Democrats are to Republicans as Rihanna is to Jay-Z: Detecting and Removing Political Biases in Word Embeddings

Julia Buffinton, April Kim, Jennifer Podracky

July 23, 2018

Abstract

Text data used for machine learning and natural language processing are rampant with the biases of the human authors of that text. Thus, blindly generating word embeddings from the data will carry these biases through to the applications these word embeddings are used in. Inspired by Bolukbasi et al.'s 2016 paper, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," in which they define metrics to quantify gender bias in embeddings and propose algorithms to remove that bias, we extend their approach to political bias. We apply their approach to word embeddings trained on different corpora to replicate and bolster their results, and then apply this approach to political bias. Using human judgments and the gender bias results as benchmarks, we demonstrate that this approach can also identify and remove complex biases such as political bias. These debiased embeddings can be used in a variety of traditional NLP applications without amplifying the bias of the original embeddings.

1 Introduction

Word embeddings are a commonly used tool in natural language processing to turn text-based content into a set of features that can be fed to any number of powerful machine learning algorithms. The process of turning English words into vectors of numbers (referred to as "embeddings") requires a large body of text to be fed into a machine, which then determines how many unique words exist in its vocabulary. The machine then assigns each word a unique vector representation in such a way that words that appear in similar contexts are located near each other in vector space. This concept is exciting to AI and machine learning researchers, as this improves performance in such areas as machine translation.

Unfortunately, however, the word embeddings themselves are dependent on the body of text that is used to generate those embeddings, and is privy to any biases or prejudices that exist in those bodies of text. Previous papers have addressed the existence of gender biases in these embeddings, as well as their negative impacts; for example, as noted in "Semantics derived automatically from language corpora contain human-like biases" [2], "Google Translate converts these Turkish sentences with gender-neutral pronouns: "O bir doktor. O bir hems¸ire." to these English sentences: "He is a doctor. She is a nurse." What has not been addressed in detail is the existence and prevalence of political biases in word embeddings, especially those generated from news corpora, and how those built-in biases manifest themselves in embedding applications. This paper will attempt to identify those biases, and then address and correct them.

2 Background

This paper was inspired by "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings" [1], a paper published in 2016 by a group of researchers at Microsoft and Boston University. The focus of that work was to examine the machine-generated vector embeddings of words associated with gender, and to find word embeddings of words that should *not* have included a gender influence but were geometrically represented in the gender subspace. The researchers then took actions to correct and regenerate the word embeddings, excluding those biases.

Democrat	Republican
Working-class	Rich
Unionized	Non-Unionize
Younger	Older
Women	Men
Secular	Evangelical
African-American	White
LGBTQ+	Heterosexual
Northern	Southern

Table 1: Perceived characteristics of the Democratic and Republican parties, as written in [3]

The “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings“ paper, along with many other papers that address bias in machine learning and natural language processing [4, 5], focused solely on gender bias. However, there are numerous sources of bias that could be present, including, but not limited to: race, religion, sexuality, age, and political leanings (liberal/conservative). In this paper, we've chosen to focus on political leanings, as it is well-studied and (in our opinions) the least controversial if we have to obtain crowd-sourced feedback on stereotypes, as Bolukbasi et al. did in 2016.

2.1 Origins of Political Party Bias in the U.S.

Since the genesis of the Democratic and Republican parties as we know them today, the two opposing parties that comprise the foundation of the United States' political system have been drifting further and further apart; a 2017 Pew Research Center study found that the gap between Republican-leaning individuals and Democratic-leaning individuals grew from 15 percentage points in 1994 to 36 points in at the end of 2017 [2]. The controversy surrounding the 2016 U.S. presidential election has no doubt contributed; the same Pew study found that Donald Trump's job approval ratings are the most polarized of any first-year president dating back to Dwight D. Eisenhower in 1953 [2]. As the parties and their constituents become more and more polarized, it becomes more important to examine the stereotypes and caricatures being applied to members on both sides. This is especially important when, as a 2015 survey found, each side holds these stereotypes to be true for a much larger percentage than is truly applicable [3].

Some typical dichotomies between Democrats and Republicans generally seen in literature are listed in Table 1.

2.2 Bias within algorithms

Generalizations and biases for and against each political party manifest themselves in the text that humans write, and then become immortalized in the machine-learned algorithms and embeddings trained on those texts.

The GloVe (Global Vectors for Word Embedding) dataset is a file, created by Stanford researchers, containing a set of pre-trained word embeddings built from various corpora [6]. One of its hallmarks in the NLP field is that "the Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words." This enables users of the dataset to make inferences about the relationships between words, based on the Euclidean distance between their embedding vectors.

One such application of understanding relationships between words is analogies. Using the vector representation of a set of three words, one can use simple vector math to find the most appropriate fourth word. For example, "*frog* is to *tadpole* as *cow* is to what?" could be represented as WE(result) = WE(*tadpole*) - WE(*frog*) + WE(*cow*), where WE represents the word embedding for the italicized word. Starting with two words, one can generate a second pair of words with a similar geometric relationship, and thus an inferred similar language relationship. For example, the first analogy below is interpreted as democrats:republicans::socialists:nationalists. Each automatically generated analogy was evaluated by the researchers and determined to be stereotypical, definitional, or unrelated to the original word pair "democrats-republicans." The most stereotypical analogies are listed below.

