Studying Political Bias via Word Embeddings

Josh Gordon
Marzieh Babaeianjelodar
Jeanna Matthews
Clarkson University
jogordo,babaeim,jnm@clarkson.edu

ABSTRACT

Machine Learning systems learn bias in addition to other patterns from input data on which they are trained. Bolukbasi et al. pioneered a method for quantifying gender bias learned from a corpus of text. Specifically, they compute a gender subspace into which words, represented as word vectors, can be placed and compared with one another. In this paper, we apply a similar methodology to a different type of bias, political bias. Unlike with gender bias, it is not obvious how to choose a set of definitional word pairs to compute a political bias subspace. We propose a methodology for doing so that could be used for modeling other types of bias as well. We collect and examine a 26 GB corpus of tweets from Republican and Democratic politicians in the United States (presidential candidates and members of Congress). With our definition of a political bias subspace, we observe several interesting and intuitive trends including that tweets from presidential candidates, both Republican and Democratic, show more political bias than tweets from other politicians of the same party. This work models political bias as a binary choice along one axis, as Bolukbasi et al. did for gender. However, most kinds of bias - political, racial and even gender bias itself - are much more complicated than two binary extremes along one axis. In this paper, we also discuss what might be required to model bias along multiple axes (e.g. liberal/conservative and authoritarian/libertarian for political bias) or as a range of points along a single axis (e.g. a gender spectrum).

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

KEYWORDS

political bias, natural language processing, Twitter dataset

ACM Reference Format:
Josh Gordon, Marzieh Babaeianjelodar, and Jeanna Matthews. 2020. Studying Political Bias via Word Embeddings. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3366424.3383560

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3383560

1 INTRODUCTION

In 2016, Bolukbasi et al. [1] published an influential paper demonstrating a method to quantify and remove gender bias that a machine learning (ML) model learned from a corpus of human text. This is important for revealing bias in human text as well as reducing the impact biased ML systems can have on hiring, housing, and credit. In this paper, we explore how we might use a similar methodology to model other kinds of bias such as political bias.

As with gender, we begin as Bolukbasi et al. did with attempting to model political bias as simply two binary extremes along a single axis. Neither gender nor political bias is as simple in the real world as two points on a single axis, but we wanted to see how useful this model could be in the case of political bias. We observe that in the US there are two major political parties (Republican and Democratic), and we start with trying to measure the degree of political bias in a corpus of Twitter data using these two binary points along a single axis. We do not, as Bolukbasi et al. did, attempt to use this method to remove bias from a text. We are using it to model and describe bias, not to debias.

The first big challenge in applying the Bolukbasi et al. methodology to political bias is that we need a defining set of word pairs. A set of definitional word pairs for gender bias (she/he, guy/gal, woman/man) are easier to find than a set of definitional word pairs for political attitudes. In other words, definitionally Republican words versus definitionally Democratic words are harder to identify and are more specific to a given political discourse/community (US Politics in 2019) than to a whole language (English). In this paper, we propose a specific methodology for identifying word pairs for political bias and apply it to a specific corpus of text that we collected. We describe how this methodology could be used with other types of bias as well and consider how we could extend the methodology to 2 axes or a spectrum along a single axis.

2 OUR TWITTER DATASET

We used the Twitter API to collect tweets from 576 accounts linked to presidential candidates and members of congress in the United States. Of the 576 accounts, 258 are classified as Republican and 318 as Democratic. The total size of the data set is 26 GB. We put these accounts into 6 categories as shown in Table 1. In some cases, an account appears in multiple categories (e.g. the account SenSanders is in both the Democratic Senate category and the Democratic Presidential Candidates category). Independents are grouped with the two parties rather than handled separately and for candidates that dropped out of the race, we continue to follow them. In some cases, there are multiple accounts linked to the same individual/campaign (e.g. SenBooker and CoryBooker).

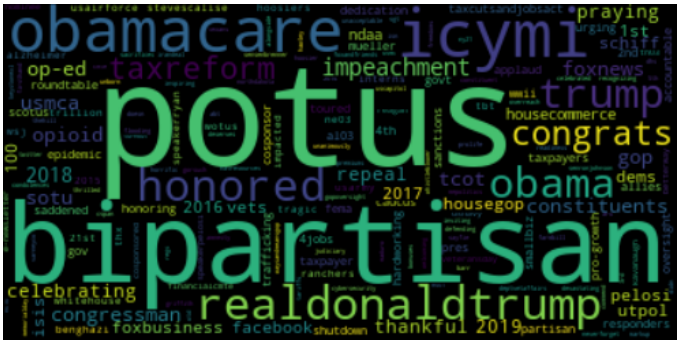

Figure 1: Word cloud of most common words in the Tweets from all Republican accounts (candidates, senators and congressional representatives

