A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings

Chris Sweeney and **Maryam Najafian**
Massachusetts Institute of Technology
Cambridge, MA, USA
{csweeney,najafian}@mit.edu

Abstract

Word embedding models have gained a lot of traction in the Natural Language Processing community, however, they suffer from unintended demographic biases. Most approaches to evaluate these biases rely on vector space based metrics like the Word Embedding Association Test (WEAT). While these approaches offer great geometric insights into unintended biases in the embedding vector space, they fail to offer an interpretable meaning for how the embeddings could cause discrimination in downstream NLP applications. In this work, we present a transparent framework and metric for evaluating discrimination across protected groups with respect to their word embedding bias. Our metric (Relative Negative Sentiment Bias, RNSB) measures fairness in word embeddings via the relative negative sentiment associated with demographic identity terms from various protected groups. We show that our framework and metric enable useful analysis into the bias in word embeddings.

1 Introduction

Word embeddings have established themselves as an integral part of Natural Language Processing (NLP) applications. Unfortunately word embeddings have also introduced unintended biases that could cause downstream NLP systems to be unfair. Recent studies have shown that word embeddings exhibit unintended gender and stereotype biases inherent in the training corpus. Bias can be defined as an unfair expression of prejudice for or against a person, a group, or an idea. Bias is a broad term, which covers a range of problems particularly relevant in natural language systems such as, discriminatory gender bias (Bolukbasi et al., 2016a; Zhao et al., 2017), bias against regionally accented speech (Najafian et al., 2016, 2017), personal or political view bias (Iyyer et al., 2014; Recasens et al., 2013), and many other examples. In our work, we restrict our definition of bias to unequal distributions of negative sentiment among demographic identity terms in word embeddings. One could also look at unequal distributions of positive sentiment, but for this work we restrict ourselves to the negative case.

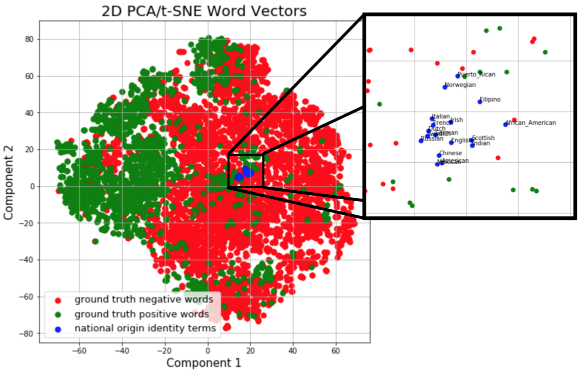


Figure 1: 2-D PCA embeddings for positive/negative sentiment words and a set of national origin identity terms. Geometrically, it is difficult to parse how these embeddings can lead to discrimination.

Sentiment analysis makes up a large portion of current NLP systems. Therefore, preventing negative sentiment from mixing with sensitive attributes (i.e. race, gender, religion) in word embeddings is needed to prevent discrimination in ML models using the embeddings. As studied in (Packer et al., 2018), unintentionally biased word embeddings can have adverse consequences when deployed in applications, such as movie sentiment analyzers or messaging apps.

Negative sentiment can be unfairly entangled in the word embeddings, and detecting this unintended bias is a difficult problem. We need clear signals to evaluate which groups are discriminated against due to the bias in an embedding model. That way we can pinpoint where to mitigate those biases. To demonstrate this need for clear signals

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1662–1667
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

of bias in word embeddings, we look at Figure 1. Figure 1 shows a 2D word embedding projection of positive sentiment (green) and negative sentiment (red) words. It would be unfair for any given demographic identity word vector (blue) to be more semantically related to negative terms than the other identities. However, many identity terms exist closer to negative words than other identity terms in the vector space. This bias may affect a downstream ML model, but the vector space has no absolute interpretable meaning, especially when it comes to whether this embedding model will lead to a unfairly discriminative algorithm. Our framework enables transparent insights into word embedding bias by instead viewing the output of a simple logistic regression algorithm trained on an *unbiased* positive/negative word sentiment dataset initialized with *biased* word vectors. We use this framework to create a clear metric for unintended demographic bias in word embeddings.