12 Stereotypical "Democrats-Republicans" analogies, automatically generated from the GloVe embedding.

- socialists-nationalists
- gorbachev-reagan
- communists-revolutionaries
- europe-america
- fascist-racist
- mommy-daddy
- fenway-yankee
- brown-white
- unemployment-incomes
- populous-wealthiest
- beer-whiskey
- allies-foes

7 Definitional "Democrats-Republicans" analogies, automatically generated from the GloVe embedding.

- democratic-republican
- democrat-gop
- d-r
- liberal-conservative
- liberals-conservatives
- liberalism-conservatism
- dnc-rnc

Another application of using embeddings to understand word relationships is to find the group of classifiers, for example, colors or professions, that are most related to a given word. By creating an axis between two words, or the two vectors that represent those words, one can see which classifiers are closer to one end of the axis than the other. This was done in the reference paper for the word pair *she-he* and professions; we repeated this examination using *progressives-conservatives* and professions, using the GloVe embedding pre-trained on Gigawords 5 and Wikipedia 2014 text corpora. Note that any occupations specifying an elected office (e.g. president) were excluded.

12 Most Extreme "Democrats" Occupations

- welder
- ballplayer
- cellist
- sportswriter
- cabbie
- sportsman
- bartender
- janitor
- patrolman
- artiste
- ballerina
- receptionist

12 Most Extreme "Republicans" Occupations

- minister
- deputy
- dean
- judge
- chancellor
- officer
- critic
- director
- attorney
- manager
- principal
- commissioner

Interestingly, the biases discussed in the earlier section are present in the word embedding, evidenced by the professions most closely related to either "progressives" (Democrats) or "conservatives" (Republicans). Under "progressives", we see three words related to working-class/blue-collar occupations: welder, cabbie, and janitor. We see ballerina and receptionist, which are two of the most extreme "she" professions. We also see artiste, cabbie, and sportsman, three of the most "black" professions. Under "conservatives", the most extreme word is "minister", a religious occupation. We see manager, one of the most "he" and "white" professions. By looking solely comparing the distances between political leanings and professions, we can see eight different types of biases that are present in U.S. society. These professions align with the human-identified stereotypes (Table 2) for each political party. Our hope is that by identifying these biases in computed embeddings, we can remove these human-created biases from those embeddings and thus any future applications that humans build on top of them.

3 Methods

The following methodology was undertaken to understand political biases present in word embeddings, with the ultimate goal of using this knowledge to apply debiasing algorithms that remove political bias from the embeddings. As in the section above, we use GloVe embedding pre-trained on Gigawords 5 and Wikipedia 2014 text corpora. To achieve this, we followed three main steps:

1. Identify geometric subspace for political bias.

2. Take steps to reduce the bias in that subspace.

3. Rebuild the word embedding without bias, while maintaining its accuracy and usability.

We first pursue this for gender bias as a baseline, to replicate Bolukbasi et al.'s findings, and then extend it to political bias.

3.1 Identifying the Geometric Subspace for Political Bias

To identify the subspace for the bias attribute of interest, we employed similar methods as Bolukbasi et al. (2016). Using several pair difference vectors (*maleVector* - *femaleVector*, or difference between extremes of other sources of bias), we computed their principal components (PCs). The PC(s) can be used to identify a bias direction $b \in \mathbb{R}^d$ that will be used to quantify bias in words and associations.

Bolukbasi et al. found a single direction that could be interpreted as the gender subspace. However, it is not necessary that this subspace that captures the attribute of interest is single-dimensional. Especially with a bias that is more complex, it may be the case that multiple dimensions capture this subspace.

A decrease in the proportion of variance captured by each PC is expected, however, the PC(s) that capture the majority of variance in the pair difference vectors should be significantly larger than the rest of the components, indicating that they can be used to explain the bulk of the difference between word pairs, namely, the bias we are examining.

To apply these methods to any sort of bias, we follow the below steps:

1. Get top 5-10 pairs of words that indicate definitional analogies, from human-generated evaluations (crowd-sourced via Mechanical Turk or otherwise).

2. Generate vector for difference between vectors in pair ('pair difference vector').

3. Compute PCs of the pair difference vectors.

4. Identify direction(s) that explain the majority of variance in these vectors.

5. Attribute this space as the bias subspace.

3.2 Reducing Bias

The subspace identified using the techniques above was used to "debias" the words that contain stereotypes. We elected to pursue the "hard" debiasing option proposed in Bolukbasi et al., where they **neutralize and equalize**. This is applied to sets of words ("equality sets) rather than pairs to allow enhancement of this approach to other types of biases and groups of words. This ensures that neutral words remain neutral in the subspace of interest and that words outside of the subspace are equalized with respect to the words in the equality set. For example, if democrats-republicans was an equality set, then after equalization, *homosexual* should be equally as close to *democrats* as it is to *republicans.* Admittedly, depending on the amount of polysemy of the words in the bias attribute subspace, this may equalize words that should not equalized, but overall this approach seems to do more good than harm.

