Validity Examinations of Social Bias Measurements and Mitigations in Word Embeddings

Lu Wang, Jina Huh-Yoo
Drexel University

ABSTRACT

With the widespread usage of word embeddings in Natural Language Processing (NLP) applications, social bias inherited by word embeddings would result in unaware discrimination and unfairness. To investigate the achievements and insufficiency of social debiasing in word embeddings, we examined the validity of measurements of social bias in word embeddings and compared the mitigation methods. We collected 146 papers from Web of Science and 168 papers from arXiv, filtered the duplicates and selected 48 papers for review based on our criteria. Results found that most existing research focused on gender bias while not much research discussed other social bias like race bias; for internal criterion-related validity, most studies applied multiple measurements to evaluate the debiasing results. However, little evidence was available to show whether they measured the same aspects of the bias. For internal construct validity, most social biases were measured through projection-based approach along one axis with two binary extremes. For external validity, evidence is needed for the effectiveness of the measurements for other datasets and word embeddings. This survey will inform the practitioners of the achievements in fairness of word embeddings and inspire future work.

INTRODUCTION

Word embeddings map words into metric vectors and capture semantic information (Mikolov, T., et al., 2013b).

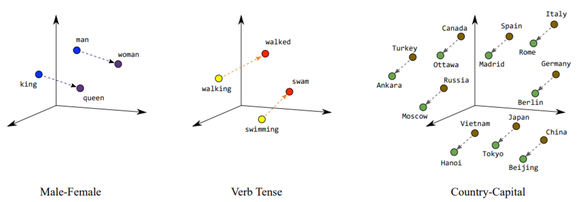


Figure 1. Semantic relationships captured by word embeddings.

Fig from Use Pre-trained Word Embedding to detect real disaster tweets | by Zeineb Ghrib | Towards Data Science

However, social bias inherited by word embeddings such as gender bias, race bias, socioeconomic status bias, and age bias. Below is an example of gender bias (Bolukbasi, T., et al., 2016) .

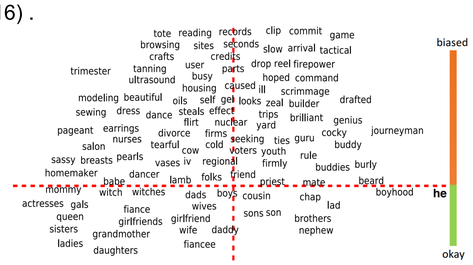


Figure 2. An example of visualized gender bias of word embeddings. From (Bolukbasi, T., et al., 2016).

To investigate the achievements and insufficiency of social debiasing in word embeddings, we examined the validity of measurements of social bias in word embeddings and compared the mitigation methods.

METHODS

On April 7th, 2022, we collected **146** papers from Web of Science through a search query: ***((((AB=(bias*)) OR AB=(debias*)) OR AB=(fairness)) OR AB=(stereotype*)) AND (AB=("word embedding*") OR AB=("word vector*"))***, and **168** papers from arXiv through a search query: ***AND abstract=bias OR biases OR debias OR debiasing OR fairness OR stereotype OR stereotypes; AND abstract="word embedding" OR "word embeddings" OR "word vector" OR "word vectors"***.
The PRISMA diagram is shown below, created through an online tool (Haddaway et al., 2022) :

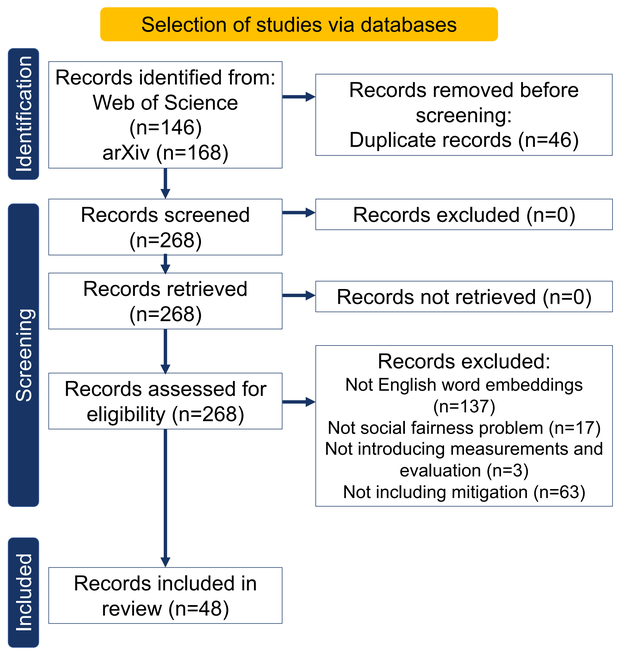
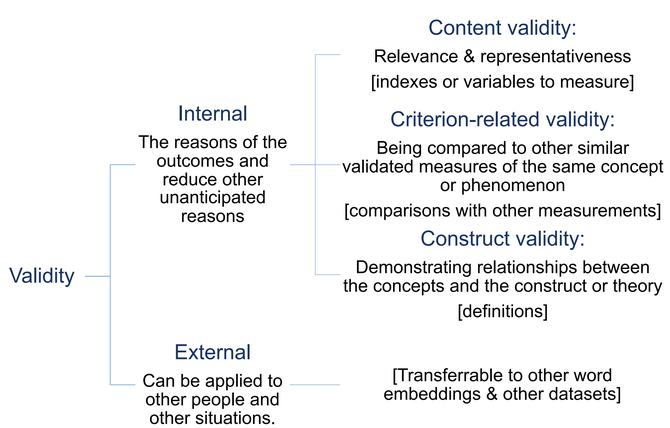


