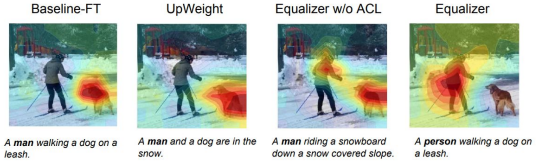
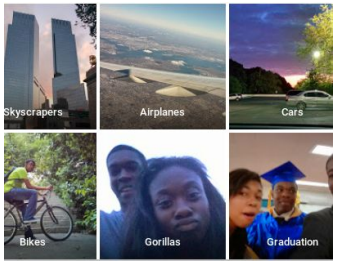
Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings

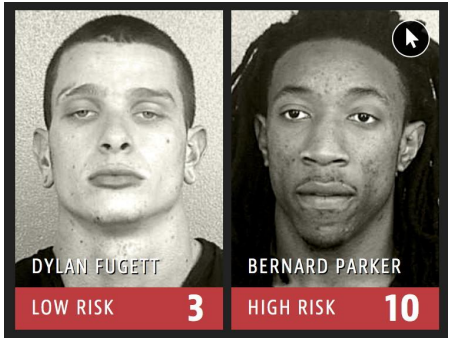
Thomas Manzini*, Yao Chong Lim*, Yulia Tsvetkov, Alan W Black

Why Care About Social Bias?

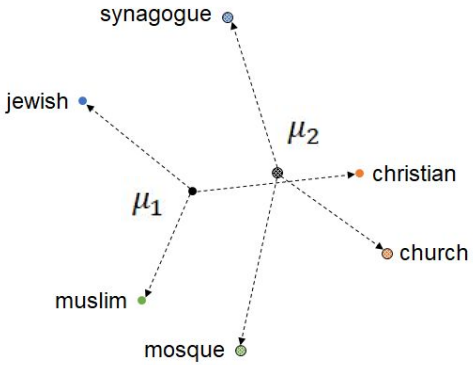




Bad Case - bias causes errors in machine learning systems with harmful consequences



Worst Case - bias causes detrimental harm to individuals, particularly for those from disadvantaged social groups



Our work explores multiclass bias (Race & Religion) in word embeddings and how to remove it

Possible Sources Of Social Bias?

The Training Data

The Model

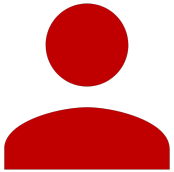
The Developer

Social bias can come from many places in a machine learning pipeline

Why Is Debiasing Important?

The Training Data

The Model

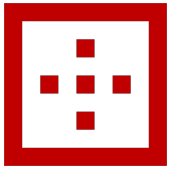


The Developer

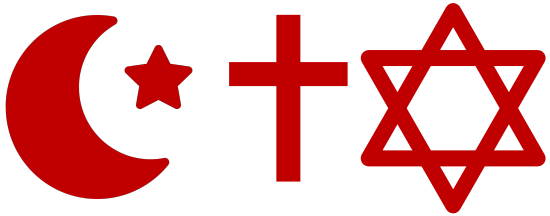
Debiasing can ***help*** these biases

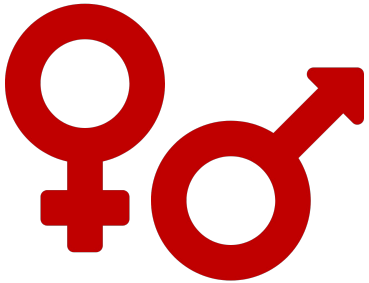
But it doesn't solve everything

Why Care About Multiclass Bias?

