Taking uncertainty seriously

A Bayesian approach to word embedding bias estimation

Alicja Dobrzeniecka & Rafal Urbaniak
(LoPSE research group, University of Gdansk)

ExpSem2021, ESSLLI

Cosine-based measures of bias

Word embeddings

- Representation of words with vectors of real numbers
- Built to predict the probability of co-occurence

word	1	2	3	4	...
woman	0.456	0.267	0.675	0.131	...
man	0.451	0.897	0.472	0.088	...

Cosine-based measures of bias

Cosine similarity & distance

$$\text{cosineSimilarity}(A, B) = \frac{A \cdot B}{||A||\,||B||} \tag{Sim}$$

$$\text{cosineDistance}(A, B) = 1 - \text{cosineSimilarity}(A, B) \tag{Distance}$$

- Geometric interpretation: direction (not length)
- cosineDistance $\in (0, 2)$
- Naive interpretation: proximity corresponds to semantic similarity (e.g. no triangle inequality)

Cosine-based measures of bias

The worry

In the learning process these models can learn implicit biases that reflect harmful stereotypical thinking

Cosine-based measures of bias

The worry

In the learning process these models can learn implicit biases that reflect harmful stereotypical thinking

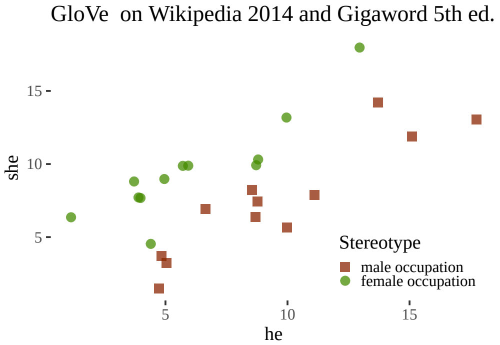
Cosine-based bias: basic intuition

Words belonging to an intuitively harmful stereotype are cosine-close to each other

Cosine-based measures of bias

A visual example

- feminine occupations: “homemaker”, “nurse”, “receptionist”, “librarian”, etc.
- masculine occupations: “maestro”, “captain”, “architect”, “boss”, etc.

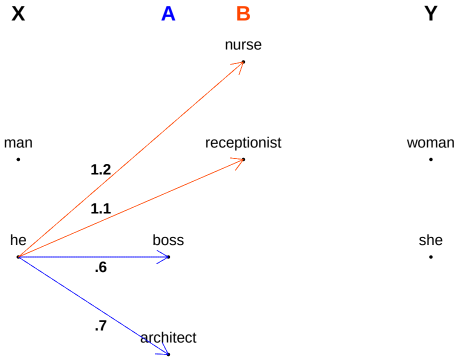


−>

−>

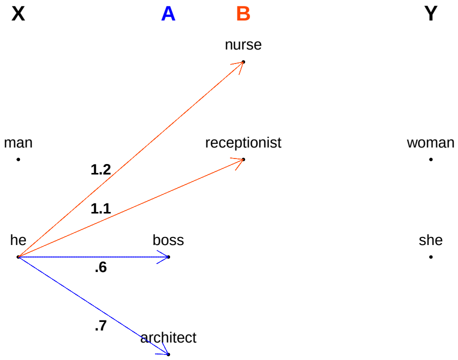
Cosine-based measures of bias

Example: Word Embedding Association Test (WEAT)



Cosine-based measures of bias

Example: Word Embedding Association Test (WEAT)



- $s_1 = s(he, A, B) = \frac{.6+.7}{2} - \frac{1.2+1.1}{2} = 0.65 - 1.15 = -0.5$
- $s_2 = s(man, A, B) = -.3$,
  $s_3 = s(woman, A, B) = .2$, $s_4 = s(she, A, B) = .3$

$$\text{WEAT}(A, B) = \frac{(s_1+s_2)/2 - (s_3+s_4)/2}{sd(\{s_1, s_2, s_3, s_4\})} \approx -1.95$$

Cosine-based measures of bias

Example: Word Embedding Association Test (WEAT)

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

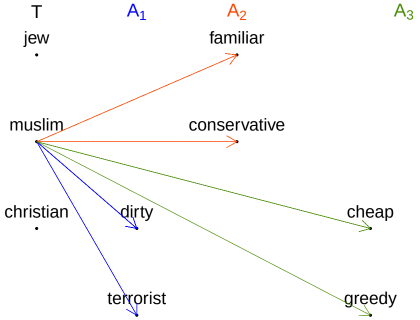
$$WEAT(A, B) = \frac{\mu\left(\{s(x, A, B)\}_{x \in X}\right) - \mu\left(\{s(y, A, B)\}_{y \in Y}\right)}{\sigma\left(\{s(w, A, B)\}_{w \in X \cup Y}\right)}$$

