 Related Work

In the field of NLP the problem of bias (and, particularly, gender bias) was covered for various tasks, such as text classification [7], machine translation [8], and recommendation systems [9]. We refer an interested reader to a comprehensive review of bias in NLP [10], and narrow the scope of the current paper to gender bias in word embeddings.

For word embeddings, the problem of bias was firstly introduced in 2015 by a blog post that highlighted the existence of gender unfairness in word vector spaces [11]. Most of the following work was trying to propose algorithms for removing bias in word embeddings: the main effort was to remove bias without hurting their semantic content (i.e., without hurting their performance on the standard benchmarks or downstream tasks).

The first work addressing the problem of measuring and removing gender bias in word embeddings was a work by [12], which proposed two metrics for removing gender bias, and a debiasing method based on the linear transformation of word vector spaces. These metrics were used for evaluation in almost
all the following studies. However, [13] noted that human-like semantic biases are reflected through ordinary language, and proposed another metric based on the association tests. [14] suggested another way to generate association tests, supposing that they should be assessed by humans instead of being used automatically.

[15] introduced a triple loss function that penalizes the model for incorrect distances of word pairs related to gender. [16] investigated cultural tracks of gender biases in word embeddings, and concluded that semantic systems of word vector space represent cultural categories rather than biases or distortions. [17] explored semantic shifts from the perspective of bias, and also concluded that the origin of such shifts can reflect the changes in cultural patterns. [18] developed a technique to trace the origin of bias in embeddings back to the original text. [19] suggested using linear projection of all words based on vectors captured by common names. [20] proposed a general framework for unbiasing word embeddings, considering other types of bias rather than only gender. [21] created a neutral-network-based encoder-decoder framework to remove gender biases from gender-stereotyped words and to preserve gender-related information in feminine and masculine words. [22,23] pointed out on the problem of gender bias in contextualized word embeddings, such as ELMo. [24] presented a debiasing method for Hindi, while [25] proposed a technique for the German language. The problem of bias in word embeddings was also investigated in [20,26–31].

2.2 Evaluation Metrics

To the best of our knowledge, there are only 3 fully automatic evaluation methods available. We do not consider non-automatic methods that require additional human assessment: for example, the “Analogies exhibiting stereotypes” score, which suggests generating analogies judged by humans [12], or clustering plots, considered as classification of biased neutral words that also need further interpretation [26].

Occupational Stereotypes (OS) [12]. This method explores whether the word vectors contain stereotypes connected to occupations, by quantifying a neutral word’s vector position on the axis between vectors corresponding to the most representative female and male words (so-called “gender axis”; it was based on the words “she” and “he” in the original study). Given a list of neutral word vectors $N$ (the authors used occupation words, e.g. “programmer”), the projection onto gender axis is computed

\[ proj_x = x \cdot \frac{(v_{female} - v_{male})}{||v_{female} - v_{male}||_2}, \]

1 Here and after we will use L2-normalized Euclidian distance in all cases where we measure the distance between the vectors: assuming that the vectors are normalized, the choice between Euclidian distance and cosine similarity does not affect their results [17].
where \( x \) is the vector of the neutral (occupation) word, \( v_{female} \) and \( v_{male} \) are vectors of gender words (e.g. “she” and “he”). The final score is computed as a mean of absolute projection values for a set of occupational words \( X \):

\[
OS = \text{mean}_{x \in X} \text{proj}_x = \sum_{x \in X} \frac{\text{proj}_x}{|X|}
\]

The lower this score, the less bias the model contains. Notably, since bias in other scores has a direct dependency, we use an inverted version of \( OS \):

\[
IOS = 1 - OS
\]

**Gender Difference (GD)** [17]. The method quantifies bias with a so-called “gender vector”, calculated as the average of the set of the most representative words of each gender (e.g. “she”, “mother”, “woman” for female):

\[
GD = \sum_{v_m \in M} ||v_m - v_1||_2 - ||v_m - v_2||_2,
\]

where \( M \) is the set of the neutral word vectors, \( v_1 \) is the averaged vector of the first gender vectors, and \( v_2 \) is the average vector of the second gender vectors. The higher the \( GD \) score, the more biased the model is.