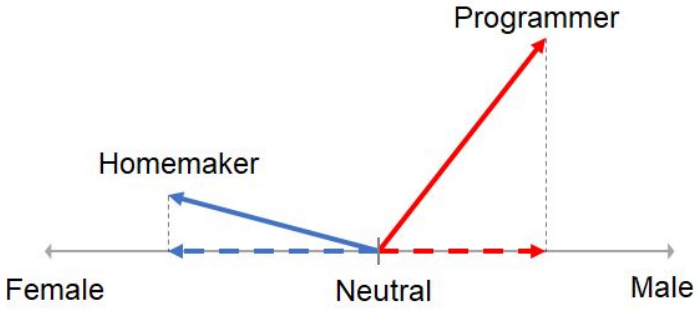


Many real world bias problems are multiclass





Bias In Word Embeddings?

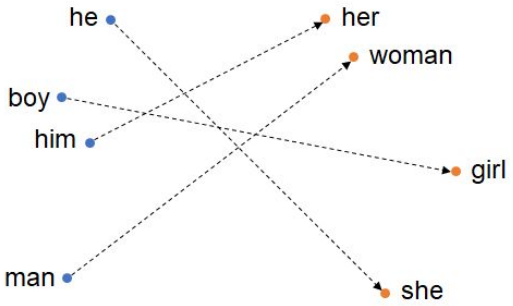


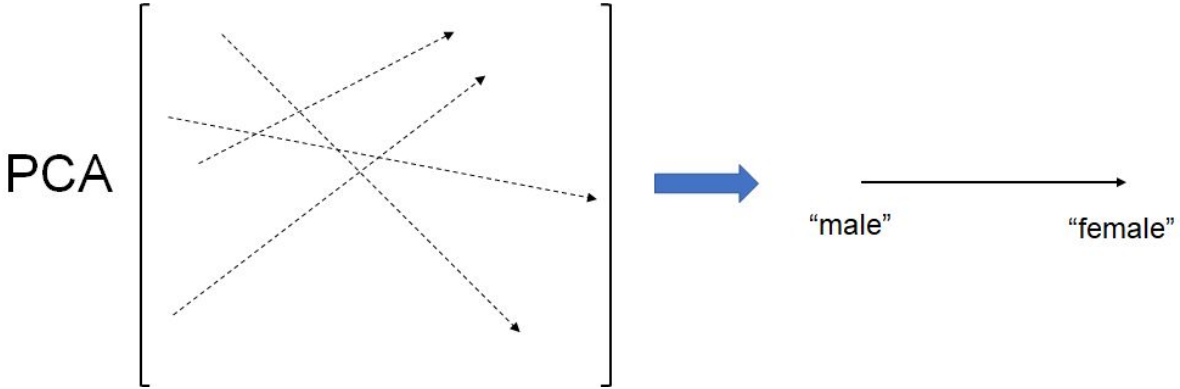
Geometric bias

(Bolukbasi et al., 2016)

Binary Debiasing of Word Embeddings

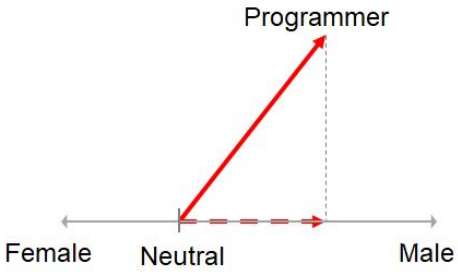
(Bolukbasi et al., 2016)



