Using Dimensions of Cultural Meaning to Score Prejudice and Debias Word Embeddings

Bhargav Srinivasa Desikan

April 2019

1 Introduction

Using Natural Language Processing techniques to analyse text has been around for a while (language translation using machines has been around since the 1960s!), but it is *now* where it has permeated almost every aspect of our digital living. Google's search algorithms, e-mail spam detection, advertisements, auto text completion, chatbots, and recommendations all tend to use textual information.

This suggests that there are certain normative concerns which should be involved when designing applications which involve these techniques. The idea that text contains inherent biases has been previously well documented Holmes and Meyerhoff [2008], so it is only natural that these biases carry on to algorithms which use text.

Indeed, ethical concerns in Machine Learning have been discussed before (Anderson and Anderson [2011], Bostrom and Yudkowsky [2014]), and particularly in text as well (Bolukbasi et al. [2016b], Caliskan et al. [2017]). Prejudice happens to be a primary source of injustice, as detailed by Fricker [2007] in her study of epistemic injustice. Dealing with prejudice would directly create justice from both an ethical and epistemic perspective. I hope to further the exploration in these fields, and particularly in scoring prejudiced articles on the extent of prejudice in them, and in removing bias from word embedding models.

2 Word Embeddings

Word embedding models allow us to capture semantic meaning in text. What this means is that these models also capture the inherent biases present in text. A popular word embedding model is Word2Vec Mikolov et al. [2013], where they demonstrate the efficacy of such models. But it is precisely these vector space models which end up capturing the most biases in text, as Bolukbasi et al. [2016b] and Caliskan et al. [2017] have shown. Word embeddings also tend to be used *before* the actual classification or regression pipelines are involved, so it is especially concerning that these vectors are not biased.

3 Scoring Prejudice and Debiasing

There have been multiple attempts to debias and identify prejudice in textual analysis (Tulkens et al. [2016], Park et al. [2018], Bolukbasi et al. [2016b], Zhao et al. [2017], Bolukbasi et al. [2016a], Hasanuzzaman et al. [2017]). They approach the problem from a variety of methods - counting words, changing the word vectors themselves, or changing the data used to train the vectors.

I will be using the idea of dimensions of cultural meaning (elaborated in Kozlowski et al. [2018]) to both debias the word embeddings, as well to score documents on the extent of bias present in them. Bolukbasi et al. [2016b] has shown how when we project words onto, for example, the gender dimension, it is possible to derive the extent to which the word has been influenced by that particular dimension. It is possible to develop techniques based on these dimensions to change the nature of the embeddings, and to score (and further classify) them. I will be looking to develop a python package which automates this process, and implements such techniques.

References

Michael Anderson and Susan Leigh Anderson. *Machine ethics*. Cambridge University Press, 2011.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*, 2016a.

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

Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, 1:316–334, 2014.

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

Miranda Fricker. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press, 2007.

Mohammed Hasanuzzaman, Gaël Dias, and Andy Way. Demographic word embeddings for racism detection on twitter. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 926–936, 2017.

Janet Holmes and Miriam Meyerhoff. *The handbook of language and gender*, volume 25. John Wiley & Sons, 2008.

Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*, 2018.

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

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*, 2016.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.