Bias in Word Embeddings

Josh Archer
UC Berkeley School of Information
jearcher@ischool.berkeley.edu

Ben Chu
UC Berkeley School of Information
ben-chu@berkeley.edu

Abstract

Current word embeddings such as Word2Vec and GloVe exhibit bias against racial, gender, and other demographic groups. While there have been architectural improvements with alternative word embeddings such as Word-Node2Vec, we do not see reported advancements directed towards measures of bias against demographic groups. This paper describes our implementation of bias reduction techniques to Word2Vec, GloVe, Word-Node2Vec, and additional alternative word embeddings to reduce systemic bias and decrease undesirable semantic connections between words. We find no proof there exists a trade-off between bias and performance. In fact, all of the debiased word embeddings performed better than the originals. These advancements will prevent erroneous bias leaking into language modeling and keep innate language relationships objective.

1 Introduction

Although there are many incremental improvements in different versions of word embeddings, we see rarely see any mention of bias adjustment or reduction. Our research will evaluate current bias metrics of various word embedding structures with the WEAT statistic, introduced by Caliskan in 2015 and apply bias reduction techniques to alter their level of bias (Caliskan et al., 2017). This paper will address bias associated with Word2Vec, GloVe, and Word-Node2Vec word embeddings with various corpuses including Twitter, Wikipedia, and Gigaword. Through evaluation of previous papers presented at the Association for Computational Linguistics, we will experiment with current bias metrics and reduce their level of bias with new reduction methods.

2 Related Works

Word Embedding Association Test (WEAT), of Caliskan et al. (2017), quantifies the level of bias within a word embedding without the use of a human subject. It is an advancement from the Implicit Association Test (IAT) which measures reaction time from a human pairing two words to evaluate implicit biases in association (Greenwald, 1998). WEAT captures the covariance between the target word and the attribute word embeddings. Through utilization of the WEAT statistic, we will have an universal measurement of evaluation for biases within word embeddings and observe how our bias reduction techniques may be able to improve advanced word embeddings.

Word2Vec, a natural language processing algorithm, utilizes a neural network model to learn word associations from text-based data and creates a vector representation for each word in a dataset. Global Vectors (GloVe), the unsupervised machine learning algorithm using a count-based model, trains on global word-word co-occurrence statistics and represents word relationships with linear substructures of the word vector space. While both have good performance depending on the dataset, GloVe adopts global statistics (word co-occurences) instead of relying on local contexts such as Word2Vec. With more recent developments, Word-Node2Vec has improved performance on both of these word embeddings through modeling relationships with graph-based word embeddings that capture both local and non-local occurrences in a dataset (Sen et al., 2019). While these embeddings are successful at word-pair similarity prediction, word-analogy, or concept categorization, it is still not standard to acknowledge and evaluate bias. Word-Node2Vec was one of many new pre-trained word embeddings introduced in 2019 that did not address bias, which we will address in

this paper (Gupta et al., 2019; vor der Brück and Pouly, 2019; Amiri and Mohtarami, 2019).

Previous research on debiasing methods have targeted racial bias with multiclass debiasing and gender bias in word embeddings(Manzini et al., 2019). Following developments in debiasing binary labeling, multiclass debiasing utilizes hard or soft debiasing techniques to remove subspace components from word embeddings (Karve et al., 2019; Bolukbasi et al., 2016) . These methods will be considered when debiasing advanced word embeddings such as Word-Node2Vec.

3 Experiment Setup

Our research encompasses designing baseline observations of each word embedding, implementing debiasing techniques, and testing the debiased word embeddings against their baseline observations. We will visualize the differential and impact of the debiasing techniques to cross-reference how they can improve bias reduction and objective relationships with race, gender, and religion.

We will utilize the WEAT statistic as a universal metric to quantify current bias within Word2Vec, GloVe, and Word-Node2Vec embeddings. WEAT measure the different distributions of the associations of target words (i.e. 'man' and 'woman') with attribute words (i.e. 'lawyer' or 'homemaker'). We will report the WEAT *effect size*, which is a normalized measure. The null hypothesis is that the distributions are not very different, yielding a WEAT effect size of 0. Thus, throughout this paper, a larger WEAT score signifies a larger amount of bias.