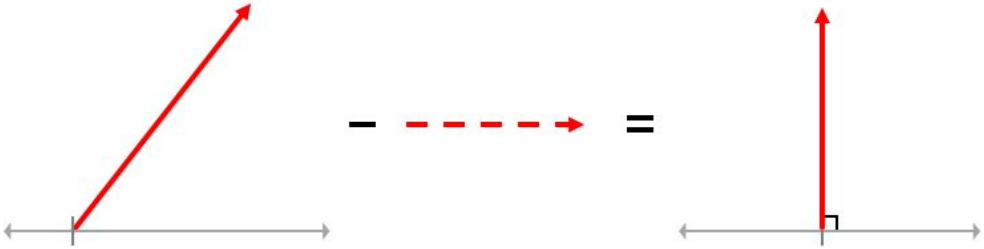


Binary Debiasing of Word Embeddings

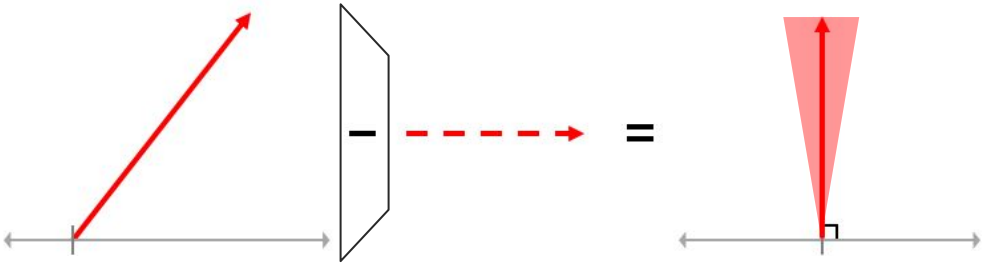
(Bolukbasi et al., 2016)



Hard debiasing

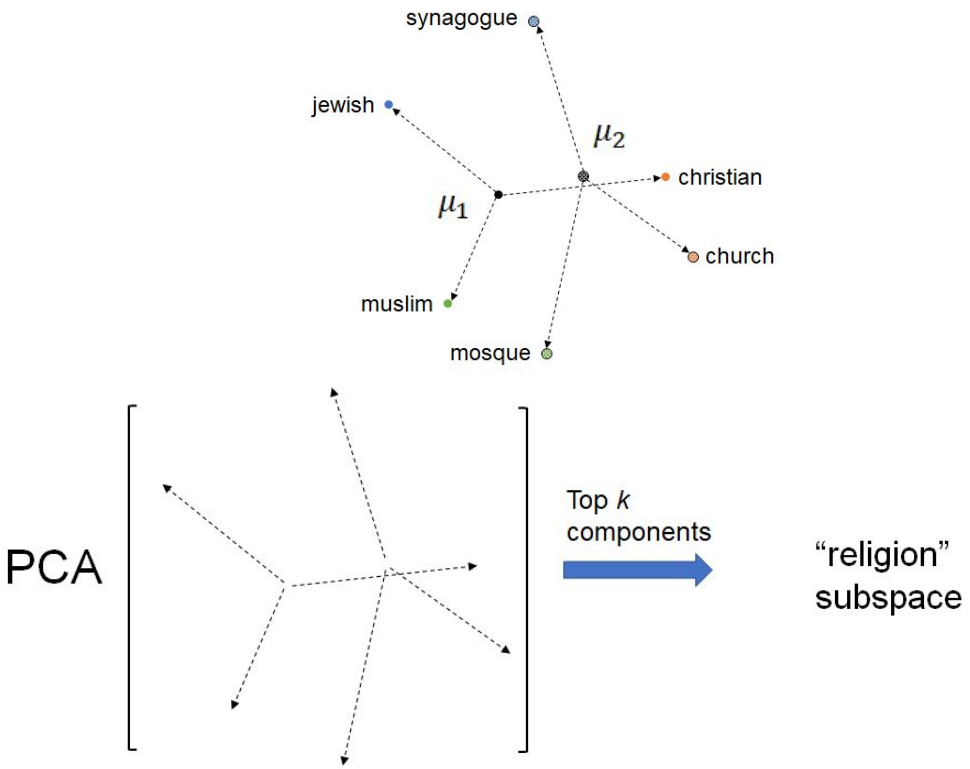


Soft debiasing



Learns a projection which balances hard debiasing with preserving inner products between embeddings.

Multiclass Debiasing of Word Embeddings



Evaluations

We Evaluate In Two Ways

Does the debiasing procedure actually decrease bias?

If we don't decrease bias… what's the point?

