WEAT is not neat: on the reliability and validity of the Word Embedding Association Test

607 and **kerkeruil**
Made for an assignment as part of the MSc Artificial Intelligence at the University of Amsterdam.

Abstract

The field of natural language processing has recently seen several measures proposed to detect biases in a language model. One of these is the Word-Embedding Association Test (WEAT). WEAT has been used in many research projects, but its reliability and validity have been called into question. Here, we assess the reliability and validity of WEAT. To study reliability, we use WEAT several times on the same pretrained embeddings, varying what subsets of collected word lists we used. We show that the reported effect size varies greatly, especially when using short word lists. Thus, we conclude that WEAT is not a reliable measure. We suggest that WEAT requires larger word lists than those used in the original paper. To study validity, we test for convergent validity with another bias measure, called subspace projection, by testing on five different pairs of target sets. We find that the results of both measures do not correlate. We suggest further research to ascertain that this is a fault of WEAT, and not due to subspace projection being unreliable and/or invalid.

1 Introduction

The study of bias in machine learning is an ongoing issue (Mehrabi et al., 2021). The field of natural language processing (NLP) also has seen a lot of research into bias in recent years, although the motivation and methodology of relevant papers have been found to often be vague and inconsistent (Blodgett et al., 2020).

One direction in bias-related research in NLP concerns constructing measures to assess bias present in a language model (Sun et al., 2019). The Word-Embedding Association Test (WEAT), proposed in Caliskan et al. (2017), is such a measure. WEAT measures bias in word embeddings by considering two sets of target words (eg. male names vs. female names) and calculating their relative similarity to two sets of attribute words (eg. career words vs. family words).

In this paper, we look into the reliability (*are the results of the measure consistent?*) and validity (*are we measuring what we want to measure?*) of WEAT. To study reliability, we use WEAT multiple times on the same pretrained embeddings, taking different subsets of target and attribute word lists each time. We find that the obtained effect size varies greatly, and that it varies more if shorter word lists are used. This is significant, because it means that WEAT might not be a good measure to use when no curated and big dataset is available to evaluate it on. To study validity, we compare to another method to measure bias in embeddings, inspired by Ravfogel et al. (2020). We find that both measures do not correlate.

2 Background

WEAT is inspired by the Implicit Association Test, which is a test from Psychology that measures biases in humans using reaction time as the dependent variable (Greenwald et al., 1998). Many research projects have relied on WEAT (eg. Wambsganss et al., 2022; An et al., 2022; Kaneko et al., 2022). However, the reliability and validity of WEAT have been called into question (eg. Ethayarajh et al., 2019; Loon et al., 2022; Schröder et al., 2021).

Out of the many group stereotypes that are present in society, more than half of the papers investigating bias in NLP considered in a 2021 review look (solely) into gender (Garrido-Muñoz et al., 2021). In the same study, more than half of the papers trained embeddings on English language only. In this paper, we aim to improve diversity of the studied contexts by looking into stereotypes related to religion (specifically, Islam vs. Christianity) in Dutch language as our case study.

3 Approach

We implemented WEAT from scratch, following the formulas provided by Caliskan et al. (2017).

For a WEAT analysis, we consider two sets of target words X, Y and two sets of attribute sets A, B. In both our experiments, A consists of positive words and B consists of negative words. We use different target sets across our experiments. WEAT measures the difference in association between sets X and Y with sets A and B. It is calculated as

$$s(X, Y, A, B) = \sum_{\mathbf{x} \in X} s(\mathbf{x}, A, B) - \sum_{\mathbf{y} \in Y} s(\mathbf{y}, A, B)$$

with

$$s(\mathbf{w}, A, B) = \text{mean}_{\mathbf{a} \in A} \cos(\mathbf{w}, \mathbf{a}) - \text{mean}_{\mathbf{b} \in B} \cos(\mathbf{w}, \mathbf{b}).$$

For example, if $S(X, Y, A, B)$ is a large positive value, that means that X is more associated with A than with B than Y. In the experiments, we report the effect size, which is given by

$$\frac{\text{mean}_{\mathbf{x} \in X} s(\mathbf{x}, A, B) - \text{mean}_{\mathbf{y} \in Y} s(\mathbf{y}, A, B)}{\text{std\_dev}_{w \in X \cup Y} s(\mathbf{w}, A, B)}.$$

In this paper, we measure the reliability and validity of WEAT. We do this on a pre-trained word2vec model trained on a Dutch corpus.

To investigate the reliability, we check whether the method is internally consistent. This is done by running WEAT multiple times using different subsets of the used target and attribute word lists. If the method is reliable, we should get a similar effect size for each run.

