Rethinking Debiasing Methods for Word Embeddings

Abstract

Word embeddings, a popular framework to represent text data, encode social bias and may as a result produce biased systems. Prior work has proposed myriad ways to debias word embeddings to yield fairer models. However, by following a proof from Kleinberg et al., and surveying what attributes these debiasing techniques remove, we show that using a debiased word embedding does not reduce bias, and may even exacerbate unfairness. Further, while studies of fairness from ML can inform studies of fairness in NLP, we acknowledge that the complexity inherent in language tasks warrants a careful examination of definitions of fairness. By studying legal perspectives on anti-classification and anti-subordination, we recommend a shift from focusing on making fairer word embeddings to fairer models defined in terms of downstream impact and adjustments for historical and systematic biases.

1 Introduction

Non-contextualized word embeddings provide a lightweight and robust way of representing text for NLP applications. Popular word embeddings such as GloVe (Pennington et al., 2014) and word2vec (Mikolov et al., 2013) are widely used in performing NLP tasks that have broad and significant impact, from Internet searches to information extraction.

However, word embeddings have been shown to reflect societal biases. These biases range from binary gender (Bolukbasi et al., 2016) to race and religion (Garg et al., 2018; Manzini et al., 2019). Qualitatively, Bolukbasi et al., showed how performing analogy tasks in word embeddings demonstrates stereotypes such as "man is to computer programmers as woman is to homemaker". Quantitatively, popular measures to evaluate bias include WEAT (Caliskan et al., 2017) that adopts from the Implicit Association Test[1] and other measures (Garg et al., 2018; Bolukbasi et al., 2016) that involve sums and differences of word embeddings from word lists of protected groups.

Along with these efforts to measure bias are efforts to remove biases, both in post-processing (Bolukbasi et al., 2016; Manzini et al., 2019) and during training (Zhao et al., 2018). While effectively reducing bias in accordance with the proposed quantitative measures, it is shown that the "removed" bias is still recoverable (Gonen and Goldberg, 2019). Further, we demonstrate that these methods of removing bias is a form of simplification, where distinct feature vectors are classified similarly by ignoring certain features, and the proof from Kleinberg et al., poses a concern about how simplicity transforms protected attributes into bias against the disadvantaged group (Kleinberg and Mullainathan, 2019).

These attempts to remove biases in word embeddings seek to achieve fairness. We start by a definition of fairness as individual fairness, or f(x; A) = f(x; D), where the predicted outcome for each individual sample is independent of the protected attribute - A representing the advantaged group and D representing the disadvantaged group (Gajane and Pechenizkiy, 2018). This definition is in line with Kleinberg et al., enabling us to investigate how the relationship between simplicity and fairness may apply to word embeddings.

While efforts in fairness in ML are centered around anti-classification - the use of protected attributes, we also bring attention to the notion of antisubordination - the subjugation of disadvantaged groups (Corbett-Davies and Goel, 2018). Legally, a history of debates about anti-subordination and anti-classification principles started with the case of Brown v. Board of Education (Siegel, 2004), and we witness a turn to anti-classification principles with current employment discrimination laws designed to explicitly respond to a history of discrimination (Areheart, 2012). We argue for more

[1] https://code.google.com/archive/p/word2vec/

discourse regarding the two principles in the NLP and ML community, and propose recommendations informed by studies in law.

2 Method and Experiment

In this section, we formulate debiasing word embeddings such that the proof from Kleinberg et al., can be adopted to show that there always exists a model built with the original representation that is at least as accurate and fair as a model built with the debiased representation (Bolukbasi et al., 2016; Zhao et al., 2018).

2.1 Proof Setup

The following theorem is shown in Kleinberg et al.,:

Theorem 1. *For any model M_2 built with representation r_2 where r_2 is a simplified version of r_1. Then, there always exists a model M_1 built with representation r_1 that is more efficient as M_2 and more equitable as M_2.*

This theorem is based on the setting of admissions, where we have a set of applicants and would like to admit a fraction of them. M_1 and M_2 are models that decide which applicants to admit, and r_1 and r_2 are feature vectors that represent each applicant. Each applicant belongs to exactly one of A (advantaged group) or D (disadvantaged group), and their qualifications are represented as additional features that exclude group information. That is, each applicant can be represented as a vector [x, A] or [x, B]. Kleinberg et al., assumes that a larger fraction of the people with high-scoring vectors x belong to the advantaged group A than D. They further assume that the acceptance decision should be independent of whether group identity. Concretely, the model tries to approximate a function f that maps each applicant vector to a score of how qualified they are for admission under the assumption that f(x:A) = f(x,D). Finally, "equity" is defined as the fraction of admitted applicants who come from group D and "efficiency" as the average quality of the admitted applicants.

