What's in a Name?
Reducing Bias in Bios without Access to Protected Attributes

Alexey Romanov[1], **Maria De-Arteaga**[2], **Hanna Wallach**[3],
Jennifer Chayes[3], **Christian Borgs**[3], **Alexandra Chouldechova**[2],
Sahin Geyik[4], **Krishnaram Kenthapadi**[4], **Anna Rumshisky**[1], **Adam Tauman Kalai**[3]
[1]University of Massachusetts Lowell
{aromanov,arum}@cs.uml.edu
[2]Carnegie Mellon University
mdeartea@andrew.cmu.edu,achould@cmu.edu
[3]Microsoft Research
{wallach,jchayes,Christian.Borgs,Adam.Kalai}@microsoft.com
[4]LinkedIn
{sgeyik,kkenthapadi}@linkedin.com

Abstract

There is a growing body of work that proposes methods for mitigating bias in machine learning systems. These methods typically rely on access to protected attributes such as race, gender, or age. However, this raises two significant challenges: (1) protected attributes may not be available or it may not be legal to use them, and (2) it is often desirable to simultaneously consider multiple protected attributes, as well as their intersections. In the context of mitigating bias in occupation classification, we propose a method for discouraging correlation between the predicted probability of an individual's true occupation and a word embedding of their name. This method leverages the societal biases that are encoded in word embeddings, eliminating the need for access to protected attributes. Crucially, it only requires access to individuals' names at training time and not at deployment time. We evaluate two variations of our proposed method using a large-scale dataset of online biographies. We find that both variations simultaneously reduce race and gender biases, with almost no reduction in the classifier's overall true positive rate.

1 Introduction

In recent years, the performance of machine learning systems has improved substantially, leading to the widespread use of machine learning in many domains, including high-stakes domains such as healthcare, employment, and criminal justice (Chalfin et al., 2016; Miotto et al., 2017; Chouldechova, 2017). This increased prevalence has led many people to ask the question, "accurate, but for whom?" (Chouldechova and G'Sell, 2017).

When the performance of a machine learning system differs substantially for different groups of people, a number of concerns arise (Barocas and Selbst, 2016; Kim, 2016). First and foremost, there is a risk that the deployment of such a method may harm already marginalized groups and widen existing inequalities. Recent work highlights this concern in the context of online recruiting and automated hiring (De-Arteaga et al., 2019). When predicting an individual's occupation from their online biography, the authors show that if occupation-specific gender gaps in true positive rates are correlated with existing gender imbalances in those occupations, then those imbalances will be compounded over time—a phenomenon sometimes referred to as the "leaky pipeline." Second, the correlations that lead to performance differences between groups are often irrelevant. For example, while an occupation classifier should predict a higher probability of software engineer if an individual's biography mentions coding experience, there is no good reason for it to predict a lower probability of software engineer if the biography also mentions softball.

Prompted by such concerns about bias in machine learning systems, there is a growing

"What's in a name? That which we call a rose by any other name would smell as sweet." – *William Shakespeare, Romeo and Juliet.*

Accepted at NAACL 2019.

body of work on fairness in machine learning. Some of the foundational papers in this area highlighted the limitations of trying to mitigate bias using methods that are "unaware" of protected attributes such as race, gender, or age (e.g., Dwork et al., 2012). As a result, subsequent work has primarily focused on introducing fairness constraints, defined in terms of protected attributes, that reduce incentives to rely on undesirable correlations (e.g., Hardt et al., 2016; Zhang et al., 2018). This approach is particularly useful if similar performance can be achieved by slightly different means—i.e., fairness constraints may aid in model selection if there are many near-optima.

In practice, though, any approach that relies on protected attributes may stand at odds with anti-discrimination law, which limits the use of protected attributes in domains such as employment and education, even for the purpose of mitigating bias. And, in other domains, protected attributes are often not available (Holstein et al., 2019). Moreover, even when they are, it is usually desirable to simultaneously consider multiple protected attributes, as well as their intersections. For example, Buolamwini (2017) showed that commercial gender classifiers have higher error rates for women with darker skin tones than for either women or people with darker skin tones overall.

We propose a method for reducing bias in machine learning classifiers without relying on protected attributes. In the context of occupation classification, this method discourages a classifier from learning a correlation between the predicted probability of an individual's occupation and a word embedding of their name. Intuitively, the probability of an individual's occupation should not depend on their name—nor on any protected attributes that may be inferred from it. We present two variations of our method—i.e., two loss functions that enforce this constraint—and show that they simultaneously reduce both race and gender biases with little reduction in classifier accuracy. Although we are motivated by the need to mitigate bias in online recruiting and automated hiring, our method can be applied in any domain where individuals' names are available at training time.

Instead of relying on protected attributes, our method leverages the societal biases that are encoded in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017). In particular, we build on the work of Swinger et al. (2019), which showed that word embeddings of names typically reflect the societal biases that are associated with those names, including race, gender, and age biases, as well encoding information about other factors that influence naming practices such as nationality and religion. By using word embeddings of names as a tool for mitigating bias, our method is conceptually simple and empirically powerful. Much like the "proxy fairness" approach of Gupta et al. (2018), it is applicable when protected attributes are not available; however, it additionally eliminates the need to specify which biases are to be mitigated, and allows simultaneous mitigation of multiple biases, including those that relate to group intersections. Moreover, our method only requires access to proxy information (i.e., names) at training time and not at deployment time, which avoids disparate treatment concerns and extends fairness gains to individuals with ambiguous names. For example, a method that explicitly or implicitly infers protected attributes from names at deployment time may fail to correctly infer that an individual named Alex is female and, in turn, fail to mitigate gender bias for her. Methodologically, our work is also similar to that of Zafar et al. (2017), which promotes fairness by requiring that the covariance between a protected attribute and a data point's distance from a classifier's decision boundary is smaller than some constant. However, unlike our method, it requires access to protected attributes, and does not facilitate simultaneous mitigation of multiple biases.

We present our method in Section 2. In section 3, we describe our evaluation, followed by results in Section 4 and conclusions in Section 5.

2 Method

Our method discourages an occupation classifier from learning a correlation between the predicted probability of an individual's occupation and a word embedding of their name. In this section, we present two variations of our method—i.e., two penalties that can be added to an arbitrary loss function and used when training any classifier.

We assume that each data point corresponds to an individual, with a label indicating that individual's occupation. We also assume access to the names of the individuals represented in the training set. The first variation, which we call Cluster Constrained Loss (CluCL), uses k-means to cluster word embeddings of the