Understanding the Origins of Bias in Word Embeddings

Marc-Etienne Brunet[1 2] Colleen Alkalay-Houlihan[1] Ashton Anderson[1 2] Richard Zemel[1 2]

Abstract

Popular word embedding algorithms exhibit stereotypical biases, such as gender bias. The widespread use of these algorithms in machine learning systems can thus amplify stereotypes in important contexts. Although some methods have been developed to mitigate this problem, how word embedding biases arise during training is poorly understood. In this work, we develop a technique to address this question. Given a word embedding, our method reveals how perturbing the training corpus would affect the resulting embedding bias. By tracing the origins of word embedding bias back to the original training documents, one can identify subsets of documents whose removal would most reduce bias. We demonstrate our methodology on Wikipedia and New York Times corpora, and find it to be very accurate.

1. Introduction

As machine learning algorithms play ever-increasing roles in our lives, there are ever-increasing risks for these algorithms to be systematically biased (Zhao et al., 2018; 2017; Kleinberg et al., 2016; Dwork et al., 2012; Hardt et al., 2016). An ongoing research effort is showing that machine learning systems can not only reflect human biases in the data they learn from, but also magnify these biases when deployed in practice (Sweeney, 2013). With algorithms aiding critical decisions ranging from medical diagnoses to hiring decisions, it is important to understand how these biases are learned from data.

In recent work, researchers have uncovered an illuminating example of bias in machine learning systems: Popular word embedding methods such as word2vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014) acquire stereotypical human biases from the text data they are trained on. For example, they disproportionately associate male terms with science terms, and female terms with art terms (Angwin et al., 2016; Caliskan et al., 2017). Deploying these word embedding algorithms in practice, for example in automated translation systems or as hiring aids, thus runs the serious risk of perpetuating problematic biases in important societal contexts. This problem is especially pernicious because these biases can be difficult to detect—for example, word embeddings were in broad industrial use before their stereotypical biases were discovered.

Although the existence of these biases is now established, their *origins*—how biases are learned from training data—are poorly understood. Ideally, we would like to be able to ascribe how much of the overall embedding bias is due to any particular small subset of the training corpus—for example, an author or single document. Naïvely, this could be done directly by removing the document in question, retraining an embedding on the perturbed corpus, then comparing the bias of the original embedding with the bias of the retrained embedding. The change in bias resulting from this perturbation could then be interpreted as the document's contribution to the overall bias. But this approach comes at a prohibitive computational cost; completely retraining the embedding for each document is clearly infeasible.

In this work, we develop an efficient and accurate method for solving this problem. Given a word embedding trained on some corpus, and a metric to evaluate bias, our method approximates how removing a small part of the training corpus would affect the resulting bias. We decompose this problem into two main subproblems: measuring how perturbing the training data changes the learned word embedding; and measuring how changing the word embedding affects its bias. Our central technical contributions solve the former subproblem (the latter is straightforward for many bias measures). Our method provides a highly efficient way of understanding the impact of *every* document in a training corpus on the overall bias of a word embedding; therefore, we can rapidly identify the most bias-influencing documents in the training corpus. These documents may be used to manipulate the word embedding's bias through highly selective pruning of the training corpus, or they may be analyzed in conjunction with metadata to identify particularly biased

[1] Department of Computer Science, University of Toronto, Toronto, Canada [2] Vector Institute for Artificial Intelligence, Toronto, Canada. Correspondence to: Marc-Etienne Brunet <mebrunet@cs.toronto.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

subsets of the training data.

We demonstrate the accuracy of our technique with experimental results on both a simplified corpus of Wikipedia articles in broad use (Wikimedia, 2018), and on a corpus of New York Times articles from 1987–2007 (Sandhaus, 2008). Across a range of experiments, we find that our method's predictions of how perturbing the input corpus will affect the bias of the embedding are extremely accurate. We study whether our results transfer across embedding methods and bias metrics, and show that our method is much more efficient at identifying bias-inducing documents than other approaches. We also investigate the qualitative properties of the influential documents surfaced by our method. Our results shed light on how bias is distributed throughout the documents in the training corpora, as well as expose interesting underlying issues in a popular bias metric.

2. Related Work

Word embeddings are compact vector representations of words learned from a training corpus, and are actively deployed in a number of domains. They not only preserve statistical relationships present in the training data, generally placing commonly co-occurring words close to each other, but they also preserve higher-order syntactic and semantic structure, capturing relationships such as *Madrid* is to *Spain* as *Paris* is to *France*, and *Man* is to *King* as *Woman* is to *Queen* (Mikolov et al., 2013b). However, they have been shown to also preserve problematic relationships in the training data, such as *Man* is to *Computer Programmer* as *Woman* is to *Homemaker* (Bolukbasi et al., 2016).

