Intersectional Bias in Word Embeddings

Morgan Wajda-Levie & Rebecca Kleinbart

December 2021

Github: `https://github.com/morganwl/csci795_intersections-of-bias`

1 Problem

1.1 Introduction

Natural language processing algorithms have made tremendous strides in their ability to learn semantic and syntactic patterns from a corpus of documents. Unfortunately, part of the patterns that machines learn is the underlying bias contained in the training corpus. Word embeddings are one such NLP tool that is susceptible to sterotypes and prejudice. In fact, it has been shown that word embeddings not only capture, but amplify, existing biases in training documents [4]. When we consider that word embeddings contribute to many fields such as translation, hiring [6], and medical algorithms it is clear that these biases can cause harm, especially to people who are part of marginalized groups. How can we *measure* and *remove bias* from training data? How does identifying *intersectional* bias differ from bias towards only one group? What can the eliminated documents in the training corpus teach us about intersectional bias overall?

1.2 What are word embeddings?

Word embeddings are representations of words as real-valued vectors. There are several types of word embeddings, but the version we will focus on is generated from the co-occurrence of nearby words, stemming from the idea that a word's semantic and syntactic meaning can be generated from its context within a document [3]. The theoretical basis of word embeddings is the linguistics field of statistical semantics, and in particular the 'Distributional Hypothesis' which states that words with similar meanings share similar contexts [7] and that the context in which a word is found contributes to its meaning. For example, the probability that "ice" and "solid" co-occur is high but the probability that "ice" and "gas" co-occur is low, which indicates that it is likely that ice shares the properties of solids and not gases [12].

1.3 How are word embeddings generated?

There are several methods of generating word embeddings, including calculating the angle or distance between word vectors. We will focus on GloVe, an unsupervised algorithm that uses a weighted least-squares model and a log bi-linear model to count global co-occurrences in order to generate vectors that represent each word [12]. To understand GloVe, let's define a few terms. X_{ij} is the number of times word j occurs in the context of word i. Continuing with our ice example, that could be the number of times *ice* co-occurs with *solid.* The cost [1] for GloVe word embeddings is:

$$\hat{J} = \sum_{i,j} f(X_{ij})(w_i^T \tilde{w}_j - log(X_{ij}))^2 \tag{1}$$

where $f(X_{ij})$ is the power law function, representing the weighting function.

$$f(x) = \begin{cases} (x/x_{max})^{\alpha} & if x < x_{max} \\ 1 & otherwise \end{cases} \tag{2}$$

If the weight co-occurrence is greater than or equal to a particular threshold, the weight will be 1. If not, it will be smaller using the indicated ratio. [12]

1.4 Examples of word embedding bias

Word embeddings reproduce and amplify biases and stereotypes that exist in their training documents. Caliskan, Bryson, and Narayanan [4] demonstrate that many of the biases that already exist in society [2] can be found in word embeddings learned by the GloVe algorithm. Some of these biases are inoffensive, like *flowers* being more "pleasant" than *insects* while others are pernicious like *European-American names* being more "pleasant" than *African-American names.* Additionally *female names and other female-related words* were more associated with "family" as opposed to "career" while *male names and other male-related words* were more associated with "career" as opposed to "family". More specifically, word embeddings can automatically generate analogies and some of the analogies generated by Bolukbasi et al. [2] were "**she** is to **sewing** as **he** is to **carpentry**", "**man** is to **computer programmer** as **woman** is to **homemaker**", and "**father** is to **doctor** as **mother** is to **nurse**".

1.5 Problem Description and Goals

Our goal is to expand our understanding of bias beyond the one-dimensional area of binary gender and into a broader, intersectional space of bias and discrimination towards multiple categories simultaneously such as race and gender or age and nationality. By looking at multiple forms of bias simultaneously, we hope to

[1] There are many steps to derive the cost function for for $\hat{J}$ but the overall idea is that we are modeling the function that represents P_{ik}/P_{jk}. In our example that would be ratio of the probability of "solid" given "ice" to the probability of "gas" given "ice". We want to encode this ratio using the using vectors representing each word in our corpus, w_i, w_j, $\tilde{w_k}$. The final cost function is a way of taking these three vector inputs and outputing a scalar which will represent our ratio. To do this we will need the dot product and a manipulated exponential function which becomes the the logarithm we see in the final equation.

[2] Caliskan et al. use results from the Implicit Association Test (IAT) to assess which biases already exist in society. In the IAT, participants see several terms and identify which pairs are similar. Faster response time correlates with stronger association between terms. For example if a person more quickly associates negative terms like "violent" with Arab Muslims than with other groups they may have an implicit bias against Arab Muslims.[8]

find commonalities and correlations that would allow us to cast a wide net for understanding, potentially eliminating more bias in our trained model than a uni-dimensional de-biasing would.[3]

Our experiment will start from the work of Brunet, Alkalay-Houlihan, Anderson and Zemel in their paper, "Understanding the Origins of Bias in Word Embeddings." in which they use an effective Word Embedding Association Test (WEAT) for measuring stereotypical gender associations in GloVe word embeddings, and a technique for estimating the impact that an individual document will have on the final bias of a word embedding[3]. Once the impact of a given document is estimated, the corpus can be *perturbed* by witholding that document when training the model, thereby reducing the resultant overall bias in the trained word embeddings. We will use WEAT to measure a large number of other bias features and then, after exploring the data directly, attempt to perform a regression on the bias features to estimate the area of effect on *intersectional bias*. We will evaluate the effectiveness of that estimate in allowing us to reduce a broad range of biases in our word embeddings, without adversely impacting the embeddings' performance.

2 Team Member Roles

Morgan focused on updating and modifying the python and Julia code, while Rebecca focused on obtaining training corpora, researching additional word embedding bias metrics, and incorporating that information into the project and the paper. Direction and analysis was performed in collaboration.

3 Related Work

Understanding underlying biases in word embeddings is an active area of research, with particular focus in binary gender bias. Researchers have proposed a number of ways to evaluate gender bias in word embeddings, and some techniques for reducing the level of bias[4, 3, 10, 1].

There are at least two approaches to removing bias:

1. De-biasing the source, in other words, the training data,

2. Removing bias once the word embedding vectors have been generated.

We will share brief overviews of each approach in the next two subsections.

[3] Any discussion of intersectionality is indebted to and dependent on the work of Crenshaw [5] who is author of the foundational scholarship on multidimensional discrimination. She writes "With Black women as the starting point, it becomes more apparent how dominant conceptions of discrimination condition us to think about subordination as disadvantage occurring along along a single categorical axis... This focus on the most privileged group members marginalizes those who are multiply-burdened and obscures claims that cannot be understood as resulting from discrete sources of discrimination.[5] In light of Crenshaw's critique on the study of uni-dimensional discrimination, we believe that machine learning de-biasing attempts must take intersectionality into account or suffer from over-simplicity and distortion.

3.1 De-biasing training data

3.1.1 Differential Bias and Bias Gradient

One clear source of bias in word embeddings is the training data. Word embedding algorithms such as GloVe and word2vec are trained on millions of documents, many of which include gender, racial, and other biases. Brunet et al. [3] focus on removing the most biased sources from the training data. They note that a naive approach would remove bias by training the embeddings on all sources except one, measure the bias, and then repeat the retraining on a new data set, each time removing only one source. Comparing the bias of each of these experiments could point to the training data sources that contribute the most to the overall bias of the embeddings. However this naive approach is not tenable due to the time and computation demands it would require. For this reason they were motivated to create an algorithm that they claim approximates this naive approach, with manageable time complexity.[3]

Brunet et al. [3] measure how perturbing the training data by removing particular documents changes the bias of the word embedding. Two of the significant concepts from this work are *differential bias* (Equation 3) and *bias gradient* (Equation 4). Differential bias, $\Delta_p B$, is the difference between the bias of the word embedding learned from the entire original corpus minus the bias of the word embedding learned from the perturbed corpus, where p is one of the parts of the partitioned corpus, w is the word embedding, $\tilde{w}$ is the word embedding learned from the altered or perturbed corpus, and *B(w)* is the bias metric[4].

$$\Delta_p B = B(w) - B(\tilde{w}) \tag{3}$$

If B(w), the bias metric, is a differentiable function and if we can approximate a word vector as a differentiable function with input X (the co-occurence matrix), then we can represent the bias gradient of word vector w as the derivative of the bias metric using the chain rule.

$$\nabla_X B(w(X)) = \nabla_w B(w) \nabla_X w(X) \tag{4}$$

This bias gradient can also be used to linearly approximate the bias of the corpus when it is perturbed (a certain document is removed) and also the differential bias of that particular document.