Claim 1. *For any model M_2 built with the debiased embeddings where the debiased embeddings is a simplified version of the original embedding. Then, there always exists a model M_1 built with the original embedding that is more efficient as M_2 and more equitable as M_2.*

We want to show claim 1 by having r_2 be the debiased word embeddings and r_1 be the original embedding. Showing this claim allows us to question the importance of debiasing word embeddings if it does not necessarily allow for a more accurate nor fair model.

2.2 Debiasing as a Simplification

In order to prove claim 1, we must show that debiasing is a simplification. A simplification means that one is treating some distinct vectors $x_1, .., x_n$ that correspond to applicant 1 through n as if they have the same qualification score. This is done by ignoring or "gluing together" certain features of the vector such that $x'_1, .., x'_n$ are all the same vector. The representation that contains $x'_1, .., x'_n$ are called a simplification from the original representation that contains $x_1, .., x_n$.

We will discuss how two popular debiased word embeddings, GN-GloVe (Zhao et al., 2018) and hard-w2v (Bolukbasi et al., 2016) are simplifications. Zhao et al., learns representations from gender-debiasing that takes GloVe 300d and make 300d vectors where the first 299 dimensions are gender-agnostic representations and the last one contains the gender information. One directly ignores the last dimension to find the debiased word embeddings, GN-GloVe. This directly applies to the definition of simplification through feature deletion. Suppose the last dimension for some applicants 1 through n are distinct, $g_1, ..., g_n$, while the non-gendered $x_1, \ldots, x_n$ are the same, GN-GloVe is a simplification from the original representation that contains $[x_1, g_1], \ldots, [x_n, g_n]$ to the simplified representation that contains $x_1, ..., x_n$.

A similar case can be made for hard-w2v, Bolukbasi et al., debiases by finding a gender direction **g** and projecting the original w2v 300-d such that they are orthogonal to **g**. Let a change-of-basis matrix M take a word embedding and rewrite it as the k components in the direction of **g** and other components in the (300 - k) other directions that form an orthogonal basis with respect to **g**. A model using the original w2v, r_1 is the same as one using $M^{-1}Mr_1$ while a model using the debiased w2v, r_2, is the same as one using Mr_1 with the first k elements deleted. This again fits the definition of simplification through feature deletion.

2.3 Does debiasing capture the protected subspace exactly?

One complication must be acknowledged is that Theorem 1 does not apply to trivial simplifications that ignore the protected attribute features, that is,

Embedding	Attributes	Target	WEAT	Garg Cosine	Garg Euclidean
Glove	race	race	-0.0014	7.2376	0.0437
GN-Glove	race	race	-0.0957	3.8520	0.0413
Glove	race	pleasantness	0.7530	7.1095	0.0434
GN-Glove	race	pleasantness	0.8092	3.0236	0.0406
Glove	religion	pleasantness	0.5118	5.7696	0.0291
GN-Glove	religion	pleasantness	0.5518	2.2404	0.0284
w2v	race	race	-0.2019	0.1784	0.0624
hard-w2v	race	race	-0.1759	0.1787	0.0172
w2v	race	pleasantness	0.5012	0.1287	0.0482
hard-w2v	race	pleasantness	0.5166	0.1278	0.0152
w2v	religion	pleasantness	0.2534	0.2583	0.0341
hard-w2v	religion	pleasantness	0.7504	0.2690	0.0342

Table 1: Debias Effects on Attributes Not Debiased for

Embedding	Attributes	Target	Delta WEAT	Delta Garg Cosine	Delta Garg Euclidean	Average Delta
Glove vs GN-Glove	race	race	6959.88%	46.78%	5.64%	2337.43%
Glove vs GN-Glove	race	pleasantness	7.46%	57.47%	6.40%	23.78%
Glove vs GN-Glove	religion	pleasantness	7.81%	61.17%	2.43%	23.80%
w2v vs hard-w2v	race	race	12.91%	0.16%	72.51%	28.53%
w2v vs hard-w2v	race	pleasantness	3.08%	0.71%	68.52%	24.10%
w2v vs hard-w2v	religion	pleasantness	196.09%	4.17%	0.48%	66.91%

Table 2: Magnitude of Debias Effects on Attributes Not Debiased for

deleting features that strictly encode A or D. In such cases, only theorem 2 applies.