To investigate the validity, we check whether WEAT correlates with another measure to detect bias in embeddings. Thus, we look for convergent validity. We use the technique of subspace projection, inspired by Ravfogel et al. (2020). We train a support vector machine on two sets of attribute words and take the subspace orthogonal to its decision boundary. We then project a word onto this subspace to get the bias score for it. To get a total bias to compare to the WEAT effect size, we obtain

$$s(X, Y, A, B) = \text{mean}_{\mathbf{x} \in X} s(\mathbf{x}, A, B) - \text{mean}_{\mathbf{y} \in Y} s(\mathbf{y}, A, B),$$

where $s(\mathbf{w}, A, B)$ is the projection of $\mathbf{w}$ onto the subspace obtained. We use WEAT and subspace projection on five different pairs of target sets. If both WEAT and subspace projection are valid measures of bias in embeddings, the scores obtained from both on the different pairs of target sets should correlate.

4 Experiments and results

Our first experiment aims to assess the reliability of WEAT. We used pre-trained word embeddings: specifically, a Word2Vec model trained on a large Dutch corpus, mostly comprised of social media messages, but also including some more formal sources.[1] To validate the embeddings, we prompted the model for the closest vectors to that for the word *aardappel* (*potato*) and got as the closest three vectors *aardappels*, *bloemkool* and *aardappelen* (resp. *potatoes*, *cauliflower*, *potatoes*), with the rest of the top 20 being mostly made up out of other vegetables.

We then used WEAT to estimate bias in a novel setting. As target sets, we used a list of Christian names suggested for Dutch children and a list of Islamic names suggested for Dutch children. As attribute sets, we used a list of Dutch positive words and a list of Dutch negative words.[2] We stripped the lists to exclude words that were not included in the embeddings, which left us with 26 Islamic girl names, 44 Islamic boy names, 73 Christian girl names, 73 Christian boy names, 453 positive words and 349 negative words. If WEAT is a reliable measure, we should get comparable results when we run WEAT on different subsets of these word lists. We used 10 random subsamples of the names sets, taking 4 names from each of the 4 sets. We kept an equal division between girl names and boy names for both sets to exclude any possible interference of gender bias. For each of the subsamples of the target words, we did runs on each of 8 different chunks of the attribute words, with 8 positive words and 8 negative words each. Thus we used 2×8 target words and 2×8 attribute words, which is in line with half of the experiments from Caliskan et al. (2017). We then repeated the experi-

[1] https://github.com/coosto/dutch-word-embeddings

[2] Target word lists were obtained from https://www.oudersvannu.nl/zwanger/babynamen/bijbelse-namen/ and https://www.ikbenzwanger.com/arabisch-marokkaans-islamitisch-namen.php, attribute word lists were obtained from https://purestarters.nl/positieve-woorden-alfabet/ and https://purestarters.nl/negatieve-woorden-alfabet/. All four web pages have been saved in the Web Archive.

ment with 2×28 target words and 2×28 attribute words, which is slightly bigger than the maximum set sizes used in Caliskan et al. (2017).

A visualization of the results of the experiment can be found in Figure 1. We find that the effect size found by WEAT varies greatly, and varies considerably more on the smaller word set sizes. Even on the larger word set sizes, the Cohen's d score varies between finding a large positive effect size and finding none at all.

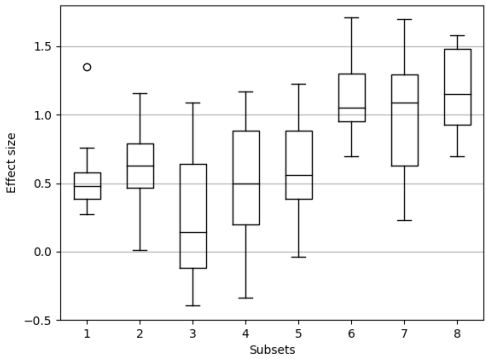
Our second experiment aims to assess the validity of WEAT. Specifically, we look for convergent validity, by comparing WEAT effect size against the score obtained from subspace projection. In an attempt to minimize the effect of the potential unreliability of both measures, we used attribute sets containing 349 words each, and target sets containing 25 words each. We validated our subspace by calculating the average bias of five overtly positive words (*nice*, *goed*, *voordelig*, *plezant*, *hoera*) and five overtly negative words (*verdorie*, *benauwend*, *gemeen*, *shit*, *teleurstellend*) that were not in the dataset of attribute words used to train the SVM. This resulted in an average bias of -0.89 and 1.3, respectively, giving us reason to believe that the concept of positivity and negativity is well represented in the subspace.

For the convergent validity experiment, we used 5 different pairs of target sets: Christian names vs. Islamic names, boy names vs. girl names, Dutch words vs. English words[3], insects vs. plants and instruments vs. weapons. For the first two pairs, we used the same datasets as for the first experiment. The other used target sets were constructed by us with inspiration from web sources, and can be found in Appendix A. We applied WEAT and subspace projection to all five dataset pairs. The resulting scores can be found in Figure 2. We calculated correlation using Pearson's r, obtaining $r = -0.12$.

