A bayesian method of cosine-based word2vec bias estimation

A considerable amount of literature exists on bias detection and mitigation in NLP models, especially word2vec embeddings, which represent words as vectors of real numbers (see e.g. [1] and [4] and references therein). The most common method used compares cosine similarity between words from protected groups and attributes that are considered to be stereotypical or harmful in some way, and this method will be in our focus.

In one well-known approach, [2] 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) [7] used in psychology and uses almost the same word sets, allowing for a *prima facie* sensible comparison with bias in humans. The association difference for a term t is $s(t,A,B)$, and the effect size is computed by normalizing the difference in means as follows (f is a similarity measure, such as cosine similarity):

$$\mathsf{s}(t,A,B) = \frac{\sum_{a\in A} f(t,a)}{|A|} - \frac{\sum_{b\in B} f(t,b)}{|B|} \qquad \mathsf{bias}(A,B) = \frac{\mu(\{s(x,A,B)\}_{x\in X}) - \mu(\{s(y,A,B)\}_{y\in Y})}{\sigma(\{s(w,A,B)\}_{w\in X\cup Y})} \tag{WEAT}$$

[2] show that significant biases—thus measured— similar to the ones discovered by IAT can be discovered in word embeddings. [5] extends the methodology to a multilingual and cross-lingual setting; a similar methodology is employed by [3], who use word embeddings trained on corpora from different decades. [6] modify WEAT to a multi-class setting, introducing Mean Average Cosine (MAC) similarity as a measure of bias (in fact, in the paper they report distances rather than similarities). Let $T = \{t_1,\ldots,t_k\}$ be a class of protected word embeddings, and let each $A_j \in A$ be a set of attributes stereotypically associated with a protected word). Then:

$$\mathsf{S}(t_i,A_j) = \frac{1}{|A_j|}\sum_{a\in A_j} \cos(t,a) \qquad \mathsf{MAC}(T,A) = \frac{1}{|T|\,|A|}\sum_{t_i\in T}\sum_{A_j\in A} S(t_i,A_j) \tag{MAC}$$

That is, for each protected word and each attribute class, they first take the mean for this protected word and all attributes in a given attribute class, and then they take the mean of thus obtained means for all the protected words.

Such methods are statistically problematic. Two issues, specific to [6], are that (i) no distinction is made based on whether a class of attributes is stereotypically associated with a given protected word or with a different protected word, and (ii) no re-scaling by standard deviation (typical in effect size measures such as Cohen's d) is used, so MAC is not a good measure of effect size. A more general problem, is that all the authors ignore the step of comparing their results with control groups, especially control groups of stereotype-neutral human attributes. That such a move is important is suggested for instance by Figure 1 prepared using the original word list for religion-related protected words extended with control attributes, where such human attributes are also closer to protected classes than neutral non-human words.

Another serious problem is that all the authors calculate means of means and run statistical tests on sets of means. Unfortunately, by pre-averaging the data they throw away information about sample sizes, and they remove variation, and so pre-averaging tends to manufacture false confidence.

To illustrate, let's employ the formulas used by [2] in a simple simulation. All such approaches 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.

X to A	X to B	Y to A	Y to B	σ	WEAT
0	0	0	0	0.05	1.82
0	0	0	0	0.001	-1.93
.1	0	0	.1	0.05	1.49
.1	0	0	.1	.1	1.22

We have randomly drawn protected classes, $X = \{t_1, t_2\}$ and $Y = \{t_3, t_4\}$, and two five-element attribute sets A and B. One simulation draws normally distributed values for two situations in which the underlying mean similarities are in fact equal to zero, in two other cases the means are different, with a range of choices of common standard deviation accross the cases (code available upon request).

The following observations are worthwhile. (1) For points randomly drawn from distributions where there is no difference in means the calculated effect size can easily be 1.82, whereas the largest effect size reported by [2] is 1.81. (2) For samples from distributions where the means are different, the (absolute) effect sizes can easily be lower than in the first two simulations. As Figure 2. illustrates, quite some uncertainty is involved, far more than what systematically low mean-based p-values reported in the papers might suggest. Part of the problem is random variation unaccounted for in the original approach (see Figures 3 and 4 for an example), and part of the problem is that non-negligible changes in effect size can result from a shift in the standard deviation of the original process, because with the decrease of standard deviation the numerator in (WEAT) decreases.