A recent line of work has begun to develop measures to document these biases as well as algorithms to correct for them. Caliskan et al. (2017) introduced the Word Embedding Association Test (WEAT) and used it to show that word embeddings trained on large public corpora (e.g., Wikipedia, Google News) consistently replicate the known human biases measured by the Implicit Association Test (Greenwald et al., 1998). For example, female terms (e.g., "her", "she", "woman") are closer to family and arts terms than they are to career and math terms, whereas the reverse is true for male terms. Bolukbasi et al. (2016) developed algorithms to de-bias word embeddings so that problematic relationships are no longer preserved, but unproblematic relationships remain. We build upon this line of work by developing a methodology to understand the sources of these biases in word embeddings.

Stereotypical biases have been found in other machine learning settings as well. Common training datasets for multilabel object classification and visual semantic role labeling contain gender bias and, moreover, models trained on these biased datasets exhibit greater gender bias than the training datasets (Zhao et al., 2017). Other types of bias, such as racial bias, have also been shown to exist in machine learning applications (Angwin et al., 2016).

Recently, Koh & Liang (2017) proposed a methodology for using influence functions, a technique from robust statistics, to explain the predictions of a black-box model by tracing the learned state of a model back to individual training examples (Cook & Weisberg, 1980). Influence functions allow us to efficiently approximate the effect on model parameters of perturbing a training data point. Other efforts to increase the explainability of machine learning models have largely focused on providing visual or textual information to the user as justification for classification or reinforcement learning decisions (Ribeiro et al., 2016; Hendricks et al., 2016; Lomas et al., 2012).

3. Background

3.1. The GloVe word embedding algorithm

Learning a GloVe (Pennington et al., 2014) embedding from a tokenized corpus and a fixed vocabulary of size V is done in two steps. First, a sparse co-occurrence matrix $X \in \mathbb{R}^{V \times V}$ is extracted from the corpus, where each entry X_{ij} represents a weighted count of the number of times word j occurs in the context of word i. Gradient-based optimization is then used to learn the optimal embedding parameters w^*, u^*, b^*, and c^* which minimize the loss:

$$J(X, w, u, b, c) = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(w_i^T u_j + b_i + c_j - \log X_{ij})^2 \tag{1}$$

where $w_i \in \mathbb{R}^D$ is the vector representation (embedding) of the ith word in the vocabulary, $1 \leq i \leq V$. The embedding dimension D is commonly chosen to be between 100 and 500. The set of $u_j \in \mathbb{R}^D$ represent the "context" word vectors[1]. Parameters b_i and c_j represent the bias terms for w_i and u_j, respectively. The weighting function $f(x) = \min((x/x_{max})^\alpha, 1)$ is used to attribute more importance to common word co-occurrences. The original authors of GloVe used $x_{max} = 100$ and found good performance with $\alpha = 0.75$. We refer to the final learned emebedding as $w^* = \{w_i^*\}$ throughout.

3.2. Influence Functions

Influence functions offer a way to approximate how a model's learned optimal parameters will change if the training data is perturbed. We summarize the theory here.

Let $R(z, \theta)$ be a convex scalar loss function for a learn-

[1] When the context window is symmetric, the two sets of vectors are equivalent and differ only based on their initializations.

ing task, with optimal model parameters θ^* of the form in Equation (2) below, where $\{z_1, ..., z_n\}$ are the training data points and $L(z_i, \theta)$ is the point-wise loss.

$$R(z,\theta) = \frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) \qquad \theta^* = \operatorname*{argmin}_{\theta} R(z,\theta) \qquad (2)$$

We would like to determine how the optimal parameters θ^* would change if we perturbed a small subset of points in the training set; i.e., when $z_k \rightarrow \tilde{z}_k$ for all k in the set of perturbed indices δ. It can be shown that the perturbed optimal parameters, which we denote $\tilde{\theta}$, can be written as:

$$\tilde{\theta} \approx \theta^* - \frac{1}{n} H_{\theta^*}^{-1} \sum_{k\in\delta} \left[\nabla_\theta L(\tilde{z}_k, \theta^*) - \nabla_\theta L(z_k, \theta^*)\right] \quad (3)$$

where $H_{\theta^*} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 L(z_i, \theta^*)$ is the Hessian of the total loss, and it is assumed $|\delta| \ll n$. Note that we have extended the equations presented by Koh & Liang (2017) to address multiple perturbations. This is explained in the supplemental materials.

3.3. The Word Embedding Association Test

The *Word Embedding Association Test* (WEAT) measures bias in word embeddings (Caliskan et al., 2017). It considers two equal-sized sets $\mathcal{S}$, $\mathcal{T}$ of *target words*, such as $\mathcal{S} = \{$*math, algebra, geometry, calculus*$\}$ and $\mathcal{T} = \{$*poetry, literature, symphony, sculpture*$\}$, and two sets $\mathcal{A}$, $\mathcal{B}$ of *attribute words*, such as $\mathcal{A} = \{$*male, man, boy, brother, he*$\}$ and $\mathcal{B} = \{$*female, woman, girl, sister, she*$\}$.