Theorem 2. *For any model M_2 built with representation r_2 where r_2 is a simplified version of r_1. Then, there always exists a model M_1 built with representation r_1 that is at least as efficient as M_2 and at least as equitable as M_2.*

Thus, if we want to show claim 1, we must also show that the debiased word embeddings does not delete the true protected information exactly. If the debiased word embeddings does indeed capture the true protected information exactly, then only theorem 1 holds.

To test whether debiasing deletes the true gender information exactly, we collect original word2vec [2] and Glove [3] embeddings as well as hard-w2v[4] from Bolukbasi et al., and GN-Glove[5]. The debiased embeddings are debiased for the gender attribute. However, we measure and compare the bias in race or religion attributes. Specifically, we measure the bias on attribute words by probing its relation to target qualities. Attribute word lists include race (white and hispanic last names) and religion (christianity and islam) while target word lists include common adjectives/biases associated with race (white and hispanic) and pleasantness. The full wordlists are recorded in Appendix A. Three methods of measuring bias are collected and implemented: WEAT (May et al., 2019), Garg Cosine, and Garg Euclidean (Garg et al., 2018). The scores are normalized by the average cosine or euclidean distance in the particular word embeddings, a larger score indicates more degrees of bias. Table 1 records the measures of biases while Table 2 summarizes the change in bias compared to the original word embedding.

With the assumption that the true gender information should be relatively orthogonal to the true race or religion information, if a noticeable change in score applies, it is evident that debiasing for one attribute affects the bias in other attributes. Thus, it cannot be that the debiased word embeddings captures the true gender information only and exactly and theorem 1 can still apply.

We observe a sizable change in delta across the board for both GN-Glove and hard-w2v when compared to the original word embedding, averaging around a 20% change. Further, from Table 1, there is a general decrease in Garg Cosine and Garg Euclidean scores when testing on the debiased word embeddings instead of the original, opposite to that of WEAT scores. This shows that the representations of attribute words not explicitly debiased for are changed sizably, usually seeing a decrease

[2] https://code.google.com/archive/p/word2vec/

[3] https://nlp.stanford.edu/projects/glove/

[4] https://drive.google.com/drive/folders/0B5vZVlu2WoS5dkRFY19YUXVIU2M

[5] https://github.com/uclanlp/gn_glove

in bias scores. As a result, we argue that the two debiasing techniques we have investigated do not remove the gender information exactly but rather it also removes other information.

2.4 Return to Proof

We have shown that the two popular debiasing techniques (Bolukbasi et al., 2016; Zhao et al., 2018) do not remove the protected attribute information exactly, and that they are a simplification of the original word embeddings. Thus, we can conclude claim 1 - "for any model M_2 built with the debiased embeddings where the debiased embeddings is a simplified version of the original embedding. Then, there always exists a model M_1 built with the original embedding that is more efficient as M_2 and more equitable as M_2."

3 Open Problems

We acknowledge several major problems with our proof of claim 1: from the use of word lists to the weak relevance of admission rate in NLP problems. In our experiments, the calculation of measures of bias depend on the choice of word lists. Similarly, the debiasing techniques used in word embeddings also rely on word lists. Thus, we effectively assumed that the subspace of attributes such as gender, race, and religion are accurately captured under these word lists. However, the choice of word list is subjected to the curator's bias and cannot be said to accurately represent the true attribute space.

Even if we grant the proof of claim 1, the applicability of this claim to fair NLP is not immediate. Kleinberg et al., defined equity to be equal admission rates. However, admission rates are seldom the focus of NLP tasks. NLP systems usually support tasks that are difficult to define equity in a clear-cut way, such as translation and question answering. Thus, it is unclear whether the sense of equity in Kleinberg et al., or in general ML research is meaningful in an NLP context. We must look further to refine what fairness and bias means for NLP models.

4 Discussion

As technology and law becomes more intertwined through recent efforts in regulating fairness in ML applications (Raji and Buolamwini, 2019), we discuss ways that legal perspectives can inform and apply to fairness in NLP, and specifically, debiased word embeddings.

