Capturing Bias in the Liberal and Conservative News Sources

Anonymous ACL submission

Abstract

Many people try to become aware of the important events happening in the world through reading news articles from well-known news outlets. Many of the big events and decisions, directly or indirectly affects our lives, so it is important to know about them and make reactions if needed. Unfortunately most of the news sources are politically biased and they convey the news in a way to give the reader an opinion close to themselves. The inherited bias in NLP models such as word embedding is a critical problem affecting all social applications of them. However, due to vague notion of the bias, and black-box nature of many NLP methods, detecting the bias is proven to be difficult. In this work, we investigate the political bias in the word embedding of news corpora. In this work the bias in liberal news sources and conservative news sources is shown to be captured by a geometrical direction in the word embeddings of them. We first embed each side's corpora separately and then align the word embeddings. We show that with aligning of the word embedding spaces we can compare the vectors. We use metrics to show the bias between the left-leaning and right-leaning news sources. We talk about how we capture the direction of bias with three methods, first looking at the entities, second looking at the positive adjectives and third looking at the negative adjectives.

1 Introduction

As Natural language processing algorithms penetrate every aspect of our social life, there is a growing risk for these algorithms to be systematically biased(Mehrabi et al., 2019). Studies show that these algorithms implicitly learning the human biases that are embedded in their training data (Caliskan et al., 2017; Mehrabi et al., 2019) and they even also magnifying them when deployed in practice. There have been extensive research on detecting and measuring the gender bias in word embeddings (Caliskan et al., 2017; Bolukbasi et al., 2016; Brunet et al., 2018) that captures the bias with respect to occupation words. These studies measure bias by simply computing the pairwise distances among two groups of words: neutral and target. Target words are set of words where all words in each word-set represent a similar concept. Such as a specific gender. Similarly, neutral words are defined as set of word sets, where all words in each word set fulfill two criteria: (1) represent similar concepts (e.g. occupation, human characteristics) (2) Must be perceived neutral with respect to target words. For example a work for analysing gender bias can have this setting for target sets A and B, and neutral words S and T:

$$S = \{math, algebra, geometry, calculus\}, \quad (1)$$

$$T = \{poetry, literature, symphony, sculpture\} \quad (2)$$

$$A = \{male, man, boy, brother, he\}, \quad (3)$$

$$B = \{female, woman, girl, sister, she\} \quad (4)$$

In our approach, we use entities, positive adjectives and also negative adjectives as target words. Our approach for capturing the bias in the news articles is based on word embedding of each of the left-leaning and right-leaning news sources. we have one vector for target words in left-leaning news sources embedding and another vector for right-leanings news sources embedding.

Research on word embeddings has drawn significant interest in machine learning and natural language processing. There have been hundreds of papers written about word embeddings and their applications. Empirical results using these methods have shown they are successful at learning the meaning of words. In fact, the resulting embed-

ding space seems to have directions of semantic and syntactic meaning that can be exposed through simple operations on word vectors. For example we can look at many analogies like:

$$vector(King) - vector(Man) + vector(Woman) \simeq vector(Queen)$$

We are going to calculate our words embedding using word2vec method(**?**). The current popular word embedding methods inherit the stereotypical biases from their train data. This is a problem because the widespread use of these algorithms in machine learning systems can thus amplify stereotypes in important contexts.

In this work we are going to address political bias in the news articles from left leaning and right leaning news sources and visualize some of the existing biases towards well-known politicians or sensitive political context like immigration.

2 Related Works

For the methodology to address bias in word embeddings we are inspired by the work of (Bolukbasi et al., 2016). In their work their target words are two sets that defines the binary gender. The formulation 1 shows summary of their target words and neutral words. They show that even word embeddings trained on Google News[1] articles exhibit female/male gender stereotypes to a disturbing extent. They calculate the differences of words in A and B and calculate the PCA of those differences. The calculated PCs show the main directions that explain the semantic difference between the two genders. Then they project the words on this direction and see that the neutral words are very polarized rather than being placed in the middle. They observe that science occupation are closer to men and art occupations are closer to female.

Another challenge is how to align the embeddings to make them comparable. Some work have been done focusing on the evolvement of temrs during time. In a work by Garg *et al.*, they integrate word embeddings trained on 100 years of text data with the US Census and develop metrics based on word embeddings to characterize how gender stereotypes and attitudes toward ethnic minorities in the United States evolved during the 20th and 21st centuries starting from 1910 (Garg et al., 2018). They compute the average embedding distance between words that represent women—e.g., she, female and a group of gender neutral words like occupations, also compute the average embedding distance between words that represent men and the same occupation words. They have used the intuitive and natural metric for the embedding bias which is the average distance for women minus the average distance for men A group of works concentrate on the evolving of word semantics during time. They have captured interesting biases looking at the metrics in different years. There is no embedding alignment in their work, they get the static word embedding for each year and calculate the metrics for that year and show the gradual change of numbers in plots. The data and code related to their paper are available on GitHub [2]. Using the similar idea for our work we need to get two set of words representing each of political sides and also a set of political neutral interesting words. Finding those sets of words that are also frequent in our dataset is challenging. Another drawback is that calculating euclidean differences in embedding spaces is not a very robust metric.