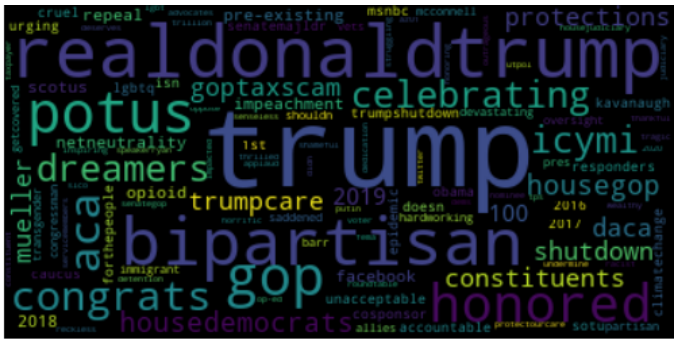

Figure 2: Word cloud of most common words in the Tweets from all Democratic accounts (candidates, senators and congressional representatives

We collect pages of tweets from each account in our list as far back as there are tweets. We filter out retweets to capture the original speech of politicians. We perform preprocessing such as removing hyperlinks and tokenizing into words. After that, we count the frequencies of every word in the original tweets and remove the 10,000 most common words in English such as "the", "and", etc.

We generated word clouds from this data where the size of font for each word corresponds to the frequency of use for that word. The Republican and Democratic word clouds are shown in Figures 1 and 2 respectively.

3 MODELING GENDER BIAS VS. MODELING POLITICAL BIAS

Bolukbasi et al. computed gender bias in their paper "Man is to computer programmer as woman is to homemaker? Debiasing Word Embeddings" using a definitional set of 10 gender word pairs: she-he, her-his, woman-man, mary-john, herself-himself, daughter-son, mother-father, gal-guy, girl-boy, female-male) [1]. They use these definitional word pairs to define a gender subspace of the word embedding.

They view these words as vectors and calculate the center of each definitional pair. For example, to calculate the center of the pair she/he, they average the vector for "she" with the vector for "he" and then calculate the distance of each word in the definitional pair from the center by subtracting the center from each word in the pair (e.g. "she" - center). Using these pairs, they compute a vector $w \in \mathbb{R}^{768}$ which captures the overall direction of gender in the word vectors; male words will be on one end of the space and female words will be on the other end about this direction. The interesting part is then measuring the cosine similarity of supposedly gender neutral words with this bias direction to measure the gender bias these other words exhibit in the embedding.

In order to apply this same methodology to political bias, we first need a set of defining words pairs. A defining set for gender seems more straightforward and primarily contains antonyms or words that are opposites from the perspective of gender. It was not immediately obvious how to do this for political bias. Here we propose and follow a specific methodology that could also be applied to other types of bias.

We proposed and utilized the following methodology. First, identify a set of documents characteristic of each bias you wish to explore. In our case, this is the set of tweets from Republican politicians in the US and the set of Democratic politicians in the US. Second, from these sets of documents, extract lists of the most commonly used words. Third, look for tuples of corresponding words that include the most frequently used words, ideally a related or corresponding pair of frequently used words. In our case, this is often different words that Republicans and Democrats might use to talk about the same idea. Fourth, use this list as inspiration for other similar word pairs.

Figures 1 and 2 show word clouds generated from all the Republican accounts and Democratic accounts. As described, we used these word clouds to inspire a set of proposed defining words to compute bias on the space of Republican and Democratic speech as shown in Table 2. We observe that this process relied on some domain specific knowledge, in our case, a knowledge of US politics. This domain specific knowledge was necessary both to identify the two corpora of text reflecting the bias we wished to explore and to identify candidate word pairs.

It is important to note that, unlike with the list of 10 gender-word pairs in Bolukbasi et al. these words are not chosen to be direct antonyms. Instead, they are different words that politicians on both sides might use to describe the same concept or to describe a parallel concept. We propose that this generalization of the approach in Bolukbasi et al. as a way to model broader categories of bias which do not have as obvious a representation in the underlying language as gender does in English (and in many other languages as well). Many languages do have antonyms across a binary gender axis, but many would not have clear antonyms across other binary bias axes, such as with political bias. In other words, our language about gender is more often binary than our language about politics. However, in both cases, we could choose to reduce a more complicated underlying reality to a simple binary model for the purposes of discussion or description. Later, in this paper, we discuss how we might extend this to a more complex model.