- t is a term, A, B are sets of stereotype attribute words, X, Y are protected group words
- 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
- s-values are used as datapoints in statistical significance tests

(Caliskan, Bryson, & Narayanan, 2017) with extensions in (Lauscher & Glavas, 2019) and applications in (Garg, Schiebinger, Jurafsky, & Zou, 2018)

Cosine-based measures of bias

Our main target: Mean Average Cosine Similarity (MAC)



$$s_1 = s(muslim, A_1) = \frac{cos(muslim, dirty) + cos(muslim, terrorist)}{2}$$

$$s_2 = s(muslim, A_2) = \frac{cos(muslim, familiar) + cos(muslim, conservative)}{2}$$

$$\vdots$$

$$MAC(T, A) = \text{mean}(\{s_i | i \in 1, \dots, k\})$$

Cosine-based measures of bias

Our main target: Mean Average Cosine Similarity (MAC)

$$S(t_i, A_j) = \frac{1}{|A_j|} \sum_{a \in A_j} \cos(t, a)$$

$$MAC(T, A) = \frac{1}{|T|\,|A|} \sum_{t_i \in T} \sum_{A_j \in A} S(t_i, A_j)$$

- $T = \{t_1, \ldots, t_k\}$ is a class of protected words
- each $A_j \in A$ is a set of attributes stereotypically associated with a protected word
- The t-tests they employ are run on average cosines used to calculate MAC

(Manzini, Lim, Tsvetkov, & Black, 2019)

Cosine-based measures of bias

Our main target: Mean Average Cosine Similarity (MAC)

Table 2: Rows the religion dataset.

protectedWord	wordToCompare	cosineDistance	cosineSimilarity
jew	greedy	0.6947042	0.3052958
rabbi	greedy	1.0306175	-0.0306175
rabbi	conservative	0.7175887	0.2824113
christian	uneducated	0.5081939	0.4918061
christianity	cheap	1.2816164	-0.2816164
muslim	terrorist	0.2726106	0.7273894

Cosine-based measures of bias

Known challenges

- Gender-direction might be an indicator of bias, but is insufficient. After debiasing other non-gendered words can remain in biased relations (Gonen & Goldberg, 2019)
- Methods which involve analogies and their evaluations by human users on Mechanical Turk are unreliable (Nissim, Noord, & Goot, 2020)
- Sensitive to the choice of protected words, capitalization, distance measures and embedding methods (Zhang, Sneyd, & Stevenson, 2020)

Some methodological problems

Word list choice is unprincipled

We run with it for comparison

Some methodological problems

Word list choice is unprincipled

We run with it for comparison

No design considerations to sample size

We statistically gauge the uncertainty that arises from raw sample sizes

Some methodological problems

No word class distinction and no control group

We make the subclasses clear, add human neutral predicates and neutral predicates for control. We used L2-Reddit corpus and GoogleNews (we present the results for Reddit for brevity).

Table 3: Rows from extended religion dataset.

protectedWord	wordToCompare	wordClass	cosineDistance	cosineSimilarity	connection
torah	hairy	jewish	1.170	-0.170	associated
christian	dirty	muslim	0.949	0.051	different
judaism	cheap	jewish	1.232	-0.232	associated
christianity	familial	christian	0.645	0.355	associated
mosque	approve	neutral	0.995	0.005	none
imam	carry	human	0.993	0.007	human
mosque	merging	neutral	0.868	0.132	none
muslim	nationalized	neutral	0.870	0.130	none

Some methodological problems

Outliers and surprisingly dissimilar words

We study those by visualizations and uncertainty estimates

Some methodological problems

Outliers and surprisingly dissimilar words

We study those by visualizations and uncertainty estimates

No principled interpretation

Religion Debiasing	MAC (distance)
Biased	0.859
Hard Debiased	0.934
Soft Debiased ($\lambda = 0.2$)	0.894

What values are sufficient for the presence of bias and what differences are sign of real improvement? Low *p*-values are not high effect indicators! We compare HPDIs.