We propose a new evaluation metric called the Mean Average Cosine similarity (MAC score) in order to evaluate this.

Does the debiasing procedure preserve semantic utility?

If debiasing destroys the meaning… what's the point?

We evaluate our embeddings on the CoNLL 2003 shared task for POS tagging, NER tagging, and POS chunking

Dataset & Biased Analogies

Reddit L2 Corpus

(Rabinovich et al., 2018)

Racial Analogies
Black → Homeless, Caucasian → Servicemen
Caucasian → Hillbilly, Asian → Suburban
Asian → Laborer, Caucasian → Landowner
Religious Analogies
Jew → Greedy, Muslim → Powerless
Christian → Familial, Muslim → Warzone
Muslim → Uneducated, Christians → Intellectually

MAC Score - Definition

T = Set of word Embeddings which inherently contain social bias {Church, Synagogue, Mosque}

(Motivated by Caliskan et al. 2017)

A = Sets of attributes $A_1, A_2, ..., A_N$ containing words that shouldn't be associated with words T (e.g. {violent, liberal, conservative}).

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

Average Set Similarity across all T and A

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

Average Cosine Similarity of t for A_j

$$cos(\mathbf{u}, \mathbf{v}) = 1 - \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}$$

Cosine Similarity (1 = No similarity)

MAC Score - Results

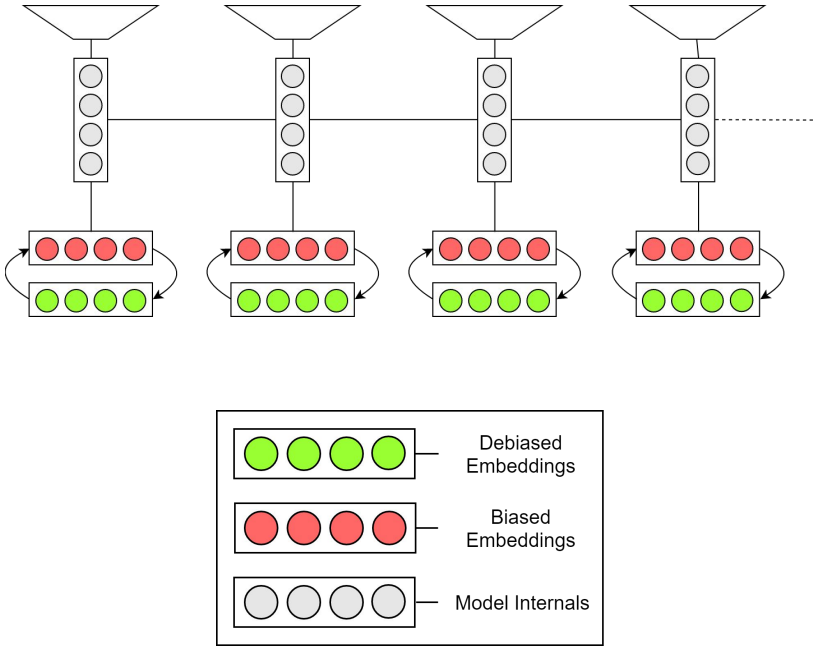
Gender Debiasing	MAC	p-Value
Biased	0.623	N/A
Hard Debiased	1.000	1.582e-14
Soft Debiased ($\lambda = 0.2$)	0.747	1.711e-12
Race Debiasing	MAC	p-Value
Biased	0.892	N/A
Hard Debiased	1.009	7.235e-04
Soft Debiased ($\lambda = 0.2$)	0.985	6.217e-05
Religion Debiasing	MAC	p-Value
Biased	0.859	N/A
Hard Debiased	1.004	3.006e-07
Soft Debiased ($\lambda = 0.2$)	0.894	0.007

MAC Scores categorically increase following the application of both hard and soft debiasing techniques.