Of the 28 word pairs originally selected by the authors, 6 contained multiple word phrases like tax reform/tax scam. Those could be supported by summing the vectors of the individual words [5], but we have not done that for this paper. Two other proposed word

Table 1: Twitter Accounts Studied

Category	Example Accounts	Number of Accounts	Number of Tweets
Republican Presidential Candidates	realdonaldtrump, GovBillWeld	4	3821
Republican Senators	senatemajldr,marcorubio	53	136653
Republican Members of the House of Representatives	RepMattGaetz,RepStefanik	201	363772
Democratic Presidential Candidates	JoeBiden, AndrewYang, PeteButtigieg	25	64892
Democratic Senators	SenSchumer, SenWarren	47	129803
Democratic Members of the House of Representatives	SpeakerPelosi, RepAOC	246	495036

pairs contained words not found in the dataset (e.g. the word potus is a common word in the data set, but impotus was not found).

4 POLITICAL BIAS RESULTS

Using our final list of 20 word pairs, we computed a bias subspace in each 100 dimension word embedding on groups of politicians. Once we had done that, we were able to compute a direct bias metric for political bias as shown in Figure 3 as a bar chart and in Figure 4 as a log scale heat map.

We see interesting patterns in these results. Trump's tweets in the Republican candidates category give that category by far the highest bias score we see of 0.97. In general, tweets from presidential candidates show higher bias than for other politicians of the same party and Republicans overall have a slightly higher political bias score. We had hypothesized that we would see candidates being more extreme (more bias) than congress because more extreme views can be effective in driving political momentum.

We are also able to ask other questions of the word embeddings we produced. For example, we can ask the top 10 words related to a given word by closeness in the vector space. Figures 5 and 6 show the word clouds of the top 10 closest words to the word immigrant for all Republican accounts and Democratic accounts respectively. The differences are striking. The Republican cloud for the word immigrant contains the words aliens, arrested and even murder, while the Democratic cloud for the word immigrant contains the words refugees, undocumented and indefinitely. This is an interesting application of word embeddings for studying political bias on its own and helped give us confidence in the value of the methodology we have proposed.

5 MODELING BIAS BEYOND A BINARY ALONG A SINGLE AXIS

Gender in practice is more complicated than a binary variable in one dimension. For example, in 2014, Facebook updated its interface to allow users to describe their gender beyond male or female to reflect cultural and societal changes [6]. These same cultural and societal changes will manifest bias in machine learning models, and the need to detect and fix it only grows.

Many other types of bias are multidimensional in nature as well; we wanted to push this methodology into new territory and validate the efficacy of direct bias on other types of bias.

We began by asking ourselves if there is a useful way to model political bias as a binary choice along one axis. In many political conversations, there are two opposing parties or candidates. As with gender, in most cases it is not that simple, but we wanted to see if modeling it as a binary could yield some useful results as it did in the case of gender bias in Bolukbasi et al.

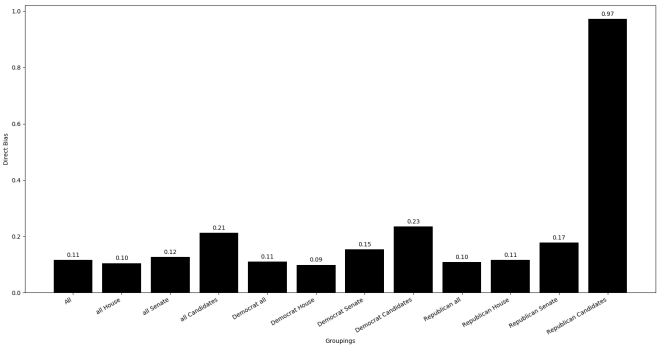

Figure 3: Direct political bias for groupings of politicians

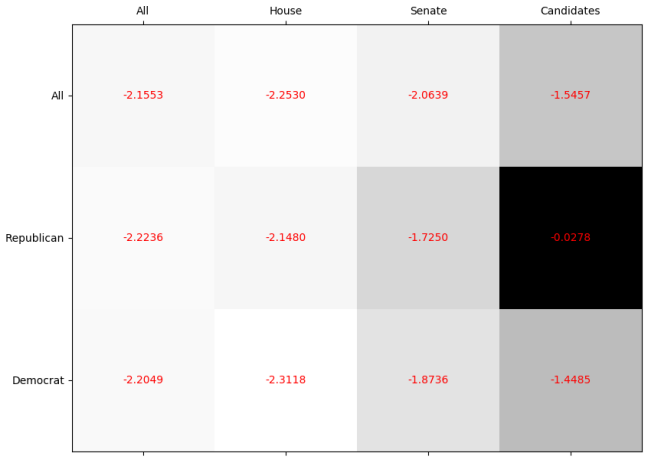

Figure 4: Log scale of direct political bias for groupings of politicians

Another model of political bias would be a two-dimensional representation, with one axis for left and right, and another axis for authoritarian and libertarian, as modeled by the Political Compass Test [3, 4]. The bias subspace there would be two dimensional, and defining sets would require two word pairs. This suffers from the same problem of determining what the "ground truth" is to compare words to; the English language doesn't have definitionally

