The Journal of Open Source Software

sweater: Speedy Word Embedding Association Test and Extras Using R

Chung-hong Chan[1]

1 Mannheimer Zentrum für Europäische Sozialforschung, Universität Mannheim

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Statement of need

The goal of this R package is to detect (implicit) biases in word embeddings. The importance of detecting biases in word embeddings is twofold. First, pretrained, biased word embeddings deployed in real-life machine learning systems can pose fairness concerns (Boyarskaya, Olteanu, & Crawford, 2020; Packer, Mitchell, Guajardo-Céspedes, & Halpern, 2018). Second, biases in word embeddings reflect the biases in the original training material. Social scientists, communication researchers included, have exploited these methods to quantify (implicit) media biases by extracting biases from word embeddings locally trained on large text corpora (e.g. Kroon, Trilling, & Raats, 2020; Knoche, Popović, Lemmerich, & Strohmaier, 2019; Sales, Balby, & Veloso, 2019). Biases in word embedding can be understood through the implicit social cognition model of media priming (Arendt, 2013). In this model, implicit stereotypes are defined as the "strength of the automatic association between a group concept (e.g., minority group) and an attribute (e.g., criminal)." (Arendt, 2013, p. 832) All of these bias detection methods are based on the strength of association between a concept (or a target) and an attribute in embedding spaces.

Previously, the software of these methods is only scatteredly available as the addendum of the original papers and was implemented in different languages (Java, Python, etc.). `sweater` provides several of these bias detection methods in one unified package with a consistent R interface (R Core Team, 2021). Also, some provided methods are implemented in C++ for speed and interfaced to R using the `Rcpp` package (Eddelbuettel, 2013).

In the usage section below, we demonstrated how the package can be used to detect biases and reproduce some published findings.

Usage

Word Embeddings

The input word embedding w is a dense $m \times n$ matrix, where m is the total size of the vocabulary in the training corpus and n is the vector dimension size.

`sweater` supports two types of w. For locally trained word embeddings, word embedding outputs from the R packages `word2vec` (Wijffels, 2021), `rsparse` (Selivanov, 2020) and `text2vec` (Selivanov et al., 2020) are directly supported.[1] For pretrained word embeddings obtained online,[2] they are usually provided in the so-called "word2vec" file format and the function `read_word2vec` reads those files into the supported matrix format.

[1] The vignette of `text2vec` provides a guide on how to locally train word embeddings using the GLoVE algorithm (Pennington, Socher, & Manning, 2014). https://cran.r-project.org/web/packages/text2vec/vignettes/glove.html

[2] For example, the pretrained GLoVE word embeddings, pretrained word2vec word embeddings and pretrained fastText word embeddings.

Query

`sweater` uses the concept of *query* (Badilla, Bravo-Marquez, & Pérez, 2020) to study the biases in w. A query contains two or more sets of seed words with at least one set of *target words* and one set of *attribute words*. `sweater` uses the $\mathcal{STAB}$ notation from Brunet, Alkalay-Houlihan, Anderson, & Zemel (2019) to form a query.

Target words are words that **should** have no bias. They are denoted as wordsets $\mathcal{S}$ and $\mathcal{T}$. All methods require $\mathcal{S}$ while $\mathcal{T}$ is only required for WEAT. For instance, the study of gender stereotypes in academic pursuits by Caliskan, Bryson, & Narayanan (2017) used $\mathcal{S} = \{math, algebra, geometry, calculus, equations, computation, numbers, addition\}$ and $\mathcal{T} = \{poetry, art, dance, literature, novel, symphony, drama, sculpture\}$.

Attribute words are words that have known properties in relation to the bias. They are denoted as wordsets $\mathcal{A}$ and $\mathcal{B}$. All methods require both wordsets except Mean Average Cosine Similarity (Manzini, Lim, Tsvetkov, & Black, 2019). For instance, the study of gender stereotypes by Caliskan et al. (2017) used $\mathcal{A} = \{he, son, his, him, ...\}$ and $\mathcal{B} = \{she, daughter, hers, her, ...\}$. In some applications, popular off-the-shelf sentiment dictionaries can also be used as $\mathcal{A}$ and $\mathcal{B}$ (e.g. Sweeney & Najafian, 2020). That being said, it is up to the researchers to select and derive these seed words in a query. However, the selection of seed words has been shown to be the most consequential part of the entire analysis (Antoniak & Mimno, 2021; Du, Fang, & Nguyen, 2021). Please read Antoniak & Mimno (2021) for recommendations.