This is done using inverse Hessians of the total loss when we perturb the training set. Brunet et al. apply the differential bias formula (Equation 3) and this loss function, modifying it for efficiency in the following way:

$$\tilde{w_i} \approx w_i^* - \frac{1}{V} H_{w_i}^{-1} [\nabla_{w_i} L(\tilde{X}_i, w) - \nabla_{w_i} L(X_i, w)] \tag{5}$$

[4] The notation here aligns with Pennington, Socher, and Manning [12]'s GloVe algorithm

when $L(z_i, \theta)$ is the point-wise loss, V is dimension of the co-occurrence matrix and w^* is the learned embedding. The perturbed embedding can be approximated using Equation 5 by subtracting the vector representing the learned word embedding from the loss generated in perturbation. [3]

3.1.2 Word Embedding Association Tests

Brunet et al. use the *Word Embedding Association Test* to measure bias in training data[3, 4], and we will do the same. The WEAT test begins with two sets of equally sized target words $\mathcal{S}$ and $\mathcal{T}$ and two sets of attribute words, $\mathcal{A}$ and $\mathcal{B}$. [3] [5] Using cosine similarity, we can quantify the similarity of two words, a and b in word embedding w: $cos(w_a, w_b)$.

Using this cosine similarity of word embeddings as a measure of similarity and building off of the Implicit Association Test from the field of psychology, Caliskan et al. compare the cosine similarity of one target word, c to an attribute word in one set and then that same word to a word in the other attribute set and subtract these to get the *differential association* (Equation 5).They argue that the distance between two vectors is similar to the reaction time in the IAT test.

$$g(c, \mathcal{A}, \mathcal{B}, w) = a \overset{\text{mean}}{\in} \mathcal{A} cos(w_c, w_a) - b \overset{\text{mean}}{\in} \mathcal{B} cos(w_c, w_b) \tag{6}$$

To measure the overall effect size the normalized equation, Equation 6 is used to calculate the differential associations between the two sets of target words and attribute words.

$$B_{weat}(w) = \frac{mean_{s \in \mathcal{S}} g(s, \mathcal{A}, \mathcal{B}, w) - mean_{t \in \mathcal{T}} g(t, \mathcal{A}, \mathcal{B}, w)}{stddev_{c \in \mathcal{S} \cup \mathcal{T}} g(c, \mathcal{A}, \mathcal{B}, w)} \tag{7}$$

3.2 De-biasing learned vectors and its criticism

Bolukbasi et al. [2]'s paper, "Man is to computer programmer as woman is to homemaker? Debiasing Word Embeddings", begins with word embeddings from the algorithm word2vec that are trained the Google News dataset. They introduce two related de-biasing algorithms: Neutralize and Equalize and Soften. Neutralize and Equalize modifies gender neutral words' vectors so that they are equidistant to pairs of gendered words. For example, "babysit" should be exactly between grandmother and grandfather. Soften is similar to Neutralize and Equalize, but does not modify the neutral word's embedding as much. The authors claim that their algorithms significantly reduce bias "while preserving the utility of the embedding" [2].

Gonen and Goldberg [10] take issue with Bolukbasi et al.'s approach, calling it a "party trick" as they claim that it obscures gender bias but does not eliminate it. They demonstrate that gender neutral words

[5] The original Caliskan [4] paper uses $\mathcal{X}$ and $\mathcal{Y}$ to represent the target sets but we are following the Brunet paper's notation [3] and using $\mathcal{S}$ and $\mathcal{T}$.

	Simple English WikiCorpus	**NYT Annotated Corpus**
Corpus		
Minimum Document Length	200	100
Maximum Document Length	10,000	30,000
Number of Documents	29,344	1,412,846
Number of Tokens	17,033,637	975,624,317
Vocabulary		
Token Minimum Count	15	15
Vocabulary Size	44,806	213,687

Table 1: Corpora Statistics

which had been biased, in the sense that they were closer to one gendered word than another (as nurse is closer to gal than guy), frequently remain biased after the post-processing ”de-biasing” step. While the vectors may have change in one dimension, Gonen and Goldberg show that words that had formerly been biased still cluster together. For example hairdresser, nurse, and receptionist maintain gender implicit-bias as they all are closer to each other to male-associated words like captain.

4 Approach

We have seen two approaches to de-biasing word embeddings, and our work follows the first: de-biasing the training data. We believe that an intersectional approach will require us to address issues in the training data, and furthermore Gonen and Goldberg [10] have poked holes in Bolukbasi et al. [2]'s post-processing de-biasing, so we find a different approach to be warranted.

Our work expands upon Brunet et al.'s by using their Approximating Differential Bias algorithm (Equation 5) to test bias beyond binary gender bias. We will use WEAT tests based on Implicit Association Tests to measure and remove intersectional bias, with a particular focus on the intersection between gender and race.

4.1 Dataset

4.1.1 Datasets and their modifications

We started with two primary data sets: the Simple English WikiCorpus, which is a simplified Wikipedia with fewer articles and simplified vocabulary [15] and the New York Times Annotated Corpus, which contains over 1 million New York Times articles written between January 1, 1987 and June, 19, 2007 [13] (more details can be found in Table 1). [6]

[6] As mentioned above, we are using the same corpora as Brunet et al. [3]. This table is organized similarly to their table organization in Supplemental Material B.

These datasets align with the corpora chosen by Brunet et al. [3] and also reflect two sets of texts that cover many diverse topics and are used in a great deal of commercial and academic experimentation, so they are appropriately typical for training our word embeddings.

Due to its immense size, running the differential bias algorithm on the complete NYT corpus took a very long time: Estimating the differential bias for a single WEAT set on a single document took roughly .24 seconds. With over 1.4 million documents in the corpus, this resulted in roughly 4 days of uninterrupted processing time for one WEAT set. To reduce runtime and allow the experiment to finish, we used a subset of the entire corpus containing roughly 20% of the articles in the NYT Annotated Corpus which we have labeled ”Select 0.2” in the subsequent tables.

Articles and their corresponding metadata were sampled using a deterministically random hash function, much like partitioning a training set and test set so that multiple subsampling of the same corpus would produce the same subcorpus. Working with a substantially smaller corpus resulted in a much shorter runtime. Using our reduced data set (New York Times Annotated Corpus Select 0.2), differential bias estimation for a single WEAT set on a single document took roughly .1 seconds. With fewer documents to estimate, we could estimate every document in the corpus in roughly 8 hours.Therefore we were able to run almost all of our WEAT tests on all documents in the corpus, dozens more than previous experiments.

While the median change in effect size from the full corpus to the reduced corpus, excluding WEAT tests with WEAT values greater than .1, was 15%, some effect sizes measured a change as high as 35%. As we will see later, a very small portion of documents in the corpus account for the bulk of the bias effect size, so by reducing the size of the corpus, it is possible that we failed to analyze a significant portion of the documents with the greatest effect. Furthermore, fewer documents means smaller vocabularies, including some of the words used in our WEAT sets. To preserve parity between paired sets, when a set contained words that did not exist in an embedding's vocabulary, we dropped an equal number of words from the paired set, starting with the least frequently occurring.