To evaluate performance, The debiased word embeddings will be tested with their cosine similarity for various classes and categories of words that can retain bias in modeling. Then we will compare the debiased embeddings performace with the original embeddings.

The WEAT test statistic measures the differential association of the two sets of target words with the attribute. The "effect size" is a normalized measure of how separated the two distributions are.

Saket Karve, Lyle Ungar, and João Sedoc introduced the method of *Conceptor Debiasing* in 2019 as a multi-class debiasing method (Karve et al., 2019). They introduce conceptor negation as a soft damping of the principal components of a subset of words being debiased. They define the bias subspace with lists of target words referencing a protected class (i.e. gendered, racial, or religious terms). To utilize this method, first we compute the conceptor that represents the space of maximum bias, then use the complement of this subspace to debias the word embeddings. For a well-written and detailed description of the process, please review the original paper Karve et al. (2019).

4 Experimentation and Evaluation

4.1 Evaluating Bias using WEAT

From our initial observations of our word embeddings, we found significant baseline progress with our debiasing techniques. With Word2Vec we found an initial WEAT statistic of 1.15872 and a value of 0.65675 after debiasing the Word2Vec. When applying our methods to GloVe on a Twitter Data corpus, we uncovered a baseline WEAT statistic of 0.11733 and a debiased GloVe statistic of -0.14090, which created negative bias in the embedding. However, with GloVe applied on Wikipedia 2014 and Gigaword 5 corpuses, we found an initial WEAT statistic of 1.15968 and a value of 1.56147 after debiasing the GloVe. This result is surprising as it suggests the conceptor debiasing technique actually *increased* bias in the embedding. This was the only embedding in which bias increased. In the future we would like to inspect this further.

The newly proposed Word-Node2Vec actually had the most bias to begin with– our baseline test resulted in a metric of 1.18091 and a debiased Word-Node2Vec outcome of 1.01030. Since this is one of many newly proposed word embeddings meant to replace Word2Vec and GloVe, we believe it is important these embeddings are evaluated on bias. While Word-Node2Vec purports to have more accuracy than older methods, there is no discussion of its relatively large bias.

Word Embedding	WEAT Statistic
Word2Vec	1.15872
Word2Vec (Debiased)	0.65675
GloVe (T)	0.11733
GloVe (T Debiased)	-0.14090
GloVe (WG)	1.15968
GloVe (WG Debiased)	-0.14090
Word-Node2Vec	1.18091
Word-Node2Vec (Debiased)	1.01030

