Debiasing Career Recommendations with Neural Fair Collaborative Filtering

Rashidul Islam
University of Maryland, Baltimore County
Baltimore, MD, USA
islam.rashidul@umbc.edu

Kamrun Naher Keya
University of Maryland, Baltimore County
Baltimore, MD, USA
kkeya1@umbc.edu

Ziqian Zeng
Hong Kong University of Science and Technology
Kowloon, Hong Kong
zzengae@cse.ust.hk

Shimei Pan
University of Maryland, Baltimore County
Baltimore, MD, USA
shimei@umbc.edu

James Foulds
University of Maryland, Baltimore County
Baltimore, MD, USA
jfoulds@umbc.edu

ABSTRACT

A growing proportion of human interactions are digitized on social media platforms and subjected to algorithmic decision-making, and it has become increasingly important to ensure fair treatment from these algorithms. In this work, we investigate gender bias in collaborative-filtering recommender systems trained on social media data. We develop neural fair collaborative filtering (*NFCF*), a practical framework for mitigating gender bias in recommending career-related sensitive items (e.g. jobs, academic concentrations, or courses of study) using a pre-training and fine-tuning approach to neural collaborative filtering, augmented with bias correction techniques. We show the utility of our methods for gender de-biased career and college major recommendations on the MovieLens dataset and a Facebook dataset, respectively, and achieve better performance and fairer behavior than several state-of-the-art models.

CCS CONCEPTS

• **Information systems** → **Collaborative filtering**; • **Computing methodologies** → **Neural networks**; • **Applied computing** → **Law, social and behavioral sciences**.

KEYWORDS

Fairness in AI, AI & society, ethical issues, social media, recommender systems, collaborative filtering, career recommendation

ACM Reference Format:
Rashidul Islam, Kamrun Naher Keya, Ziqian Zeng, Shimei Pan, and James Foulds. 2021. Debiasing Career Recommendations with Neural Fair Collaborative Filtering. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3449904

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449904

1 INTRODUCTION

There is increasing awareness that machine learning (ML) algorithms can affect people in unfair ways with legal or ethical consequences when used to automate decisions [2, 3], for example, exhibiting discrimination towards certain demographic groups. Systemic bias, which has long been the concern of civil rights and feminist scholars and activists [1, 14, 15, 42, 49], in turn affects data, and hence ML algorithms trained on data [3]. The need to connect the fairness and bias demonstrated in ML algorithms with the broader context of fairness and bias in society is increasingly well understood [37, 46]. Structural disadvantages and systems of oppression in our society such as sexism and racism can lead individuals from marginalized groups to perform below their true potential. For example, these issues can reduce the available cognitive bandwidth required for academic success [50] or increase the probability and length of incarceration [1, 16] for minority groups. It is important to ensure that these patterns are not replicated or amplified by ML models which are used to make consequential decisions [12].

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 ML fairness issues in social media can potentially cause substantial societal harm. Recommender systems are the primary method for a variety of ML tasks for social media data, e.g. suggesting targets of advertisements, products, friends, web pages, and potentially consequential suggestions such as romantic partners or even career paths.

Despite the practical challenges from labor market dynamics [36], professional networking sites' job recommendations [4, 23, 24] are helpful for job seekers and employers. However, biases inherent in social media data can potentially lead recommender systems to produce unfair suggestions [54]. Many studies have been conducted which demonstrated the demographic biases in the different aspects of the job market. For example, racial discrimination was shown in the recruitment process of the labor market [6]. A similar study [47] was conducted to confirm the presence of discrimination in a job market in Canada with respect to race as well as ethnicity. A recent study on a job platform, XING, similar to LinkedIn demonstrated that it ranks less qualified male candidates higher than more qualified female candidates [41]. Recommendations in educational

and career choices are another important application for fair recommender systems. Students' academic choices can have significant impacts on their future careers and lives. An earlier study illustrated that the screening process of a medical school in London was highly biased [44] against women and members of ethnic minorities. In 2010, women accounted for only 18% of the bachelor's degrees awarded in computer science [9], and interventions to bridge this gap are crucial to support the economic competitiveness and level of innovation of the United States [5].