The similarity of words a and b in word embedding w is measured by the cosine similarity of their vectors, $cos(w_a, w_b)$. The differential association of word c with the word sets $\mathcal{A}$ and $\mathcal{B}$ is measured with:

$$g(c,\mathcal{A},\mathcal{B},w) = \operatorname*{mean}_{a\in\mathcal{A}} cos(w_c,w_a) - \operatorname*{mean}_{b\in\mathcal{B}} cos(w_c,w_b)$$

For a given $\{\mathcal{S},\mathcal{T},\mathcal{A},\mathcal{B}\}$, the *effect size* through which we measure bias is:

$$B_{\text{weat}}(w) = \frac{\operatorname*{mean}_{s\in\mathcal{S}} g(s,\mathcal{A},\mathcal{B},w) - \operatorname*{mean}_{t\in\mathcal{T}} g(t,\mathcal{A},\mathcal{B},w)}{\operatorname*{std\text{-}dev}_{c\in\mathcal{S}\cup\mathcal{T}} g(c,\mathcal{A},\mathcal{B},w)} \quad (4)$$

Where *mean* and *std-dev* refer to the arithmetic mean and the sample standard deviation respectively. Note that B_{weat} only depends on the set of word vectors $\{w_i | \; i \in \mathcal{S}\cup\mathcal{T}\cup\mathcal{A}\cup\mathcal{B}\}$.

4. Methodology

Our technical contributions are twofold. First, we formalize the problem of understanding bias in word embeddings, introducing the concepts of *differential bias* and *bias gradient*. Then, we show how the differential bias can be approximated in word embeddings trained using the GloVe algorithm. We address how to approximate the bias gradient in GloVe in the supplemental material.

4.1. Formalizing the Problem

Differential Bias. Let $w = \{w_1, w_2, ..., w_V\}$, $w_i \in \mathbb{R}^D$ be a word embedding learned on a corpus C. Let $B(w)$ denote any bias metric that takes as input a word embedding and outputs a scalar. Consider a partition of the corpus into many small parts (e.g. paragraphs, documents), and let p be one of those parts. Let $\tilde{w}$ be the word embedding learned from the perturbed corpus $\tilde{C} = C \setminus p$. We define the *differential bias* of part $p \subset C$ to be:

$$\Delta_p B = B(w) - B(\tilde{w}) \qquad (5)$$

Which is the incremental contribution of part p to the total bias. This value decomposes the total bias, enabling a wide range of analyses (e.g., studying bias across metadata associated with each part).

It is natural to think of C as a collection of individual documents, and think of p as a single document. Since a word embedding is generally trained on a corpus consisting of a large set of individual documents (e.g., websites, newspaper articles, Wikipedia entries), we use this framing throughout our analysis. Nonetheless, we note that the unit of analysis can take an arbitrary size (e.g., paragraphs, sets of documents), provided that only a relatively small portion of the corpus is removed. Thus our methodology allows an analyst to study how bias varies across documents, groups of documents, or whichever grouping is best suited to the domain.

Co-occurrence perturbations. Several word embedding algorithms, including GloVe, operate on a co-occurrence matrix rather than directly on the corpus. The co-occurrence matrix X is a function of the corpus C, and can be viewed as being constructed additively from the co-occurrence matrices of the n individual documents in the corpus, where $X^{(k)}$ is the co-occurrence matrix for document k. In this manner, we can view X as $X = \sum_{k=1}^{n} X^{(k)}$. We then define $\tilde{X}$ as the co-occurrence matrix constructed from the perturbed corpus $\tilde{C}$. If $\tilde{C}$ is obtained by omitting document k, we have $\tilde{X} = X - X^{(k)}$.

Bias Gradient. If a word embedding w is (or can be approximated by) a differentiable function of the co-occurrence matrix X, and the bias metric $B(w)$ is also differentiable, we can consider the *bias gradient*:

$$\nabla_X B(w(X)) = \nabla_w B(w) \nabla_X w(X) \qquad (6)$$

Where the above equality is obtained using the chain rule.

The bias gradient has the same dimension as the co-occurrence matrix X. While $V \times V$ is a daunting size, if the bias metric is only affected by a small subset of the words in the vocabulary, as is the case with the WEAT bias metric, the gradient will be very sparse. It may then be feasible to compute and study. Since it "points" in the direction of maximal bias increase, it provides insight into the co-occurrences most affecting bias.

The bias gradient can also be used to linearly approximate how the bias will change due to a small perturbation of X. It can therefore be used to approximate the differential bias of document k. Again letting $\tilde{X} = X - X^{(k)}$, we start from a first order Taylor approximation of $B(w(\tilde{X}))$ around X:

$$B(w(\tilde{X})) \approx B(w(X)) - \nabla_X B(w(X)) \cdot X^{(k)}$$

We then rearrange, and apply the chain rule, obtaining:

$$B(w(X)) - B(w(\tilde{X})) \approx \nabla_w B(w) \nabla_X w(X) \cdot X^{(k)}$$

Where $w(\tilde{X})$ is equivalent to the $\tilde{w}$ of Equation (5).