**Word Embedding Association Test (WEAT)** [13]. The method uses vector similarity to measure the association of two given sets of neutral target words (e.g. occupation words), with two sets of gender attributes words (“she”, “mother”, “woman”, etc.). The score is computed as a probability that a random permutation of the attribute words would produce the observed (or greater) difference in sample means:

\[
WEAT = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B),
\]

where \( X \) and \( Y \) are two sets of target words of equal size, and the \( s(w, A, B) \) statistic measures the association of \( w \) with the two sets of attribute words \( A \) and \( B \):

\[
s(w, A, B) = \text{mean}_{a \in A} ||w - a||_2 - \text{mean}_{b \in B} ||w - b||_2,
\]

The higher the \( WEAT \) score, the more biased the model.

### 3 Comparison

To empirically compare the aforementioned methods, we evaluate them with a set of pre-trained word embeddings. We used a Word2Vec (continuous skip-gram) model [32] trained on British National Corpus (BNC, 160k tokens), Gigaword (300k tokens), Google News (3M tokens), and Wikipedia (300k tokens), as well
as GloVe [33] trained on Gigaword and Wikipedia [34]. For OS we used the
dataset of 320 “neutral” words of occupations, along with a dataset of 6 male-
female word pairs from [12]. Notably, the “gender” pairs set used in our study is
a modified version of the original dataset, since we dropped all pronouns which
were out-of-vocabulary in some of our models. Also, unlike the original study,
we used projection on the “female”-“male” axis instead of the “she”-“he” axis due
to the same reason. These 2 datasets were also used to compute the GD score.
For the WEAT score we used the original dataset proposed by the authors of
this metric without any changes [13].

Following the experimental setting of [12], we used 4 standard word embed-
ddings evaluation benchmarks: RG [35], RW [36], WS-353 [37] and MSR [38].
The first three measure whether related words have similar embeddings, while
the latter measures how well the embedding performs in analogy tasks [39].

The results of our evaluation experiments are presented in Table 1 (best
results are highlighted in bold). Surprisingly, the variation in the scores for GD
is quite low, unlike the scores for 6 other benchmarks. Considering that a “good”
model should have good scores on the standard benchmarks, we note that the
amount of bias does not depend on the performance of the model. Henceforth,
we can conclude that even a model that is good, from the perspective of standard
evaluation benchmarks, can have prejudices and stereotypes.

The results also suggest that while models tend to generalize better on a
larger amount of data, it does not mean that a model trained on larger corpora
will contain more bias: for instance, according to the WEAT score, the lowest bias
have models trained on BNC (the smallest corpora) and GoogleNews (the largest
corpora). The amount of bias also does not depend on the training algorithm,
and even two different methods for measuring bias can have different scores (the
two sets of scores for OS and WEAT have Spearman rank correlation coefficient
of −0.2).

The code for reproducing the experiments can be found at https://gitlab.com/bakarov/fair-embeddings.

|                  | OS  | WEAT | GD  | RG  | RW  | WS  | MSR |
|------------------|-----|------|-----|-----|-----|-----|-----|
| GloVe, Gigaword  | 0.89| 0.67 | 0.05| 0.74| 0.41| 0.64| 0.08|
| GloVe, Wikipedia | 0.91| 0.54 | 0.05| 0.81| 0.4 | 0.72| 0.1 |
| Word2Vec, BNC    | 0.89| 0.37 | 0.05| 0.75| 0.37| 0.75| 0.01|
| Word2Vec, Gigaword| **0.92**| 0.6 | 0.05| 0.65| **0.53**| 0.75| 0.05|
| Word2Vec, GoogleNews | **0.92**| 0.35| **0.06**| 0.75| 0.51| 0.74| 0.04|
| Word2Vec, Wikipedia | 0.9 | 0.48 | 0.05| **0.81**| 0.51| **0.79**| 0.07|