Recommender systems can reinforce this disparity, or –potentially– help to mitigate it. We envision an ML-based career counseling tool which makes personalized data-driven recommendations regarding important career choices such as profession, college major, certifications, or jobs to apply for, while ensuring that the recommendations do not perpetuate systemic bias or harmful stereotypes which are damaging both for our society and for the individuals who use the system. Such a tool could support young people in consequential life decisions in partnership with their parents and counselors, as well as professionals who aim to make smart career moves. Social media data is readily available to support personalized recommendations, as long as bias issues are adequately countered.

We propose a practical technique to mitigate gender bias in sensitive item (e.g. college major or career path) recommendations. Our approach, which we call *neural fair collaborative filtering (NFCF)*, achieves accurate predictions while addressing ***sensitive data sparsity*** (e.g., users typically have only one or two college majors or occupations) by pre-training a deep neural network on big implicit feedback data for non-sensitive items (e.g. "liked" Facebook pages, movies or music), and then fine-tuning the neural network for sensitive item recommendations. We perform two bias corrections, to address *(1) bias in the* ***input embeddings*** *due to the non-sensitive items*, and *(2) bias in the* ***prediction outputs*** *due to the sensitive items*. An ablation study shows that ***both*** *interventions are important for fairness*. We demonstrate the utility of our method on two datasets: *MovieLens* (non-sensitive *movie ratings* and sensitive *occupations*), and a *Facebook* dataset (non-sensitive *Facebook page "likes"* and sensitive *college majors*). Our main contributions include:

- We develop a pre-training + fine-tuning neural network method for fair recommendations on social media data.
- We propose two de-biasing methods for this task: 1) debiasing latent embeddings, and 2) learning with a fairness penalty. We also develop two simpler model variants.
- We perform extensive experiments showing both fairness and accuracy benefits over baselines on two datasets.

2 BACKGROUND

In this section we formalize the problem, and discuss collaborative filtering with implicit data, and fairness metrics.

2.1 Problem Formulation

Let M and N denote the number of users and items, respectively (see Table 1 for relevant notation). Suppose we are given a user-item interaction matrix $\mathbf{Y} \in \mathbb{R}^{M \times N}$ of *implicit feedback* from users, e.g. social media "likes," defined as

$$y_{ui} = \begin{cases} 1, & \text{if } u \text{ interacts with } i \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Here, $y_{ui} = 1$ when there is an interaction between user u and item i, e.g. when u "likes" Facebook page i. In this setting, a value of 0 does not necessarily mean u is not interested in i, as it can be that the user is not yet aware of it, or has not yet interacted with it. While interacted entries reflects users' interest in items, the unobserved entries may just be missing data. Therefore, there is a natural scarcity of strong negative feedback.

The collaborative filtering (*CF*) problem with implicit feedback is formulated as the problem of predicting scores of unobserved entries, which can be used for ranking the items. The *CF* model outputs $\hat{y}_{ui} = f(u, i|\Theta)$, where $\hat{y}_{ui}$ denotes the estimated score of interaction y_{ui}, Θ denotes model parameters, and f denotes the function that maps model parameters to the estimated score. If we constrain $\hat{y}_{ui}$ in the range of [0,1] and interpret it as the probability of an interaction, we can learn Θ by minimizing the following negative log-likelihood objective function:

$$L = -\sum_{(u,i) \in \chi \cup \chi^-} y_{ui} \log \hat{y}_{ui} + (1 - y_{ui}) \log(1 - \hat{y}_{ui}) \,, \quad (2)$$