Supported methods

Table 1 lists all methods supported by sweater. The function `query` is used to conduct a query. The function `calculate_es` can be used for some methods to calculate the effect size representing the overall bias of w from the query.

Table 1: All methods supported by sweater

Method	Target words	Attribute words
Mean Average Cosine Similarity (Manzini et al., 2019)	$\mathcal{S}$	$\mathcal{A}$
Relative Norm Distance (Garg, Schiebinger, Jurafsky, & Zou, 2018)	$\mathcal{S}$	$\mathcal{A}$, $\mathcal{B}$
Relative Negative Sentiment Bias (Sweeney & Najafian, 2020)	$\mathcal{S}$	$\mathcal{A}$, $\mathcal{B}$
SemAxis (An, Kwak, & Ahn, 2018)	$\mathcal{S}$	$\mathcal{A}$, $\mathcal{B}$
Normalized Association Score (Caliskan et al., 2017)	$\mathcal{S}$	$\mathcal{A}$, $\mathcal{B}$
Word Embedding Association Test (Caliskan et al., 2017)	$\mathcal{S}$, $\mathcal{T}$	$\mathcal{A}$, $\mathcal{B}$

Example 1

Relative Norm Distance (RND) (Garg et al., 2018) is calculated with two sets of attribute words. The following analysis reproduces the calculation of "women bias" values in Garg et al. (2018). The publicly available word2vec word embeddings trained on the Google News corpus is used (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Words such as "nurse", "midwife" and "librarian" are more associated with female, as indicated by the positive relative norm distance (Figure 1).

```r
library(sweater)
data(googlenews)
```

The Journal of Open Source Software

```r
S1 <- c("janitor", "statistician", "midwife", "bailiff", "auctioneer",
       "photographer", "geologist", "shoemaker", "athlete", "cashier",
       "dancer", "housekeeper", "accountant", "physicist", "gardener",
       "dentist", "weaver", "blacksmith", "psychologist", "supervisor",
       "mathematician", "surveyor", "tailor", "designer", "economist",
       "mechanic", "laborer", "postmaster", "broker", "chemist",
       "librarian", "attendant", "clerical", "musician", "porter",
       "scientist", "carpenter", "sailor", "instructor", "sheriff",
       "pilot", "inspector", "mason", "baker", "administrator",
       "architect", "collector", "operator", "surgeon", "driver",
       "painter", "conductor", "nurse", "cook", "engineer", "retired",
       "sales", "lawyer", "clergy", "physician", "farmer", "clerk",
       "manager", "guard", "artist", "smith", "official", "police",
       "doctor", "professor", "student", "judge", "teacher", "author",
       "secretary", "soldier")
A1 <- c("he", "son", "his", "him", "father", "man", "boy", "himself",
        "male", "brother", "sons", "fathers", "men", "boys", "males",
        "brothers", "uncle", "uncles", "nephew", "nephews")
B1 <- c("she", "daughter", "hers", "her", "mother", "woman", "girl",
       "herself", "female", "sister", "daughters", "mothers", "women",
       "girls", "females", "sisters", "aunt", "aunts", "niece", "nieces")
res_rnd_male <- query(w = googlenews, S_words = S1,
                      A_words = A1, B_words= B1,
                      method = "rnd")
plot_bias(res_rnd_male)
```

Example 2

Word Embedding Association Test (WEAT) (Caliskan et al., 2017) requires all four wordsets of $\mathcal{S}$, $\mathcal{T}$, $\mathcal{A}$, and $\mathcal{B}$. The method is modeled after the Implicit Association Test (IAT) (Nosek, Greenwald, & Banaji, 2005) and it measures the relative strength of $\mathcal{S}$'s association with $\mathcal{A}$ to $\mathcal{B}$ against the same of $\mathcal{T}$. The effect sizes calculated from a large corpus, as shown by Caliskan et al. (2017), are comparable to the published IAT effect sizes obtained from volunteers.

