Mitigating Demographic Biases in Social Media-Based Recommender Systems

Rashidul Islam, Kamrun Naher Keya, Shimei Pan, James Foulds*
{islam.rashidul,kkeya1,shimei,jfoulds}@umbc.edu
Department of Information Systems
University of Maryland, Baltimore County

ABSTRACT

As a growing proportion of our daily human interactions are digitized and subjected to algorithmic decision-making on social media platforms, it has become increasingly important to ensure that these algorithms behave in a fair manner. In this work, we study fairness in collaborative-filtering recommender systems trained on social media data. We empirically demonstrate the prevalence of demographic bias in these systems for a large Facebook dataset, both in terms of encoding harmful stereotypes, and in the impact on consequential decisions such as recommending academic concentrations to the users. We then develop a simple technique to mitigate bias in social media-based recommender systems, and show that this results in fairer behavior with only a minor loss in accuracy.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**;
• **Applied computing** → *Law, social and behavioral sciences.*

KEYWORDS

fairness in machine learning, social media analytics, recommender systems

ACM Reference Format:
Rashidul Islam, Kamrun Naher Keya, Shimei Pan, James Foulds. 2019. Mitigating Demographic Biases in Social Media-Based Recommender Systems. In *KDD '19: Social Impact Track, August 04–08, 2019, Anchorage, Alaska.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

There is increasing awareness that machine learning algorithms can impact people in unfair ways with legal or ethical consequences when used to automate decisions in areas such as insurance, credit scoring, loan assessment, hiring, and crime prediction [2, 3]. As social media platforms are a major contributor to the number of automated data-driven decisions that we as individuals are subjected to, it is clear that such fairness issues in social media can potentially also cause substantial societal harm. Recommender systems [1, 12] are the workhorse method for a variety of machine learning tasks for social media data, e.g. suggesting advertisements, products, friends, pages, and potentially, consequential suggestions such as romantic partners or even career paths. In this paper, we therefore investigate the "unfairness" of social media-based recommender systems. Our fundamental research questions are: *Do different demographic groups experience mistreatment in the form of bias from recommender systems trained on data from online social networks? If so, how can we quantify and mitigate these biases?*

Users	Likes	Pairs	Concentrations
16,619	143,303	3,501,203	1,380

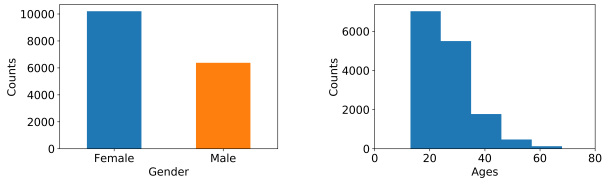

Figure 1: Summary of the Facebook dataset.

2 BACKGROUND AND RELATED WORK

The recommender system research community has begun to consider issues of fairness in recommendation. Fair recommendation systems have been proposed, e.g. penalizing disparate distributions of prediction error [13], and making recommended items independent from protected attributes such as gender, race, and age [10]. [5, 6] taxonomize fairness objectives and methods based on which set of stakeholders in the recommender system are being considered, since it may be meaningful to consider fairness among many groups in recommender systems. Unlike previous work, we specifically study fairness for recommender systems trained on social media data, recommending pages to "like," and academic concentrations.

3 DATA AND EXPERIMENTAL METHODS

The Facebook data we analyzed was collected from 2007 to 2012 as part of the myPersonality project [11], a popular Facebook app. It offered psychometric tests to its users and returned feedback on their performance [8]. We consider several kinds of information of Facebook users such as demographic profile, eg. gender and age, their academic concentrations, and user-like pairs. The "likes" and academic concentrations are the items we aim to predict in this study to evaluate recommender systems on the basis of fairness. See Figure 1 for a summary of the dataset after pre-processing.

Our starting point is a neural network collaborative filtering model, summarized in Figure 2, for predicting the pages a user

*The first two authors contributed equally to this work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '19:, August 04–08, 2019, Anchorage, Alaska
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

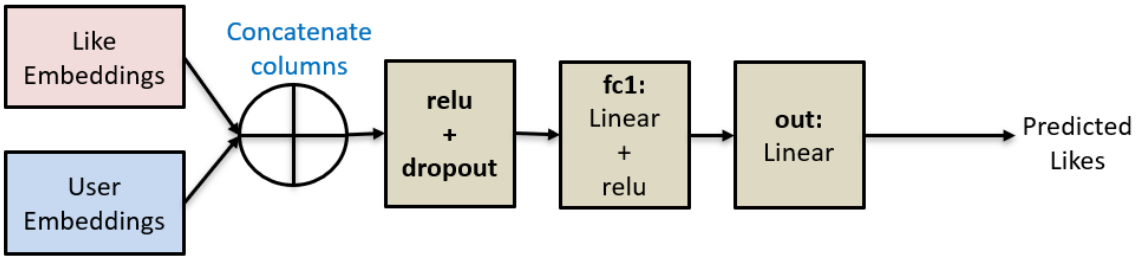

Figure 2: Schematic diagram of neural network for collaborative filtering.