where χ represents the set of interacted user-item pairs, and χ^- represents the set of negative instances, which can be all (or a sample of) unobserved interactions. In our setting, we further suppose that items i are divided into non-sensitive items (i_n) and sensitive items (i_s). For example, the i_n's can be *Facebook pages* where user preferences may reasonably be influenced by the protected attribute such as gender, and the user's "likes" of the pages are the implicit feedback. Since each user u can (and often does) "like" many pages, u's observed non-sensitive data (u-i_n) is typically large. On the other hand, i_s may be the users' *occupation* or *academic concentration* provided in their social media profiles. We desire that the recommendations of i_s to new users should be unrelated to the users' gender (or other protected attribute). Since each user u may typically be associated with only a single occupation (or

u-i	Interacted user and item (non-sensitive and sensitive)
u-i_n	Interacted user and non-sensitive item
u-i_s	Interacted user and sensitive item
p_u	User vector
q_{i_n}	Non-sensitive item vector
q_{i_s}	Sensitive item vector
v_{female}	Bias direction for female users
v_{male}	Bias direction for male users
v_B	Gender bias vector
p'_u	De-biased user vector
$\boldsymbol{W}$	Neural model's parameters
$L_{\chi \cup \chi^-}$	Loss function for interacted and not-interacted pairs
R_χ	Fairness penalty function
λ	Tuning parameter for fairness and accuracy trade-off
ϵ_{i_s}	Differential fairness measure for a sensitive item i_s
ϵ_{mean}	Mean differential fairness measure
U_{abs}	Absolute unfairness measure

Table 1: Summary of notation.

other sensitive personal data rarely disclosed), the data sparsity in the observed sensitive item interactions (u-i_s) is a major challenge. Typical collaborative filtering methods can suffer from overfitting in this scenario that often amplifies unfairness or demographic bias in the data [22, 57]. Alternatively, the non-sensitive interactions u-i_n can be leveraged, but these will by definition encode biases that are unwanted for predicting the sensitive items. For example, liking the *Barbie doll* Facebook page may be correlated with being female and negatively correlated with *computer science*, thus implicitly encoding societal bias in the career recommendations.

2.2 Neural Collaborative Filtering

Matrix factorization (*MF*) models [39] map both users and items to a joint latent factor space of dimensionality v such that user-item interactions are modeled as inner products in that space. Each item i and user u are associated with a vector $q_i \in R^v$ and $p_u \in R^v$, with

$$\hat{y}_{ui} = q_i^T p_u + \mu + b_i + b_u \ , \tag{3}$$

where μ is the overall average rating, and b_u and b_i indicate the deviations of user u and item i from μ, respectively.

Neural collaborative filtering (*NCF*) [27] replaces the inner products in MF with a deep neural network (*DNN*) which learns the user-item interactions. In the input layer, the users and items are typically one-hot encoded, then mapped into the latent space with an embedding layer. *NCF* combines the latent features of users p_u and items q_i by concatenating them. Complex non-linear interactions are modeled by stacking hidden layers on the concatenated vector, e.g. using a standard multi-layer perceptron (*MLP*).

2.3 Fairness Metrics

We consider several existing fairness metrics which are applicable for collaborative filtering problems.

2.3.1 Differential Fairness. The differential fairness [21, 22] metric aims to ensure equitable treatment for all protected groups, and it provides a privacy interpretation of disparity, and economic guarantees. Let $M(x)$ be an algorithmic mechanism (e.g. a recommender system) which takes an individual's data x and assigns them an outcome y (e.g. a class label or whether a user-item interaction is present). The mechanism $M(x)$ is *ϵ-differentially fair (DF)* with respect to (A, Θ) if for all $\theta \in \Theta$ with $x \sim \theta$, and $y \in \text{Range}(M)$,

