A bayesian method of cosine-based word2vec bias

Alicja Dobrzeniecka and Rafal Urbaniak

test

1 Cosine-based measures of bias

One of the first measures in the discussion has been developed by Bolukbasi, Chang, Zou, Saligrama, & Kalai (2016). There, the gender bias of a word w is understood as its projection on the gender direction $\vec{w}\cdot(\overrightarrow{he} - \overrightarrow{she})$ (the gender direction is the top principal compontent of ten gender pair difference vectors). The underlying idea is that no bias is present if non-explicitly gendered words are in equal distance to both elements in all explicitly gendered pairs. Given the (ideally) gender netural words N and the gender direction g the direct direct gender bias is defined as the average distance of the words in N from g (c is a parameter determining how strict we want to be):

$$\mathsf{directBias_c}(\mathsf{N},\mathsf{g}) = \frac{\sum_{w\in N} |\cos(\vec{w},g)|^c}{|N|} \tag{1}$$

The use of projections has been ciriticized for instance by Gonen & Goldberg (2019), who point out that while gender-direction might be an indicator of bias, it is only one possible manifestation of it, and reducing a projection of words might be insufficient. For instance, “math” and “delicate” might be in equal distance to both explicitly gendered words while being closer to quite different stereotypical attribute words. Further, the authors point out that most word pairs preserve similarity under debiasing meant to minimize projection-based bias.[1]

To measure bias in word embeddings, Caliskan, Bryson, & Narayanan (2017) proposed the Word Embedding Association Test (WEAT). The idea is that the measure of biases between two sets of target words, X and Y, (we call them protected words) should be quantified in terms of the cosine similarity between the protected words and attribute words coming from two sets of stereotype attribute words, A and B (we’ll call them attributes). For instance, X might be a set of male names, Y a set of female names, A might contain stereotypically male-related career words, and B stereotypically female-related family words. WEAT is a modification of the Implicit Association Test (IAT) (Nosek, Banaji, & Greenwald, 2002) used in psychology and uses almost the same word sets, allowing for a *prima facie* sensible comparison with bias in humans. Let f be a distance or similarity measure (usually, cosine similarity). The association difference for a term t is:

$$s(t,A,B) = \frac{\sum_{a\in A} f(t,a)}{|A|} - \frac{\sum_{b\in B} f(t,b)}{|B|} \tag{2}$$

then, the association difference between A a B is:

$$s(X,Y,A,B) = \sum_{x\in X} s(x,A,B) - \sum_{y\in Y} s(y,A,B) \tag{3}$$

[1] Bolukbasi et al. (2016) use also another method which involves analogies and their evaluations by human users on Mechanical Turk. It is discussed and criticized in (Nissim, Noord, & Goot, 2020).

$s(X,Y,A,B)$ is the statistic used in the signifcance test, and the p-value obtained by bootstrapping: it is the frequency of $s(X_i,Y_i,A,B) > s(X,Y,A,B)$ for all equally sized partitions X_i,Y_i of $X \cup Y$. The effect size is computed by normalizing the difference in means as follows:

$$bias(A,B) = \frac{\mu(\{s(x,A,B)\}_{x \in X}) - \mu(\{s(y,A,B)\}_{y \in Y})}{\sigma(\{s(w,A,B)\}_{w \in X \cup Y})} \tag{4}$$

Caliskan et al. (2017) show that significant biases—thus measured— similar to the ones discovered by IAT can be discovered in word embeddings. Lauscher & Glavas (2019) extended the methodology to a multilingual and cross-lingual setting, arguing that using Euclidean distance instead of similarty does not make much difference, while the bias effects vary greatly accross embedding models (interestingly, with social media-text trained embeddings being less biased than those based on Wikipedia).

A similar methodology is employed by Garg, Schiebinger, Jurafsky, & Zou (2018), who employ word embeddings trained od corpora from different decades to study the shifts in various biases. For instance, to compute the occupational embeddings bias for women the authors first compute the average vector of vector emeddings of words that represent women (e.g. “she”, “female”), then calculate the Euclidean distance between this mean vector and words for occupations. Then they take the mean of these distances and subtract from it the analogously obtained mean for the average vector of vector embeddings of words that represent men. Formally they take the relative norm distance between X and Y to be:

$$\text{relative norm distance} = \sum_{v_m \in M} ||v_m - v_X||_2 - |v_m - v_Y||_2 \tag{5}$$

where the norm used is Euclidean, and v_X and x_Y are average vectors for sets X and Y respectively.