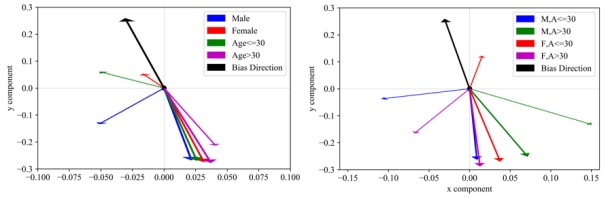

Figure 3: PCA projection of Debiased Model-2. Dotted vectors represent vectors from typical CF.

	Overall Mean Squared Error (MSE)		
	Typical CF	Debiased Model 1	Debiased Model 2
	0.0320	0.0322	0.0329

	Mean Absolute Error (MAE)		
Groups/Sub-groups	Typical CF	Debiased Model 1	Debiased Model 2
M	0.146	0.146	0.148
F	0.147	0.147	0.149
A<=30	0.146	0.146	0.148
A>30	0.149	0.149	0.151
M, A<=30	0.145	0.145	0.147
M, A>30	0.151	0.151	0.153
F, A<=30	0.146	0.147	0.149
F, A>30	0.148	0.148	0.150

Figure 4: Performance comparison of different model in terms of like predictions.

"likes," encoded as 1, or does not "like," encoded as 0, which are analogous to ratings. The embedding size for user and like embeddings was set to 100. The user and like embeddings are combined by concatenation, rather than a dot product, as we found that this improved performance. One hidden layer with 10 linear units is used along with dropout regularization of probability 0.1 followed by a linear output layer. Finally, we train the model by optimizing MSE loss using *Adam* in batch mode with a learning rate of 0.01.

3.1 Debiasing Methods

Our debiasing approach adapts very recent work on attenuating bias in word vectors [7] to the problem of collaborative filtering. [7] propose to debias word vectors by a linear projection of all words w orthogonally to the bias vector v_B as follows:

$$w' = w - (w \cdot v_B)v_B \ . \tag{1}$$

The main challenge here is to find the proper bias direction, which is application dependent, and differs from the word embedding case for collaborative filtering. First, we obtain bias directions for each protected group. For women, we obtain bias vectors as

$$v_F = \frac{f_1 + f_2 + \ldots}{||f_1 + f_2 + \ldots||} \tag{2}$$

where, $f_1, f_2, \ldots$ are vectors for particular female users, and similarly for men v_M, and so on. The male-female bias direction is $v_F - v_M$. Following [9] and [4], we further aim to protect the *intersections* of the groups. We obtain *intersectional bias* directions by adding the bias directions per protected attribute, e.g. for *gender* × *age*:

$$v_B = \frac{(v_F - v_M) + (v_{A\leq 30} - v_{A>30})}{||(v_F - v_M) + (v_{A\leq 30} - v_{A>30})||} \ . \tag{3}$$

As well as Equation 1, we consider a hard debiasing approach, *Debiased Model 2*, by neglecting the sign of the dot product as follows so that all the vectors of different groups remain closer:

$$w' = w - (|w \cdot v_B|)v_B \ . \tag{4}$$

3.2 Experimental Results

For visualization, we apply PCA on the user embeddings (Figure 3). We found that while Debiased Model 1 (Equation 1, not shown for space) shifts protected groups and intersectional groups away from the bias direction, Debiased Model 2 further shifts all the groups and intersectional groups to approximately a 180-degree angle from the bias direction under the PCA projection. Thus, all groups end up having a similar direction, which more strongly ensures fairness.

To show the impact of demographic biases and their mitigation, we study the use of the embeddings to suggest academic concentrations, using cosine distance based k-nearest neighbors with $k = 10$. For each demographic group, we generate recommendation of academic concentrations from all the model based on the 10 similar users. We found that both debiased models mitigate the bias of the typical CF algorithm by suggesting similar concentrations to the protected groups, while Debiased Model 2 mitigates the bias further and suggests very similar concentrations to all of the groups. For example, the typical CF model recommends psychology to men and nursing to women, while Debiased Model 2 recommends both of these to both genders. We will show more examples for the different groups in the presentation. Finally, we study the effect of debiasing on the recommendation performance (Figure 4). The debiased models only slightly increase mean squared error (MSE) and mean absolute error (MAE) metrics, validating that our proposed methods remain effective while achieving fair behavior.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering* 6 (2005), 734–749.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May* 23 (2016).

[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[4] J. Buolamwini and T. Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT**. 77–91.

[5] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

[6] Robin Burke, Nasim Sonboli, Masoud Mansoury, and Aldo Ordoñez-Gauger. 2017. Balanced neighborhoods for fairness-aware collaborative recommendation. (2017).

[7] Sunipa Dev and Jeff M. Phillips. 2019. Attenuating Bias in Word Vectors. *CoRR* abs/1901.07656 (2019). arXiv:1901.07656 http://arxiv.org/abs/1901.07656

[8] Tao Ding, Warren K Bickel, and Shimei Pan. 2017. Multi-view unsupervised user feature embedding for social media-based substance use prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2275–2284.

[9] James Foulds, Rashidul Islam, Kamrun Keya, and Shimei Pan. 2018. Bayesian Modeling of Intersectional Fairness: The Variance of Bias. *arXiv preprint arXiv:1811.07255* (2018).

[10] Toshihiro Kamishima and Shotaro Akaho. 2017. Considerations on Recommendation Independence for a Find-Good-Items Task. (2017).

[11] Michal Kosinski, Sandra C Matz, Samuel D Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist* 70, 6 (2015), 543.

[12] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 1–35.

[13] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*. 2921–2930.