4.2 Experimental Bias Metric

Following Brunet et al. [3] and Caliskan, Bryson, and Narayanan [4] we use Word Embedding Association Tests (WEAT) to measure bias. As mentioned in Section 5.2, these tests are meant to reflect the Implicit Association Tests that illustrate human implicit bias. We went well beyond Brunet et al. and Caliskan et al., however, in that we ran more than twenty WEAT sets, where Brunet only tested two [7], and specificaly considered intersectional or multi-dimensional bias, where Caliskan ony test uni-dimensional bias. We also

[7] In Brunet's WEAT1 test they used the target word sets *science* and *arts* and the attribute sets *male terms* and *female terms.* In WEAT2 they used the target sets *musical instruments* and *weapons* and the attribute sets *pleasant* and *unpleasant*[3]

	WEAT A [14]	**WEAT B** [14]
Target Sets		
$\mathcal{S}$	male	European-American Names
$\mathcal{T}$	female	African-American Names
Attribute Sets		
$\mathcal{A}$	Likeable, Not Hostile	Likeable, Not Hostile
$\mathcal{B}$	Unlikeable, Hostile	Unlikeable, Hostile
effect size		
Simple English Wikipedia	1.007294807 (p = 0.0208)	1.225371199 (p = 0)
NYT Annotated Corpus (Select)	0.9507512039 (p = 0.0262)	1.310189047 (p = 0)
NYT Annotated Corpus (Select 0.2)	1.059083535 (p = 0.0125)	1.504042256 (p = 0)

Table 2: Pleasantness WEAT tests

	WEAT C [14]	**WEAT D** [14]	**WEAT E** [14]	**WEAT F**
Target Sets				
$\mathcal{S}$	White Female Names	White Male Names	White Male Names	European-American Names
$\mathcal{T}$	Black Female Names	Black Male Names	Black Female Names	African-American Names
Attribute Sets				
$\mathcal{A}$	Pleasant	Pleasant	Pleasant	Pleasant
$\mathcal{B}$	Unpleasant	Unpleasant	Unpleasant	Unpleasant
effect size				
Simple English Wiki	1.56881508 (p = 0.0019)	1.593806989 (p = 0)	1.666436008 (p = 0)	1.593407274 (p = 0)
NYT (Sel.)	1.337088663 (p = 0.0001)	1.53724698 (p = 0)	1.478455673 (p = 0)	1.175534449 (p = 0)
NYT (Sel. 0.2)	1.039770228 (p = 0.0085)	1.15362045 (p = 0.0021)	1.10823558 (p = 0.0067)	1.359190787 (p = 0)

Table 3: Likeability WEAT tests

included additional types of bias including bias relating to age, race, and mental illness.

In Tables 2, 3, 4, and 5, we have included the list of some of the WEAT sets that we measured.

5 Experiments and Evaluation

5.1 Experimental Procedure

Having prepared our 20% subsampling of the New York Times corpus, we trained a GloVe word embedding to serve as our baseline for evaluating WEAT effect sizes and estimating differential biases of individual documents. We chose GloVe hyperparameters based on those used in Brunet et al's research, though we increased the minimum vocabulary from 15 to 20 and reduced the number of iterations, because these increased performance at a tolerable cost[3]. On a 2018 iMac, this took about 5 hours.

Using our modified version of the `diff_bias` program from Brunet et al, we estimated the differential bias on a number of WEAT sets for every document in our corpus, yielding results for 11 WEAT sets on

	WEAT H [14]	**WEAT I** [14]
Target Sets		
$\mathcal{S}$	Male Names (white)	Male Names (white and Black)
$\mathcal{T}$	Female Names (white)	Female Names (white and Black)
Attribute Sets		
$\mathcal{A}$	Career	Male Occupation
$\mathcal{B}$	Family	Female Occupation
effect size		
Simple English Wikipedia	1.633191877 (p = 0.0002)	1.247440846 (p = 0)
NYT Annotated Corpus (Select)	1.842213489 (p = 0)	1.233416076 (p = 0)
NYT Annotated Corpus (Select 0.2)	1.732417501 (p = 0.0001)	1.255035318 (p = 0)

Table 4: Career and Occupation WEAT tests

	WEAT J [14]	**WEAT K** [14]
Target Sets		
$\mathcal{S}$	Male Names (white)	European-American Names (multiple genders)
$\mathcal{T}$	Female Names (white)	African-American Names (multiple genders)
Attribute Sets		
$\mathcal{A}$	Competent, Achievement Oriented	Competent, Achievement Oriented
$\mathcal{B}$	Incompetent, Not Achievement Oriented	Incompetent, Not Achievement Oriented
effect size		
Simple English Wikipedia	0.9804834197 (p = 0.0255)	1.589065333 (p = 0)
NYT Annotated Corpus (Select)	1.270389026 (p = 0.0031)	1.298326608 (p = 0)
NYT Annotated Corpus (Select 0.2)	1.35317625 (p = 0.0017)	1.540194784 (p = 0)

Table 5: Career and Occupation WEAT tests

Table 6: comparison of full and reduced NY Times corpus

index	delta_effect_sizes	delta_p_values	effect_sizes_ratio	p_values_ratio
WEAT A	0.108332	0.0137	0.102289	1.096000
WEAT B	0.193853	0.0000	0.128888	0.000000
WEAT C	0.297318	0.0084	0.285946	0.988235
WEAT D	0.383627	0.0021	0.332541	1.000000
WEAT E	0.370220	0.0067	0.334063	1.000000
WEAT F	0.183656	0.0000	0.135122	0.000000
WEAT FF	0.361930	0.0059	0.387186	1.000000
WEAT H	0.109796	0.0001	0.063377	1.000000
WEAT I	0.021619	0.0000	0.017226	0.000000
WEAT J	0.082787	0.0014	0.061180	0.823529
WEAT K	0.241868	0.0000	0.157037	0.000000
WEAT L	0.587134	0.2841	0.718016	5.727823
WEAT M	0.180225	0.0127	0.151617	0.808917
WEAT N	0.149681	0.0000	0.105932	0.000000

283,744 documents. Performed in multiple batches on a 2018 iMac, this totaled in all around 24 hours of computation.

We calculated WEAT effect sizes for these 11 sets, as well as 15 others. We disregarded differential bias estimates and effect sizes for all WEAT sets with a p-value greater than 0.05, leaving us with estimated differential biases for 7 WEAT sets and 7 more effect sizes without corresponding differential biases. Likewise, we calculated these WEAT effect sizes for an extant word embedding made from the entire NYT corpus, to serve as a comparison with our reduced corpus. It is worth noting that, unlike ours, this embedding was trained with a minimum vocab count of 15 and 150 iterations[3].

5.2 Evaluation

5.2.1 Evaluating the Reduced Corpus

While the size of our reduced corpus was not chosen algorithmically, we did some evaluation afterwards to judge the impact of the smaller corpus on our WEAT effect sizes. By comparing the effect sizes and p-values as measured on the complete New York Times corpus and our reduced corpus, we observed a significant, but workable difference in effect sizes and p-values: the median change in effect size was 14.3%, with one dramtic outlier of 71.8%. While p-values changed by a much higher ratio (a median of 90.1%), only one set changed enough to move it in or out of our threshold — the same set that say a 71.8% change in effect size.

Given more time, a more thorough approach would have attempted to optimize the corpus size by algorithmically measuring effect sizes on variously-sized corpora, minimizing error against a loss function of corpus size, in a manner somewhat akin to gradient descent. Unfortunately, training a single embedding on a corpus of this size takes hours, so this search would easily have consumed several days of computation.

Table 7: differential bias contributed by documents in select percentiles

index	mean	std	min	25%	50%	75%	max
0.90	0.921532	0.932433	0.985476	0.966592	0.936197	0.896796	0.802273
0.99	0.488092	0.891427	0.620132	0.555852	0.501013	0.432359	0.319078

Table 8: correlation between documents in 90th percentile of WEAT tests

index	WEAT B	WEAT F	WEAT H	WEAT J	WEAT K	WEAT L	WEAT M	composite_score
WEAT B	1.000	0.231	0.056	0.050	0.212	0.079	0.038	0.241
WEAT F	0.231	1.000	0.053	0.054	0.135	0.108	0.052	0.298
WEAT H	0.056	0.053	1.000	0.101	0.049	0.040	0.031	0.360
WEAT J	0.050	0.054	0.101	1.000	0.182	0.049	0.037	0.141
WEAT K	0.212	0.135	0.049	0.182	1.000	0.061	0.037	0.183
WEAT L	0.079	0.108	0.040	0.049	0.061	1.000	0.059	0.248
WEAT M	0.038	0.052	0.031	0.037	0.037	0.059	1.000	0.093
composite_score	0.241	0.298	0.360	0.141	0.183	0.248	0.093	1.000

5.2.2 Evaluating the Differential Biases

The documents with a significant impact on bias are so sparse that we needed to focus on the 99th percentile with three digits of precision to see meaningful changes in estimated differential biases. In fact, the 99th percentile accounts for an average of 48.8% of cumulative estimated bias among our results. The 90th percentile accounts for an average of 92.2%. For a corpus of 283,744 documents, this amounts to 28,374 and 2,837 documents, respectively.