Once the subspace was identified, we used this space as well as a list of words to neutralize ($N \subseteq W$) and a family of equality sets ($\varepsilon = E_1, E_2, ..., E_m, E_i \subseteq W$). Each word $w \in N$ is re-embedded to be

$$\vec{w} := (\vec{w} - \vec{w}_B / \|\vec{w} - \vec{w}_B\|)$$

Additionally, for each set $E \in \varepsilon$

$$\mu := \sum_{w \in E} w/|E|$$

$$v := \mu - \mu_B$$

$$\vec{w} := v + \sqrt{1 - \|v\|^2}\frac{\vec{w} - \mu_B}{\|\vec{w} - \mu_B\|}$$

(SOMETHING ELSE ABOUT CENTERING?)

4 Evaluation

4.1 Bias attribute subspace

To evaluate our methods, we used our baseline of gender bias to confirm that we achieve similar performance and results as Bolukbasi et al. This ensures that our methods work as expected, so we are confident in results achieved when expanding to other sources of bias.

To validate the list of word pairs used for identifying the subspace of the bias attributes of interest, we obtained crowd-sourced ratings of how closely the proposed word pairs correspond to human understanding of the bias attribute classifier (i.e., *democrats-republicans*).

4.2 Debiasing

To confirm that words have been correctly debiased, we tested that only stereotypical words have been debiased and definitional words remain unaffected. So, we ensured that the bias attribute has not been removed from definitional words, and that definitional analogies have not changed, such as *democrats:liberalism::republicans::conservatism*.

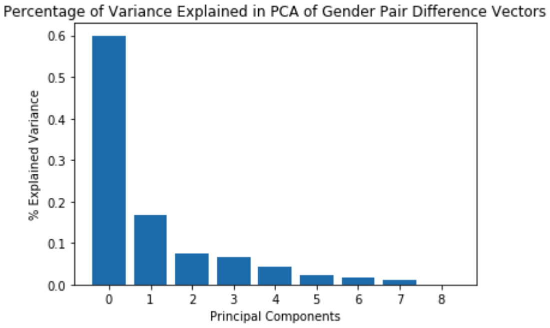
Then, we examined whether the bias component is removed from other stereotypical word embeddings. We tested the debiased embeddings to measure whether related words have similar embeddings and that they still perform well in analogy tasks. Bolukbasi et al. use RG [7], WS [8], MSR-analogy [9] metrics, so we used the same. The biased and debiased word embeddings should perform similarly on all of these metrics.

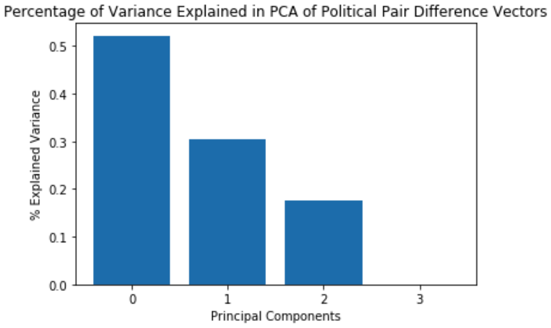
5 Results and Discussion

5.1 Identifying Bias Subspace

Employing similar methods as Bolukbasi et al., we used PCA to identify the gender subspace based on the same list of gender pairs (excluding *Mary-John* because Mary was not found in the corpus): *she-he*, *her-his*, *woman-man*, *herself-himself*, *daughter-son*, *mother-father*, *gal-guy*, *girl-boy*, *female-male*. It is similarly clear that there is a single direction that explains the majority (approx. 60%) of the variance in the gender pair difference vectors. This supports the hypothesis proposed by Bolukbasi et al. that this top PC captures the gender subspace.

However, it is not necessary that this subspace that captures the attribute of interest is single-dimensional, as seen in the PCs of the political difference vectors. We used the following list of political pairs to generate difference vectors: *democrat-republican*, *liberal-conservative*, *liberalism-conservatism*, *left-wing-right-wing*. The top 3 pair difference vectors achieve similar results for the variance they capture, suggesting that the political subspace may be more complex, captured by several dimensions. Most interestingly, all of the variance in these words is captured in these three vectors.





5.2 Debiasing

(TODO)

6 Next Steps

(TODO)

References

[1] Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., and Kalai, A.T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NIPS.

[2] Doherty, C. (2017, October 05). Key takeaways on Americans' growing partisan divide over political values. Retrieved from http://www.pewresearch.org/fact-tank/2017/10/05/takeaways-on-americans-growing-partisan-divide-over-political-values

[3] Ahler, D. J., and Sood, G. (2018). The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences. The Journal of Politics, 80(3), 964-981. doi:10.1086/697253

[4] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). doi:10.18653/v1/n18-2003

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

[7] H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. Communications of the ACM, 8(10):627–633, 1965.

[8] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. In WWW. ACM, 2001.

[9] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In HLT-NAACL, pages 746–751, 2013.