To improve on the situation, we build Bayesian models to estimate the biases involved using the raw datapoints, actually using control groups, distinguishing the connection types, and taking the uncertainty involved seriously. For the general impact of being associated, we build models using hamiltonian Monte Carlo (STAN) according to the following specification (linear categorical model and the priors):

$$cosineDistance_i \sim dnorm(\mu_i, \sigma) \qquad \mu_i = m_{pw} + co_{con} \qquad \text{(Model)}$$

$$m_{pw} \sim dnorm(1, .5) \qquad co_{con} \sim dnorm(0, 1) \qquad \sigma \sim dcauchy(0, 1) \qquad \text{(Priors)}$$

That is, given a protected word we assume its cosine distance to attributes from a given group is normally distributed around a mean, which is determined by the mean assigned to this protected word and by the connection type of that group of attributes (stereotypically associated/ associated with a different predicate/neutral human attribute/random word), and we use weak regularizing priors for the parameters. The task of the model is to estimate the mean for a given protected word and the coefficients of change in that mean corresponding to different connection types.

The resulting coefficients for the religion dataset based on Reddit embeddings are in Figure 5. While there is some difference in the means, the 89% highest posterior density invervals are quite wide and include 0s for all the coefficients. Then, motivated by large differences between the states of different protected words, we build bayesian models with separate coefficients for protected classes, with analogous regularizing priors. The results for the religion dataset based on Reddit embeddings are in Figure 6.

We build analogous models for other topic groups (race, gender), and different embeddings (we use three embeddings: GoogleNews-vectors-negative300, and Reddit Corpus, and the hard-debiased Reddit embedding used by [6]). The general conclusion is that from this bayesian perpective the situation is much less obvious. The word list sizes are small, sample sizes are small, and so posterior density intervals are wide. Some stronger bias can be discerned in the gender case. Moreover, sometimes the differences between associated, different and human, are not very impressive. The cost of debiasing turns out sometimes to be that neutral human predicates get closer to protected classes (in religion), change in gender after debiasing is really minor with zero still out out of HPDI range, and debiasing improvement for race is a small improvement achieved at the price of moving neutral terms closer to protected words (there is no space to include these results in this abstract).

The bottom line is that if we want to take bias seriously, so should we approach the uncertainty involved in our estimations. There is no replacement for proper statistical evaluation that does not discard information about the uncertaintly involved, larger word lists are needed, and visualisation of the results for particular protected classes provides much better guidance than chasing a single metric based on a means of means.

Figure 1.
Empirical distribution of cosine distances (religion, Reddit)

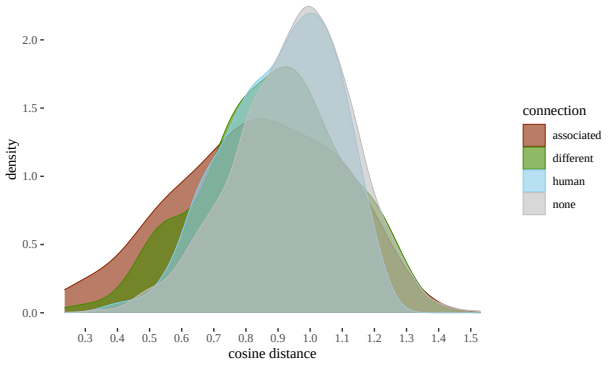


Figure 2.
10k biases for different means and sd =.05 (case 3 in the table)

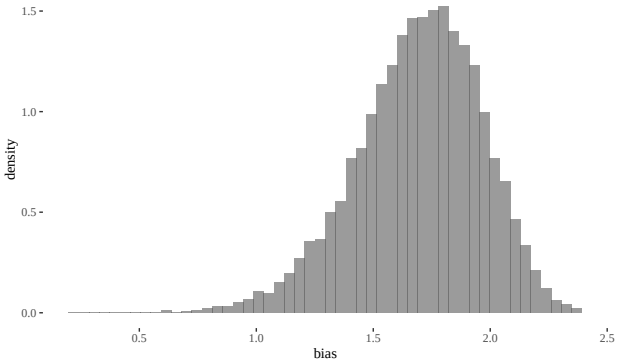


Figure 3.
Cosine distances for active attributes (musim, Reddit)