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(x) = y | s_i, \theta)}{P_{M,\theta}(M(x) = y | s_j, \theta)} \leq e^{\epsilon} \ , \tag{4}$$

for all $(s_i, s_j) \in A \times A$ where $P(s_i|\theta) > 0$, $P(s_j|\theta) > 0$. Here, s_i, $s_j \in A$ are tuples of all protected attribute values, e.g. male and female, and Θ, the set of data generating distributions, is typically a point estimate of the data distribution. If all of the $P_{M,\theta}(M(x) = y|s, \theta)$ probabilities are equal for each group s, across all outcomes y and distributions θ, $\epsilon = 0$, otherwise $\epsilon > 0$. [22] proved that a small ϵ guarantees similar utility per protected group, and ensures that protected attributes cannot be inferred based on outcomes. For gender bias in our recommender (assuming a gender binary), we can estimate ϵ-DF per sensitive item i by verifying that:

$$e^{-\epsilon} \leq \frac{\sum_{u:A=m} \hat{y}_{ui} + \alpha}{N_m + 2\alpha} \frac{N_f + 2\alpha}{\sum_{u:A=f} \hat{y}_{ui} + \alpha} \leq e^{\epsilon} \ ,$$
$$e^{-\epsilon} \leq \frac{\sum_{u:A=m} (1 - \hat{y}_{ui}) + \alpha}{N_m + 2\alpha} \frac{N_f + 2\alpha}{\sum_{u:A=f} (1 - \hat{y}_{ui}) + \alpha} \leq e^{\epsilon} \ , \tag{5}$$

where scalar α is each entry of the parameter of a symmetric Dirichlet prior with concentration parameter 2α, i is an item and N_A is the number of users of gender A (m or f).

2.3.2 Absolute Unfairness. The absolute unfairness (U_{abs}) metric for recommender systems measures the discrepancy between the predicted behavior for disadvantaged and advantaged users [54]. It measures differences in absolute estimation error across user types:

$$U_{abs} = \frac{1}{N} \sum_{j=1}^{N} ||(E_D[\hat{y}_{ui}]_j - E_D[r]_j)| - |(E_A[\hat{y}_{ui}]_j - E_A[r]_j)|| \tag{6}$$

where, for N items, $E_D[\hat{y}_{ui}]_j$ is the average predicted score for the j-th item for disadvantaged users, $E_A[\hat{y}_{ui}]_j$ is the average predicted score for advantaged users, and $E_D[r]_j$ and $E_A[r]_j$ are the average score for the disadvantaged and advantaged users, respectively.

3 NEURAL FAIR CF

Due to biased data which encode harmful human stereotypes in our society, typical social media-based collaborative filtering (*CF*) models can encode gender bias and make unfair decisions. In this section, we propose a practical framework to mitigate gender (or other demographic) bias in *CF* recommendations, which we refer to as *neural fair collaborative filtering* (*NFCF*) as shown in Figure 1. The main components in our *NFCF* framework are as follows: an *NCF model*, *pre-training user and non-sensitive item embeddings*, *debiasing pre-trained user embeddings*, and *fine-tuning with a fairness penalty*. We use *NCF* as the *CF* model because of its flexible network structure for pre-training and fine-tuning. We will show the value of each component below with an ***ablation study*** (Table 4). Similarly to [27], the *DNN* under the *NCF* model can be defined as:

$$z_1 = \phi_1(p_u, q_i) = \begin{bmatrix} p_u \\ q_i \end{bmatrix} , z_2 = \phi_2(z_1) = a_2(W_2^T z_1 + b_2) \ , \ \dots \ ,$$
$$\phi_L(z_{L-1}) = a_L(W_L^T z_{L-1} + b_L) \ , \hat{y}_{ui} = \sigma(h^T \phi_L(z_{L-1})) \tag{7}$$

where z_l, ϕ_l, W_l, b_l. and a_l denote the neuron values, mapping function, weight matrix, intercept term, and activation function for the l-th layer's perceptron, respectively. The DNN is applied to z_1 to learn the user-item latent interactions.

In the first step of our *NFCF* method, *pre-training user and item embeddings*, *NCF* is trained to predict users' interactions with *non-sensitive* items (e.g. "liked" social media pages) via back-propagation. This leverages plentiful non-sensitive social media data to learn user embeddings of the user's preference or profile and network weights, but may introduce **demographic bias due to correlations between non-sensitive items and demographics**. E.g., liking the *Barbie doll* page typically correlates with user gender. These correlations are expected to result in systematic differences in the embeddings for different demographics, which in turn can lead to systematic differences in sensitive item recommendations.