We were able to do some minimal investigation into the correlation between respective WEAT tests. By comparing which documents are in the 95th percentile of differential bias for each WEAT test, we were able to observe some (greater than 10%) correlation between WEAT sets in the same general area of bias (anti-Blackness and gender bias), and slight (roughly 5%) correlation across some bias areas.

5.2.3 Evaluating the Estimation Algorithm

Having estimated the differential biases, our next step would have been to compare our estimates to real word embeddings and evaluate the quality of our estimates. To do this, we would have trained a new embedding with the most bias affecting documents removed and compared the new effect size with the cumulative estimated change in effect sizes for those documents. As a control, we would train an embedding with the same number of documents selected at random and removed

To determine whether withholding documents had negatively impacted the overall performance of our embedding, we would also evaluate the embeddings on traditional benchmarks. If the benchmarks were comparable with the original embedding, and the new embeddings saw improvements in bias effect sizes against the original embedding and the control, we would consider our experiment a success. The closer those improvements matched our estimate, the greater the success.

5.2.4 Evaluating the Most Significant Documents

We would also like to retrieve the documents with the greatest impact on the effect size, approximately the 99th percentile, and perform sentiment analysis and topic detection on them, so that we can learn more about the kinds of documents that are contributing most to the overall bias in an embedding.

Furthermore, while changes in metrics are important, we feel that approaching bias in natural language has to include a qualitative element. A hasty review of the three most impactful documents for a few datasets gave some suggestion of what we might find. The top two documents for *WEAT F*, a set that measures bias towards pleasant and unpleasant words between traditionally European-American names and traditionally African-American names, were articles about violent crime in New York City, though, ironically enough, the most impactful article was actually a critique of the way crimes committed by Black Americans are covered as opposed to crimes committed by white Americans[11, 9]. Meanwhile, for *WEAT B*, a set measuring the bias towards likable and unlikable words towards European-American and African-American names, the most impactful document was a series of glowing capsule movie reviews, all direct by and starring white people. While the preponderance of negative representation of Black Americans can clearly lead to anti-Black bias in NLP models, this also suggests that the bias could just as easily be fueled by an overabundance of positive representation for white Americans.

We believe that a great deal can be learned by attempting to see what kinds of documents contributed the most to bias.

6 Discussion

Our core observations are in the effect sizes and estimated differential biases for each WEAT set. We saw significant effect sizes (that is, greater than 1) on all but one of our statistically significant tests for anti-Blackness, comparable to the effects that Brunet et al measured for gender bias[3]. This demonstrates that anti-Blackness, as measured in a number of ways, is embedded in word embeddings trained on a typical corpus, warranting further study in that area.

Likewise, our estimated differential biases seem to uphold the hypothesis of Brunet et al, that a relatively small number of documents contribute most to the overall bias, and that witholding these documents from training corpora could help to mitigate bias in word embeddings trained on those corpora[3]. While time constraints did not allow us to dig as deeply into the correlations between different biases as we would have liked, the low correlations that we did observe were somewhat discouraging. That said, we still believe that attempting to train a model to identify documents that might contribute to a broad range of measurable biases, and attempting to optimize the recall, might still yield meaningful information. If witholding 1% of all documents in a corpus could result in a 50% reduction of bias in one given WEAT set, how many documents would we need to withold to mitigate a number of different biases? And can we select those documents by estimating a manageable number of WEAT sets?

Our minimal qualitative review of documents only provides anecdotal encouragement as to an understandable link between a document's content and its measurable effect on bias, though that effect might run counter to a document's intention, as in the case of "Youths Criticize Media on Coverage of Children" by Dennis Hevesi[11]. Understanding the ways in which content contributes to bias as learned by NLP models could help us improve the quality of content fed to those models in the future.

7 Machine Learning Lessons Learned

1. **Machine learning tools are only as good, or as unbiased, as their training data.** While this is certainly not an idea that we are the first to discover, our work confirms the importance of minimizing bias and shows that Machine Learning algorithms contain, and even amplify, human biases.

2. **Machine learning is not just one approach; it can provide quite different ways of solving a problem (even one it creates).** Researchers approaching de-biasing word embeddings at different steps (training data, once word embeddings are learned, etc.) really drives home the point that machine learning strategies are not monolithic and standardized; with creativity a proliferation of tools and algorithms may emerge.

3. **Working backwards may generate fruitful machine learning strategies and algorithms.** Knowing the form of the desired solution, for example we sought a scalar that represented the ratio of probabilities and were going to start with vector inputs, can provide a hint to the form of the formula we are seeking. (In the case of differential bias this led to using the dot product and manipulated exponential function to get our desired outcome.)

4. **Understanding the real processing needs of an experiment, from start to finish, is essential.** Calculating total processing times on a toy dataset, and estimating how times scale with larger datasets and different hyperparameters was a strategy that supported us in completing (some) parts of our project.

5. **Faster training could save time manyfold.** Take reduction of training and processing time seriously, and be wary of diminishing returns from larger datasets, more iterations and other hyperparameters that might increase the time needed for training or experiments.

6. **When possible, organize and chunk longer projects so that parts can be done in parallel.** While this is not a machine learning-specific lesson, it is one that we took away from this project. Especially with partner or group work, dividing tasks in such a way that tasks aren't only able to be done in sequence would have supported us in able to complete more of our goals and stretch-goals.

8 Challenges

8.1 Adapting legacy code

We initially encountered some challenges with updating and running, the legacy code used in Brunet et al. in order to build on on their Approximating Differential Bias Algorithm. We planned to build off of this code in order to spend most of our time implementing our own novel approach. Unfortunately, the code required many updates of deprecated language features, spread across tens of source files written in 3 programming languages, including an older form of Julia, a language we had not been familiar with.

The code was undocumented. Simply reproducing the steps to generate the results from the Brunet et al. paper required thorough study and it was tightly coupled, requiring restructuring several source files to allow for running only select WEAT sets and specifying which WEAT sets to run from the command line. While the code included some tests, they were minimal, making it difficult to confirm that adjustments were working as expected. We were successful in altering this code but it took much longer than anticipated.

8.2 Dataset and Processing Time

We sought large data sets that are as authentic as possible in the sense that they would be used in generating word embeddings in commercial or academic setting. Some such datasets were not readily available but we did obtain access to the NYT Annotated Dataset as well as the Simple English Wikipedia training data. While these sets had sufficient data to generate significant results, they took an immense amount of time to process. Even the "toy" dataset required hours to run to completion. In order to get any data in time, we needed to reduce the size of our training corpus dramatically (as described in section 5.1). Even after reducing the size of the corpus, our personal computers were prohibitively slow in processing experiments; we borrowed a more powerful desktop computer to get any results on time.

8.3 Obtaining WEAT sets

The Implicit Association Test underpins the WEAT test. While the IAT has been written about since 1998, IAT vocabulary sets, which are also used for WEAT, are still not widely distributed. Creating novel WEAT vocabulary sets requires its own psychological and sociological research for which we are not qualified. We are grateful to Drs Yi and Celis for publishing their IAT vocab sets in an easily digestible and adaptable manner.[14] While we did explore multiple areas of bias, most of our data comes from race and gender, with a few test analyzing bias relating to age and disability. We also attempted to analyze relgion-based bias, like antisemitism and anti-Islam, but our experiments yielded results with very high p-values so we could not make conclusions about differential bias in those cases.

9 Opportunities for Future Work

There are many directions that we could take our work in order to build on our findings and strengthen and improve the tests and data we used. We will describe just a few of the possible next steps we may take.

9.1 Alternate Additional Corpora

With significantly more time or computing power we could re-run our test on the entire New York Times Annotated Corpus in order to more accurately analyze the effect size, bias differential, and the most bias-ing documents. We could also choose a data set that has articles from more recent times or identify popular corpora that are used in important NLP algorithms like hiring and test those data sets for bias.

