A User Study on a De-biased Career Recommender System

Clarice Wang
The Harker School, San Jose, CA
e22claricew@students.harker.org

Kathryn Wang
The Dalton School, New York, NY
wang.kathryn.wang@gmail.com

Andrew Bian
River Hill High School, Clarksville, MD
abian1960@inst.hcpss.org

Rashidul Islam, Kamrun Naher Keya
James R. Foulds, Shimei Pan
UMBC, Baltimore, MD
islam.rashidul,kkeya1,jfoulds,shimei@umbc.edu

1 Introduction

AI is increasingly being used in making consequential decisions such as determining whether someone is granted parole or not (Angwin et al., 2016). Unfortunately, there have been a wide range of recent discoveries of biased AI systems that are prejudiced against certain groups of people (Dastin, 2018; Noble, 2018; Angwin et al., 2016). In this research, we focus on developing new techniques that mitigate gender biases in automated career recommendation systems. Since biases are typically inherent in AI systems trained on data influenced by our society, an AI recommender must be "de-biased" to avoid reinforcing harmful stereotypes (e.g., recommending *computer programming* to boys and *nursing* to girls) (Bolukbasi et al., 2016; Yao and Huang, 2017). Although it is technically possible to remove biases from an AI system, it is unclear whether intended users prefer such a system. We conduct a user study to investigate this.

2 AI-based Career Recommendation

We implemented two variations of a career recommender system: *gender-aware* and *gender-debiased*. The gender-aware system makes career recommendations based on the choices by the people of the same gender (e.g., recommending to girls based on the career choices of other girls) while the gender-debiased system (Islam et al., 2019) mitigates the influence of gender. We train the systems using the "likes" and declared career concentrations of 15,000 people. We first train a neural collaborative filtering (NCF) model (He et al., 2017) to learn a vector representation of users and "likes" (i.e. user- and like-embeddings). We then use a logistic regression classifier to suggest career concentrations based on the user- and like-embeddings. For the gender-debiased recommender, there is an additional de-biasing step where we adapt a recent work on attenuating bias in word vectors (Dev and Phillips, 2019). First, we obtain a *male-female bias direction* as $v_B = (v_F - v_M)/(||v_F - v_M||)$, where v_F and v_M are the average vector over the embeddings of female and male users. We then de-bias user embeddings (p_u) by removing their component along the bias direction: $p'_u = p_u - (p_u \cdot v_B)v_B$. To evaluate the system performance, we employed both accuracy and fairness-based measures such as Hit Rate (HR), Normalized Discounted Cumulative Gain (NDCG) and Non-parity Unfairness (Yao and Huang, 2017). Our results demonstrate that the de-biased recommender achieved better fairness without losing any prediction accuracy.

3 User Study

To investigate whether users prefer a de-biased recommender or not, we designed a customized online survey using SurveyJS. Each user is randomly assigned to interact with either a "gender-aware" or a "gender-debiased" system. For each participant, we collect data on their interests/likes (e.g., whether they like certain music, brands and hobbies). Based on these "likes," the system recommends three career concentrations. For each recommended concentration, a user is asked to indicate whether they consider it as a possible future career choice. If the answer is "yes," the system receives 1 point. It receives 0 points if the user said "no" and 0.5 if the user says "I don't know." Based on an independent two-sample t-test, the mean acceptance rate for the gender-debiased system is 0.27 while that for the gender-aware system is 0.38. Although the difference is only marginally significant ($p < 0.08$), a lower mean acceptance rate for the gender debiased system suggests that users on average do not prefer the de-biased recommender.

4 Conclusion and Discussion

Despite the general belief that gender de-biasing AI systems is important, users on average do not prefer the de-biased recommender. One possible explanation is human bias (e.g., people may unconsciously prefer careers that conform to gender stereotypes). This has significant implications on the design and evaluation of de-biased AI systems.

References

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May*, 23.

T. Bolukbasi, K.W. Chang, J.Y. Zou, V. Saligrama, and A.T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

J. Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, page 173–182.

Rashidul Islam, Kamrun Naher Keya, Shimei Pan, and James Foulds. 2019. Mitigating demographic biases in social media-based recommender systems. In *KDD (Social Impact Track)*.

S.U. Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

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