The problem with pre-averaging

Key conceptual issues

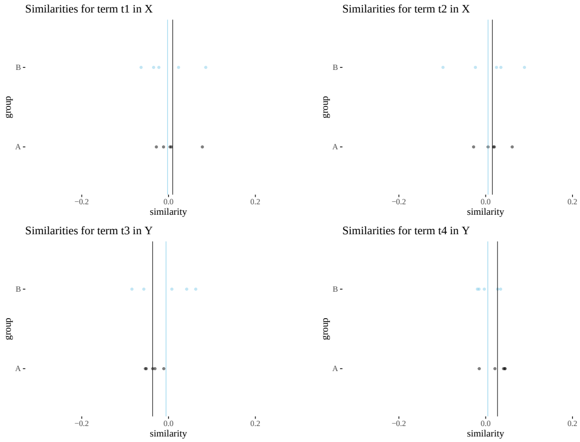
- It throws away information about sample sizes
- It ignores variation in the raw data, which leads to false confidence

Our simulation

Suppose all similarities for two classes are randomly drawn from the same distribution, Normal$(0, .05)$, you still can get really high WEAT!

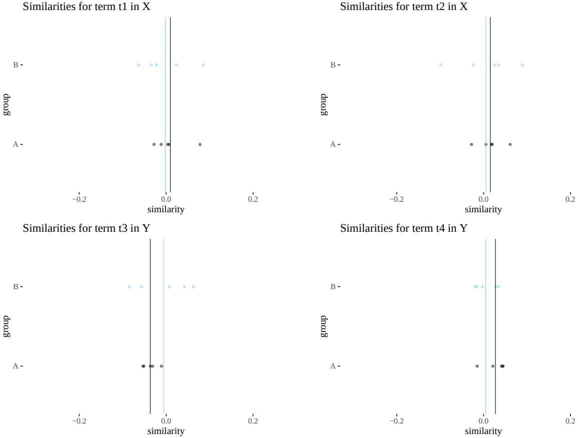
The problem with pre-averaging

One simulation



The problem with pre-averaging

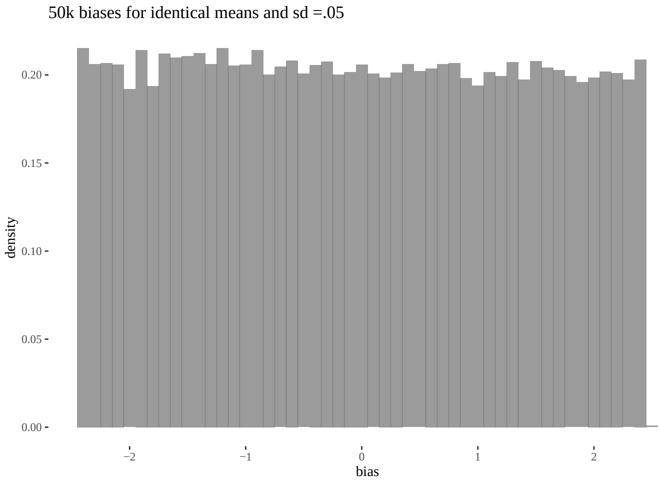
One simulation



- Raw sd in data is 0.045
- The sd of means is 0.023
- The WEAT score is 1.825
- The largest effect size reported by Caliskan et al. (2017) is 1.81!

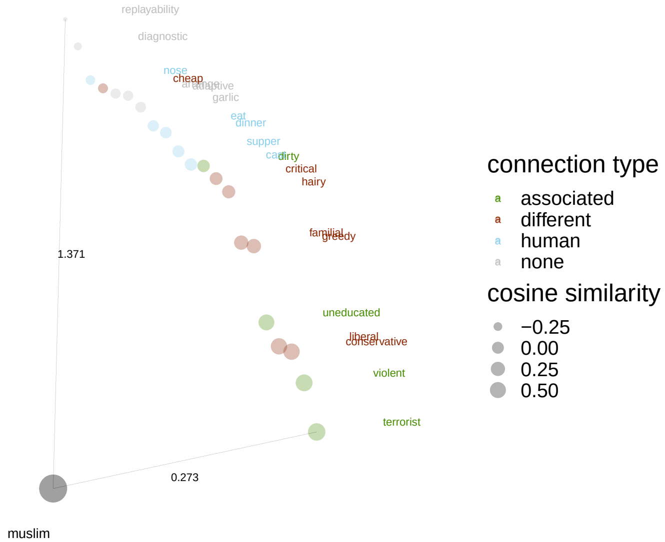
The problem with pre-averaging

50k simulations (same parameters)



