BOLD and Beyond: Better Automated Metrics for Word Embedding Bias

Katie Lu
kqlu@mit.edu

Suleeporn Sujichantararat
ssuji@mit.edu

Rawisara Lohanimit
rloha@mit.edu

Abstract

Metrics for bias in word embeddings has been widely investigated in prior work. Our work expands the gender polarity metric of Dhamala et al. (2021), which measures male-female gender bias in word embeddings, to function on other domains that are not binary, such as race and an expanded set of gender/sexuality categories. To take advantage of previous work, we apply One-vs-One and One-vs-Rest methods that convert a multi-class classification problem into multiple binary classification problems. Then, validation on the Gender-Neutral GloVe word embeddings shows a decrease in bias compared to the original GloVe word embeddings. We also apply the metrics to show the bias trend across history through the Histwords word embeddings trained on the corpora from different decades in history.

1 Introduction

Natural language processing (NLP) systems built on word embeddings have the potential to drive technology that can change people's lives for the better, but at the same time, bias in these systems can lock in or even worsen inequities already present in society as their usage becomes more widespread. Researchers and developers should have access to a set of effective tools to assess bias at every stage of development to ensure that NLP systems do not harm marginalized communities when they are deployed for real-world use.

Word embeddings, in particular, are a key contributor to the success of deep learning-based NLP systems, as they provide a dense and meaningful low-dimensional representation of words (and/or subwords and phrases). The most common techniques for training word embeddings involve unsupervised prediction of (Mikolov et al., 2013) and/or utilizing the word counts of (Pennington et al., 2014) real-world text corpora. The use of these methods create word embeddings which replicate biases present in real-world text data (Bolukbasi et al., 2016), because doing so increases their predictive power. These biases may contribute to the continued emergence of biases in downstream tasks, such as text generation (Dhamala et al., 2021; Sheng et al., 2021).

Recent surveys of the field have shown that while numerous papers claim to address bias in NLP, they often fail to ground themselves in the relevant social science literature about bias, fail to clearly state their motivations, confound different types of bias in their stated motivations and actual methods, and/or do not properly consider the effect of NLP systems on relevant demographics (Blodgett et al., 2020).

The type of "bias" in word embeddings we seek to address causes *representational harms* in the form of *stereotyping*, or propagating undesirable generalizations about particular social groups. For example, the word embeddings for professions which are considered to be 'female' professions, such as teacher, nurse, and secretary, are closer in embedding space to words that are 'female' by definition, such as woman, she, and mother, and the opposite is true for 'male' professions and 'male' words (Bolukbasi et al., 2016). This behavior is undesirable, because ideal word embeddings should represent only the inherent meaning of the word and not stereotypical connections to gender, race, and other protected identity classes. As language takes an important role in labeling social groups and transmitting information about them, representational harms themselves are important to address to prevent further reinforcing social inequalities (Blodgett et al., 2020). Even if word embeddings affect populations of people only indirectly through NLP applications, addressing stereotyping on the word embedding level may eliminate one source of bias in the final NLP application.

Previous work on "bias" in word embeddings often quantified bias using polarity metrics for binary categories such as male and female (Bolukbasi et al., 2016). However, real-world populations are often separated into more than two categories, and we seek to create a bias metric that can provide a useful quantification of stereotyping in those situations. Our contributions are as follows:

- Analysis to emphasize the importance of going beyond binary class methods of quantifying bias
- Design and implementation of two novel metrics to automatically evaluate word embeddings for multiclass bias: one-vs-one and one-vs-rest models
- Demonstration of performance of novel metrics on "gender-neutral" pretrained embeddings and in displaying trends of bias across history

2 Related Works

2.1 Gender bias in word embeddings

Numerous studies have been done to study and debias the gender stereotype in word embeddings. (Bolukbasi et al., 2016) points out that the word embeddings trained on Google News articles express male and female stereotypes. In word embeddings space, the gender bias can be captured geometrically by calculating the relative direction of the embeddings to the embeddings of male- or female-identified words. Although the direction and distance of the word embeddings can help to understand the semantic meaning and the relationship between them, they can diagnose the implicit sexism in the embeddings. For example, $\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{king} - \overrightarrow{queen}$ and $\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{computer\ programmer} - \overrightarrow{homemaker}$ (Bolukbasi et al., 2016). The first relationship is reasonable as king and queen are the titles to call royalty based on gender which is proved by the differences between men and women in word embeddings from the equation. However, both computer programmer and homemaker are occupations that have no gender limitation. Thus, ideally, they should be gender-neutral and should not be close to either gender. This relation exhibits gender bias in occupations.