4.2. Computing the Differential Bias for GloVe

The naive way to compute the differential bias for a document is to simply remove the document from the corpus and retrain the embedding. However, if we wish to learn the differential bias, of every document in the corpus, this approach is clearly computationally infeasible. Instead of computing the perturbed embedding $\tilde{w}$ directly, we calculate an approximation of it by applying a tailored version of influence functions. Generally, influence functions require the use of $H_{\theta^*}^{-1}$, as in Equation (3). In the case of GloVe this would be a $2V(D+1)$ by $2V(D+1)$ matrix, which would be much too large to work with.

The need for a new method. To overcome the computational barrier of using influence functions in large models, Koh & Liang (2017) use the LiSSA algorithm (Agarwal et al., 2017) to efficiently compute inverse Hessian vector products. They compute influence in roughly $\mathrm{O}(np)$ time, where p is the number of model parameters and n is the number of training examples. However, our analysis and initial experimentation showed that this method would still be too slow for our needs. In a typical setup, GloVe simply has too many model parameters ($2V(D+1)$), and most corpora of interest cause n to be too large. One of our principal contributions is a simplifying assumption about the behavior of the GloVe loss function around the learned embedding w^*. This simplification causes the Hessian of the loss to be block diagonal, allowing for the rapid and accurate approximation of the differential bias for every document in a corpus.

Tractably approximating influence functions. To approximate $\tilde{w}$ using influence functions, we must apply Equation (3) to the GloVe loss function from Equation (1). In doing so, we make a simplifying assumption, treating the GloVe parameters u, b, and c as constants throughout the analysis. As a result, the parameters θ consist only of w (i.e., u, b, and c are excluded from θ). The number of points n is V, and the training points $z = \{z_i\}$ are in our case $X = \{X_i\}$, where X_i refers to the ith row of the co-occurrence matrix (not to be confused with the co-occurrence matrix of the ith document, denoted as $X^{(i)}$). With these variables mapped over, the point-wise loss function for GloVe becomes:

$$L(X_i, w) = \sum_{j=1}^{V} V f(X_{ij})(w_i^T u_j + b_i + c_j - \log X_{ij})^2$$

and the total loss is then $J(X, w) = \frac{1}{V} \sum_{i=1}^{V} L(X_i, w)$, now in the form of Equation (2).

Note that our embedding w^* is still learned through dynamic updates of all of the parameters. It is only in deriving this influence function-based approximation for $\tilde{w}$ that we treat u, b, and c as constants.

In order to use Equation (3) to approximate $\tilde{w}$ we need an expression for the gradient with respect to w of the point-wise loss, $\nabla_w L(X_i, w)$, as well as the Hessian of the total loss, H_w. We derive these here, starting with the gradient.

Recall that $w = \{w_1, w_2, ..., w_V\}$, $w_k \in \mathbb{R}^D$. We observe that $L(X_i, w)$ depends only on w_i, u, b_i, and c; no word vector w_k with $k \neq i$ is needed to compute the point-wise loss at X_i. Because of this, $\nabla_w L(X_i, w)$, the gradient with respect to w (a vector in $\mathbb{R}^{VD}$), will have only D non-zero entries. These non-zero entries are the entries in $\nabla_{w_i} L(X_i, w)$, the gradient of the point-wise loss function at X_i with respect to only word vector w_i. Visually, this is as follows:

$$\nabla_w L(X_i, w) = \Bigg(\underbrace{\overbrace{0, ..., 0}^{D(i-1)}, \overbrace{\nabla_{w_i} L(X_i, w)}^{D}, \overbrace{0, ..., 0}^{D(V-i)}}_{VD \text{ dimensions}} \Bigg) \quad (7)$$

where the D-dimensional vector given by $\nabla_{w_i} L(X_i, w)$ is:

$$\sum_{j=1}^{V} 2V f(X_{ij})(w_i^T u_j + b_i + c_j - \log X_{ij})\, u_j$$

From Equation (7), we see that the Hessian of the point-wise loss with respect to w, $\nabla_w^2 L(X_i, w)$ (a $VD \times VD$-dimensional matrix), is extremely sparse, consisting of only a single $D \times D$ block in the ith diagonal block position. As a result, the Hessian of the total loss, $H_w = \frac{1}{V} \sum_{i=1}^{V} \nabla_w^2 L(X_i, w)$ (also a $VD \times VD$ matrix), is block diagonal, with V blocks of dimension $D \times D$. Each $D \times D$ diagonal block is given by:

$$H_{w_i} = \nabla_{w_i}^2 L(X_i, w) = \sum_{j=1}^{V} 2V f(X_{ij}) u_j u_j^T$$

which is the Hessian with respect to only word vector w_i of the point-wise loss at X_i.