2 Prior Work

Researchers have found a variety of ways in which dangerous unintended bias can show up in NLP applications (Blodgett and O'Connor, 2017; Hovy and Spruit, 2016; Tatman, 2017). Mitigating such biases is a difficult problem, and researchers have created many ways to make fairer NLP applications. Much of the focus for mitigating unintended bias in NLP is either targeted at reducing gender stereotypes in text (Bolukbasi et al., 2016b,a; Zhao et al., 2017; Zhang et al., 2018), or inequality of sentiment or toxicity for various protected groups (Caliskan-Islam et al., 2016; Bakarov, 2018; Dixon et al.; Garg et al., 2018; Kiritchenko and Mohammad, 2018).

More specifically, word embeddings has been an area of focus for evaluating unintended bias. (Bolukbasi et al., 2016b) defines a useful metric for identifying gender bias and (Caliskan-Islam et al., 2016) defines a metric called the WEAT score for evaluating unfair correlations with sentiment for various demographics in text.

Unfortunately metrics like these leverage vector space arguments between only two identities at a time like *man* vs *woman* (Bolukbasi et al., 2016a), or *European American names* vs. *African American names* (Caliskan-Islam et al., 2016). Though geometrically intuitive, these tests do not have a direct relation to discrimination in general. Our framework and RNSB metric enable a clear evaluation of discrimination with respect to word embedding bias for a whole class of demographics.

3 Methods

We present our framework for understanding and evaluating unintentional demographic bias in word embeddings. We first describe the flow of our framework. Then, we address which datasets/models were chosen for our approach. Finally, we show how our framework can enable analysis and new metrics like RNSB.

3.1 Framework

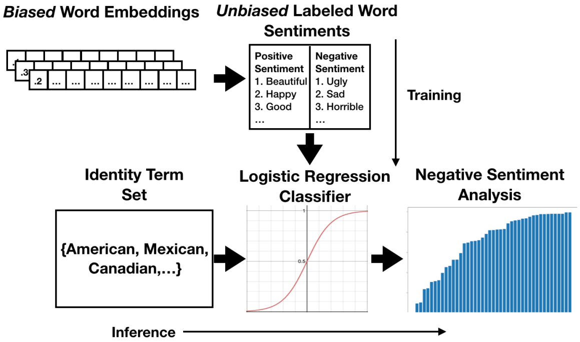


Figure 2: We isolate unintended bias to the word embeddings by training a logistic regression classifier on a *unbiased* positive/negative word sentiment dataset (initialized with the biased word embeddings). We measure word embedding bias by analyzing the predicted probability of negative sentiment for identity terms.

Our framework enables the evaluation of unintended bias in word embeddings through the results of negative sentiment predictions. Our framework has a simple layout. Figure 2 shows the flow of our system. We first use the embedding model we are trying to evaluate to initialize vectors for an unbiased positive/negative word sentiment dataset. Using this dataset, we train a logistic classification algorithm to predict the probability of any word being a negative sentiment word. After training, we take a set of neutral identity terms from a protected group (i.e. national origin) and predict the probability of negative sentiment for each word in the set. Neutral identity terms that are unfairly entangled with negative sentiment in the word embeddings will be classified like their neighboring sentiment words from the sentiment dataset. We leverage this set of negative sentiment probabilities to summarize unintended demographic bias using RNSB.

3.2 Models and Data

We evaluate three pretrained embedding models: GloVe (Pennington et al., 2014), Word2vec (Mikolov et al., 2013) (trained on the large Google News corpus), and ConceptNet (Speer et al., 2017). GloVe and Word2vec embeddings have been shown to contain unintended bias in (Bolukbasi et al., 2016a; Caliskan-Islam et al., 2016). ConceptNet has been shown to be less biased than these models (Speer, 2017) due to the mixture of curated corpora used for training. As part of our pipeline, we also use a labeled positive/negative sentiment training set (Hu and Liu, 2004). This dataset has been shown to be a trustworthy lexicon for negative and positive sentiment words (Pang et al., 2008; Liu, 2012; Wilson et al., 2005). We trust these labels to be unbiased so that we may isolate the unintended biases entering our system to the word embeddings. Finally, we use a simple logistic regression algorithm to predict negative sentiment. Although the choice of ML model can have an impact on fairness for sentiment applications as shown in (Kiritchenko and Mohammad, 2018), we choose a simple ML model to limit the possible unintended biases introduced downstream from our word embeddings.