A similar method is introduced in (Dhamala et al., 2021). The paper introduces the idea of 'gender polarity.' The gender polarity of the word w_i is defined by

$$b_i = \frac{\overrightarrow{w_i} \cdot \overrightarrow{g}}{\| \overrightarrow{w_i} \| \| \overrightarrow{g} \|}, \quad (1)$$

where $\overrightarrow{g} = \overrightarrow{she} - \overrightarrow{he}$. If b_i is closer to 1, $\overrightarrow{w_i}$ is female-aligned. If b_i is closer to -1, $\overrightarrow{w_i}$ is male-aligned. Otherwise, if b_i is closer to 0, $\overrightarrow{w_i}$ is neutral. This method will help us identify the gender bias within the word: whether it is more closely related with female or male.

2.2 Multi-class bias in word embeddings

Some of the domains, for instance, races, occupations, and even gender beyond male and female, have more than 2 subgroups to consider. Therefore, it is challenging to identify the bias compared to the male-female aspect.

The Word Embeddings Association Test (WEAT), proposed by (Caliskan et al., 2017), is a state-of-the-art method to measure the biases in word embeddings. WEAT uses the cosine similarity between word embeddings by considering two sets of target words and two sets of attribute words. The WEAT test statistic measures the differential association of the two sets of target words with the attribute. However, this method only applies for 2 subgroups comparison. Thus, (Manzini et al., 2019) expands WEAT by introducing Mean average cosine similarity (MAC) to quantify multi-class bias. Although this method can effectively determine the biases, it requires 4 sets of words instead of one word of interest. Thus we propose a metric to evaluate the bias in the word embeddings for multi-class by expanding the idea of gender polarity used for 2 classes evaluation.

3 Methodology

We used the existing 2-class bias metric as our baseline model. Then we expand this baseline model into one-vs-one and one-vs-rest model to support multi-class bias with similar approaches in multi-class classification using binary classifier. Even though one-vs-one model is more accurate than one-vs-rest model, it is more computational intensive. This is the reason why we used one-vs-one model to calculate average bias while one-vs-rest model is used to calculate maximum bias among multiple classes.

3.1 Baseline model

Our baseline model is based on the gender polarity metric from Dhamala et al. (2021). Their work used gender polarity to evaluate sentences or whole corpora by summing up the gender polarity of each word in the text. We also used a similar idea of binary polarity as a baseline model for quantifying 2-class bias in a set of word embeddings by taking an average of absolute polarity values over the set of occupation/profession words used by Dhamala et al. (2021).[1] Ideally, unbiased word embeddings for occupation words should not contain gender, race, or other information about protected identity classes which means that these words should be neutral and have a polarity score closer to 0. So, a lower score by this method should mean the embeddings are less biased.

3.2 One-vs-one model

To expand the functionality of the baseline matrix to multi-class domains, one approach is to apply the gender polarity multiple times on different binary pairs of genders. This approach was successfully used to solve multi-class classification problems in Support Vector Machines (SVM) as shown in the work by Hsu and Lin (2002) where the authors referred to this approach as one-against-one. The idea is a multi-class classifier was constructed by combining several binary classifiers together. The intuition behind this idea is that for every pair we can use gender polarity to tell whether a particular word is biased or not based on the score: positive and negative scores signify the bias while zero scores imply no bias. Thus, summing all the scores in every possible pair and dividing by the number of pairs will tell us the average bias in the domain of interest. For example, assuming we have 5 genders including man, woman, gay, lesbian and queer, we can construct $5 \cdot 4 = 20$ binary classification problems with each problem enclosed in parentheses as listed below.