9.2 In-depth Biasing Documents Textual Analysis

One extension to this project, that had been an original stretch goal of ours but that we did not fully achieve, is to use Natural Language Processing to analyze the most and least biasing articles or paragraphs of the data. We could, for example, use sentiment analysis or topic analysis in order to identify patterns in these documents or train models to classify "biasing" and "not biasing" articles. This analysis could have many applications: it would allow us to better understand what makes a document contribute more to a bias differential which would help with future debiasing of training sets and ideally lead to less biased word embeddings. Additionally it could teach us lessons about journalism and even bias in general from a sociological perspective: what does biased writing look like? What sort of articles should editors of the New York Times seek to include, if any, or publish fewer of?

While we did not complete all of the goals, we learned a great deal about machine learning in general and word embedding bias in particular, and there are many rich next steps for continued exploration and analysis.

References

[1] Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. "WEFE: The Word Embeddings Fairness Evaluation Framework". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20}. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 430–436. ISBN: 978-0-9992411-6-5. DOI: 10.24963/ijcai.2020/60. URL: https://www.ijcai.org/proceedings/2020/60 (visited on 10/12/2021).

[2] Tolga Bolukbasi et al. *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. _eprint: 1607.06520. 2016.

[3] Marc-Etienne Brunet et al. "Understanding the Origins of Bias in Word Embeddings". In: *arXiv:1810.03611 [cs, stat]* (June 7, 2019). arXiv: 1810.03611. URL: http://arxiv.org/abs/1810.03611 (visited on 10/10/2021).

[4] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (Apr. 14, 2017), pp. 183–186. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aal4230. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.aal4230 (visited on 10/11/2021).

[5] Kimberle Crenshaw. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics". In: (), p. 31.

[6] "Differential Hiring using a Combination of NER and Word Embedding". In: *International Journal of Recent Technology and Engineering* (2020).

[7] John R. Firth. "The technique of semantics". In: *Papers in linguistics, 1934-1951*. Oxford: Oxford University Press, 1957, pp. 7–33.

[8] *Frequently Asked Questions*. URL: https://implicit.harvard.edu/implicit/faqs.html (visited on 01/27/2022).

[9] Anita Gates. "CRITIC'S CHOICE; Movies". In: *The New York Times* (May 29, 2005). ISSN: 0362-4331. URL: https://www.nytimes.com/2005/05/29/arts/television/critics-choice-movies.html (visited on 01/27/2022).

[10] Hila Gonen and Yoav Goldberg. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". In: *arXiv:1903.03862 [cs]* (Sept. 24, 2019). arXiv: 1903.03862. URL: http://arxiv.org/abs/1903.03862 (visited on 10/11/2021).

[11] Dennis Hevesi. "Youths Criticize Media On Coverage of Children". In: *The New York Times* (Nov. 19, 1990). ISSN: 0362-4331. URL: https://www.nytimes.com/1990/11/19/nyregion/youths-criticize-media-on-coverage-of-children.html (visited on 01/27/2022).

[12] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

[13] Evan Sandhaus. *The New York Times Annotated Corpus*. Artwork Size: 3250585 KB Pages: 3250585 KB Type: dataset. Oct. 17, 2008. DOI: 10.35111/77BA-9X74. URL: https://catalog.ldc.upenn.edu/LDC2008T19 (visited on 10/18/2021).

[14] Yi Chern Tan and L. Elisa Celis. "Assessing Social and Intersectional Biases in Contextualized Word Representations". In: *arXiv:1911.01485 [cs, stat]* (Nov. 4, 2019). arXiv: 1911.01485. URL: http://arxiv.org/abs/1911.01485 (visited on 01/26/2022).

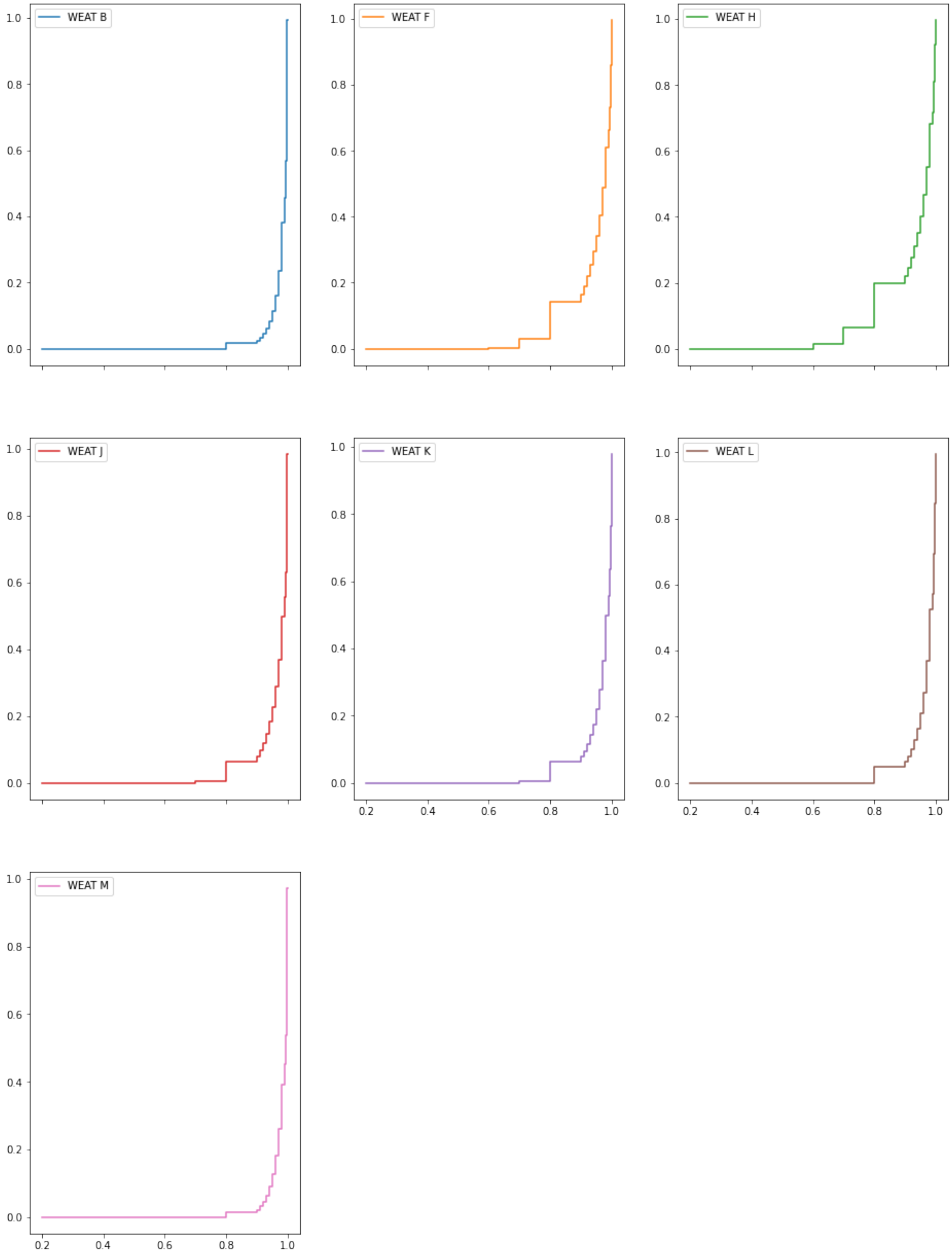
[15] Wikimedia. *Simplewiki:database download, 2021*. Oct. 1, 2021. URL: https://dumps.wikimedia.org/simplewiki/20211001/ (visited on 10/19/2021).