This block-diagonal structure allows us to solve for each $\tilde{w_i}$ independently. Moreover, $\tilde{w_i}$ will only differ from w_i^* for the tiny fraction of words whose co-occurrences are affected by the removal of the selected document for the corpus perturbation. We can approximate how any word vector will change due to a given corpus perturbation with:

$$\tilde{w_i} \approx w_i^* - \frac{1}{V} H_{w_i}^{-1} \left[\nabla_{w_i} L(\tilde{X_i}, w^*) - \nabla_{w_i} L(X_i, w^*) \right] \tag{8}$$

An efficient algorithm. Combining Equation (8) with Equation (5), we can approximate the differential bias of every document in the corpus. Notice in Equation (8) that $\tilde{w_i} = w_i^*$ for all i where $\tilde{X_i} = X_i$. Also recall that B_{weat} only depends on a small set of WEAT words $\{\mathcal{S}, \mathcal{T}, \mathcal{A}, \mathcal{B}\}$. Therefore, when approximating the differential bias for a document, we only need to compute $\tilde{w_i}$ for the WEAT words in that document. This is outlined in Algorithm 1.

Algorithm 1 Approximating Differential Bias

```
input  Co-occ Matrix: $X$, WEAT words: $\{\mathcal{S}, \mathcal{T}, \mathcal{A}, \mathcal{B}\}$
  $w^*$, $u^*$, $b^*$, $c^*$ = GloVe($X$) # Train embedding
  for doc in corpus do
    $\tilde{X} = X - X^{(k)}$ # Subtract coocs from doc k
    for word $i$ in doc $\cap (\mathcal{S} \cup \mathcal{T} \cup \mathcal{A} \cup \mathcal{B})$ do
      # Only need change in WEAT word vectors
      $\tilde{w_i} = w_i^* + H_{w_i}^{-1} \left[ \nabla_{w_i} L(\tilde{X_i}, w^*) - \nabla_{w_i} L(X_i, w^*) \right]$
    end for
    $\Delta_{\text{doc}} B \approx B_{\text{weat}}(w^*) - B_{\text{weat}}(\tilde{w})$
  end for
```

5. Experimentation

Our experimentation has several objectives. First, we test the accuracy of our differential bias approximation. We then compare our method to a simpler count-based baseline. We also test whether the documents which we identify as bias influencing in GloVe embeddings affect bias in word2vec. Finally, we investigate the qualitative properties of the influential documents surfaced by our method. Our results shed light on how bias is distributed throughout the documents in the training corpora, and expose interesting underlying issues in the WEAT bias metric.

5.1. Experimental Setup

Choice of corpus and hyperparameters. We use two corpora in our experiments, each with a different set of GloVe hyperparameters. This first setup consists of a corpus constructed from a Simple English Wikipedia dump (2017-11-03) (Wikimedia, 2018) using 75-dimensional word vectors. These dimensions are small by the standards of a typical word embedding, but sufficient to start capturing syntactic and semantic meaning. Performance on the TOP-1 analogies test shipped with the GloVe code base was around 35%, lower than state-of-the-art performance but still clearly capturing significant meaning.

Our second setup is more representative of the academic and commercial contexts in which our technique could be applied. The corpus is constructed from 20 years of New York Times (NYT) articles (Sandhaus, 2008), using 200-dimensional vectors. The TOP-1 analogy performance is approximately 54%. The details of these two configurations are tabulated in the supplemental material.

Choice of experimental bias metric. Throughout our experiments, we consider the *effect size* of two different WEAT biases as presented by Caliskan et al. (2017). Recall that these metrics have been shown to correlate with known human biases as measured by the Implicit Association Test. In WEAT1, the target word sets are *science* and *arts* terms, while the attribute word sets are *male* and *female* terms. In WEAT2, the target word sets are *musical instruments* and *weapons*, while the attribute word sets are *pleasant* and *unpleasant* terms. A full list of the words in these sets can be found in the supplemental material. They are summarized in Table 1. These sets were chosen so as to include one societal bias that would be widely viewed as problematic, and another which would be widely viewed as benign.

5.2. Testing the Accuracy of our Method

Experimental Methodology. To test the accuracy of our methodology, ideally we would simply remove a single document from a word embedding's corpus, train a new embedding, and compare the change in bias with our differential bias approximation. However, the cosine similarities between small sets of word vectors in two word embeddings trained on the same corpus can differ considerably simply because of the stochastic nature of the optimization (Antoniak & Mimno, 2018). As a result, the WEAT biases vary between training runs. The effect of removing a single document, which is near zero for a typical document, is hidden in this variation. Fixing the random seed is not a practical approach. Many popular word embedding implementations also require limiting training to a single thread to fully eliminate randomness. This would make experimentation prohibitively slow.

In order to obtain measurable changes, we instead remove sets of documents, resulting in larger corpus perturbations. Accuracy is assessed by comparing our method's predictions to the actual change in bias measured when each document set is removed from the corpus and a new embedding is trained on this perturbed corpus. Furthermore, we make all

Table 1. WEAT Target and Attribute Sets

		WEAT1	WEAT2
Target Sets	$\mathcal{S}$	science	instruments
	$\mathcal{T}$	arts	weapons
Attribute Sets	$\mathcal{A}$	male	pleasant
	$\mathcal{B}$	female	unpleasant

Table 2. Baseline WEAT Effect Sizes

	WEAT1	WEAT2
Wiki	0.957 (± 0.150)	0.108 (± 0.213)
NYT	1.14, (± 0.124)	1.32, (± 0.056)

predictions and assessments using several embeddings, each trained with the same hyperparameters, but differing in their random seeds.