Table 1: Results of WEAT Tests
T: Twitter Data, WG: Wikipedia 2014 + Gigaword 5

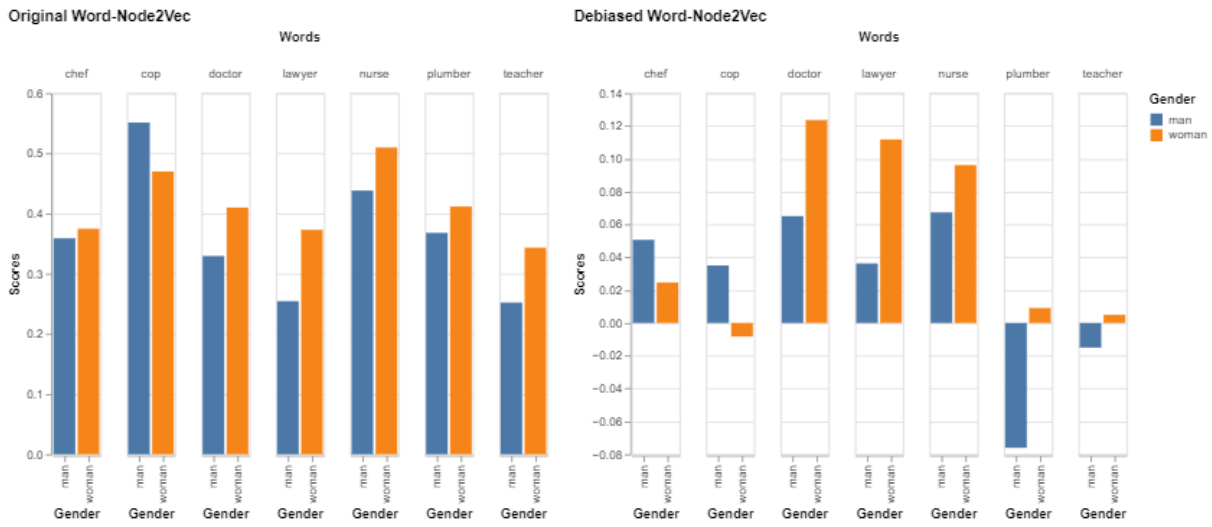

Figure 1: Gender Identifiers with Occupational Positions

The Conceptor debiasing technique reduces the relatedness of target words such as 'man' and 'woman' to attribute words such as occupations. Figure 1 illustrates the cosine-similarity between target and attribute words before and after debiasing Word-Node2Vec. It is clear that the cosine similarities are much smaller in the debiased version, however this does not reduce performance as explained in the following subsection.

Figure 2 displays cosine similarites between race identifiers and positive and negative sentiment for Word-Node2Vec. Once again, the debiased have much smaller magnitudes of cosine similarity. In some cases, the ranking for each race is the same. For example, hispanic is closest to 'pleasant' among the biased and debiased version.

4.2 Evaluating Performance with Word Similarity

To evaluate and compare the performance of each word embedding, we used SimLex-999 and WordSim-353. These evaluate how well each word embedding captures word similarity and relatedness. SimLex-999 evaluates the word similarity whereas WordSim-353 evaluates the word relatedness. While *clothes* and *closet* are related, they are not similar in that they are made from different materials. (Hill et al., 2014; Finkelstein et al., 2001).

Figure 3 plots each word embedding's WEAT effect size up against either the Word Relatedness score or Word Similarity score. In each case, the debiased version performed better than the original versions. However, it must be noted taht the 'debiased' version of GloVe (Wikipedia 2014 + Gigaword 5) actually has more bias (and better performance) than the original embedding.

Given we find no evidence of a trade-off between bias and performance, we find it crucial that evaluating and debiasing word embeddings become standard in any future literature or word embeddings.

4.3 Replication

This project is entirely reproducable. Each notebook to debias is found in Github https://github.com/jearcher/bias-in-word-embeddings/tree/colab_style along with notebooks to evaluate word embeddings. Each notebook includes a badge to open in Google Colab. Due to the size of the embeddings and process of debiasing, these notebooks perform better in Google Colab rather than in a jupyter notebook environment.

All of the embeddings, both original and debiased, are hosted publicly via Google Cloud Storage. For example the debiased Word-Node2Vec can be found here: https://storage.googleapis.com/word-embedding-bias/Word-Node2Vec_2019/Word-node2Vec_debias.txt

Each notebook in Colab will import directly from our Github and the google cloud, so no local downloads are necessary to reporduce this project.

Please contact either of the authors of this paper for any questions.