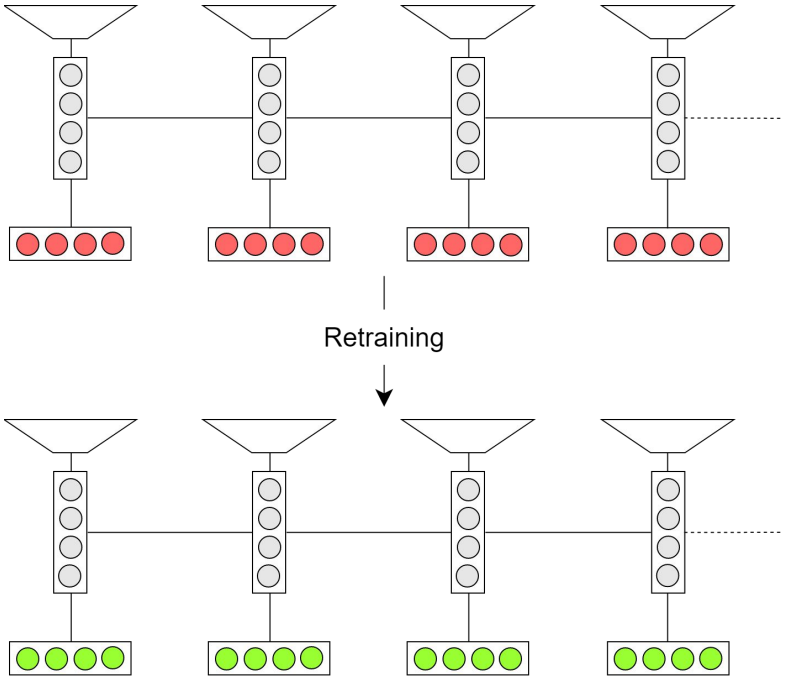
P-values for these MAC scores all decrease indicating statistical significance.

Downstream Evaluation

Embedding Matrix Replacement



Model Retraining



Downstream Evaluation

Embedding Matrix Replacement									
	Hard Gender Debiasing			Hard Racial Debiasing			Hard Religious Debiasing		
	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking
Biased F1	0.9954	0.9657	0.9958	0.9948	0.9668	0.9958	0.9971	0.9665	0.9968
Δ F1	+0.0045	-0.0098	+0.0041	+0.0051	-0.0117	+0.0041	+0.0103	-0.0345	+0.0120
Δ Precision	0.0	-0.0177	0.0	0.0	-0.0208	0.0	0.0	-0.0337	0.0
Δ Recall	+0.0165	-0.0208	+0.0156	+0.0186	-0.0250	+0.0155	+0.00286	-0.0174	+0.0031
	Soft Gender Debiasing			Soft Racial Debiasing			Soft Religious Debiasing		
	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking
Biased F1	0.9952	0.9614	0.9950	0.9946	0.9612	0.9946	0.9964	0.9616	0.9961
Δ F1	+0.0047	-0.0102	+0.0049	+0.0053	-0.0107	+0.0053	+0.0128	-0.0242	+0.0148
Δ Precision	0.0	-0.0202	0.0	0.0	-0.0223	0.0	0.0	-0.0199	0.0
Δ Recall	+0.0169	-0.0198	+0.0187	+0.0193	-0.0197	+0.0202	+0.0035	-0.0112	+0.0038

For each task we train a model on biased embeddings, replace the embeddings matrix

We only evaluate samples with tokens that were affected by debiasing

We are simply measuring the changes in outputs by changing inputs.

This doesn't debias the model...