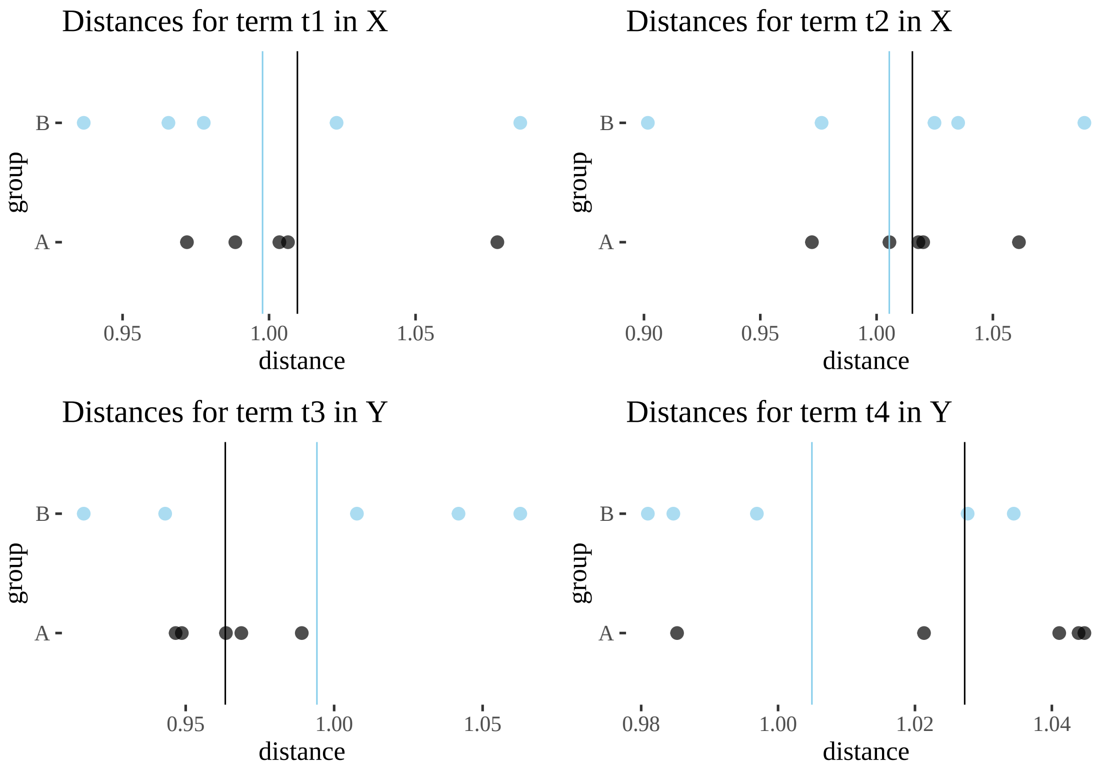
There is a problem with taking means of means in the calculations, though. By pre-averaging data we throw away information about sample size. For instance, think about proportions: 10 out of 20 and 2 out of 4 give the same mean, but you would obtain more information by making the former observation rather than by making the latter. Connectedly, when we pre-average, we remove variation, and so pre-averaging tend to manufacture false confidence.

To illustrate let’s employ the formulas used by Caliskan et al. (2017) and go through an example. Conceptually, all such tests come up with rather short lists of protected words and rather short lists of stereotypical attributes. Clearly, these are not complete list. So let’s treat them as samples from richer pools of stereotypical predicates and let’s take the uncertainty and variation involved seriously.

Consider a situation in which there are two protected classes, $X = \{t_1,t_2\}$ and $Y = \{t_3,t_4\}$ and two five-element sets A and B.

First, we play around with a scenario in which all the protected terms are on average equally distant from both sets ($\mu = 1$) with standard deviation of .05. First, let’s randomly draw distances and plot the results with group means plotted as vertical lines.

```
set.seed(123)
t1 <- data.frame(A  = rnorm(5,1,0.05), B = rnorm(5,1,0.05))
t2 <- data.frame(A  = rnorm(5,1,0.05), B = rnorm(5,1,0.05))
t3 <- data.frame(A  = rnorm(5,1,0.05), B = rnorm(5,1,0.05))
t4 <- data.frame(A  = rnorm(5,1,0.05), B = rnorm(5,1,0.05))
```



When you look at the datapoints, do you have the impression of a strong bias here? We wouldn't think so. But now let's run the calculations from (Caliskan et al., 2017).