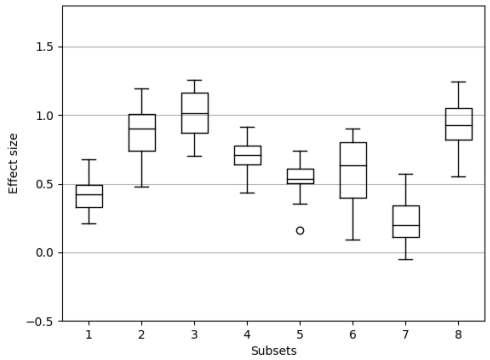
5 Discussion

We conducted a study on the reliability and validity of WEAT. Based on the results of our first experiment, shown in Figure 1, we conclude that WEAT is not a reliable measure. One factor in this is the size of the attribute set. In Caliskan et al. (2017), the sets of male and female names consist of eight words each, which by our results causes low reli-

[3]Our word2vec model was trained on text from Dutch sources, but we found that English words were also well-represented.



(a) 2×8 target words, 2×8 attribute words.



(b) 2×28 target words, 2×28 attribute words.

Figure 1: **Results of reliability study.** The x-axis shows the different chunks of the attribute words. The boxes show the different subsamples of the target words. On the y-axis, WEAT effect size is reported. In general, WEAT finds that the Islamic names, compared to the Christian names, are more associated with the negative words than with the positive words. However, we find that the effect size found by WEAT varies a lot, and varies considerably more on the smaller word set sizes. Even on the larger word set sizes, the Cohen's d score varies from large (> 0.8) to none at all (0.0). On the smaller word set sizes, an effect size in the opposite direction is found on occasion.

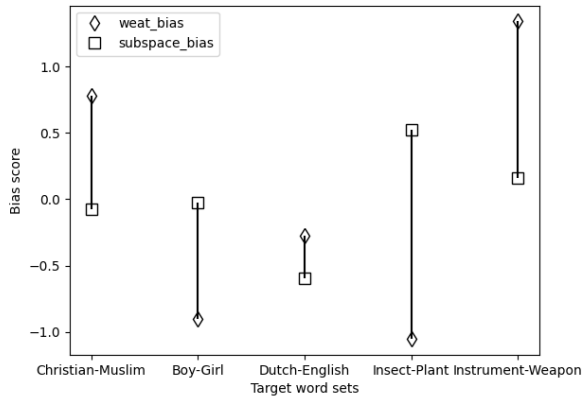


Figure 2: **Results of validity study**. The x-axis shows what word pair was used for the bias calculations. The y-axis shows two different bias scores, indicated by the different shapes. The line between the points shows the distance between the bias measures. As can be seen in the figure the distance fluctuates heavily. The distance between the first word pair (Christian - Muslim) is ~0.8 while the second pair (Boy - Girl) scores ~-0.8. The third pair has again positive difference but this time considerably smaller. The trend remains unpredictable, as is reflected in the Pearson's correlation coefficient, which gives $r = -0.12$.

ability. A small set means that the bias measure will only be able to capture certain aspects of a group and these aspects can be different depending on the set. This claim is supported by the results of Figure 1b which shows that with a larger set the variability of the results per subset of attribute words decreases. In an ideal world, however, the mean of all the subsets would be around the same value.

It is unclear whether using even larger target and attribute set sizes would stabilize the results. It could be that the method of comparing a word to a subset of words that have a known bias is too simple because it is not guaranteed that the bias is the only thing captured this way. A suggestion to solve this issue would be to find some way of preprocessing the set of attributes further to ensure it only represents the bias.

However, even in the case that using larger target and attribute set sizes would allow for a more accurate result, this would impose a significant restriction on the use of WEAT. It would require a lot of data to use WEAT, which is often costly and/or time-consuming to acquire.

The results of our second experiment, shown in Figure 2, show that WEAT and subspace projection do not correlate. In fact, a Pearson's r of -0.12 suggests a small negative correlation, but we find this unlikely, and contribute this result to unreliability of one or both measures, which was not balanced out by using more target set pairs. Both methods produce very different results even though the same sets of target words and attribute words were used. The sets used had a greater or equal size compared to the sets used in Caliskan et al. (2017), which should lead to a more valid result, since more words should theoretically better define the boundaries of a concept. It could be beneficial to increase the size of the target sets further. Further research is required to assess if this could lead to a higher correlation. It can be costly and/or time-consuming to obtain more data, though. Furthermore, it is important to note that the lists might have to be curated for a valid measure. For example, our *Weapons* word set included *goedendag*[4], which can refer to either a spiked ball attached to a stick or to a common greeting. Such semantic ambiguities might impact different bias measures differently.