A Differential Bias Cumulative Distribution

Table 9: cumulative density of effect size vs quantile

quantile	WEAT B	WEAT F	WEAT H	WEAT J	WEAT K	WEAT L	WEAT M
0.100	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.200	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.300	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.400	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.500	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.600	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.700	0.000000	0.001534	0.012980	0.000000	0.000000	0.000000	0.000000
0.800	0.000011	0.028509	0.063790	0.005821	0.004817	0.000018	0.000000
0.900	0.017607	0.142457	0.197727	0.063952	0.063803	0.049208	0.014524
0.910	0.024488	0.164343	0.220662	0.078590	0.077956	0.063359	0.022045
0.920	0.033634	0.189611	0.246636	0.096604	0.095012	0.080776	0.032171
0.930	0.045740	0.218946	0.276404	0.118975	0.115911	0.102432	0.045871
0.940	0.061841	0.253188	0.311029	0.146972	0.141952	0.129654	0.064706
0.950	0.083550	0.293950	0.352180	0.182435	0.175012	0.164415	0.091057
0.960	0.114003	0.343433	0.402374	0.227709	0.218280	0.210604	0.128559
0.970	0.159694	0.405670	0.465869	0.287321	0.277080	0.274468	0.182002
0.980	0.234738	0.488251	0.551879	0.370214	0.362277	0.368296	0.260963
0.990	0.379868	0.609909	0.680922	0.498987	0.497035	0.525373	0.391262
0.991	0.402231	0.626223	0.698156	0.516627	0.515305	0.547489	0.409721
0.992	0.427143	0.643805	0.716561	0.535589	0.534955	0.571547	0.429829
0.993	0.455201	0.662731	0.736394	0.556013	0.556221	0.597613	0.451953
0.994	0.487097	0.683442	0.758082	0.578391	0.579538	0.626207	0.476545
0.995	0.524005	0.706703	0.782061	0.603214	0.605415	0.657681	0.504020
0.996	0.567748	0.733147	0.809183	0.631120	0.634589	0.693105	0.536050
0.997	0.620815	0.764284	0.839965	0.663506	0.668430	0.733958	0.573850
0.998	0.689560	0.804238	0.876162	0.702791	0.709118	0.782684	0.621282
0.999	0.783873	0.860330	0.921582	0.755021	0.763095	0.844872	0.685868
1.000	0.993802	0.997186	0.997507	0.984820	0.978332	0.994945	0.972183

Figure 1: Cumulative Distribution as percentage of total effect size vs percentile