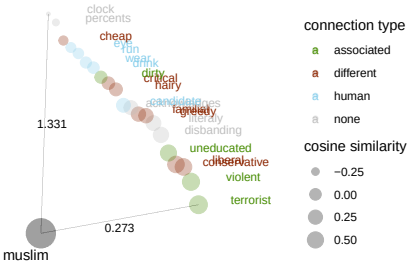
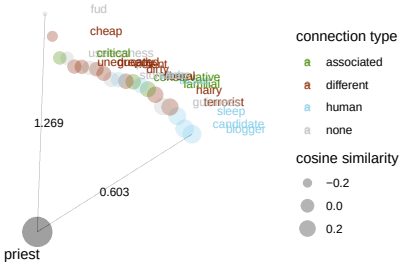
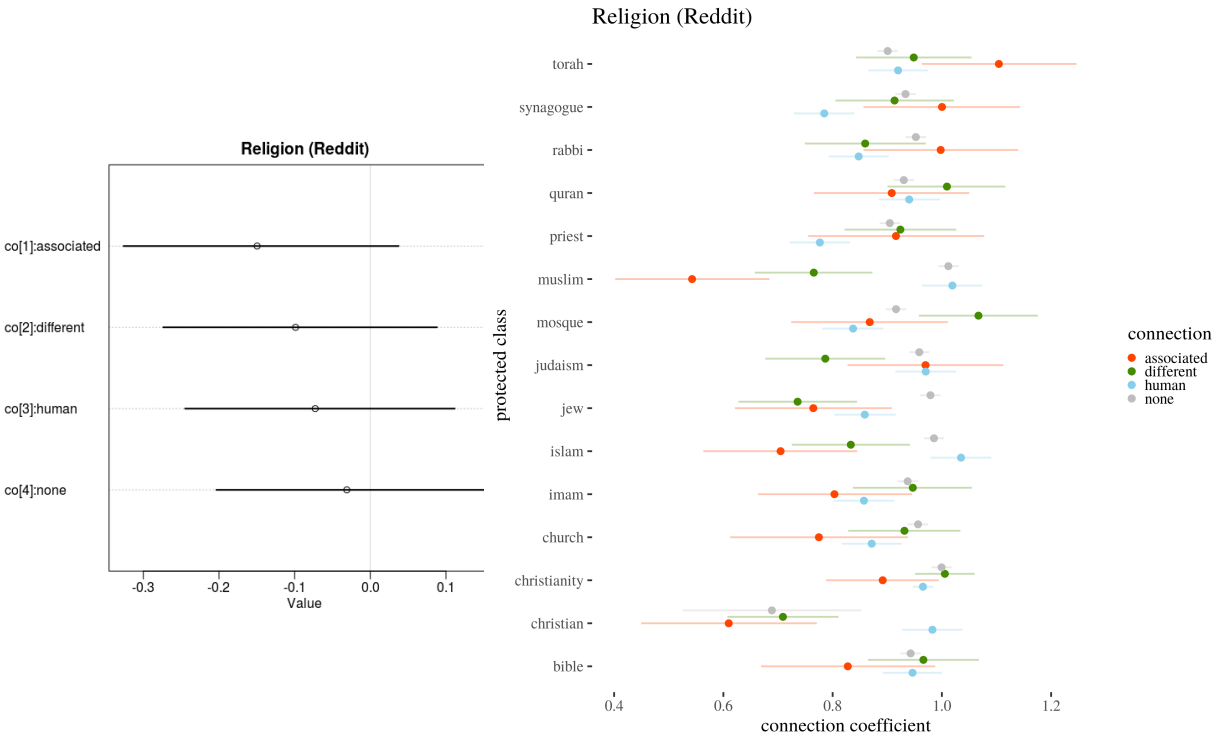


Figure 4.
Cosine distances for active attributes (priest, Reddit)





[1] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Retrieved from http://arxiv.org/abs/1607.06520

[2] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (April 2017), 183–186. DOI:https://doi.org/10.1126/science.aal4230

[3] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (April 2018), E3635–E3644. DOI:https://doi.org/10.1073/pnas.1720347115

[4] Hila Gonen and Yoav Goldberg. 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 609–614. DOI:https://doi.org/10.18653/v1/N19-1061

[5] Anne Lauscher and Goran Glavas. 2019. Are we consistently biased? Multidimensional analysis of biases in distributional word vectors. *CoRR* abs/1904.11783, (2019). Retrieved from http://arxiv.org/abs/1904.11783

[6] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. Retrieved from http://arxiv.org/abs/1904.04047

[7] Brian A. Nosek, Mahzarin R. Banaji, and Anthony G. Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice* 6, 1 (2002), 101–115. DOI:https://doi.org/10.1037/1089-2699.6.1.101