In this example, the publicly available GLoVE embeddings made available by the original Stanford Team (Pennington et al., 2014) were used. In the following example, the calculation of "Math. vs Arts" gender bias in Caliskan et al. (2017) is reproduced. In this example, the positive effect size indicates the words in the wordset $\mathcal{S}$ are more associated with males than $\mathcal{T}$ associated with males.

```r
data(glove_math) # a subset of the original GLoVE word vectors
S2 <- c("math", "algebra", "geometry", "calculus", "equations",
        "computation", "numbers", "addition")
T2 <- c("poetry", "art", "dance", "literature", "novel", "symphony",
        "drama", "sculpture")
A2 <- c("male", "man", "boy", "brother", "he", "him", "his", "son")
B2 <- c("female", "woman", "girl", "sister", "she", "her", "hers",
        "daughter")
sw <- query(w = glove_math,
            S_words = S2, T_words = T2,
            A_words = A2, B_words = B2)
calculate_es(sw)
```

```
## [1] 1.055015
```

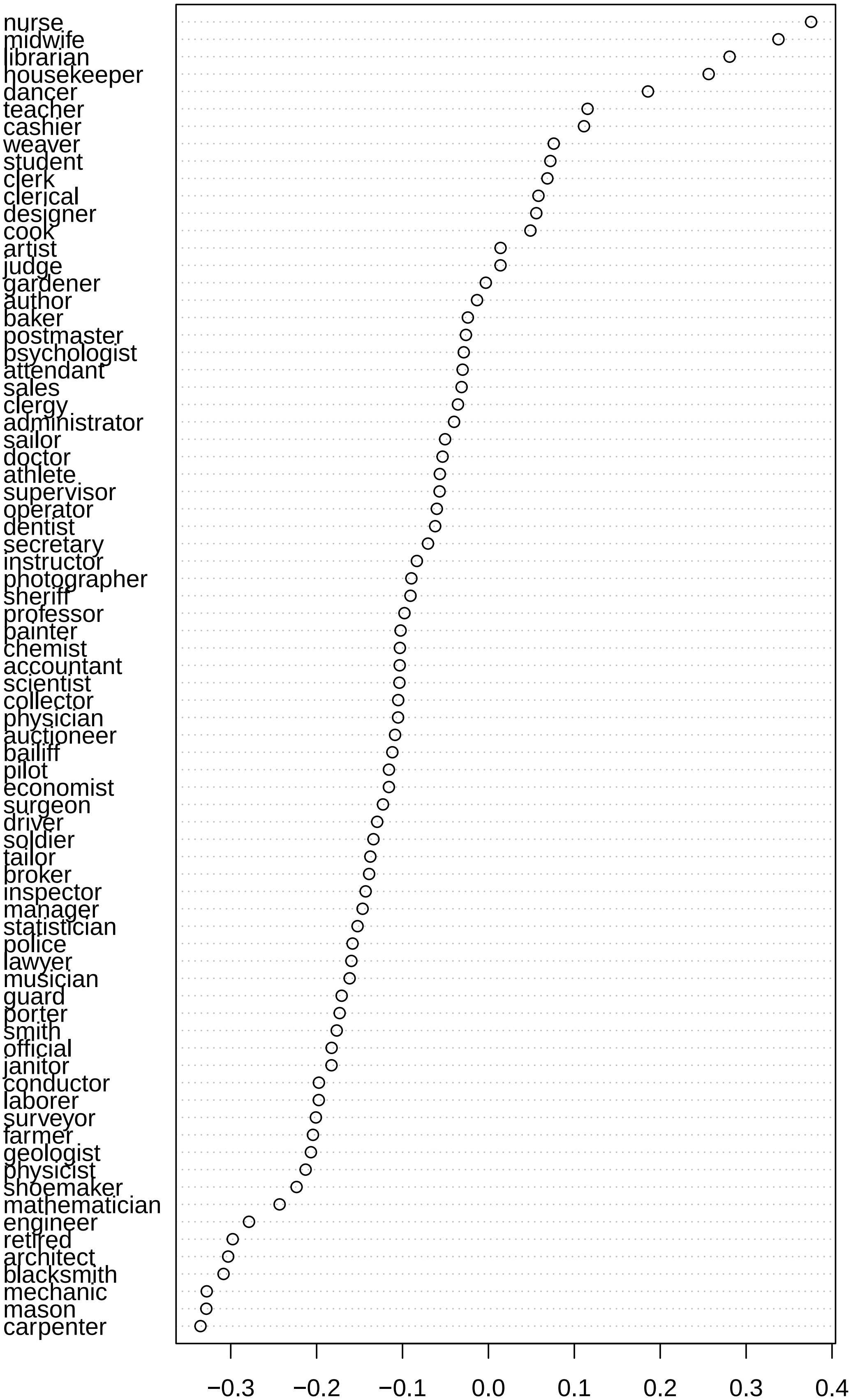

Figure 1: Bias of words in the target wordset according to relative norm distance