B Weat Vocab Sets

```
{
    "WEAT A": {
        "attr1": {
            "category": "LikableNotHostile",
            "vocab": [
                "agreeable",
                "fair",
                "honest",
                "trustworthy",
                "selfless",
                "accommodating",
                "likable",
                "liked"
            ]
        },
        "attr2": {
            "category": "UnlikableHostile",
            "vocab": [
                "abrasive",
                "conniving",
                "manipulative",
                "dishonest",
                "selfish",
                "pushy",
                "unlikable",
                "unliked"
            ]
        },
        "link": "https://github.com/tanyichern/social-biases-contextualized",
        "source": "Tan et al, 2019",
        "targ1": {
            "category": "Male",
            "vocab": [
                "John",
                "Paul",
                "Mike",
                "Kevin",
                "Steve",
                "Greg",
                "Jeff",
                "Bill"
            ]
        },
        "targ2": {
            "category": "Female",
            "vocab": [
                "Amy",
                "Joan",
                "Lisa",
                "Sarah",
                "Diana",
                "Kate",
                "Ann",
                "Donna"
            ]
        }
    },
    "WEAT B": {
        "attr1": {
            "category": "LikableNotHostile",
            "vocab": [
                "agreeable",
                "fair",
                "honest",
                "trustworthy",
                "selfless",
            "accommodating",
            "likable",
            "liked"
        ]
    },
    "attr2": {
        "category": "UnlikableHostile",
        "vocab": [
            "abrasive",
            "conniving",
            "manipulative",
            "dishonest",
            "selfish",
            "pushy",
            "unlikable",
            "unliked"
        ]
    },
    "link": "https://github.com/tanyichern/social-biases-contextualized",
    "source": "Tan et al, 2019",
    "targ1": {
        "category": "EuropeanAmericanNames",
        "vocab": [
            "Adam",
            "Harry",
            "Josh",
            "Roger",
            "Alan",
            "Frank",
            "Justin",
            "Ryan",
            "Andrew",
            "Jack",
            "Matthew",
            "Stephen",
            "Brad",
            "Greg",
            "Paul",
            "Jonathan",
            "Peter",
            "Amanda",
            "Courtney",
            "Heather",
            "Melanie",
            "Katie",
            "Betsy",
            "Kristin",
            "Nancy",
            "Stephanie",
            "Ellen",
            "Lauren",
            "Colleen",
            "Emily",
            "Megan",
            "Rachel"
        ]
    },
    "targ2": {
        "category": "AfricanAmericanNames",
        "vocab": [
            "Alonzo",
            "Jamel",
            "Theo",
            "Alphonse",
            "Jerome",
            "Leroy",
            "Torrance",
            "Darnell",
                "Lamar",
                "Lionel",
                "Tyree",
                "Deion",
                "Lamont",
                "Malik",
                "Terrence",
                "Tyrone",
                "Lavon",
                "Marcellus",
                "Wardell",
                "Nichelle",
                "Shereen",
                "Ebony",
                "Latisha",
                "Shaniqua",
                "Jasmine",
                "Tanisha",
                "Tia",
                "Lakisha",
                "Latoya",
                "Yolanda",
                "Malika",
                "Yvette"
            ]
        }
    },
    "WEAT C": {
        "attr1": {
            "category": "Pleasant",
            "vocab": [
                "caress",
                "freedom",
                "health",
                "love",
                "peace",
                "cheer",
                "friend",
                "heaven",
                "loyal",
                "pleasure",
                "diamond",
                "gentle",
                "honest",
                "lucky",
                "rainbow",
                "diploma",
                "gift",
                "honor",
                "miracle",
                "sunrise",
                "family",
                "happy",
                "laughter",
                "paradise",
                "vacation"
            ]
        },
        "attr2": {
            "category": "Unpleasant",
            "vocab": [
                "abuse",
                "crash",
                "filth",
                "murder",
                "sickness",
                "accident",
                "death",
                "grief",
                "poison",
                "stink",
                "assault",
                "disaster",
                "hatred",
                "pollute",
                "tragedy",
                "bomb",
                "divorce",
                "jail",
                "poverty",
                "ugly",
                "cancer",
                "evil",
                "kill",
                "rotten",
                "vomit"
            ]
        },
        "link": "https://github.com/tanyichern/social-biases-contextualized",
        "source": "Tan et al, 2019",
        "targ1": {
            "category": "WhiteFemaleNames",
            "vocab": [
                "Allison",
                "Anne",
                "Carrie",
                "Emily",
                "Jill",
                "Laurie",
                "Kristen",
                "Meredith",
                "Molly",
                "Amy",
                "Claire",
                "Katie",
                "Madeline"
            ]
        },
        "targ2": {
            "category": "BlackFemaleNames",
            "vocab": [
                "Aisha",
                "Ebony",
                "Keisha",
                "Lakisha",
                "Latoya",
                "Tamika",
                "Imani",
                "Shanice",
                "Aaliyah",
                "Precious",
                "Nia",
                "Deja",
                "Latisha"
            ]
        }
    },
    "WEAT D": {
        "attr1": {
            "category": "Pleasant",
            "vocab": [
                "caress",
                "freedom",
                "health",
                "love",
                "peace",
            "cheer",
            "friend",
            "heaven",
            "loyal",
            "pleasure",
            "diamond",
            "gentle",
            "honest",
            "lucky",
            "rainbow",
            "diploma",
            "gift",
            "honor",
            "miracle",
            "sunrise",
            "family",
            "happy",
            "laughter",
            "paradise",
            "vacation"
        ]
    },
    "attr2": {
        "category": "Unpleasant",
        "vocab": [
            "abuse",
            "crash",
            "filth",
            "murder",
            "sickness",
            "accident",
            "death",
            "grief",
            "poison",
            "stink",
            "assault",
            "disaster",
            "hatred",
            "pollute",
            "tragedy",
            "bomb",
            "divorce",
            "jail",
            "poverty",
            "ugly",
            "cancer",
            "evil",
            "kill",
            "rotten",
            "vomit"
        ]
    },
    "link": "https://github.com/tanyichern/social-biases-contextualized",
    "source": "Tan et al, 2019",
    "targ1": {
        "category": "WhiteMaleNames",
        "vocab": [
            "John",
            "Paul",
            "Mike",
            "Kevin",
            "Steve",
            "Greg",
            "Jeff",
            "Brad",
            "Brendan",
            "Geoffrey",
            "Brett",
                "Matthew",
                "Neil"
            ]
        },
        "targ2": {
            "category": "BlackMaleNames",
            "vocab": [
                "Darnell",
                "Hakim",
                "Jermaine",
                "Kareem",
                "Jamal",
                "Leroy",
                "Rasheed",
                "DeShawn",
                "DeAndre",
                "Marquis",
                "Terrell",
                "Malik",
                "Tyrone"
            ]
        }
    },
    "WEAT E": {
        "attr1": {
            "category": "Pleasant",
            "vocab": [
                "caress",
                "freedom",
                "health",
                "love",
                "peace",
                "cheer",
                "friend",
                "heaven",
                "loyal",
                "pleasure",
                "diamond",
                "gentle",
                "honest",
                "lucky",
                "rainbow",
                "diploma",
                "gift",
                "honor",
                "miracle",
                "sunrise",
                "family",
                "happy",
                "laughter",
                "paradise",
                "vacation"
            ]
        },
        "attr2": {
            "category": "Unpleasant",
            "vocab": [
                "abuse",
                "crash",
                "filth",
                "murder",
                "sickness",
                "accident",
                "death",
                "grief",
                "poison",
                "stink",
                "assault",
                "disaster",
                "hatred",
                "pollute",
                "tragedy",
                "bomb",
                "divorce",
                "jail",
                "poverty",
                "ugly",
                "cancer",
                "evil",
                "kill",
                "rotten",
                "vomit"
            ]
        },
        "link": "https://github.com/tanyichern/social-biases-contextualized",
        "source": "Tan et al, 2019",
        "targ1": {
            "category": "WhiteMaleNames",
            "vocab": [
                "John",
                "Paul",
                "Mike",
                "Kevin",
                "Steve",
                "Greg",
                "Jeff",
                "Brad",
                "Brendan",
                "Geoffrey",
                "Brett",
                "Matthew",
                "Neil"
            ]
        },
        "targ2": {
            "category": "BlackFemaleNames",
            "vocab": [
                "Aisha",
                "Ebony",
                "Keisha",
                "Lakisha",
                "Latoya",
                "Tamika",
                "Imani",
                "Shanice",
                "Aaliyah",
                "Precious",
                "Nia",
                "Deja",
                "Latisha"
            ]
        }
    },
    "WEAT F": {
        "attr1": {
            "category": "Pleasant",
            "vocab": [
                "caress",
                "freedom",
                "health",
                "love",
                "peace",
                "cheer",
                "friend",
                "heaven",
                "loyal",
            "pleasure",
            "diamond",
            "gentle",
            "honest",
            "lucky",
            "rainbow",
            "diploma",
            "gift",
            "honor",
            "miracle",
            "sunrise",
            "family",
            "happy",
            "laughter",
            "paradise",
            "vacation"
        ]
    },
    "attr2": {
        "category": "Unpleasant",
        "vocab": [
            "abuse",
            "crash",
            "filth",
            "murder",
            "sickness",
            "accident",
            "death",
            "grief",
            "poison",
            "stink",
            "assault",
            "disaster",
            "hatred",
            "pollute",
            "tragedy",
            "bomb",
            "divorce",
            "jail",
            "poverty",
            "ugly",
            "cancer",
            "evil",
            "kill",
            "rotten",
            "vomit"
        ]
    },
    "link": "https://github.com/tanyichern/social-biases-contextualized",
    "source": "Tan et al, 2019",
    "targ1": {
        "category": "EuropeanAmericanNames",
        "vocab": [
            "Adam",
            "Harry",
            "Josh",
            "Roger",
            "Alan",
            "Frank",
            "Justin",
            "Ryan",
            "Andrew",
            "Jack",
            "Matthew",
            "Stephen",
            "Brad",
            "Greg",
            "Paul",
                "Jonathan",
                "Peter",
                "Amanda",
                "Courtney",
                "Heather",
                "Melanie",
                "Katie",
                "Betsy",
                "Kristin",
                "Nancy",
                "Stephanie",
                "Ellen",
                "Lauren",
                "Colleen",
                "Emily",
                "Megan",
                "Rachel"
            ]
        },
        "targ2": {
            "category": "AfricanAmericanNames",
            "vocab": [
                "Alonzo",
                "Jamel",
                "Theo",
                "Alphonse",
                "Jerome",
                "Leroy",
                "Torrance",
                "Darnell",
                "Lamar",
                "Lionel",
                "Tyree",
                "Deion",
                "Lamont",
                "Malik",
                "Terrence",
                "Tyrone",
                "Lavon",
                "Marcellus",
                "Wardell",
                "Nichelle",
                "Shereen",
                "Ebony",
                "Latisha",
                "Shaniqua",
                "Jasmine",
                "Tanisha",
                "Tia",
                "Lakisha",
                "Latoya",
                "Yolanda",
                "Malika",
                "Yvette"
            ]
        }
    },
    "WEAT FF": {
        "attr1": {
            "category": "Pleasant",
            "vocab": [
                "joy",
                "love",
                "peace",
                "wonderful",
                "pleasure",
                "friend",
                "laughter",
                "happy"
            ]
        },
        "attr2": {
            "category": "Unpleasant",
            "vocab": [
                "agony",
                "terrible",
                "horrible",
                "nasty",
                "evil",
                "war",
                "awful",
                "failure"
            ]
        },
        "link": "https://github.com/tanyichern/social-biases-contextualized",