We construct three types of perturbation sets: *increase*, *random*, and *decrease*. The targeted (increase, decrease) perturbation sets are constructed from the documents whose removals were predicted (by our method) to cause the greatest differential bias, e.g., the documents located in the tails of the histograms in Figure 1. The random perturbation sets are simply documents chosen from the corpus uniformly at random. For a more detailed description, please refer to the supplemental material. Most of the code used in the experimentation has been made available online[2].

Experimental Results. Here we present a subset of our experimental results, principally from NYT WEAT1 (science vs. arts). Complete sets of results from the four configurations ({NYT, Wiki} × {WEAT1, WEAT2}) can be found in the supplemental materials.

The baseline WEAT effect sizes (± 1 std. dev.) are shown in Table 2. It is worth noting that the WEAT2 (weapons vs. instruments) bias was not significant in our Wiki setup. However, our analysis does not require that the bias under consideration fall within any particular range of values.

A histogram of the differential bias of removal for each document in our NYT setup (WEAT1) can be seen in Figure 1. Notice the log scale on the vertical axis, and how the vast majority of documents are predicted to have a very small impact on the differential bias.

We assess the accuracy of our approximations by measuring how they correlate with the ground truth change in bias (as measured by retraining the embedding after removing a subset of the training corpus). Recall these ground truth changes are obtained using several retraining runs with different random seeds. We find extremely strong correlations ($r^2 \geq 0.985$) in every configuration, for example Figure 2.

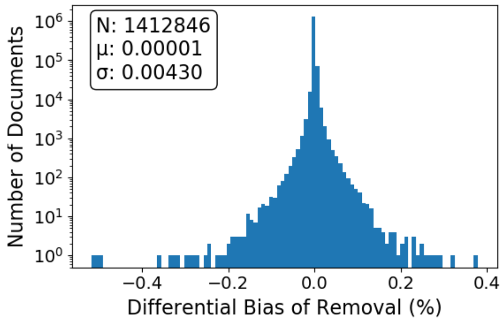

Figure 1. Histogram of the approximated differential bias of removal for every document in our NYT setup, considering WEAT1, measured in percent change from the baseline mean.

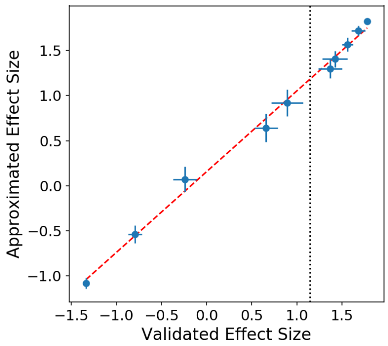

Figure 2. Approximated and ground truth WEAT bias effect size due to the removal of various perturbation sets for our NYT corpus, considering WEAT1. Each point describes the mean effect size of one set; error bars depict one standard deviation; the baseline (unperturbed) mean is shown with a vertical dotted line.

We further compare our approximations to the ground truth in Figure 3. We see that while our approximations underestimate the magnitude of the change in effect size when the perturbation causes the bias to invert, relative ranking is nonetheless preserved. There was no apparent change in the TOP-1 analogy performance of the perturbed embeddings.

We ran a Welch's t-test comparing the perturbed embeddings' biases with the baseline biases measured in the original (unperturbed) embeddings. For 36 random perturbation sets, only 2 differed significantly ($p < 0.05$) from the baseline. Both of these sets were perturbations of the smaller Wiki corpus and they only caused a significant difference for WEAT2. This is in strong contrast to the 40 targeted perturbation sets, where only 2 did *not* significantly differ from their respective baselines. In this case, both were from the smallest (10 document) perturbation sets.

5.3. Comparison to a PPMI Baseline

We have shown that our method can be used to identify bias-influencing documents and accurately approximate the

[2] Code at https://github.com/mebrunet/understanding-bias

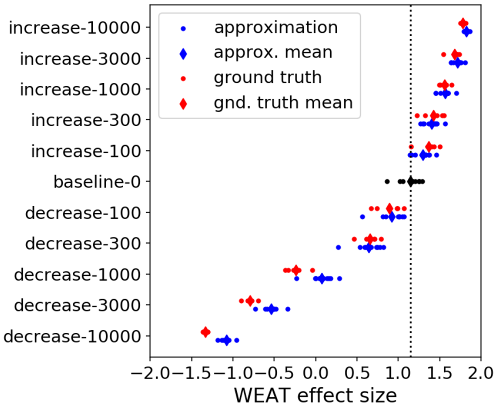

Figure 3. Approximated and ground truth differential bias of removal for every perturbation set. Results for different perturbation sets arranged vertically, named as *type - size (number of documents removed)*. (NYT - WEAT1)