Some of the related works are focusing on bilingual word embedding which builds semantic embeddings associated across two languages. The work of (Zou et al., 2013) introduces an unsupervised neural model to learn bilingual semantic embedding. The result of this work might not be very interesting for our task because it embeds our two different set of corpus (left and right) in a way that the corresponding words that have the same meanings will end up very close in the vector space. Another disadvantage of this method is its slowness; it took 19 days for their model to train on a 8-core system. This paper is old and they have compared their methods like naive and pruned tf-idf and we don't have comparison of it with contemporary state of the art models.

We want to be able to separately embed the words from the corpora corresponding to each of the right and the left side news sources and then align the vector spaces. The work of (Hamilton et al., 2016) use orthogonal Procrustes in order to align word embeddings across time-periods. This method searches for the best rotational alignment and preserves cosine similarities. They use two measures to evaluate their results: synchronic ac-

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

[2] https:// github.com/nikhgarg/EmbeddingDynamicStereotypes

curacy (i.e., ability to capture word similarity) and diachronic validity (i.e., ability to quantify semantic changes over time) which they do in two ways: detecting known shifts and also discovering shifts from data. This method can be applied to our problem because we are trying to find the alignment between embeddings of left-wing news corpora and right-wing news corpora. We also can look at the embeddings of all news corpora during time spans and another interesting question is whether the similarity of the words changes over time in compare to left terms and right terms. A drawback of their method can be that they only look at rotational alignment and don't capture the changes in the cosine similarities between the words. They have their code available on github.[3]

Later than Hamilton's work, there is another work(Yao et al., 2018) that instead of aligning different static embeddings simultaneously learns time-aware embeddings. Previous techniques usually do not consider temporal factors, and assume that the word is static across time. They are interested in computing time-aware embedding of words. They have used qualitative and quantitative methods to evaluate temporal embeddings for evolving word semantics. Their work can be modified for our problem setting to obtain political-aware embeddings.

3 Data-set

In this work we have used "All the News" data-set from Kaggle[4]. It contains 143,000 articles from 15 American News Outlets.
The liberal news sources in this dataset are: New York Times, CNN, Atlantic, Buzzfeed News, New York Post, Guardian, NPR, Vox, and Washington Post.
The conservative news sources are Breitbart, Fox News, National Review, and New York Post.
There are also Reuters and Business Insider which are central-leaning news sources and we do not work with them. In the figure 1 we see the distribution of articles among different news sources.

We have 85,551 articles for left news sources and 34,338 articles for right news sources respectively 71% and 29% of data. That means we need to balance the number of articles for each side. For that we randomly sample from the bigger side.

[3]https://github.com/williamleif/histwords
[4]https://www.kaggle.com/snapcrack/all-the-news

	source	count	%
0	Breitbart	23781	0.17
1	New York Post	17493	0.12
2	NPR	11992	0.08
3	CNN	11488	0.08
4	Washington Post	11114	0.08
5	Reuters	10710	0.08
6	Guardian	8681	0.06
7	New York Times	7803	0.05
8	Atlantic	7179	0.05
9	Business Insider	6757	0.05
10	National Review	6203	0.04
11	Talking Points Memo	5214	0.04
12	Vox	4947	0.03
13	Buzzfeed News	4854	0.03
14	Fox News	4354	0.03

Figure 1: Distribution of articles in the news sources

4 Methodology

4.1 Attribute Words

We want to identify the subspace that conveys the subject of the bias. For calculating that subspace, first we need sets for attribute words that gives a characteristic to that subspace. We have three experiments with different definitions of the attribute words.

In the first experiment we focus on the entities. In that case the attribute words that represent each group are the most frequent common entities in each of the left-leaning news sources and right-leaning news sources. In this case we are trying to define the subspace that in its extreme values we can see how left-leaning and right-leaning news sources perceive the frequent entities (which are mostly the names of well known people, name of places, and also organisations) After embedding each side separately, we have the attribute sets A and B defined as:
A = The vectors of common entities in the word embedding of left-leaning news sources
B = The vectors of common entities in the word embedding of right-leaning news sources

In the second and third experiment, the attribute

words that we chose are positive-meaning and negative-meaning sets of adjectives respectively. A few examples of elements of these sets are as below:
pos_adj = {affluent, agreeable, amazing,...}
neg_adj = {abysmal, adverse, anxious, awful,...}
In these cases we are focusing on how each side talks about good and bad adjectives and we further analyse which of the well-known people do they associate with each of these sets.