Table 1. Evaluation of a set of word embeddings using gender bias metrics (3 first columns) and standard evaluation metrics (4 last columns).
4 Conclusions and Future Work

Our work analyzed the recent methods for measuring gender-oriented biases in word embeddings, and compared the evaluation scores with word embeddings performance on the standard evaluation benchmarks. The results report that the amount of bias does not depend on the performance against these benchmarks, as well as of the training corpora and training algorithms. Results for the different gender bias evaluation methods also do not correlate with each other, so we can conclude that currently, we are unable to unequivocally say which word embeddings model is more biased, and moreover, we cannot reveal the amount of bias from the corpora size or the training algorithm. Any shifts in the embedding model are possible for several reasons. First, shifts are incorporated in the data itself, including collocations of frequency words or cultural phenomena, and second, the training method could be sensitive to certain correlations. Therefore, our study has supported the idea that we cannot capture these correlations (particularly, related to gender) by analyzing external model factors (e.g. training corpora size), and need specific (gender bias detection) methods for this task.

In the future, we suggest extending our research of these methods to the development of new ones, as well as to evaluation experiments in other settings (e.g. other corpora, other training algorithms, or other evaluation benchmarks). Also, in the current study, we used a set of pre-trained models, for which we cannot track the dependence of the results on the random initialization of weights during the training. We plan to train our models and measure the stability of the metric during several training iterations of the same model on the same corpus. Finally, we want to test the dependence of the results of the choices of the gender-specific words, and plan to extend our experiments to other languages (particularly, Russian) by creating datasets for them.

Acknowledgments. The reported study was funded by the Russian Foundation for Basic Research project 20-37-90153 “Development of framework for distributional semantic models evaluation”.

References

1. Rogers, A., Hosur Ananthakrishna, S., Rumshisky, A.: What’s in your embedding, and how it predicts task performance. In: Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, Association for Computational Linguistics, pp. 2690–2703, August 2018
2. Senel, L.K., Utlu, I., Yucesoy, V., Koc, A., Cukur, T.: Semantic structure and interpretability of word embeddings. arXiv preprint arXiv:1711.00331 (2017)
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
4. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1521–1528. IEEE (2011)
5. Hardt, M., Price, E., Srebro, N., et al.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, pp. 3315–3323 (2016)
6. Gordon, J., Van Durme, B.: Reporting bias and knowledge extraction (2013)
7. Wagner, C., Garcia, D., Jadidi, M., Strohmaier, M.: It’s a man’s Wikipedia? Assessing gender inequality in an online encyclopedia. In: ICWSM, pp. 454–463 (2015)
8. Font, J.E., Costa-jussà, M.R.: Equalizing gender biases in neural machine translation with word embeddings techniques. arXiv preprint arXiv:1901.03116 (2019)
9. Mishra, A., Mishra, H., Rathee, S.: Examining the presence of gender bias in customer reviews using word embedding. arXiv preprint arXiv:1902.00496 (2019)
10. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: a critical survey of “bias” in NLP. arXiv preprint arXiv:2005.14050 (2020)
11. Schmidt, B.: Rejecting the gender binary: a vector-space operation (2015)
12. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in Neural Information Processing Systems, pp. 4349–4357 (2016)
13. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science 356(6334), 183–186 (2017)
14. Swinger, N., De-Arteaga, M., NeilThomasHeffernan, I., Leiserson, M.D.M., Kalai, A.T.: What are the biases in my word embedding? CoRR abs/1812.08769 (2018)
15. Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.W.: Learning gender-neutral word embeddings. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4847–4853 (2018)
16. Kozlowski, A.C., Taddy, M., Evans, J.A.: The geometry of culture: analyzing meaning through word embeddings. arXiv preprint arXiv:1803.09288 (2018)
17. Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc. Nat. Acad. Sci. 115(16), E3635–E3644 (2018)
18. Brunet, M.E., Alkalay-Houlihan, C., Anderson, A., Zemel, R.: Understanding the origins of bias in word embeddings. In: International Conference on Machine Learning, pp. 803–811 (2019)
19. Dev, S., Phillips, J.: Attenuating bias in word vectors. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 879–887 (2019)
20. Lauscher, A., Glavaš, G., Ponzetto, S.P., Vulić, I.: A general framework for implicit and explicit debiasing of distributional word vector spaces. arXiv preprint arXiv:1909.06092 (2019)
21. Kaneko, M., Bollegala, D.: Gender-preserving debiasing for pre-trained word embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1641–1650 (2019)
22. Hoyle, A.M., Wolf-sonkin, L., Wallach, H., Augenstein, I., Cotterell, R.: Unsupervised discovery of gendered language through latent-variable modeling. In: 57th Annual Meeting of the Association for Computational Linguistics Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 1706–1716 (2019)
23. Basta, C., Costa-jussà, M.R., Casas, N.: Extensive study on the underlying gender bias in contextualized word embeddings. Neural Comput. Appl. 1–14 (2020). https://doi.org/10.1007/s00521-020-05211-z
24. Pujari, A.K., Mittal, A., Padhi, A., Jain, A., Jadon, M., Kumar, V.: Debiasing gender biased Hindi words with word-embedding. In: Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, pp. 450–456 (2019)
25. Papakyriakopoulos, O., Hegelich, S., Serrano, J.C.M., Marco, F.: Bias in word embeddings. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 446–457 (2020)