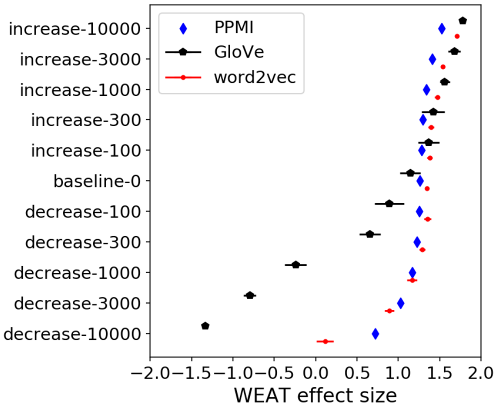

Figure 4. The effects of removing the different perturbation sets (most impactful documents as identified by our method) on the WEAT bias in: our GloVe embeddings, the PPMI representation, and word2vec embeddings with comparable hyper-parameters; error bars represent one standard deviation. (NYT - WEAT1)

impact of their removal, but how does it compare to a more naive, straightforward approach? The positive point-wise mutual information (PPMI) matrix is a count-based distributed representation commonly used in natural language processing (Levy et al., 2015). We compare the WEAT effect size in our NYT GloVe embeddings versus when measured in the corpus' PPMI representation (on 2000 randomly generated word sets). As expected, there is a clear correlation ($r^2 = 0.725$). It is therefore sensible to use the change in PPMI WEAT effect size to predict how the GloVe WEAT effect size will change.

A change in the PPMI representation due to a co-occurrence perturbation (e.g. document removal) can be computed rapidly. This allows us to scan the whole corpus for the most bias influencing documents. However, we find that the documents identified in this way have a much smaller impact on the bias than those identified by our method. For example in our Wiki setup (WEAT1) removing the 10 documents identified as most bias increasing by the PPMI method reduced the WEAT effect size by 4%. In contrast, the 10 identified by our method reduced it by 40%. Further comparisons are tabulated in the supplemental material.

5.4. Impact on Word2Vec and Other Bias Metrics

The documents identified as influential by our method clearly have a strong impact on the WEAT effect size in GloVe embeddings. Here we explore how those same documents impact the bias in word2vec embeddings, as well as other bias metrics.

We start by training five word2vec emebeddings with comparable hyperparameters[3] for each perturbation set, and measure how their removals affect the bias. Figure 4 shows how the WEAT effect size changes in GloVe, the PPMI, and word2vec for each set (NYT-WEAT1). We see that while the response is weaker, both the PPMI representation and the word2vec embeddings show a clear change in effect size due to the perturbations. For example, the baseline WEAT effect size in word2vec is 1.35 in the unperturbed corpus, but after removing *decrease-10000* (the 10k most bias contributing documents for GloVe), the effect size drops to 0.11. This means we have nearly neutralized the bias in word2vec through the removal of less than 1% of the corpus (and there is no significant change in TOP-1 analogy performance).

We also see a change as measured by other bias metrics in our perturbed GloVe embeddings. The metric proposed by Bolukbasi et al. (2016) involves computing a single dimensional gender subspace using a definitional sets of words. One can then project test words onto this axis and measure how the embedding implicitly genders them. We explore this in our NYT setup by using the WEAT 1 attribute word sets (male, female) to construct a gender axis, then projecting the target words (science, arts) onto it. In Figure 5 we show the baseline projections and compare them to the projections after having removed the 10k most bias increasing and bias decreasing documents. We see a strong response to the perturbations in the expected directions.

5.5. Qualitative Analysis

We've demonstrated that removing the most influential documents identified by our methodology significantly impacts the WEAT, a metric that has been shown to correlate with known human biases. But can the *semantic content* of these documents be intuitively understood to affect bias?

[3] We use a CBOW architecture with the same vocabulary, vector dimensions, and window size as our GloVe embeddings.

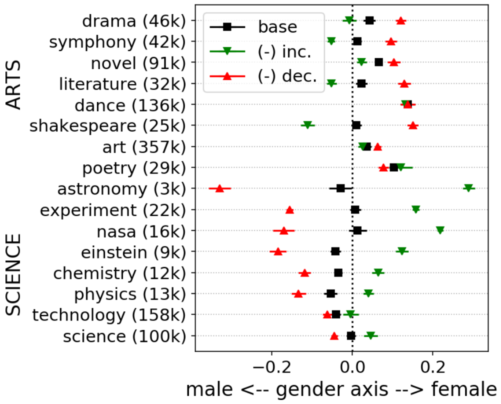

Figure 5. The effect of *removing* the 10k most bias increasing and bias decreasing documents as identified by our method on the projection of the target words onto the gender axis vs. unperturbed corpus (base); error bars show one std dev; corpus word frequency noted in parentheses. (NYT - WEAT1)

We comment here on the 50 most bias influencing documents in the New York Times corpus, considering the WEAT 1 bias metric ($\{male, female\}$, $\{science, arts\}$). This list is included in the supplemental materials. We indeed found that most of these documents could be readily understood to affect the bias in the expected semantic sense. For example, the second most bias *decreasing* document is entitled *"For Women in Astronomy, a Glass Ceiling in the Sky"*, which investigates the pay and recognition gap in astronomy. Many of the other bias decreasing documents included interviews with female doctors or scientists.