- ({man} vs. {woman}), ({man} vs. {gay}), ({man} vs. {lesbian}), ({man} vs. {queer})
- ({woman} vs. {man}), ({woman} vs. {gay}), ({woman} vs. {lesbian}), ({woman} vs. {queer})
- ({gay} vs. {man}), ... , ({gay} vs. {queer})
- ({lesbian} vs. {man}), ... , ({lesbian} vs. {queer})
- ({queer} vs. {man}), ({queer} vs. {woman}), ({queer} vs. {gay}), ({queer} vs. {lesbian})

Then we calculate gender polarity score of each pair by adjusting equation (1) which supports 2-class bias into equation (2), (3) and (4) which support multi-class bias using one-vs-one approach.

$$b_{ijk} = \frac{\overrightarrow{w_i} \cdot \overrightarrow{g_{jk}}}{\| \overrightarrow{w_i} \| \| \overrightarrow{g_{jk}} \|}, \tag{2}$$

$$\overrightarrow{g_{jk}} = \overrightarrow{c_j} - \overrightarrow{c_k}, \tag{3}$$

$$b_i = \frac{\sum_{jk, j \neq k} | b_{ijk} |}{N \cdot (N-1)}, j, k = 1, ..., N \tag{4}$$

Since we are interested in measuring gender bias in occupation, we let $\overrightarrow{w_i}$ represents word embedding of occupation word i. The typical $\overrightarrow{g}$ in binary gender polarity metric from equation (1) is expanded into $\overrightarrow{g_{jk}}$ in equation (3) which represents word embedding as differences between the gender word $\overrightarrow{c_j}$ and $\overrightarrow{c_k}$. Finally, we sum binary bias across all pairs of b_{ijk} and divide by the number of total pairs. With N genders, we can formulate $N \cdot (N-1)$ one-vs-one pairs in total. The magnitude of b_i will demonstrate the level of average bias towards the occupation word $\overrightarrow{w_i}$ among all N gender words.

3.3 One-vs-rest model

Since the gender polarity metric from Dhamala et al. (2021) only support 2 classes of genders, we adopt the one-vs-rest approach which was proved by (Rifkin and Klautau, 2004) as a successful approach to perform multi-class classification using a binary classifier. The idea is to change a multi-class classification problem into multiple binary classification problems. The predicted class is the class with the highest score among all of these binary classification problems. With the same example illustrated in one-vs-one approach, we have 5 genders including man, woman, gay, lesbian and queer where we can construct 5 binary classification problems with one-vs-rest approach. For each problem, we assign one gender to a single class and group the remaining 4 classes into another class as listed below.

- ({man} vs. {woman, gay, lesbian, queer})

[1] https://github.com/amazon-science/bold/tree/main/prompts

- ({woman} vs. {man, gay, lesbian, queer})
- ({gay} vs. {man, woman, lesbian, queer})
- ({lesbian} vs. {man, woman, gay, queer})
- ({queer vs. {man, woman, gay, lesbian})

Then we calculate gender polarity score of each subgroup by adjusting equation (1) which supports 2-class bias into equation (5), (6) and (7) which support multi-class bias using one-vs-rest approach.

$$b_{ij} = \frac{\overrightarrow{w_i} \cdot \overrightarrow{g_j}}{\| \overrightarrow{w_i} \| \| \overrightarrow{g_j} \|}, \tag{5}$$

$$\overrightarrow{g_j} = \overrightarrow{c_j} - \frac{\sum_{k \neq j}^{N} \overrightarrow{c_k}}{N-1}, \tag{6}$$

$$b_i = \max_j b_{ij}, j = 1, ..., N \tag{7}$$

Similar to one-vs-one approach, we let $\overrightarrow{w_i}$ represents word embedding of occupation word i because we are interested in measuring gender bias in occupation. However, $\overrightarrow{g_j}$ is calculated differently than that of the one-vs-one approach and $\overrightarrow{g_j}$ represents word embedding of the j^{th} pair of gender grouping with one-vs-rest approach where N is also the total number of genders. In order to calculate $\overrightarrow{g_j}$, we calculate the distance between the gender word $\overrightarrow{c_j}$ (i.e. one in the one-vs-rest model) and the average across all the remaining gender words $\overrightarrow{c_k}$ (i.e. rest in the one-vs-rest model). For example, to figure out if an occupation word $\overrightarrow{nurse}$ has bias towards one of the 5 genders mentioned above or not, we let $\overrightarrow{w_i}$ = $\overrightarrow{nurse}$ and $\{\overrightarrow{c_1}, \overrightarrow{c_2}, ... , \overrightarrow{c_5}\}$ = $\{\overrightarrow{man}, \overrightarrow{woman}, \overrightarrow{gay}, \overrightarrow{lesbian}, \overrightarrow{queer}\}$. Finally, the multi-class bias b_i is the maximum bias among all N pairs of occupation word $\overrightarrow{w_i}$ and one-vs-rest grouping of gender word $\overrightarrow{g_j}$. Assuming the occupation word $\overrightarrow{nurse}$ has the maximum bias towards the gender word $\overrightarrow{woman}$, then $b_i = b_{i2}$ where $\overrightarrow{g_2}$ = $\overrightarrow{woman}$ - (($\overrightarrow{man}$ + $\overrightarrow{gay}$ + $\overrightarrow{lesbian}$ + $\overrightarrow{queer}$) / 4)).