26. Gonen, H., Goldberg, Y.: Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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), pp. 609–614 (2019)

27. Shin, S., Song, K., Jang, J., Kim, H., Joo, W., Moon, I.C.: Neutralizing gender bias in word embedding with latent disentanglement and counterfactual generation. arXiv preprint arXiv:2004.03133 (2020)

28. Gyamfi, E.O., Rao, Y., Gou, M., Shao, Y.: deb2viz: debiasing gender in word embedding data using subspace visualization. In: Eleventh International Conference on Graphics and Image Processing (ICGIP 2019), vol. 11373, p. 113732F. International Society for Optics and Photonics (2020)

29. Wang, T., Lin, X.V., Rajani, N.F., McCann, B., Ordonez, V., Xiong, C.: Double-hard debias: tailoring word embeddings for gender bias mitigation. arXiv preprint arXiv:2005.00965 (2020)

30. Kumar, V., Bhotia, T.S., Kumar, V., Chakraborty, T.: Nurse is closer to woman than surgeon? Mitigating gender-biased proximities in word embeddings. Trans. Assoc. Comput. Linguist. 8, 486–503 (2020)

31. Rios, A., Joshi, R., Shin, H.: Quantifying 60 years of gender bias in biomedical research with word embeddings. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, pp. 1–13 (2020)

32. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)

33. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

34. Kutuzov, A., Fares, M., Oepen, S., Velldal, E.: Word vectors, reuse, and replicability: towards a community repository of large-text resources. In: Proceedings of the 58th Conference on Simulation and Modelling, pp. 271–276. Linköping University Electronic Press (2017)

35. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Commun. ACM 8(10), 627–633 (1965)

36. Luong, M.T., Socher, R., Manning, C.D.: Better word representations with recursive neural networks for morphology. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pp. 104–113 (2013)

37. Agirre, E., Alfonseca, E., Hall, K., Kravalová, J., Pasca, M., Soroa, A.: A study on similarity and relatedness using distributional and wordnet-based approaches. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 19–27 (2009)

38. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 746–751 (2013)

39. Bakarov, A.: A survey of word embeddings evaluation methods. arXiv preprint arXiv:1801.09536 (2018)