3.3 Bias Analysis: RNSB

We now present our metric for unintended demographic bias, RNSB. For gold standard labeled positive/negative sentiment words, (x_i, y_i), in training set, S, where x_i is a word vector from a possibly biased word embedding model, we find the minimizer, $f^*(x_i) = \sigma(w^T x_i)$, for the logistic loss, l, and learned weights, w.

$$min_{w \in R^d} \sum_{i=0}^{n} l(y_i, w^T x_i) + \lambda \|w\|^2, \lambda > 0$$

Then for a set, $K = \{k_1, ..., k_t\}$, of t demographic identity word vectors from a particular protected group (i.e. national origin, religion, etc.), we define a set, P, containing the predicted negative sentiment probability via minimizer, f^*, normalized to be one probability mass.

$$P = \left\{ \frac{f^*(k_1)}{\sum_{i=1}^{t} f^*(k_i)}, ..., \frac{f^*(k_t)}{\sum_{i=1}^{t} f^*(k_i)} \right\}$$

Thus, our metric, $RNSB(P)$, is defined as the KL divergence of P from U, where U is the uniform distribution for t elements.

$$RNSB(P) = D_{KL}\left(P \| U\right)$$

We choose our set of neutral identity terms based on the most populous demographics for each protected group. However, due to the simplicity of this method, one can easily adapt it to include identity terms that suit the application in need of analysis.

Since neutral identity terms are inherently not associated with sentiment, it is unfair to have identity term with differing levels of negative sentiment. This type of discrimination can show up in many downstream sentiment analysis applications. Thus, we want no differences between negative sentiment predictions of various identity terms. Mathematically, this can be represented as a uniform distribution of negative sentiment probability for identity terms from a protected group. Our RNSB metric captures the distance, via KL divergence, between the current distribution of negative sentiment and the fair uniform distribution. So the more fair a word embedding model with respect to sentiment bias, the lower the RNSB metric.

4 Results and Discussion

We evaluate our framework and metric on two cases studies: National Origin Discrimination and Religious Discrimination. For each case study, we create a set of the most frequent identity terms from the protected groups in the Wikipedia word corpus and analyze bias with respect to these terms via our framework. First, we compare the RNSB metric for 3 pretrained word embeddings, showing that our metric is consistent with other word embedding analysis like WEAT (Caliskan-Islam et al., 2016). We then show that our framework enables an insightful view into word embedding bias.

4.1 RNSB Metric on Word Embeddings

We vary the word embeddings used in our framework and calculate the RNSB metric for each embedding. The results are displayed in Table 1. For both case studies, the bias is largest in GloVe, as shown by the largest RNSB metric. As mentioned earlier, ConceptNet is a state of the art model that mixes models like GloVe and Word2vec, creating fairer word embeddings. Through the RNSB metric, one can see that the unintended demographic

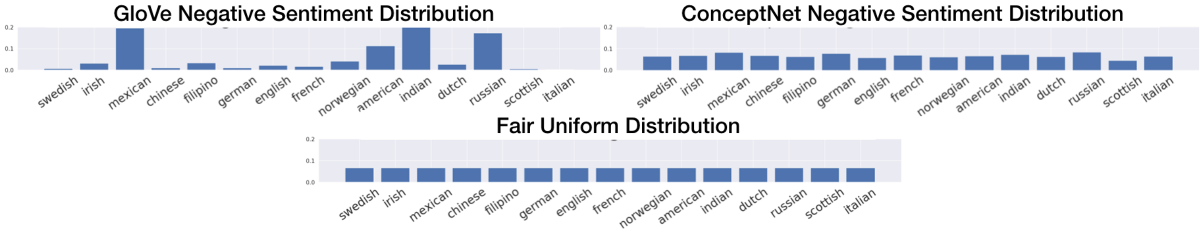


Figure 3: Histograms showing relative negative sentiment probability between national origin identity terms. The top left graph is GloVe, the top right is ConceptNet. The bottom histogram is the uniform distribution of negative sentiment in a perfect fair scenario.

bias of these word embeddings is an order of magnitude lower than GloVe or Word2vec.