4.2 Calculating PCA

To identify the subspace, we took the attribute words vectors in the embedding of each side, calculated their pairwise differences and computed its principal components (PCs). In each setting we have looked at one or two directions that explain the majority of variance in these vectors. Note that, from the randomness in a finite sample of noisy vectors, one expects a decrease in eigenvalues. Therefore we hypothesize that the top PCs, captures the subspace corresponding to the attribute words(Bolukbasi et al., 2016). In each experiment, the extreme values in the subspace (linear if we only choose the one top PC) show how each side talks about the attribute words. In our setting, the more positive values correspond to the left-leaning perspective and the more negative values correspond to the right-leaning perspective.

4.3 Training Embedding

In our methodology first of all, we train the word embedding of the vocabulary in left leaning news sources and right leaning news sources separately. We set the dimension of the embedding space equal to 300, the window size 7 and minimum count of each word equal to 10. We use gensim package[5] for training the word2vec model. Gensim contains many easy-to-use variants of word embeddings (e.g. LSI/SVD, word2vec, wordrank, ...), wrappers for using other packages like GloVe, and is very well maintained, so this method is a reasonable choice.

4.4 Procrustes matrix alignment

When we embed the two subset of data separately, due to different initial seeds they can't be very comparable. In figure 2 we show that when we align the embeddings, the cosine similarity of the vectors of the same word increases. Hence, that is a necessary step for defining measures of bias between these two separate embeddings.

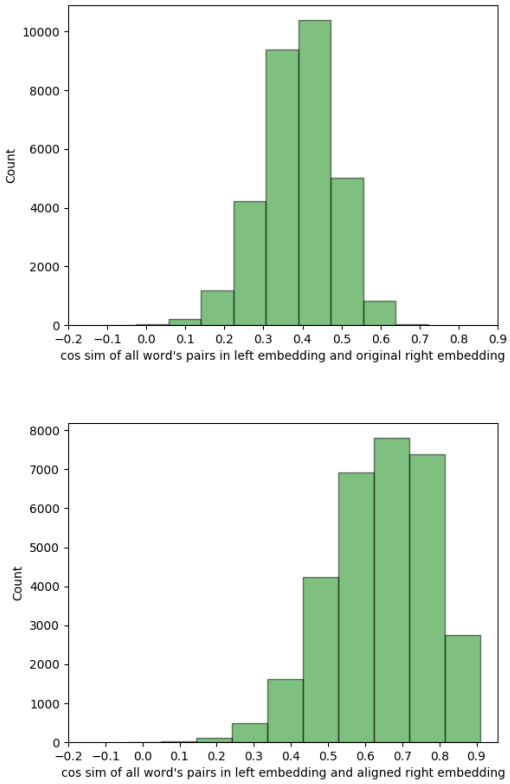

Figure 2: The cosine similarity of pairs of vectors in the two embeddings, become larger after aligning the embeddings using Procrustes matrix algorithm. After aligning we can calculate the measures on the pairs of same word's vector.

We use the Procrustes matrix alignment(Hamilton et al., 2016) algorithm, in order to align the two embeddings that we have for each of left news sources and right news sources. There is a wrapper of (Hamilton et al., 2016) published in github[6] that is a code for aligning two gensim word2vec models using Procrustes matrix. Equation 5 shows the loss function of Procrustes algorithm. From the loss function we can see that this algorithm does not change the relative positions of vectors in each embedding. It only learns a rotation matrix for aligning.

$$R^T = argmin_{Q^TQ=I}||W^TQ - W^{(T+1)}||_F \quad (5)$$

4.5 Extracting Entities

We need to focus on some words. For that reason we have extracted the entities of left-leaning news

[5]https://radimrehurek.com/gensim/

[6]https://gist.github.com/quadrismegistus/09a93e219a6ffc4f216fb85235535fa

articles and right-leaning news articles using spacy package. The types of entities that we extract are:

- Person
- Location
- Organization

5 Experiments

In our methodology after aligning the two embeddings, we look at the attribute words that appear in both of left and right leaning news sources and calculate the PCA of their differences. We chose three sets of attribute words: common entities, positive adjectives, and negative adjectives.

5.1 Common Entities

In this setting, we want to see if the principal component of direction between the difference of vectors of entities in the embedding of left side and right side, can capture some kind of bias. We first calculate the two top PCs and then project all of the words existing in the both of the embeddings to this two directions. We except to see that direction is representing the ideology of each political side. In the projected words to the direction, we except to see at the extremely positive values as the left-leaning's news sources popular words and in the extremely negative values the right-leaning's news sources popular words. Figure 3 shows the largest positive values of words projected on PC1 which clearly shows the words associated with foreign and middle east news. On the other hand figure 4 shows the smallest values which show the words that are from right sources on that direction which we can see that are about film industry. We expected to see the two different political views on each side. This result can be because of our dataset and in future works we need to implement the result on a larger news dataset.