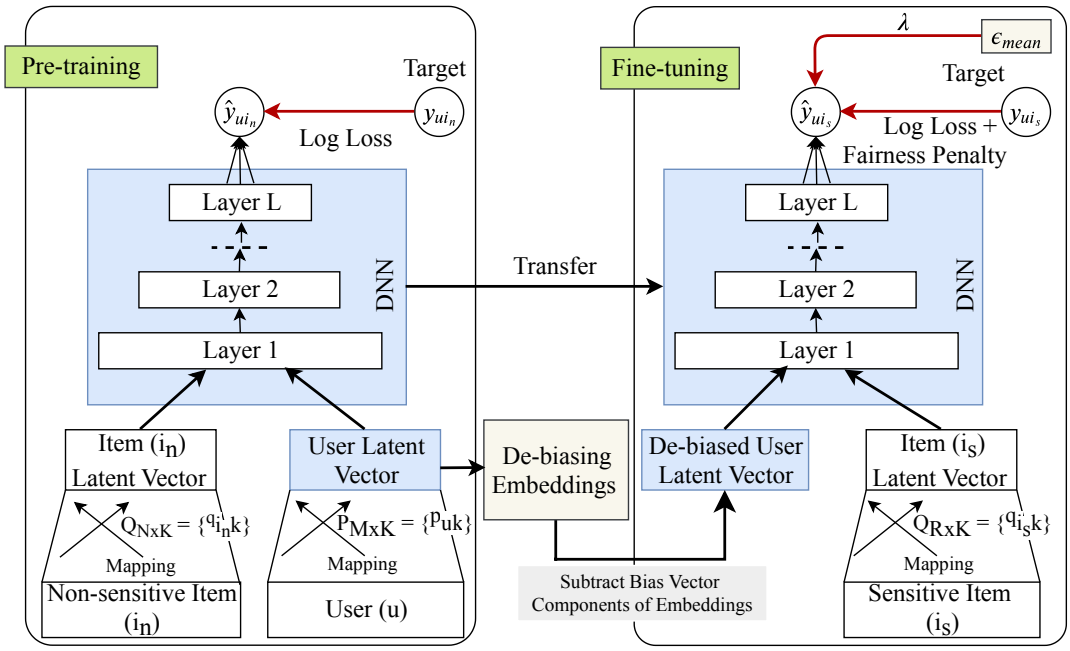


Figure 1: Schematic diagram of neural fair collaborative filtering (NFCF). Red arrows indicate back-propagation only.

Our aim is to leverage the valuable signal of the user's preferences for sensitive item recommendations, but also address the problems with it regarding bias. In step two, the user embeddings from step one are *de-biased*. Our method to de-bias user embeddings adapts a very recent work on attenuating bias in word vectors [18] to the task of collaborative filtering. Specifically, [18] propose to debias word vectors using a linear projection of each word embedding w orthogonally onto a *bias vector* v_B, which identifies the "bias component" of w. The bias component is then removed via $w' = w - (w \cdot v_B)v_B$.

To adapt this method to *CF*, the main challenge is to find the proper bias direction v_B. [18] construct v_B based on word embeddings for gender-specific names, which are not applicable for *CF*. We instead use *CF embeddings for users from each protected group*. We first compute a group-specific bias direction for female users as

$$v_{female} = \frac{1}{n_f}(f_1 + f_2 + \cdots + f_n)\ , \tag{8}$$

where $f_1, f_2, \ldots$ are vectors for each female user, and n_f is the total number of female users. We similarly compute a bias direction for male v_{male}. Finally, we compute the overall gender bias vector:

$$v_B = \frac{v_{female} - v_{male}}{||v_{female} - v_{male}||}\ . \tag{9}$$

We then de-bias each user embedding p_u by subtracting its component in the direction of the bias vector:

$$p'_u = p_u - (p_u \cdot v_B)v_B\ . \tag{10}$$

As we typically do not have demographic attributes for items, we only de-bias user embeddings and not item embeddings. In the third step, we *transfer* the de-biased user embeddings and pre-trained *DNN*'s parameters to a model for recommending *sensitive items*, which we *fine-tune* for this task. During fine-tuning, a *fairness penalty* is added to the objective function to address a second source of bias: **demographic bias in the sensitive items**. E.g., more men than women choose computer science careers [9], and this should be corrected [5]. We penalize the mean of the *per-item* ϵ's:

$$\epsilon_{mean} = \frac{1}{n_s}\sum_{i=1}^{n_s}\epsilon_i\ , \tag{11}$$

where $\epsilon_1, \epsilon_2, \ldots \epsilon_{n_s}$ are the *DF* measures for sensitive items and ϵ_{mean} is the average across the ϵ's for each item. Following [22], our learning algorithm for fine-tuning uses the fairness cost as a regularizer to balance the trade-off between fairness and accuracy. Using back-propagation, we minimize the loss function $L_{\chi\cup\chi^-}(\mathbf{W})$ from Equation 2 for model parameters $\mathbf{W}$ plus a penalty on ϵ_{mean}, weighted by a tuning parameter $\lambda > 0$:

$$\min_{\mathbf{W}}[L_{\chi\cup\chi^-}(\mathbf{W}) + \lambda R_\chi(\epsilon_{mean})] \tag{12}$$

where $R_\chi(\epsilon_{mean}) = max(0, \epsilon_{mean_{M_W(\chi)}} - \epsilon_{mean_0})$ is the fairness penalty term, and $\epsilon_{mean_{M_W(\chi)}}$ is the ϵ_{mean} for the *CF* model $M_W(\chi)$ while χ and χ^- are the set of interacted and not-interacted user-item pairs, respectively. In our experiments, we use $\epsilon_{mean_0} = 0$ to encourage demographic parity. Pseudo-code is given in Algorithm 1.

3.1 Variants of NFCF Model

We also consider two variants of our method which are simplifications of the *NFCF* model.

3.1.1 ***NFCF_embd.*** This variant only de-biases the user embeddings. In the *NFCF_embd* algorithm, we compute the bias vector v_B on the pre-trained user embeddings, fine-tune the model for sensitive item recommendations without any fairness penalty, and

Algorithm 1 Training NFCF Model

Input: pairs of user and non-sensitive item: $\mathcal{D}_n = (u, i_n)$, pairs of user and sensitive item: $\mathcal{D}_s = (u, i_s)$, and gender attribute: A
Output: Fair *CF* model $M_{\mathrm{W}}(\mathrm{x})$ for i_s recommendations

Pre-training steps:

- Randomly initialize $M_{\mathrm{W}}(\mathrm{x})$'s parameters $\boldsymbol{W}$: p_u, q_{i_n}, W_l, and b_l
- For each epoch of D_n:
  - For each mini-batch:
    - Learn $M_{\mathrm{W}}(\mathrm{x})$'s parameters $\boldsymbol{W}$ by minimizing:
      $L = -\sum_{(u,i_n)\in\chi\cup\chi^-}[\, y_{ui_n} \log \hat{y}_{ui_n} + (1-y_{ui_n}) \log(1-\hat{y}_{ui_n})\,]$

De-biasing embeddings steps:

- Compute gender bias vector v_B using Equation 8 and 9
- De-bias each user embedding using: $p'_u := p_u - (p_u \cdot v_B) v_B$

Fine-tuning steps:

- Initialize with pre-trained $M_{\mathrm{W}}(\mathrm{x})$'s parameters $\boldsymbol{W}$: W_l, and b_l
- Randomly initialize q_{i_s}, while p_u is replaced with de-biased p'_u
- For each epoch of D_s:
  - For each mini-batch:
    - Fine-tune $M_{\mathrm{W}}(\mathrm{x})$ by minimizing (while p'_u is kept fixed):
      $\min_{\mathbf{W}} [L_{\chi\cup\chi^-}(\mathbf{W}) + \lambda R_\chi(\epsilon_{mean})]$

then de-bias the held-out user embeddings using the pre-computed bias vector. Since there is no additional fairness penalty in the objective function, this algorithm converges faster. There is also no requirement to tune the λ hyperparameter.