Although the RNSB metric is not directly comparable to WEAT scores, these results are still consistent with some of the bias predicted by (Caliskan-Islam et al., 2016). The WEAT score shows that word embeddings like Word2vec and GloVe are biased with respect to national origin because European-American names are more correlated with positive sentiment than African-American names. RNSB captures the same types of biases, but has a clear and larger scope, measuring discrimination with respect to more than two demographics within a protected group.

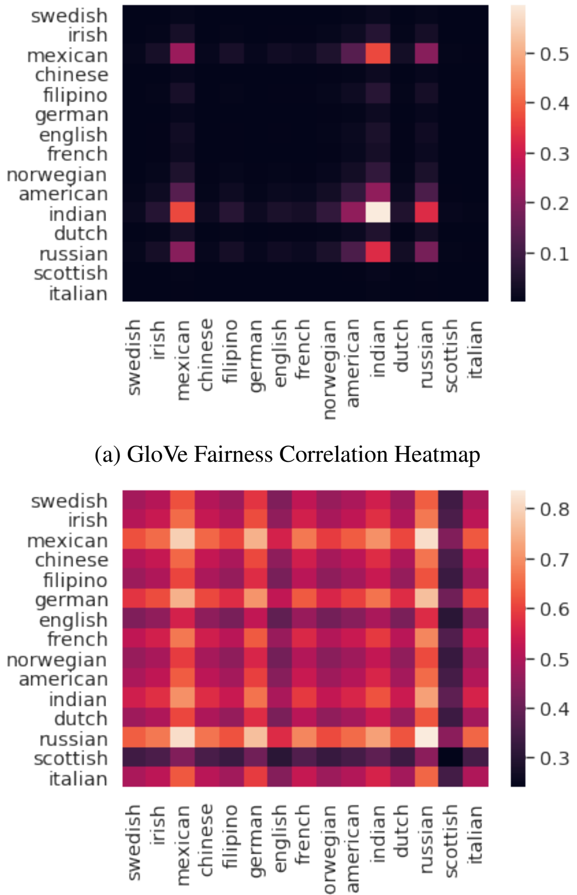
Case Study	GloVe	Word2Vec	ConceptNet
National Origin Identity	0.6225	0.1945	**0.0102**
Religion Identity	0.3692	0.1026	**0.0291**

Table 1: Table showing our RNSB metric for various word embeddings on two case studies. Our metric effectively predicts the unintended demographic bias in the presented word embeddings with respect to negative sentiment.

4.2 Analyzing Unintended Demographic Bias in Word Embeddings

Using the probability distribution of negative sentiment for the identity terms in a protected group, we can gain insights into the relative risks for discrimination between various demographics. Figure 3 shows three histograms. The bottom histogram is the uniform distribution. As described earlier, zero unintended demographic bias with respect to our definition is achieved when all the identity terms within a protected group have equal negative sentiment. The top two histograms show the negative sentiment probability for each identity normalized across all terms to be a probability distribution. The left histogram is computed using the GloVe word embeddings, and the right histogram is computed using the fairer ConceptNet embeddings. One can see that certain demographics have very high negative sentiment predictions, while others have very low predictions. The ConceptNet distribution seems to equalize much of this disparity. This type of analysis is very insightful as it enables one to see which identities are more at risk for discrimination.