4 Experiments

To demonstrate how our model assesses bias in gender embeddings, we tested it on several sets of established pre-trained word embeddings with different properties. As a baseline, we used pre-trained GloVe word embeddings from the Stanford Natural Language Processing Group (Pennington et al., 2014)[2]. To illustrate the model's utility for gender-related bias, we used the Gender-Neutral GloVe embeddings trained by Zhao et al. (2018)[3] so we could compare the scores with default GloVe. We measure the gender and race bias of the word embeddings in professions. The list of professions is taken from Dhamala et al. (2021).

In addition, we used the Histwords embeddings trained by Hamilton et al. (2016)[4] to see if our model can capture changes in bias over time as society generally moved in more progressive directions.

Histwords provides embeddings trained using three different algorithms: a point-wise mutual information (PPMI) based approach, a singular value decomposition approach to generate low-dimensional approximations of the PPMI embeddings, and skip-gram with negative sampling (SGNS), the same approach used to train word2vec. We chose to use the SGNS embeddings, which were made available as pre-trained embeddings. The SGNS embeddings for English include a set of embeddings for every decade between 1800 and 1990 inclusive. We used all of these in our experiments.

4.1 Baseline model

The baseline model doesn't have adjustable parameters, so it was run as-is on all sets of word embeddings. This experiment is intended to implement the original gender polarity to measure male- or female-aligned bias in the word embeddings.

4.2 One-vs-one model

We ran the one-vs-one multiclass model on two settings: the first was a gender and sexuality-related setting which used the set of words man, woman, gay as categories, and the second was a race-related setting which used the set of words black, white, asian, native as categories.

4.3 One-vs-rest model

We ran the one-vs-rest multiclass model on three settings: two gender and sexuality-related settings which used the sets of words man, woman, gay and man, woman, gay, lesbian, queer as categories, and a race-related setting which used the set of words black, white, asian, native as categories. We ran

[2]https://nlp.stanford.edu/projects/glove/
[3]https://github.com/uclanlp/gn_glove
[4]https://nlp.stanford.edu/projects/histwords/

one more experiment in gender and sexuality related setting in one-vs-rest but not in one-vs-one since this version didn't show very conclusive results.

5 Results

Table [1] shows our model scores on the GloVe and GN-GloVe embeddings. Results show that our metric evaluates the GN-GloVe embeddings as significantly less biased for the baseline (gender polarity) and g1 gender-based settings of the multiclass metrics. These results are a positive demonstration of our metric correctly picking out a reduction in gender bias but not race bias, as GN-GloVe was not trained to reduce racial bias.

The GN-GloVe embeddings score better on the multiclass gender metrics using man, woman, gay as categories but not for the g2 setting using man, woman, gay, lesbian, queer, which may indicate that it only does better on the g1 setting because the man, woman pair are the majority of the categories. Zhao et al. (2018) only used a single-dimensional gender feature to neutralize gender in their work, so their approach may have been insufficient to neutralize multiclass bias when more categories aside from man and woman are introduced.

Figure [1] shows our model scores for gender-related bias on the Histword embeddings for all decades between 1800 and 1990. Gender bias, as illustrated by the gender polarity and g1 multiclass metrics, has a clear downwards trend over time, which indicates that biases in language reflect increasing gender equality in society. There is not a clear trend for the g2 multiclass metric, which may be due to a number of reasons, including definitions of the words used for each class not being constant over time (gay, for example, evolved from a word that meant 'happy') and/or social biases themselves not having a clear downward trend. This is a potential avenue for further exploration.