Downstream Evaluation

Model Retraining									
	Hard Gender Debiasing			Hard Racial Debiasing			Hard Religious Debiasing		
	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking
Biased F1	0.9954	0.9657	0.9958	0.9948	0.9668	0.9958	0.9971	0.9665	0.9968
Δ F1	+0.0045	-0.0137	+0.0041	+0.0051	-0.0165	+0.0041	+0.0103	-0.0344	+0.0120
Δ Precision	0.0	-0.0259	0.0	0.0	-0.0339	0.0	0.0	-0.0287	0.0
Δ Recall	+0.0165	-0.0278	+0.0156	+0.0186	-0.0306	+0.0156	+0.00286	-0.0161	+0.0031
	Soft Gender Debiasing			Soft Racial Debiasing			Soft Religious Debiasing		
	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking
Biased F1	0.9952	0.9614	0.9950	0.9946	0.9612	0.9946	0.9964	0.9616	0.9961
Δ F1	+0.0047	+0.00178	+0.0049	+0.0053	-0.00119	+0.0053	+0.0128	-0.0098	+0.0148
Δ Precision	0.0	+0.0048	0.0	0.0	-0.00187	0.0	0.0	-0.0125	0.0
Δ Recall	+0.0169	+0.00206	+0.0187	+0.0193	-0.00264	+0.0202	+0.0035	-0.0057	+0.0038

For each task we retrain a new model with debiased embeddings.

Models can depend on bias subspaces differently depending on the task and model.

Our Contributions

Exploration into multiclass biases present in word embeddings

Methodology for computing the bias subspace enabling multiclass debiasing for word embeddings

An approach for evaluating the removal of multiclass bias word embeddings

Discussion

Reliance on human generated lists of words to define the bias subspace

Analogy task criticism by Nissim et al. (2019)
& Response by Prof. Kai-Wei Chang
http://kwchang.net/papers/response_Nissim_19.pdf

Does removing each “type” of bias change how we handle intersectional bias?

Future Work

We only address the geometric bias in this work and we do not address the cluster bias described in Gonen and Goldberg (2019).

Representation learning for words appears to be moving towards contextualized embeddings (ELMo, BERT, etc). It is unclear if these methods will remain robust for debiasing these techniques.

Most social classes in social bias are actually continuous variables. We must handle debiasing with this in mind.

Thank You!

Code & Vocabularies

https://github.com/TManzini/DebiasMulticlassWordEmbedding

Email:

Thomas.Manzini@Microsoft.com,
yaochonl@andrew.cmu.edu,
ytsvetko@cs.cmu.edu,
awb@cs.cmu.edu

Appendix: Related Work

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (Bolukbasi et al. 2016)

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al. 2017)

Word embeddings quantify 100 years of gender and ethnic stereotypes (Garg et al. 2018)

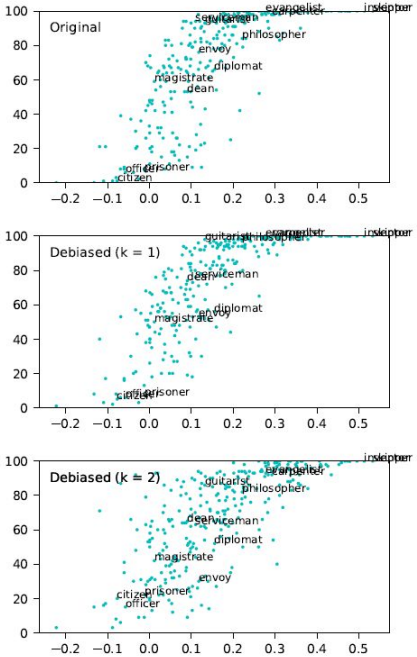
What's in a Name? Reducing Bias in Bios without Access to Protected Attributes (Romanov et al. NAACL 2019)

Appendix: Cluster Bias

Target	k	r	ρ
jew	0	0.767	0.875
	1	0.795	0.891
	2	0.718	0.756
	3	0.736	0.772
christian	0	0.925	0.947
	1	0.835	0.841
	2	0.825	0.831
	3	0.832	0.839
muslim	0	0.858	0.894
	1	0.774	0.812
	2	0.715	0.721
	3	0.712	0.718

Appendix: Cluster Bias

Neighbors to *jew*



Neighbors to *muslim*