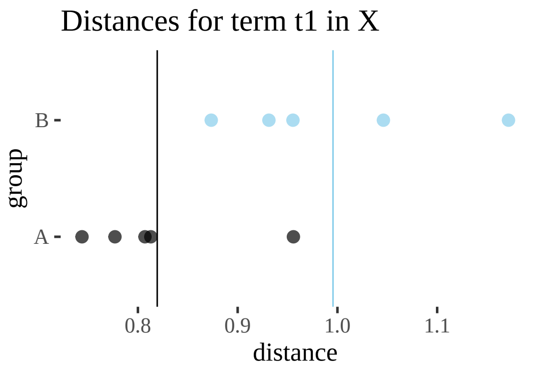
```
s <- function (table){ mean(table$A) - mean(table$B)}
factor <- sd(c(s(t1),s(t2),s(t3),s(t4)))
numerator <-  mean(s(t1),s(t2)) - mean(s(t3),s(t4))
bias <- numerator / factor
bias
```

```
## [1] 1.825005
```

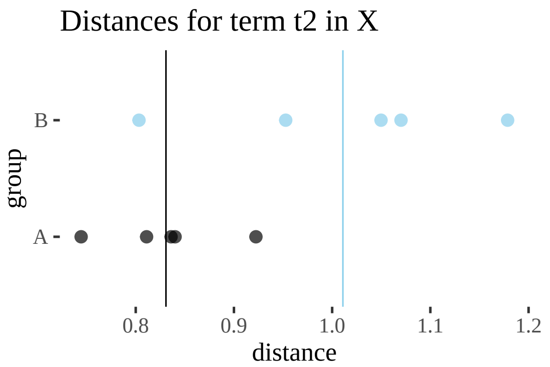
This should make us pause. We know these were points randomly drawn from distributions where there is no difference in means. Yet, the calculated effect size is 1.82, whereas the largest effect size reported by Caliskan et al. (2017) is 1.81!

Now, let's try sampling from distributions where there in fact is a difference in means. Terms from X are on average .8 away from A (and 1 from B), while terms from Y are .8 away to B (and 1 to A). The standard deviation is 0.1 in all the cases. There is a clear difference between X and Y and quick visual inspection should tell us so.

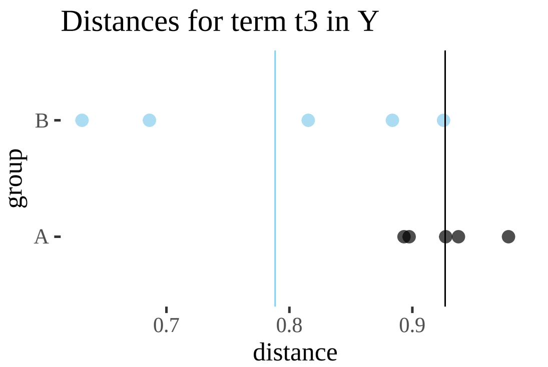
```
set.seed(123)
t1 <- data.frame(A  = rnorm(5,.8,0.1), B = rnorm(5,1,0.1))
t2 <- data.frame(A  = rnorm(5,.8,0.1), B = rnorm(5,1,0.1))
t3 <- data.frame(A  = rnorm(5,1,0.1), B = rnorm(5,.8,0.1))
t4 <- data.frame(A  = rnorm(5,1,0.1), B = rnorm(5,.8,0.1))
```



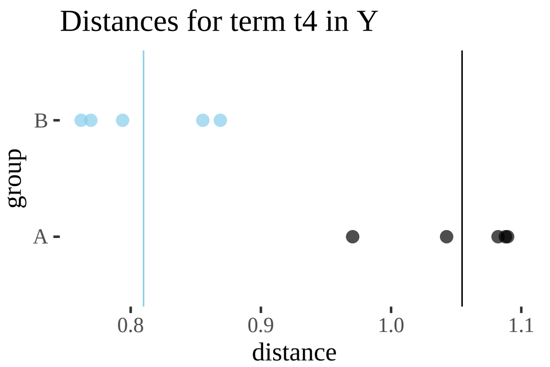
Distances for term t1 in X



Distances for term t2 in X



Distances for term t3 in Y



Distances for term t4 in Y

Is this clear difference mirrored in the calculations?

```
factor <- sd(c(s(t1),s(t2),s(t3),s(t4)))
numerator <-  mean(s(t1),s(t2)) - mean(s(t3),s(t4))
bias <- numerator / factor
bias
```

```
## [1] -1.444026
```

The absolute value of the effect size is smaller than in the previous case!

Appendix

Word lists, including human and neutral predicates

References

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. American Association for the Advancement of Science (AAAS). Retrieved from https://doi.org/10.1126/science.aal4230

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. Proceedings of the National Academy of Sciences. Retrieved from https://doi.org/10.1073/pnas.1720347115

Gonen, H., & Goldberg, Y. (2019). 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). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N19-1061

Lauscher, A., & Glavas, G. (2019). Are we consistently biased? Multidimensional analysis of biases in distributional word vectors. *CoRR*, *abs/1904.11783*. Retrieved from http://arxiv.org/abs/1904.11783

Nissim, M., Noord, R. van, & Goot, R. van der. (2020). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, *46*(2), 487–497. MIT Press - Journals. Retrieved from https://doi.org/10.1162/coli_a_00379

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115. American Psychological Association (APA). Retrieved from https://doi.org/10.1037/1089-2699.6.1.101