Figure [2] shows our model scores for race-related bias on the same set of embeddings. The one-vs-one metric shows a downward trend, but it is unclear if the one-vs-rest metric does the same. This is also a potential avenue for future exploration.

6 Conclusions and Future Work

We developed a novel set of metrics to measure bias in word embeddings in multiclass settings, such as gender and sexuality and racial bias. Our method is general and can be applied for any type of bias and set of word embeddings by supplying the corresponding bias categories. We also demonstrated some successes of our model in capturing bias in the gender and race domains.

There are some areas in which our work could be improved. We performed our experiments using pretrained embeddings, which gave us less control over the specifics of the training process that produced them. In the future, we could perform new experiments on embeddings we trained ourselves using the same methods as in GN-GloVe and Histwords to control the dataset and other hyperparameters. This would allow for more informative comparisons between sets of embeddings.

In addition, the Histwords comparisons could be improved by doing more research into how the meaning of certain words we used as bias categories evolved over time and adapting the categories as times changed.

With the success of constructing metric to measure multi-class bias in word embeddings, our next step is to figure out a way to debias corresponding NLP systems with multi-class bias. Amini et al. (2019) proposed a method to automatically identify bias and also debias machine learning models in computer vision using learned latent structure which is also applicable to NLP systems. We aim to apply a similar strategy using a variational autoencoder (VAE) to identify minority samples in the training datasets where multi-class bias is stemmed from and increase the probability that these minority samples are selected during training. By applying this approach towards gender bias in occupation, the training data of some minority genders will be more likely sampled when contructing word embeddings for occupation in NLP systems.

7 Impact statement

Quantifying and reducing bias in word embeddings is a promising first step in the lofty goal of eliminating bias from natural language processing systems. However, focusing on word embeddings alone comes with certain limitations. As with many modern machine learning methods, word embeddings are not inherently interpretable, and researchers have to draw indirect conclusions about their features by examining which words occupy different parts of the implicit feature space. Therefore, it is difficult to make definitive statements about the bias in a set of word embeddings, and ef-

Embeddings	Baseline	ovo (g1)	ovo (r)	ovr (g1)	ovr (g2)	ovr (r)
GloVe	0.0976	0.1090	0.0978	0.1103	0.1528	0.1475
GN-GloVe	**0.0519**	**0.0595**	0.0827	**0.0657**	0.1456	0.1224

Table 1: Results for various metrics on the GloVe and GN-GloVe embeddings. ovo = one-vs-one, ovr = one-vs-rest. (g1) marks the gender-related setting with man, woman, gay for the multiclass metrics, (g2) marks the gender-related setting with man, woman, gay, lesbian, queer, and (r) marks the race-related setting.

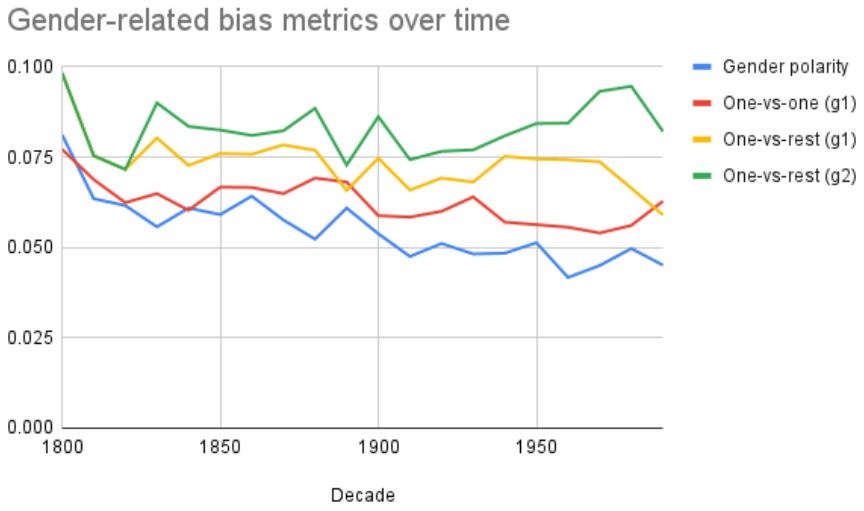

Figure 1: Evolution of gender-related bias over time, as evaluated by our gender polarity, one-vs-one, and one-vs-rest metrics.