Current efforts in debiasing word embeddings loosely follow the principle of anti-classification, by assuming that f(x,D) = f(x,A) is the golden rule. However, careful study of legal definitions show inconsistencies in principle. Legally, the intent of a decision is not what matters but the treatment or decision itself, an unintentional unfair decision warrants the same degree of prosecution as an intentional one. We make the analogy of word embeddings as "intent", as it is used in models to frame the inputs but not the actual classifier itself. The treatment or decision is made by the model for downstream tasks rather than the embeddings. In this sense, biased embeddings may still allow for a fair model, just as our implicit bias does not prevent us from making fair decisions. This resonates with claim 1, where a model that is more equitable and more efficient can be built with the original embedding. While this observation does not invalidate studies about biases in word embeddings, as having a less biased word embedding might make it easier for the model to be fair, it demands a focus on the classifier itself. Thus, we encourage an evaluation of bias not in terms of projections and clustering of the word embeddings, but rather the performance in relevant downstream tasks itself. Benchmark tasks such as WINOGender (Rudinger et al., 2018) and GAP (Webster et al., 2018) exist for coreference resolution. We hope to see benchmark tasks developed in other areas such as named entity recognition, question answering, translation, and more.

A shift in paradigm from anti-classification to anti-subordination, from treatment-focused to impact-focused, should also be considered. With anti-subordination principles, systems need to be aware of the impact of historic and systematic biases, and adjust accordingly such that the current impact of a policy is fair. Here, we propose two ideas of measuring and/or reversing this impact. First, Caliskan et al., showed that word embeddings trained on data from different times exhibit a trend in bias that roughly corresponds to the societal atmosphere of that time (Caliskan et al., 2017). One might benefit from looking into the trend of such a plot and see if there is a momentum associated with it that should be reversed by driving the bias in word embeddings toward the other direction. Another idea might be to train a classifier with two objectives: optimizing on current data and negated versions of past data. We acknowledge the uncertainty surrounding these proposals, but nonetheless encourage applying anti-subordination principles in NLP due to its legal and possibly ethical significance.

5 Conclusion

By following the proof from Kleinberg et al. with word embeddings and surveying whether debiasing captures the protected subspace exactly, we conclude that a more equitable and efficient model can be created using the original word embeddings instead of its debiased counterpart. This observation, in addition to studies of anti-classification, brings attention to the actual model and its downstream impacts rather than just having a fairer word embedding. Legal debates about anti-subordination also prompts a recommendation of redefining fairness in NLP using information on historic and systematic biases.

References

Bradley A. Areheart. 2012. The anticlassification turn in employment discrimination law. *Alabama Law Review*, (955).

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

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

Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning.

Pratik Gajane and Mykola Pechenizkiy. 2018. On formalizing fairness in prediction with machine learning.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.

Wei Guo and Aylin Caliskan. 2020. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases.

Jon Kleinberg and Sendhil Mullainathan. 2019. Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, 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 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, Doha, Qatar. Association for Computational Linguistics.

Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 429–435, New York, NY, USA. Association for Computing Machinery.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Reva B. Siegel. 2004. Equality talk: Antisubordination and anticlassification values in constitutional struggles over brown. *Harvard Law Review*, 117(5):1470–1547.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns.

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

A Appendix

The word lists used in Method and Experiments are taken from Garg et al., Guo et al., and Bertand et al. (Garg et al., 2018; Guo and Caliskan, 2020; Caliskan et al., 2017)

attributes-religion

- **Hispanic**: castillo, gomez, soto, gonzalez, sanchez, rivera, martinez, torres, rodriguez, perez, lopez, medina, diaz, garcia, castro, cruz

- **Asian**: cho, wong, tang, huang, chu, chung, ng, wu, liu, chen, lin, yang, kim, chang, shah, wang, li, khan, singh, hong

attributes-religion

- **Christianity**: baptism, messiah, catholicism, resurrection, christianity, salvation, protestant, gospel, trinity, jesus, christ, christian, cross, catholic, church

- **Islam**: allah, ramadan, turban, emir, salaam, sunni, koran,imam, sultan, prophet, veil, ayatollah, shiite, mosque, islam, sheik, muslim, muhammad

target-race

- **European**: all-American, arrogant, attractive, blond, blue-eyes, highstatus, ignorant, intelligent, overweight, patronizing, privileged, racist, red-neck, rich, tall

- **Mexican**: darkskinned, day-laborer, family-oriented, gangster, hardworker, illegal-immigrant, lazy, loud, macho, overweight, poor, short, uneducated, unintelligent

target-pleasantness

- **Pleasant**: 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

- **Unpleasant**: 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