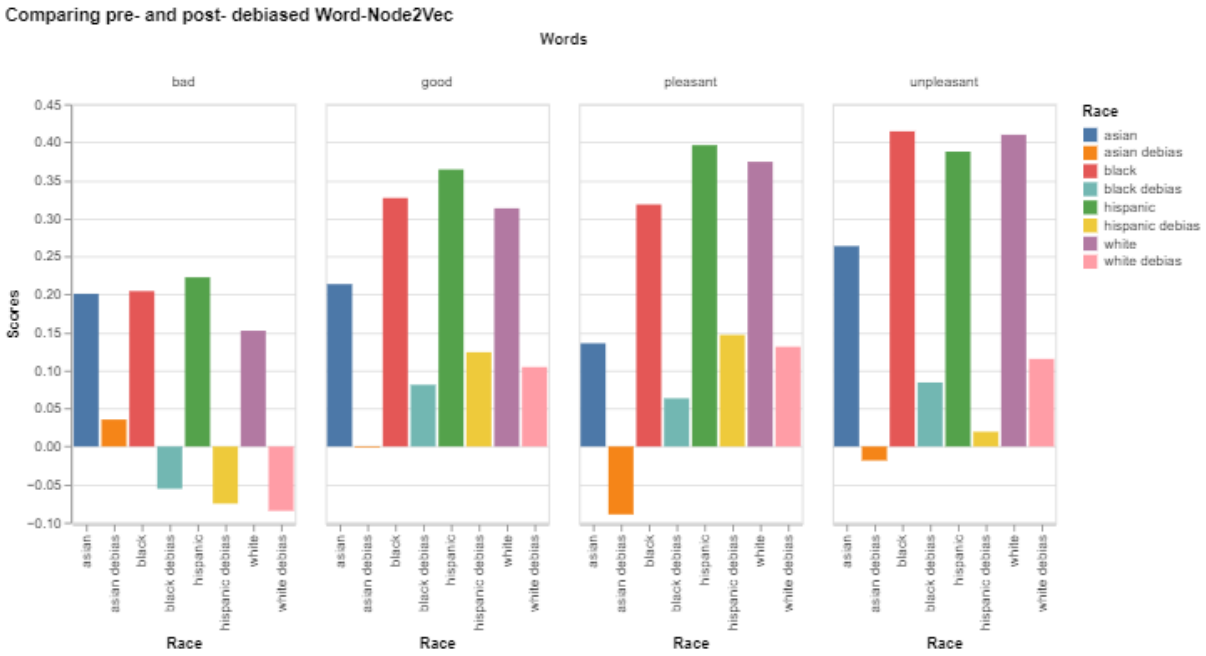

Figure 2: Race Identifiers with Positive and Negative Sentiment

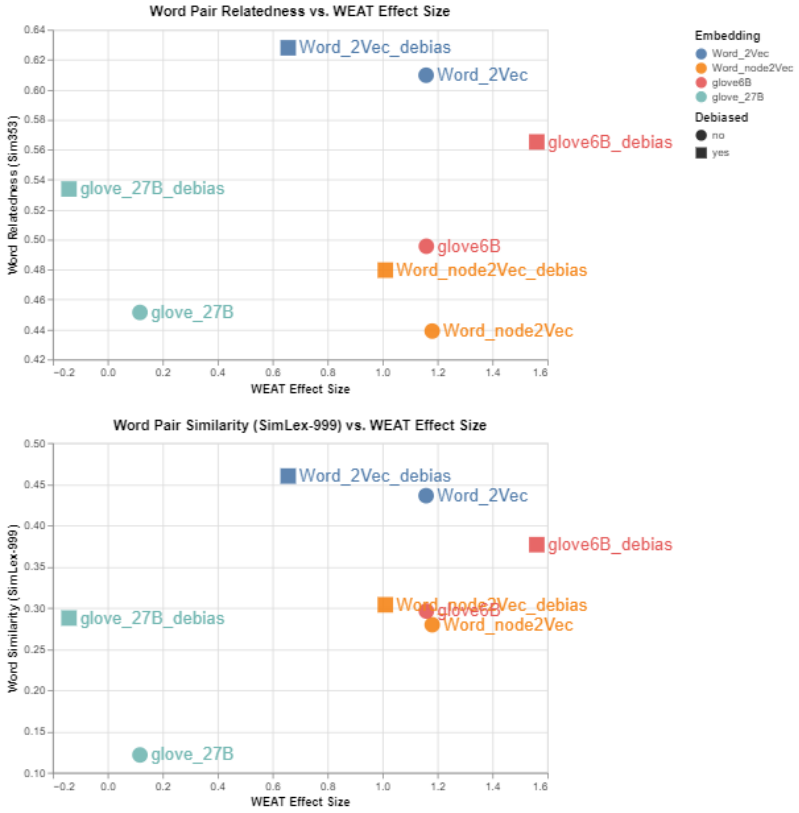

Figure 3: Word Pair Relatedness and Similarity vs. WEAT Effect Size

5 Conclusion

All factors considered, our research study discovered the effects of implementing debiasing techniques for Word2Vec, GloVe, and Word-Node2Vec on various corpuses. We found that applying bias-reduction techniques can heavily impact the level of bias displayed in word embeddings, through visualizing the relationship between words related to race, gender, and religion with no negative impact on performance. In fact, every debiased embedding

outperformed its original version. The conceptor debias technique reduced bias for 3 of the 4 embeddings we studied. We hope to continue our research to investigate why the conceptor method increased the bias in GloVe (Wikipedia 2014 + Gigaword 5). Race-identifying words related to positive or negative descriptions had a drastic decrease in bias for both sentiments. Gender-related words tied with occupational positions had a severe reduction in bias, leaning to more neutral judgement in connection. Finally, with religious classifiers and sentiment labeling we saw a heavy inclination to remain objective and eradicate strong connections to negative labeling.

After evaluating the cosine similarity between bias-conductive words, we were able to assess and verify the improvements of our word embeddings. The conceptor debiasing methods have been able to consistently remove lingering bias with word relationships in our vector space while improving general performance. Overall, designing more unbiased models for word embeddings will allow us to better objectively model the world around us.

Acknowledgments

This research is conducted by Joshua Archer and Ben Chu, current graduate students at the University of California Berkeley School of Information. We would like to acknowledge the wonderful support provided by the faculty and professors for the course W266 Natural Language Processing with Deep Learning, in particular Peter Grabowski who's help was critical for this project. We would also like to acknowledge the many previous researchers in natural language processing that have made these developments possible.

References

Hadi Amiri and Mitra Mohtarami. 2019. Vector of locally aggregated embeddings for text representation. 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 1408–1414, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Tim vor der Brück and Marc Pouly. 2019. Text similarity estimation based on word embeddings and matrix norms for targeted marketing. 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 1827–1836, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. volume 20, pages 406–414.

McGhee D. E. Schwartz J. L. K. Greenwald, A. G. 1998. Measuring individual differences in implicit cognition: The implicit association test.

Prakhar Gupta, Matteo Pagliardini, and Martin Jaggi. 2019. Better word embeddings by disentangling contextual n-gram information. 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 933–939, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48, Florence, Italy. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Procheta Sen, Debasis Ganguly, and Gareth Jones. 2019. Word-Node2Vec: Improving word embedding with document-level non-local word co-occurrences. 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 1041–1051, Minneapolis, Minnesota. Association for Computational Linguistics.