Correspondingly, the most bias *increasing* documents consisted mainly of articles describing the work of male engineers and scientists. There were several obituary entries detailing the scientific accomplishments of men, e.g., *"Kaj Aage Strand, 93, Astronomer At the U.S. Naval Observatory"*. Perhaps the most self-evident example was an article entitled *"60 New Members Elected to Academy of Sciences"*, a list of almost exclusively male scientists receiving awards.

There were, however, a few examples of articles that seemed like their semantic content should affect the bias inversely to how they were categorized. For example, an article entitled *"The Guide"*, a guide to events in Long Island, mentions that the group Woman in Science would be hosting an astronomy event, but nonetheless *increases* the bias. Only 2 or 3 documents seemed altogether unrelated to the bias' theme.

Surprisingly, some of the most bias influencing articles contained none of the science or arts WEAT terms explicitly, only synonyms (and some of the male or female terms). This shows that the impact of secondary co-occurrences can be very strong. A naive approach to understanding bias may only consider co-occurrences between WEAT words, but our method shows that this would miss some of the most bias influencing documents in the corpus.

Importantly, we also noticed a large portion of the most bias influencing documents dealt with *astronomy* or contained *hers*, the rarest words their respective WEAT subsets. Upon further investigation, we found that the log of a word's frequency is correlated with the extent to which its relative position (among WEAT words) is affected by the perturbation sets ($r^2 = 0.828$). This can be seen in Figure 5. Not surprisingly, our results indicate that the embedded representations of rare words are more sensitive to corpus perturbations. However, this leaves the WEAT metric vulnerable to exploitation through the manipulation of rarer words. The WEAT effect size is an average of cosine-similarities between the embedded representations of four subsets of words. A handful of well chosen documents can significantly alter the embeddings of a few rare words in those subsets. Therefore documents containing the rare words can have a disproportionate impact on the metric. This weakness helps explain how removing a mere 0.07% of articles can reverse the WEAT effect size in the New York Times, as is shown in Figure 3, *decrease-1000*.

6. Conclusion

In this work, we introduce the problem of tracing the origins of bias in word embeddings, and we develop and experimentally validate a methodology to solve it. We conceptualize the problem as measuring the resulting change in bias when we remove a training document (or small subset of the training corpus), and interpret this as the amount of bias contributed by the document to the overall embedding bias. Computing this naively for each training document would be infeasible. We develop an efficient approximation of this differential bias using influence functions and apply it to the GloVe word embedding algorithm. We experimentally validate our approach and find that it very accurately approximates the true change in bias that results from manually removing training documents and retraining. It performs well on tests using Simple Wikipedia and New York Times corpora and two WEAT bias metrics.

Our work represents a new approach to understanding how machine learning algorithms learn biases from training data. Our methodology could be applied to assess how the bias of a set of texts has evolved over time. For example, using publicly available datasets of newspaper articles or books, one could measure how cultural biases as measured by WEAT or other metrics have evolved over time. More broadly, our efficient method for tracing how perturbations in training data affect changes in the bias of the output is a general idea, and could be applied in many other contexts.

References

Agarwal, N., Bullins, B., and Hazan, E. Second-order stochastic optimization for machine learning in linear time. *J. Mach. Learn. Res.*, 18(1):4148–4187, January 2017. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=3122009.3176860.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: Theres software used across the country to predict future criminals. and its biased against blacks. *ProPublica*, 2016.

Antoniak, M. and Mimno, D. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119, 2018. ISSN 2307-387X. URL https://transacl.org/ojs/index.php/tacl/article/view/1202.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *30th Conference on Neural Information Processing Systems (NIPS)*, 2016.

Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Cook, R. and Weisberg, S. Characterizations of an empirical influence function for detecting influential cases in regression. 22:495–508, 11 1980.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480, 1998.

Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Hendricks, L., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. Generating visual explanations. In *European Conference on Computer Vision*, pp. 3–19. Springer, 2016.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Koh, P. W. and Liang, P. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, 2017.

Levy, O., Goldberg, Y., and Dagan, I. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics (TACL)*, 3:211–225, 2015. ISSN 2307-387X. doi: 10.1186/1472-6947-15-S2-S2. URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/570.

Lomas, M., Chevalier, R., II, E. C., Garrett, R., Hoare, J., and Kopack, M. Explaining robot actions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pp. 187–188. ACM, 2012.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013a.

Mikolov, T., t. Yih, W., and Zweig, G. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT 2013*, 2013b.

Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pp. 1532–1543, 01 2014.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016.

Sandhaus, E. The new york times annotated corpus, 2008. URL https://catalog.ldc.upenn.edu/LDC2008T19.

Sweeney, L. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

Wikimedia. Simplewiki:database download, 2018. URL https://dumps.wikimedia.org/simplewiki/.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989, 2017.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*, 2018.