Figure 3. The PRISMA diagram of the review process.

We followed the framework of validity to examine the social debiasing approaches of word embeddings.



RESULTS-1

Internal validity:

❖ **Content validity:**

- The results of Table 1 showed that most studies focused on gender bias, including 36 out of 48 studies. Eight studies investigated multiple social bias such as gender bias, race bias, and religion bias. Less studied social bias were name bias, political bias, and religion bias.

Table 1. The distributions of social bias and the targeted attributes among 48 studies

Content validity	Gender	Name	Politics	Religion	Multi-biases
Occupations (O)	14	1			
Pre-defined (P)	17		1	1	7
Sentiment (Se)	2	1			
Science (Sc)	1				
O + polarity (e.g., evil, good)					1
O+P	1				
P + Se	1				
Sum	36	2	1	1	8

❖ **Criterion-related validity:**

- 8 out of 48 studies applied single measurement or evaluation method, while 39 out of 48 studies applied multiple measurements to evaluate the social bias integrated in the word embeddings. One study designed a human-in-the-loop approach to support interactive debiasing.
- The main measurements were (1) established metrics and tasks such as word similarity tasks and analogy tasks, (2) downstream tasks such as coreference resolution and classification, (3) visualization methods to cluster and examine the neighbors of the words, (4) designed automatic evaluations and tests specific for bias detection such as Word Embedding Association Test (WEAT), and (5) human ratings.

❖ **Construct validity:**

- Researchers applied different ways to design the construct of social bias such as through the ideal equal natures of biased words to a certain word set or project a certain word set on to bias subspaces, as the Table 2 shows.

Table 2. The frequency of designed constructs to measure the social bias among 48 studies

Construct validity	N
Equality of word distances	11
Equality of downstream performance	8
Definitions for biased or appropriate words	10
Linear projection along the bias direction with neutral range	9
Linear projection along the bias direction without neutral range	3
Kernelized projection along the principal components of bias word vectors with neutral range	2
Non-linear matrix projection	1
Hyperbolic space projection	1
Causal influence	1
Linear projection along the bias direction without neutral range, equality of word distances	1
Definitions for biased or appropriate words, linear projection and kernelized projection comparisons	1

RESULTS-2

External validity:

- 31 out of 48 studies applied only one kind of word embeddings while 17 out of 48 studies compared the mitigation methods through multiple word embeddings.
- The most popular word embeddings were word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2017). And their corpora were from web crawl, Wikipedia, Gigaword, and Google News.

Mitigations:

- Methods to mitigate social bias varied from pre-processing, in-processing, and post-processing to equalize or neutralize the word embeddings.
- For pre-processing, researchers applied data augmentation, gender swap, and name intervention.
- For in-processing, researchers designed neutralized and gendered components or applied adversarial learning approach to retrain word embeddings.
- For post-processing, researchers corrected the word embeddings by removing the bias subspaces.

CONCLUSIONS

Through this study, we investigated the achievements and insufficiency of social debiasing in word embeddings from perspectives of validity. Future work could improve the validity of debiasing by investigating less studied social bias such as socioeconomic status bias, age bias and literacy bias, comparing different measurements and constructs of social bias, and testing the mitigation approach on other word embeddings and potential specialized datasets.

REFERENCES

- 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. *Advances* in neural information processing systems, 29.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.
- Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Systematic Reviews, 18(2), e1230.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.
- *Mikolov, T., Yih, W. T., & Zweig, G. (2013b, June). 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).*
- *Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).*