3.1.2 ***Projection-based CF.*** In the *Projection-based CF* algorithm, our approach is to learn an *NCF* model for non-sensitive u-i_n interactions (using Equation 7), and then debias the user embeddings using the linear projection technique in Equation 10. Finally, we learn a classifier such as k-nearest neighbors or logistic regression on the de-biased user embeddings to predict sensitive items (i_s). There is no fine-tuning to address overfitting for sensitive items or fairness penalty-based bias correction in this approach. We previously presented this simpler model as a non-archival extended abstract at a workshop [29].

Since a user usually interacts with a single sensitive item (e.g. occupation), it is tempting to use a classifier, as in the *Projection-based CF* method, to predict the sensitive items such as careers, viewing them as discrete class labels. However, our experiments will show that classification approaches including *Projection-based CF* and a deep neural network classifier are suboptimal. The intuition is that even though the output for sensitive items is like classification (a single label), the input data is like recommendation (interactions of users with other items), and the overall system hence benefits from an end-to-end collaborative filtering approach as in *NFCF*.

4 EXPERIMENTS

In this section, we validate and compare our model with multiple baselines for recommending careers and academic concentrations using social media data. Our implementation's source code is provided on GitHub.[1]

4.1 Datasets

We evaluate our models on two datasets: *MovieLens*,[2] a public dataset which facilitates research reproducibility, and a *Facebook* dataset which is larger and is a more realistic setting for a fair social media-based recommender system.

4.1.1 ***MovieLens Data.*** We analyzed the widely-used *MovieLens* dataset which contains 1 million ratings of 3,900 movies by 6,040 users who joined MovieLens [25], a noncommercial movie recommendation service operated by the University of Minnesota. We used *gender* as the protected attribute, self-reported *occupation* as the sensitive item (with one occupation per user), and *movies* as the non-sensitive items. Since we focus on implicit feedback, which is common in a social media setting (e.g. page "likes"), we follow [27, 38] to convert explicit movie ratings to binary implicit feedback, where a 1 indicates that the user has rated the item. We discarded movies that were rated less than 5 times, and users who declared their occupation as "K-12 student," "retired," "unemployed," and "unknown or not specified" were discarded for career recommendation. A summary of the pre-processed dataset is shown in Table 2.

4.1.2 ***Facebook Data.*** The *Facebook* dataset was collected as part of the myPersonality project [40]. The data for research were collected with opt-in consent. We used *gender* as the protected attribute, *college major* as the sensitive items (at most one per user), and *user-page* interaction pairs as the non-sensitive items. A user-page interaction occurs when a user "likes" a Facebook page. We discarded pages that occurred in less than 5 user-page interactions. See Table 2 for a summary of the dataset after pre-processing.

4.1.3 ***Gender Distributions for Datasets.*** In Figure 2, we show disparities in the gender distributions of 10 example careers and college majors for *MovieLens* and *Facebook* datasets, respectively. For example, 97% of the associated users for the occupation *homemaker* are women in the *MovieLens* data, while there are only 27% women among the users associated with the *computer science* major in the *Facebook* data. As a qualitative illustration, we also show the gender distribution of top-1 recommendations from our proposed *NFCF* model. *NFCF* mitigated gender bias for most of these sensitive items. In the above examples, *NFCF* decreased the percentage of women for *homemaker* from 97% to 50%, while increasing the percentage of women for *computer science* from 27% to 48%.