A more direct way to measure how certain groups receive similar unfair treatment is to compute a correlation matrix between the vectors containing negative sentiment predictions for each identity term. We compute this matrix for the same two cases: GloVe word embeddings (top) and ConceptNet word embeddings (bottom) shown in Figure 4. The GloVe word embedding correlation matrix contains a lot of dark low correlations between identities, as a lot of identities contain small amounts of negative sentiment. But this visual brings out that certain groups like *Indian*, *Mexican*, and *Russian* have a high correlation, indicating that they could be treated similarly unfairly in a downstream ML algorithm. This is a useful insight that could allow a practitioner to change to embedding training corpora to create fairer models. For the ConceptNet word embeddings, we see a much more colorful heat map, indicating there are higher correlations between more identity terms. This hints that ConceptNet contains less targeted discrimination via negative sentiment. This visual also brings out slight differences in negative sentiment prediction. Identity terms like *Scottish* have lower correlations across the board, manifesting that this identity has slightly less negative sentiment than the rest of the identities. This is important to analyze to get a broader context for how various identities could receive different amounts of discrimination stemming from the word embedding bias.



(a) GloVe Fairness Correlation Heatmap

(b) ConceptNet Fairness Correlation Heatmap

Figure 4: National origin correlation matrix for negative sentiment prediction using GloVe (a) and ConceptNet (b) word embeddings. We can use these figures to analyze how certain groups could be similarly discriminated against via their negative sentiment correlation.

5 Discussion

We showed how our framework can be used in the religious and national origin case studies. In practice, our framework should be used to measure bias among demographics of interest for the NLP application in question. Our RNSB metric is a useful signal a practitioner can use to choose the embedding model with the least amount of risk for discrimination in their application, or even to evaluate what types of unintended biases exists in their training corpora. We used our framework to evaluate unintended bias with respect to sentiment, but there exists many other types of unintended demographic bias to create clear signals for in word embeddings.

6 Conclusion

We presented a transparent framework for evaluating unintended demographic bias in word embeddings. For this work our scope was limited to unfair biases with respect to negative sentiment. In our framework, we train a classifier on an unbiased positive/negative word sentiment dataset initialized with biased word embeddings. This way, we can observe the unfairness in the word embeddings at the ML prediction level. This allows us to observe clearer signals of bias in our metric, Relative Negative Sentiment Bias (RNSB). Previous metrics and analysis into unintended bias in word embeddings rely on vector space arguments for only two demographics at a time, which does not lend itself well to evaluating real world discrimination. Our metric has a direct connection to discrimination and can evaluate any number of demographics in a protected group. Finally, our framework and metric reveal transparent analysis of the unintended bias hidden in word embeddings.

Acknowledgments

This work was made possible in part through support of the United States Agency for International Development. The opinions expressed herein are those of the authors and do not necessarily reflect the views of the United States Agency for International Development or the US Government.

References

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *FATML*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. Quantifying and reducing stereotypes in word embeddings. *ICML*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, pages 4349–4357.

Aylin Caliskan-Islam, Joanna J Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *Science*, pages 1–14.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *AAAI*.

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).

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *ACM*, pages 168–177.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *ACL*, volume 1, pages 1113–1122.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *Proceedings of the 7thJoint Conference on Lexical and Computational Se-mantics(*SEM), New Orleans, USA*.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Maryam Najafian, Wei-Ning Hsu, Ahmed Ali, and James Glass. 2017. Automatic speech recognition of arabic multi-genre broadcast media. In *ASRU*, pages 353–359.

Maryam Najafian, Saeid Safavi, John HL Hansen, and Martin Russell. 2016. Improving speech recognition using limited accent diverse british english training data with deep neural networks. In *MLSP*, pages 1–6.

Ben Packer, Yoni Halpern, Mario Guajardo-Cspedes, and Margaret Mitchell. 2018. Text embedding models contain bias. here's why that matters. Google Developers.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *FTIR*, 2(1–2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *ACL*, volume 1, pages 1650–1659.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451.

Robyn Speer. 2017. Conceptnet numberbatch 17.04: better, less-stereotyped word vectors. ConceptNet.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *EMNLP*, pages 347–354.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. *AIES*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *EMNLP*.