The statistical significance of the effect size can be evaluated using the function `weat_resampling`.

```
weat_resampling(sw)
```

```
##
##  Resampling approximation of the exact test in Caliskan et al. (2017)
##
## data:  sw
## bias = 0.024865, p-value = 0.0171
## alternative hypothesis: true bias is greater than 7.245425e-05
## sample estimates:
##       bias
## 0.02486533
```

Acknowledgements

The development of this package was supported by the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (*Bundesministerium für Familie, Senioren, Frauen und Jugend*), the Federal Republic of Germany – Research project: "*Erfahrungen von Alltagsrassismus und medienvermittelter Rassismus in der (politischen) Öffentlichkeit*".

References

An, J., Kwak, H., & Ahn, Y.-Y. (2018). Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. *arXiv preprint arXiv:1806.05521*. doi:10.18653/v1/p18-1228

Antoniak, M., & Mimno, D. (2021). Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1889–1904). doi:10.18653/v1/2021.acl-long.148

Arendt, F. (2013). Dose-dependent media priming effects of stereotypic newspaper articles on implicit and explicit stereotypes. *Journal of Communication*, *63*(5), 830–851. doi:10.1111/jcom.12056

Badilla, P., Bravo-Marquez, F., & Pérez, J. (2020). WEFE: The word embeddings fairness evaluation framework. In *IJCAI* (pp. 430–436). doi:10.24963/ijcai.2020/60

Boyarskaya, M., Olteanu, A., & Crawford, K. (2020). Overcoming Failures of Imagination in AI Infused System Development and Deployment. *arXiv preprint arXiv:2011.13416*.

Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019). Understanding the origins of bias in word embeddings. In *International conference on machine learning* (pp. 803–811).

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. doi:10.1126/science.aal4230

Du, Y., Fang, Q., & Nguyen, D. (2021). Assessing the reliability of word embedding gender bias measures. *arXiv preprint arXiv:2109.04732*.

Eddelbuettel, D. (2013). Seamless R and C++ Integration with Rcpp. doi:10.1007/978-1-4614-6868-4

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. doi:10.1073/pnas.1720347115

Knoche, M., Popović, R., Lemmerich, F., & Strohmaier, M. (2019). Identifying biases in politically biased wikis through word embeddings. In *Proceedings of the 30th ACM conference on hypertext and social media* (pp. 253–257). doi:10.1145/3342220.3343658

Kroon, A. C., Trilling, D., & Raats, T. (2020). Guilty by association: Using word embeddings to measure ethnic stereotypes in news coverage. *Journalism & Mass Communication Quarterly*, 1077699020932304. doi:10.1177/1077699020932304

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*. doi:10.18653/v1/n19-1062

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and Using the Implicit Association Test: II. Method Variables and Construct Validity. *Personality and Social Psychology Bulletin*, *31*(2), 166–180. doi:10.1177/0146167204271418

Packer, B., Mitchell, M., Guajardo-Céspedes, M., & Halpern, Y. (2018). Text embeddings contain bias. Here's why that matters. Retrieved from https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. doi:10.3115/v1/d14-1162

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sales, A., Balby, L., & Veloso, A. (2019). Media bias characterization in brazilian presidential elections. In *Proceedings of the 30th acm conference on hypertext and social media* (pp. 231–240). doi:10.1145/3345645.3351107

Selivanov, D. (2020). *Rsparse: Statistical learning on sparse matrices*. Retrieved from https://CRAN.R-project.org/package=rsparse

Selivanov, D., Bickel, M., & Wang, Q. (2020). *Text2vec: Modern text mining framework for R*. Retrieved from https://CRAN.R-project.org/package=text2vec

Sweeney, C., & Najafian, M. (2020). Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 359–368). doi:10.1145/3351095.3372837

Wijffels, J. (2021). *Word2vec: Distributed representations of words*. Retrieved from https://CRAN.R-project.org/package=word2vec