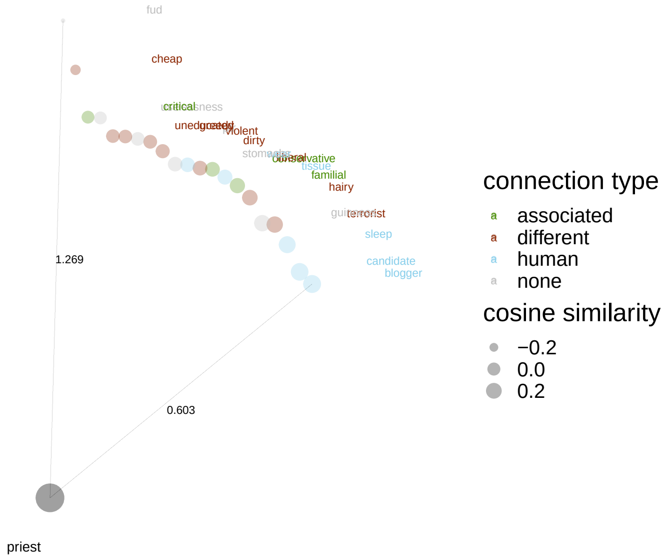
Cosine distance and connection type

Distances for “muslim”



Cosine distance and connection type

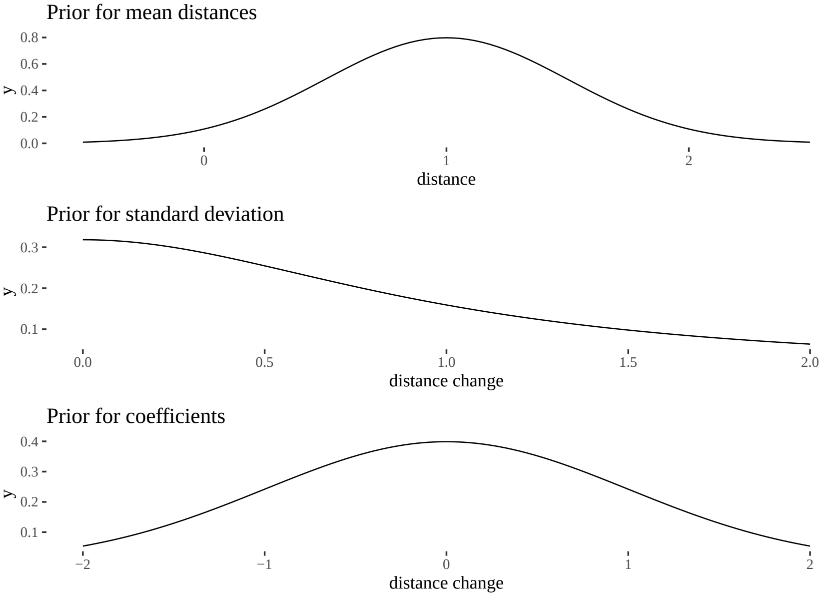
Distances for “priest”



Advantages of the Bayesian way

- Direct impact of sample sizes
- Straightforward interpretation in terms of posterior probabilities
- Freedom to choose granularity level
- More honest risk assessment and decision making

Bayesian model
Choosing priors

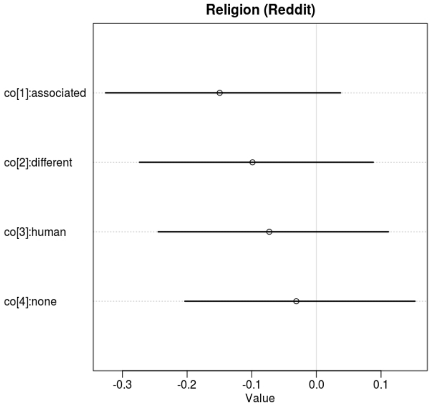


Bayesian model architecture

```
library(rethinking)
options(buildtools.check = function(action) TRUE )
religionCoefs <- ulam(
  alist(
    cosineDistance ~ dnorm(mu,sigma),
    mu <- m + co[con],
    m ~ dnorm(1,.5),
    co[con] ~ dnorm(0,.5),
    sigma ~ dcauchy(0,1)
  ),
  data = religion,
  chains=2 , iter=8000 , warmup=1000,
  log_lik = TRUE
)
```

Dataset-level coefficients

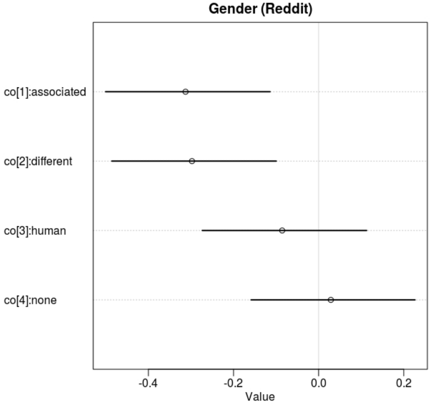
Religion with 89%-compatibility intervals (HPDI)



- All HPDIs overlap with 0
- Differences between classes are relatively small
- Coefficients for Race are similar