        "source": "Tan et al, 2019",
        "targ1": {
            "category": "EuropeanAmericanNames",
            "vocab": [
                "Brad",
                "Brendan",
                "Geoffrey",
                "Greg",
                "Brett",
                "Matthew",
                "Neil",
                "Todd",
                "Allison",
                "Anne",
                "Carrie",
                "Emily",
                "Jill",
                "Laurie",
                "Meredith",
                "Sarah"
            ]
        },
        "targ2": {
            "category": "AfricanAmericanNames",
            "vocab": [
                "Darnell",
                "Hakim",
                "Jermaine",
                "Kareem",
                "Jamal",
                "Leroy",
                "Rasheed",
                "Tyrone",
                "Aisha",
                "Ebony",
                "Keisha",
                "Kenya",
                "Lakisha",
                "Latoya",
                "Tamika",
                "Tanisha"
            ]
        }
    },
    "WEAT H": {
        "attr1": {
            "category": "Career",
            "vocab": [
                "executive",
                "management",
                "professional",
                "corporation",
                "salary",
                "office",
                "business",
                "career"
            ]
        },
        "attr2": {
            "category": "Family",
            "vocab": [
                "home",
                "parents",
                "children",
                "family",
                "cousins",
                "marriage",
                "wedding",
                "relatives"
            ]
        },
        "link": "https://github.com/tanyichern/social-biases-contextualized",
        "source": "Tan et al, 2019",
        "targ1": {
            "category": "MaleNames",
            "vocab": [
                "John",
                "Paul",
                "Mike",
                "Kevin",
                "Steve",
                "Greg",
                "Jeff",
                "Bill"
            ]
        },
        "targ2": {
            "category": "FemaleNames",
            "vocab": [
                "Amy",
                "Joan",
                "Lisa",
                "Sarah",
                "Diana",
                "Kate",
                "Ann",
                "Donna"
            ]
        }
    },
    "WEAT I": {
        "attr1": {
            "category": "MaleOccupations",
            "vocab": [
                "driver",
                "supervisor",
                "janitor",
                "mover",
                "laborer",
                "construction",
                "worker",
                "chief",
                "developer",
                "carpenter",
                "manager",
                "lawyer",
                "farmer",
                "salesperson",
                "physician",
            "guard",
            "analyst",
            "mechanic",
            "sheriff",
            "ceo"
        ]
    },
    "attr2": {
        "category": "FemaleOccupations",
        "vocab": [
            "attendant",
            "cashier",
            "teacher",
            "nurse",
            "assistant",
            "secretary",
            "auditor",
            "cleaner",
            "receptionist",
            "clerk",
            "counselor",
            "designer",
            "hairdresser",
            "writer",
            "housekeeper",
            "baker",
            "accountant",
            "editor",
            "librarian",
            "tailor"
        ]
    },
    "link": "https://github.com/tanyichern/social-biases-contextualized",
    "source": "Tan et al, 2019",
    "targ1": {
        "category": "MaleNames",
        "vocab": [
            "John",
            "Paul",
            "Mike",
            "Kevin",
            "Steve",
            "Greg",
            "Jeff",
            "Brad",
            "Brendan",
            "Geoffrey",
            "Brett",
            "Matthew",
            "Neil",
            "Darnell",
            "Hakim",
            "Jermaine",
            "Kareem",
            "Jamal",
            "Leroy",
            "Rasheed",
            "DeShawn",
            "DeAndre",
            "Marquis",
            "Terrell",
            "Malik",
            "Tyrone"
        ]
    },
    "targ2": {
        "category": "FemaleNames",
        "vocab": [
                "Allison",
                "Anne",
                "Carrie",
                "Emily",
                "Jill",
                "Laurie",
                "Kristen",
                "Meredith",
                "Molly",
                "Amy",
                "Claire",
                "Katie",
                "Madeline",
                "Aisha",
                "Ebony",
                "Keisha",
                "Lakisha",
                "Latoya",
                "Tamika",
                "Imani",
                "Shanice",
                "Aaliyah",
                "Precious",
                "Nia",
                "Deja",
                "Latisha"
            ]
        }
    },
    "WEAT J": {
        "attr1": {
            "category": "CompetentAchievementOriented",
            "vocab": [
                "competent",
                "productive",
                "effective",
                "ambitious",
                "active",
                "decisive",
                "strong",
                "tough",
                "bold",
                "assertive"
            ]
        },
        "attr2": {
            "category": "IncompetentNotAchievementOriented",
            "vocab": [
                "incompetent",
                "unproductive",
                "ineffective",
                "unambitious",
                "passive",
                "indecisive",
                "weak",
                "gentle",
                "timid",
                "unassertive"
            ]
        },
        "link": "https://github.com/tanyichern/social-biases-contextualized",
        "source": "Tan et al, 2019",
        "targ1": {
            "category": "MaleNames",
            "vocab": [
                "John",
                "Paul",
                "Mike",
                "Kevin",
                "Steve",
                "Greg",
                "Jeff",
                "Bill"
            ]
        },
        "targ2": {
            "category": "FemaleNames",
            "vocab": [
                "Amy",
                "Joan",
                "Lisa",
                "Sarah",
                "Diana",
                "Kate",
                "Ann",
                "Donna"
            ]
        }
    },
    "WEAT K": {
        "attr1": {
            "category": "CompetentAchievementOriented",
            "vocab": [
                "competent",
                "productive",
                "effective",
                "ambitious",
                "active",
                "decisive",
                "strong",
                "tough",
                "bold",
                "assertive"
            ]
        },
        "attr2": {
            "category": "IncompetentNotAchievementOriented",
            "vocab": [
                "incompetent",
                "unproductive",
                "ineffective",
                "unambitious",
                "passive",
                "indecisive",
                "weak",
                "gentle",
                "timid",
                "unassertive"
            ]
        },
        "link": "https://github.com/tanyichern/social-biases-contextualized",
        "source": "Tan et al, 2019",
        "targ1": {
            "category": "EuropeanAmericanNames",
            "vocab": [
                "Adam",
                "Harry",
                "Josh",
                "Roger",
                "Alan",
                "Frank",
                "Justin",
                "Ryan",
                "Andrew",
                "Jack",
                "Matthew",
                "Stephen",
                "Brad",
                "Greg",
                "Paul",
                "Jonathan",
                "Peter",
                "Amanda",
                "Courtney",
                "Heather",
                "Melanie",
                "Katie",
                "Betsy",
                "Kristin",
                "Nancy",
                "Stephanie",
                "Ellen",
                "Lauren",
                "Colleen",
                "Emily",
                "Megan",
                "Rachel"
            ]
        },
        "targ2": {
            "category": "AfricanAmericanNames",
            "vocab": [
                "Alonzo",
                "Jamel",
                "Theo",
                "Alphonse",
                "Jerome",
                "Leroy",
                "Torrance",
                "Darnell",
                "Lamar",
                "Lionel",
                "Tyree",
                "Deion",
                "Lamont",
                "Malik",
                "Terrence",
                "Tyrone",
                "Lavon",
                "Marcellus",
                "Wardell",
                "Nichelle",
                "Shereen",
                "Ebony",
                "Latisha",
                "Shaniqua",
                "Jasmine",
                "Tanisha",
                "Tia",
                "Lakisha",
                "Latoya",
                "Yolanda",
                "Malika",
                "Yvette"
            ]
        }
    },
    "WEAT L": {
        "attr1": {
            "category": "Pleasant",
            "vocab": [
                "joy",
                "love",
                "peace",
                "wonderful",
                "pleasure",
                "friend",
                "laughter",
                "happy"
            ]
        },
        "attr2": {
            "category": "Unpleasant",
            "vocab": [
                "agony",
                "terrible",
                "horrible",
                "nasty",
                "evil",
                "war",
                "awful",
                "failure"
            ]
        },
        "link": "https://github.com/tanyichern/social-biases-contextualized",
        "source": "Tan et al, 2019",
        "targ1": {
            "category": "YoungPeoplesNames",
            "vocab": [
                "Tiffany",
                "Michelle",
                "Cindy",
                "Kristy",
                "Brad",
                "Eric",
                "Joey",
                "Billy"
            ]
        },
        "targ2": {
            "category": "OldPeoplesNames",
            "vocab": [
                "Ethel",
                "Bernice",
                "Gertrude",
                "Agnes",
                "Cecil",
                "Wilbert",
                "Mortimer",
                "Edgar"
            ]
        }
    },
    "WEAT M": {
        "attr1": {
            "category": "Temporary",
            "vocab": [
                "impermanent",
                "unstable",
                "variable",
                "fleeting",
                "short-term",
                "brief",
                "occasional"
            ]
        },
        "attr2": {
            "category": "Permanent",
            "vocab": [
                "stable",
                "always",
                "constant",
                "persistent",
                "chronic",
                "prolonged",
                "forever"
            ]
        },
        "link": "https://github.com/tanyichern/social-biases-contextualized",
        "source": "Tan et al, 2019",
        "targ1": {
            "category": "MentalDisease",
            "vocab": [
                "sad",
                "hopeless",
                "gloomy",
                "tearful",
                "miserable",
                "depressed"
            ]
        },
        "targ2": {
            "category": "PhysicalDisease",
            "vocab": [
                "sick",
                "illness",
                "influenza",
                "disease",
                "virus",
                "cancer"
            ]
        }
    },
    "WEAT N": {
        "attr1": {
            "category": "Pleasant",
            "vocab": [
                "caress",
                "freedom",
                "health",
                "love",
                "peace",
                "cheer",
                "friend",
                "heaven",
                "loyal",
                "pleasure",
                "diamond",
                "gentle",
                "honest",
                "lucky",
                "rainbow",
                "diploma",
                "gift",
                "honor",
                "miracle",
                "sunrise",
                "family",
                "happy",
                "laughter",
                "paradise",
                "vacation"
            ]
        },
        "attr2": {
            "category": "Unpleasant",
            "vocab": [
                "abuse",
                "crash",
                "filth",
            "murder",
            "sickness",
            "accident",
            "death",
            "grief",
            "poison",
            "stink",
            "assault",
            "disaster",
            "hatred",
            "pollute",
            "tragedy",
            "divorce",
            "jail",
            "poverty",
            "ugly",
            "cancer",
            "kill",
            "rotten",
            "vomit",
            "agony",
            "prison"
        ]
    },
    "link": "https://github.com/mebrunet/understanding-bias",
    "source": "brunet et al, 2019",
    "targ1": {
        "category": "Instruments",
        "vocab": [
            "bagpipe",
            "cello",
            "guitar",
            "lute",
            "trombone",
            "banjo",
            "clarinet",
            "harmonica",
            "mandolin",
            "trumpet",
            "bassoon",
            "drum",
            "harp",
            "oboe",
            "tuba",
            "bell",
            "fiddle",
            "harpsichord",
            "piano",
            "viola",
            "bongo",
            "flute",
            "horn",
            "saxophone",
            "violin"
        ]
    },
    "targ2": {
        "category": "Weapons",
        "vocab": [
            "arrow",
            "club",
            "gun",
            "missile",
            "spear",
            "axe",
            "dagger",
            "harpoon",
            "pistol",
                "sword",
                "blade",
                "dynamite",
                "hatchet",
                "rifle",
                "tank",
                "bomb",
                "firearm",
                "knife",
                "shotgun",
                "teargas",
                "cannon",
                "grenade",
                "mace",
                "slingshot",
                "whip"
            ]
        }
    }
}
```