Republican Word	Democratic Words	Discarded?
trumpcare	obamacare	—
tax reform	tax scam	not one word
invasion	immigration	—
illegal	refugee	—
illegal	dreamer	—
illegal	asylum seeker	not one word
gop	dems	—
republicans	democrats	—
libtard	deplorable	not in vocabulary
housegop	housedemocrats	—
gun rights	gun control	not one word
2nd amendment	gun control	not one word
pro life	pro choice	not one word
witch hunt	investigation	not one word
hoax	investigation	—
redistricting	gerrymandering	—
bluelivesmatter	blacklivesmatter	—
alllivesmatter	blacklivesmatter	—
potus	impotus	not in vocabulary
fox	cnn	—
fox	msnbc	—
mcconnell	pelosi	—
capitalism	socialism	—
isolationist	globalist	—
nationalist	globalist	—
red	blue	—
right	left	—
entitlements	programs	—

Table 2: Defining Sets for Republican/Democrat Bias

Figure 5: Words closest to "immigrant" in the Republican embedding

libertarian words, but said words can be gathered by analyzing frequencies of known libertarian sources. If using multiple axes, it would be interesting to evaluate the degree to which the axes are independent or correlated with each other (e.g. a left/right political axis may not be independent from an authoritarian/libertarian axis).

The other missing component of the one dimensional binary bias model is capturing bias on a continuous variable, such as gender as

Figure 6: Words closest to "immigrant" in the Democrat embedding

a spectrum. To avoid the issues with representing continuous variables in a computer, our goal will be to deal with a one dimensional discrete variable with many possible values. When computing a bias direction for a binary variable, each defining set contains two vectors, the center of which is used. This case with two points is always co-linear. For the discrete variable, we propose fitting a least-squares line to each defining set, and using that as the representation of the defining set for use in PCA, as opposed to the difference of the vector pair.

Forms of bias on non-binary variables have been studied before, such as in [2] where racial bias was measured using skin tone along a spectrum and separated into discrete values. Using a range of skin tones along a single axis, rather than a more complex and multi-faceted concept of race was a key innovation. Other researchers have used the top-N names in a racial category as reported by census data as a way of tracking racial patterns and racial bias in professions [7, 8]. We notice mary-john in the Bolukbasi et al. list of gender word pairs [1]. Summing word vectors of names drawn from Census data could be a useful approach to defining a subspace for racial bias.

6 CONCLUSIONS

We demonstrated that direct bias as a metric can be successfully applied to a different type of bias less fundamentally present in language than gender. We used it to measure political bias in a data set of tweets from current US politicians (Republican and Democratic senators, congressional representatives and presidential candidates). We proposed a methodology for finding a defining set to apply to other bias problems and discussed extending this method beyond bias modeled as two binary points along a single axis. With our definition of a political bias subspace, we observed several interesting and intuitive trends including that tweets from presidential candidates, both Republican and Democratic, show more political bias than tweets from other politicians of the same party.

7 FUTURE WORK

Some immediate future work would be including multi-word tokens like gun rights/gun control, as well as different types of word embeddings.

We would like to apply this approach to difference political environments such as applying this to the politics of a different country or region or applying it to a different time period for the United States. It would be interesting to use it to compare political discussions for overlapping geographic areas of increased size such as city, state, national and international. We would also like to use this to quantify political bias across different types of media and use this method to explore questions such as "Is social media more biased than speeches?", "Is social media more or less biased than websites?", "Is the individual speech of candidates more or less biased than left/right leaning print media or cable news media?".

8 ACKNOWLEDGMENTS

This work could not have been started without important discussion from Izzi Grasso, Hunter Bashaw, Graham Northup, and Abigail Matthews. The dataset could not have been created and processed without the hardware and support from the Clarkson Open Source Institute.

REFERENCES

[1] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[2] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[3] The Political Compass. 2017. The Political Compass - a brief intro. (2017). https://www.youtube.com/watch?v=5u3UCz0TM5Q

[4] The Political Compass. 2017. Political Compass Test. (2017). https://www.politicalcompass.org/test

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Brandon Griggs. 2014. Facebook goes beyond 'male' and 'female' with new gender options. (2014). https://www.cnn.com/2014/02/13/tech/social-media/facebook-gender-custom/index.html

[7] Alexey Romanov, Maria De-Arteaga, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. *CoRR* abs/1904.05233 (2019). arXiv:1904.05233 http://arxiv.org/abs/1904.05233

[8] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark D. M. Leiserson, and Adam Tauman Kalai. 2018. What are the biases in my word embedding? *CoRR* abs/1812.08769 (2018). arXiv:1812.08769 http://arxiv.org/abs/1812.08769