Dataset-level coefficients

Gender with 89%-compatibility intervals (HPDI)

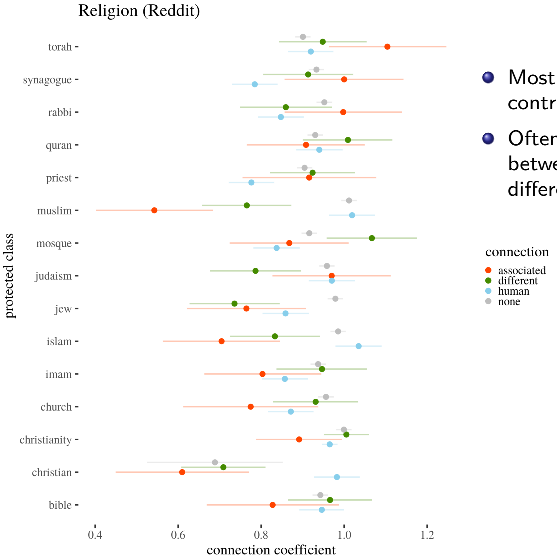


- Associated and different are away from 0
- But they were supposed to be opposites and are very close to each other (co-occurrence?)
- Differences between classes are still relatively small

Bayesian model architecture

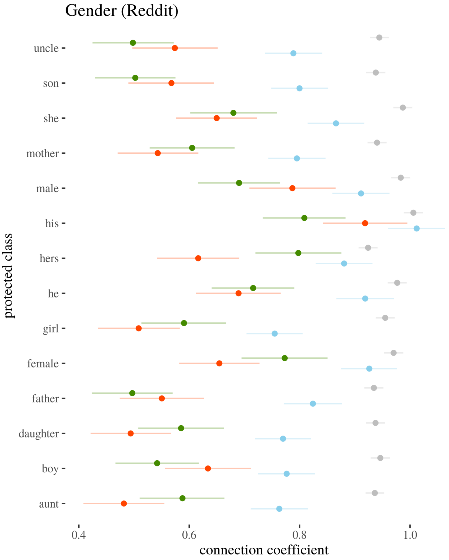
```r
library(rethinking)
options(buildtools.check = function(action) TRUE )
religionCoefs <- ulam(
  alist(
    cosineDistance ~ dnorm(mu,sigma),
    mu <- m[pw] + co[con],
    m[pw] ~ dnorm(1,.5),
    co[con] ~ dnorm(0,.5),
    sigma ~ dcauchy(0,1)
  ),
  data = religion,
  chains=2 , iter=8000 , warmup=1000,
  log_lik = TRUE
)
```

Word-level coefficients



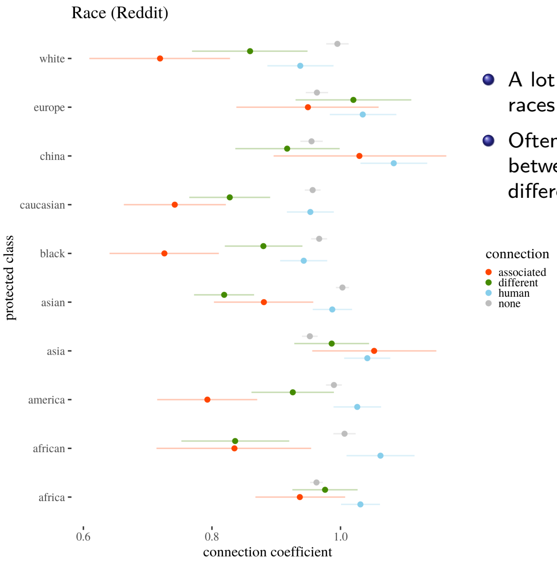
- Most intervals overlap with control groups
- Often not too much diffence between associated and different

Word-level coefficients



- Male attributes strong co-occurrence with female attributes
- Sometimes different is stronger than associated
- Almost no overlap with control groups

Word-level coefficients



- A lot of variation between races
- Often not much difference between associated and different

Further work

- Including contrasts in Bayesian calculation
- Performance cross-validation in comparison to other methods (regular linear regression, KNN, . . . ?)
- Downstream tasks and connection with intrinsic evaluation
- Testing data from the original Implicit Association Test (IAT)
- Applying uncertainty to WEAT and better word lists
- Looking at other similarity measures

References

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. 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. 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. *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)*, 609–614. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/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

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*.

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. https://doi.org/10.1162/coli_a_00379

Zhang, H., Sneyd, A., & Stevenson, M. (2020). *Robustness and reliability of gender bias assessment in word embeddings: The role of base pairs*.