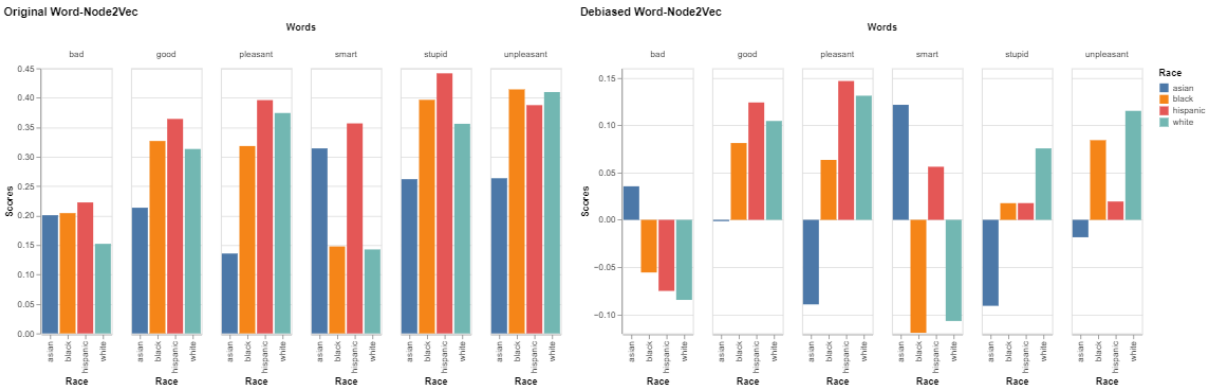

Figure 4: Race Identifiers with Positive and Negative Sentiment

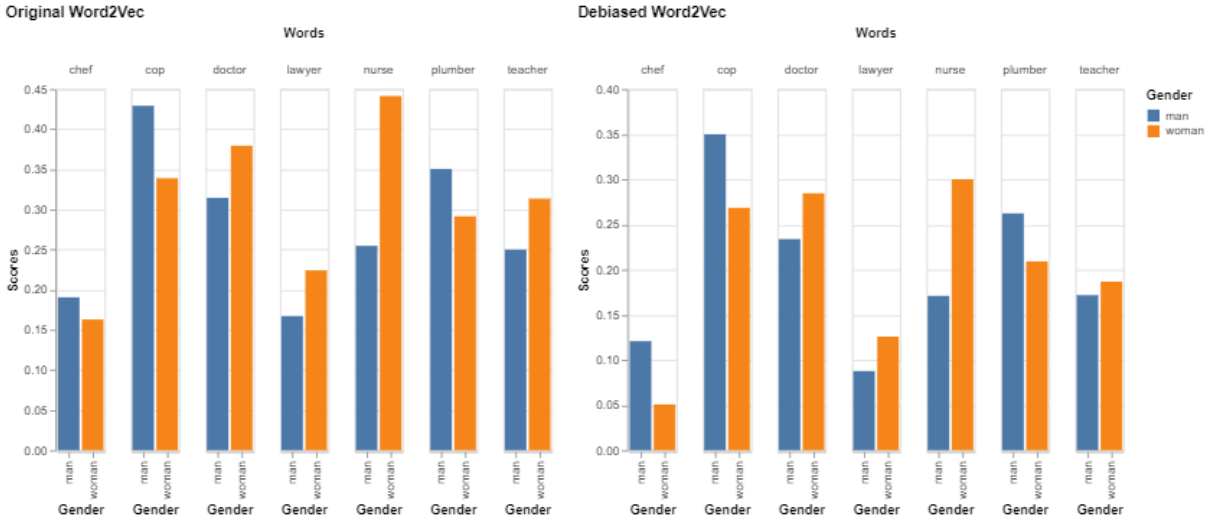

Figure 5: Word2Vec Gender Identifiers with Occupational Positions

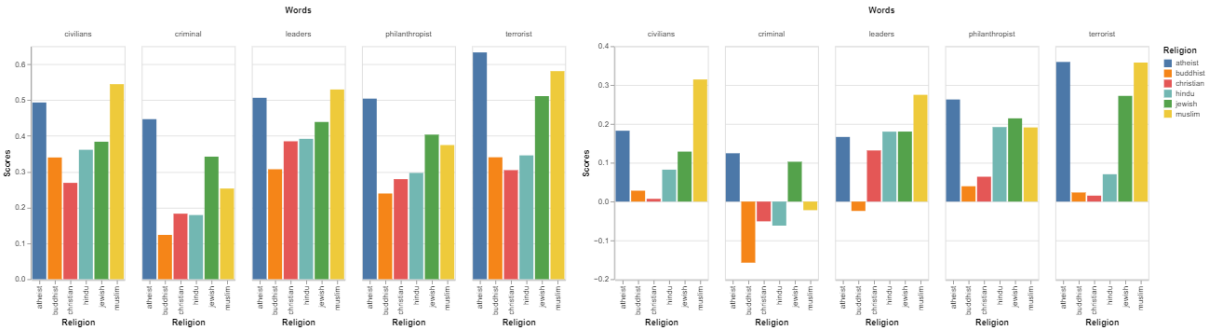

Figure 6: Religious Identifiers with Sentiment Labeling