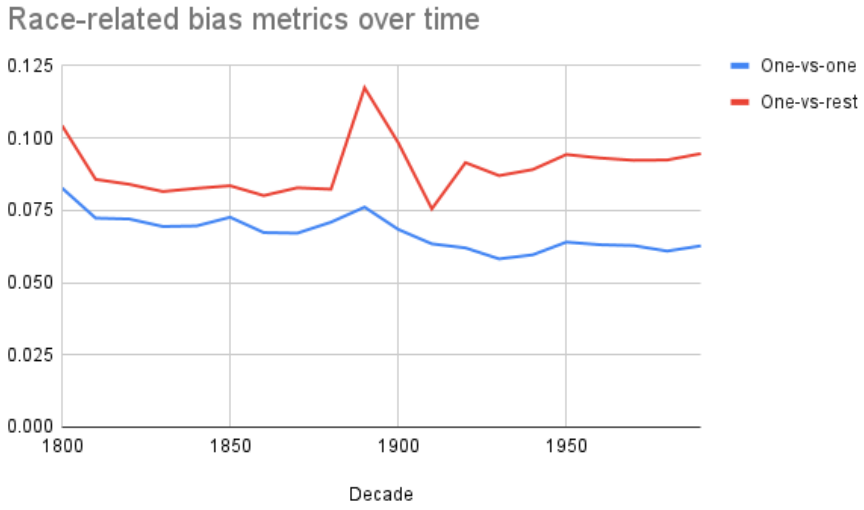

Figure 2: Evolution of race-related bias over time, as evaluated by our one-vs-one and one-vs-rest metrics.

forts to create debiased embeddings often get overturned by later research which finds other sources of hidden bias.

Even if it possible to perfectly assess bias in word embeddings and debias them accordingly, bias may be introduced at many other points downstream, and it is not obvious that debiasing word embeddings ensures unbiased performance on downstream tasks. Furthermore, the authors of this paper are not aware of a robust literature on whether and how such upstream debiasing efforts impact the lives of people who use and are impacted by NLP technology. As the ultimate goal of debiasing should be to ensure equitable impact on real people affected by NLP technology, it is important to investigate this chain of effects to determine where it is most useful to focus our efforts as researchers.

Furthermore, bias can be defined in many ways even within sociological research. Selecting any one version of 'bias' is a value judgment that places that prioritizes that particular definition. If one bias metric becomes very popular within a field, that is liable to cause value lock-in and reduce consideration for other, equally valid definitions of bias. Once a metric becomes established, it is easy for the field to commit the fallacy of optimizing for that metric rather than solving the actual problem the metric seeks to illustrate, which is the impact of biased NLP systems on real people.

An example of this can be seen in the field of machine translation, where BLEU and ROUGE, two n-gram based metrics, are highly entrenched. N-gram matching as the gold standard may make sense for genres of text that are restricted in their style such as newspapers and UN resolutions, but these metrics are not ideal for fiction, which admits a greater variety of wording choices without a translation becoming less correct. The dominance of these metrics incentivizes models which do not perform as well in text genres like fiction.

Introducing a new metric will deviate the focus from optimizing the existing popular metric. On top of that, the metric might be able to capture a different aspect of the same thing. For instance, one bias metric used on the debiased word embeddings says that the embeddings are fair. However, if the second metric is applied to the same word embeddings and signifies bias, these word embeddings might be overfitted for the first metric and thus ignore other biases that the first metric cannot capture. In other words, we may conclude that the word embeddings are still biased. Using multiple metrics will allow us to capture different things and cover other metrics' issues.

Therefore, the authors of this paper caution users of our metric to remember to consider downstream sources of bias, consider studying the impact of their NLP models on affected parties, and avoid overreliance on specific metrics in their work.

8 Appendix

To view and rerun the experiment, our codes can be found at https://tinyurl.com/BoldAndBeyond.

References

Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. 2019. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. *CoRR*, abs/2005.14050.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520.

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

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change.

Chih-Wei Hsu and Chih-Jen Lin. 2002. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ryan Rifkin and Aldebaro Klautau. 2004. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *CoRR*, abs/2105.04054.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *CoRR*, abs/1809.01496.