From the results of our second experiment, we can conclude that WEAT and subspace projection are not convergently valid. However, this does not indicate per se that WEAT is not a valid measure: the bias direction method should also be called into question. Looking at the results in Figure 2, it can be noted that subspace projection found insect names to be more positive than plant names. However, results from psychological research suggest the opposite, agreeing with the effect direction of WEAT (Greenwald et al., 1998). Also, it has not been established whether subspace projection is a reliable measure. Future research should look for convergent validity of WEAT with several more measures.

Another direction for future research into the validity of WEAT involves generating datasets that are supposed to be more or less biased. For example, one could use a hate speech dataset and vary the amount of hate speech included, assuming that a greater portion of hate speech would amount to a greater bias, and investigate whether this is detected by WEAT. However, a challenge in this direction is that hate speech datasets tend to be small and sparse (Alkomah and Ma, 2022).

[4]In fact, *goedendag* was in the word set twice, illustrating how easily datasets can contain imperfections.

References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

Haozhe An, Xiaojiang Liu, and Donald Zhang. 2022. Learning bias-reduced word embeddings using dictionary definitions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1139–1152, Dublin, Ireland. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Ismael Garrido-Muñoz , Arturo Montejo-Ráez , Fernando Martínez-Santiago , and L. Alfonso Ureña-López . 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7).

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464–1480.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in meta-embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Austin van Loon, Salvatore Giorgi, Robb Willer, and Johannes Eichstaedt. 2022. Negative associations in word embeddings predict anti-black bias across regions–but only via name frequency. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1419–1424.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6). 115.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Sarah Schröder, Alexander Schulz, Philip Kenneweg, Robert Feldhans, Fabian Hinder, and Barbara Hammer. 2021. Evaluating metrics for bias in word embeddings.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Thiemo Wambsganss, Vinitra Swamy, Roman Rietsche, and Tanja Käser. 2022. Bias at a second glance: A deep dive into bias for German educational peer-review data modeling. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1344–1356, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Target sets

The novel target sets constructed for the second experiment can be found here. Words that turned out not to have an embedding in our model are italicised.

A.1 Dutch words

tijd
persoon
jaar
manier
dag
ding
man
wereld
leven
hand
deel
kind
oog
vrouw
plaats
werk
week
geval
punt
regering
bedrijf
nummer
groep

probleem
feit

A.2 English words

time
person
year
way
day
thing
man
world
life
hand
part
child
eye
woman
place
work
week
case
point
government
company
number
group
problem
fact

A.3 Insects

mier
bij
spin
mug
luis
kever
mot
vlieg
wesp
want
trips
krekel
hommel
termiet
vlinder
rups
lieveheersbeestje
worm
oorworm
regenworm
mijt
zweefvlieg
langmug
langpoot
houtzwamkever
larven
sluipwesp
kakkerlak
tor
teek
steekvlieg
waterwant
maden

A.4 Plants

Hortensia
Rozen
Vlinderstruik
Glansmispel
Kleine maagdenpalm
Vrouwenmantel
Portugese Laurier
Buxus
Lampenpoetsersgras
Lavendel
Klimop
Liguster
Hartlelie
Blauwe regen
IJzerhard
Vetkruid
Druif
Olijfboom
Vijgenboom
Munt
Koriander
Tulpen
Basilicum
Gras
Viooltje
Geranium
Appelboom
Perenboom
Bloesem
Perzik
Eik
Berk
Spar
Kastanje
Beuk
Boswilg
Jeneverbes

Cypres
Lijsterbes
Denneboom
Populier

A.5 Instruments

Fluit
Gitaar
Piano
Tabla
Hoorn
Saxofoon
Orgel
Viool
Trompet
Triangel
Trommel
Drums
basgitaar
bas
synthesiser
cajon
Harp
Kazoo
Klavecimbel
Lier
Luit
Mandoline
Marimba
Melodica
Mondharmonica
Harmonica
Ocarina
Orkestklokken
Piccolo
Pijporgel
Windklok
Vleugel
Cello

A.6 Weapons

slinger
pijl-en-boog
kruisboog
katapult
ballista
trebuchet
speer
plumbata
werpspies
shuriken
mes
dolk
ponjaard
harpoen
floret
degen
seax
goedendag
katana
sai
tanto
knuppel
ploertendoder
zwaard
boksbeugel
vechtstok
sax
goedendag
morgenster
strijdvlegel
bijl
nunchaku
hagelgeweer
jachtgeweer
machinegeweer
aanvalsgeweer
machinepistool
gevechtsgeweer
AK-47
M4
Uzi
Diemaco
Minimi
Galil
Steyr AUG
vlammenwerper
bazooka
granaatwerper
RPG-7