Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings

Thomas Manzini[†]*, **Yao Chong Lim**[‡]*, **Yulia Tsvetkov**[‡], **Alan W Black**[‡]
Microsoft AI Development Acceleration Program[†], Carnegie Mellon University[‡]
Thomas.Manzini@microsoft.com, {yaochonl,ytsvetko,awb}@cs.cmu.edu

Abstract

Online texts—across genres, registers, domains, and styles—are riddled with human stereotypes, expressed in overt or subtle ways. Word embeddings, trained on these texts, perpetuate and amplify these stereotypes, and propagate biases to machine learning models that use word embeddings as features. In this work, we propose a method to debias word embeddings in multiclass settings such as race and religion, extending the work of (Bolukbasi et al., 2016) from the binary setting, such as binary gender. Next, we propose a novel methodology for the evaluation of multiclass debiasing. We demonstrate that our multiclass debiasing is robust and maintains the efficacy in standard NLP tasks.

1 Introduction

In addition to possessing informative features useful for a variety of NLP tasks, word embeddings reflect and propagate social biases present in training corpora (Caliskan et al., 2017; Garg et al., 2018). Machine learning systems that use embeddings can further amplify biases (Barocas and Selbst, 2016; Zhao et al., 2017), discriminating against users, particularly those from disadvantaged social groups.

(Bolukbasi et al., 2016) introduced a method to *debias* embeddings by removing components that lie in stereotype-related embedding subspaces. They demonstrate the effectiveness of the approach by removing gender bias from word2vec embeddings (Mikolov et al., 2013), preserving the utility of embeddings and potentially alleviating biases in downstream tasks. However, this method was only for *binary* labels (e.g., male/female), whereas most real-world demographic attributes, including gender, race, religion, are not binary but continuous or categorical, with more than two categories.

Gender Biased Analogies	
man → doctor	woman → nurse
woman → receptionist	man → supervisor
woman → secretary	man → principal
Racially Biased Analogies	
black → criminal	caucasian → police
asian → doctor	caucasian → dad
caucasian → leader	black → led
Religiously Biased Analogies	
muslim → terrorist	christian → civilians
jewish → philanthropist	christian → stooge
christian → unemployed	jewish → pensioners

Table 1: Examples of gender, racial, and religious biases in analogies generated from word embeddings trained on the Reddit data from users from the USA.

In this work, we show a generalization of Bolukbasi et al.'s (2016) which enables *multiclass* debiasing, while preserving utility of embeddings (§3). We train word2vec embeddings using the Reddit L2 corpus (Rabinovich et al., 2018) and apply multiclass debiasing using lexicons from studies on bias in NLP and social science (§4.2). We introduce a novel metric for evaluation of bias in collections of word embeddings (§5). Finally, we validate that the utility of debiased embeddings in the tasks of part-of-speech (POS) tagging, named entity recognition (NER), and POS chunking is on par with off-the-shelf embeddings.

2 Background

As defined by (Bolukbasi et al., 2016), debiasing word embeddings in a binary setting requires identifying the bias subspace of the embeddings. Components lying in that subspace are then removed

* Equal contributions
† Work done while at CMU and The Microsoft AI Development Acceleration Program