5.2 Positive Adjectives

In this setting we are using positive adjectives as attribute words. That makes the PC1 to have the words considered to be good from left-leaning news sources perspective on its largest values and the words considered to be good from right-leaning news sources perspective on its smallest (most negative) values. Because we project all of the words to this direction it is challenging to see the result for important words. We have chosen

pc1	pc2	vocab
6.66467568143413	-0.584121455015584	countries
6.02875078348558	2.17478926374624	nato
5.91888888874225	0.665340538419113	syria
5.90029164633812	1.79951177305829	government
5.84032244131515	0.599063224577371	isis
5.67777334036449	3.36038812319741	russia
5.51402202439615	-0.142060293022209	troops
5.30635709045007	0.828543738527833	military
5.2879929715017	-0.571001738843629	islamic
5.24079811271321	2.50454904765087	assad
5.15770969266652	-0.902921846730435	civilians
5.1515478299605	1.46026525078768	iran
5.10873486587569	-0.459533440911218	refugees
5.06645507213043	1.71516071007683	erdogan
5.04452956538674	-0.776420162797887	turkey
5.01215521405282	0.188513901119291	rebels
4.9852316666643	0.0447200465895026	forces
4.98454474643857	0.541271643886018	governments
4.9579006643415	-0.244222949661243	syrian

Figure 3: The largest values of words projected on PC1

pc1	pc2	vocab
-4.87505951916242	-0.301603200609746	actress
-4.75089203828576	-1.06414435327203	starring
-4.3365879546061	0.455640315418303	kevin
-4.30539314688961	-0.356772484590176	producer
-4.25241829416909	-0.938065017664178	mary
-4.24428529038942	-0.620360945657871	singer
-4.24224910041197	1.58399318387162	bob
-4.09458931996865	0.453536381604634	amy
-4.05654659105184	1.459014583178	chris
-4.0425695985691	0.990768564048192	brian
-4.02557525460958	-0.449106839937901	justin
-3.99819792585883	0.418126438185377	jennifer
-3.98883231346736	1.11302512355873	tim
-3.92747372625897	1.18663989759408	dan
-3.92379004932237	0.17837421086722	billy
-3.91755350632404	1.11856947470779	sarah
-3.90845720722132	-0.0652088710416961	aaron
-3.90479216296341	0.0253567269863497	jerry
-3.88494908709308	-0.205465580946916	actor

Figure 4: The smallest values of words projected on PC1

a few names of well-known people from democratic side and also republican side and see that the democratic persons have larger values (which shows that left news sources have talk about them better than the right news sources) and also some well-known republican people that are leaning along the smaller values (this shows that republican news sources have been using good adjectives about them more than liberal news sources). Table below shows some of the results. In these results we can see that especially right news sources have been talking about Bush very positively, who is a popular person among conservatives and was the president of the United States. Another interesting point is that left-leaning news sources talk more positive than the other side about immigration.

harris	6.15
obama	3.5
immigrated	2.25
biden	2.02
barackobama	1.2
clinton	0.99
bernie	0.90
hillaryclinton	0.18
immigrating	0.091
thornberry	-0.6
cuom	-2.77
pompeo	-3.2
bush	-7

5.3 Negative Adjectives

In this setting, instead of seeing which side talks in favor of who, we are looking at which side talks against who. Table below shows some of the well-known people from liberal and conservative groups. Our result is matching the view of the two political sides of these people because we see that the liberal persons get smaller values (because right news sources use negative adjectives for them more than left news sources) and conservative persons get larger values. The interesting observation in this part is that because our news dataset is from the time that Hillary Clinton was candidate for presidency from the democratic party, we can see that right news sources were using negative adjectives towards her. Also similar to what saw in the previous setting, we can see that republicans talk negatively about immigration.

mccarthy	3.86
cheney	2.42
scalise	2.26
pompeo	1.88
cuomo	1.75
biden	-1.44
immigrant	-5.05
obama	-5.08
sanders	-7.39
hillary	-9.44
clinton	-10.31
democrats	-12
immigrants	-13.97

6 Future Work

In the future work, we are going to use a large list of governers and conressmen and women from each pary to come up with an average of right and left-leaning news sources positive/negative view about them.

Moreover, there is a work that defines Word Embedding Association Test (WEAT)(Brunet et al., 2018) and we can use that definition instead of PCA of differences to run all of our experiments again and compare the results.

Finally, we should use a larger data-set to avoid noises of non-important articles.

References

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*, pages 4349–4357.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2018. Understanding the origins of bias in word embeddings. *arXiv preprint arXiv:1810.03611*.

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

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.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 673–681. ACM.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.