4.2 Baselines

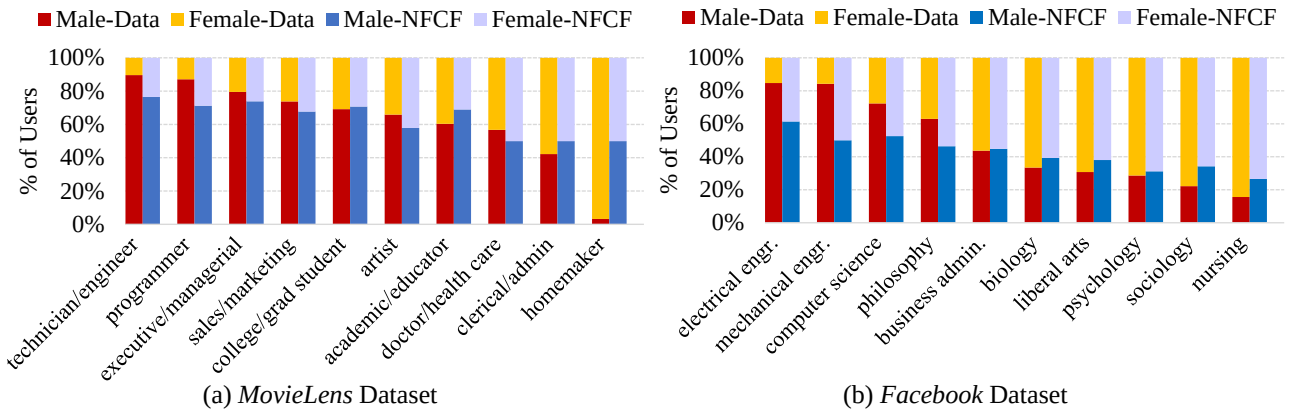
We compare our methods to the following "typical" baseline models:

- **MF w/o Pre-train** and **NCF w/o Pre-train**. Typical *MF* and *NCF* models, respectively, which are trained with the u-i interactions for i_s recommendations, where i contains both i_n and i_s.
- **MF w Pre-train** and **NCF w Pre-train**. Typical *MF* and *NCF* model, respectively, which are pre-trained with the u-i_n interactions and fine-tuned with the u-i_s interactions for i_s recommendations. Specifically, q_i, b_i, and b_u from Equation 3 and q_i, W_l, and b_l from Equation 7 are fine-tuned for *MF w*

[1] https://github.com/rashid-islam/nfcf.
[2] http://grouplens.org/datasets/movielens/1m/.

	Non-sensitive Data				Sensitive Data					
	Users	Items	Pairs	Sparsity	Users	Males	Females	Items	Pairs	Sparsity
MovieLens Dataset	6,040	3,416	999,611	95.16%	4,920	3,558	1,362	17	4,920	94.12%
Facebook Dataset	29,081	42,169	5,389,541	99.56%	13,362	5,053	8,309	169	13,362	99.41%

Table 2: Statistics of the datasets.



(a) *MovieLens* Dataset (b) *Facebook* Dataset

Figure 2: Gender distributions of example gender-biased careers and college majors for (a) *MovieLens* and (b) *Facebook* datasets. We report the distributions in the dataset (left columns), and corresponding top-1 recommendation by our NFCF model (right columns).

MovieLens Dataset				
Models	HR@10	NDCG@10	HR@25	NDCG@25
NCF	0.543	**0.306**	0.825	**0.377**
MF	**0.551**	0.304	**0.832**	0.374
Facebook Dataset				
Models	HR@10	NDCG@10	HR@25	NDCG@25
NCF	**0.720**	**0.468**	**0.904**	**0.514**
MF	0.609	0.382	0.812	0.434

Table 3: Performance of NCF and MF models for movie and Facebook page recommendations (the pre-training task).

Pre-train and *NCF w Pre-train*, respectively, while p_u are kept fixed for both models.

- **DNN Classifier**. A simple baseline where we train a *DNN*-based classifier to predict i_s given the u-i_n interactions as features (i.e. binary features, one per user-page "like" or one per user-movie "rating"). No user embeddings are learned.
- **BPMF**. Bayesian probabilistic *MF* (*BPMF*) via *MCMC* [48] is also used, since it typically has good performance with small data. *BPMF* is trained with the u-i interactions for i_s recommendations, where i contains both i_n and i_s.

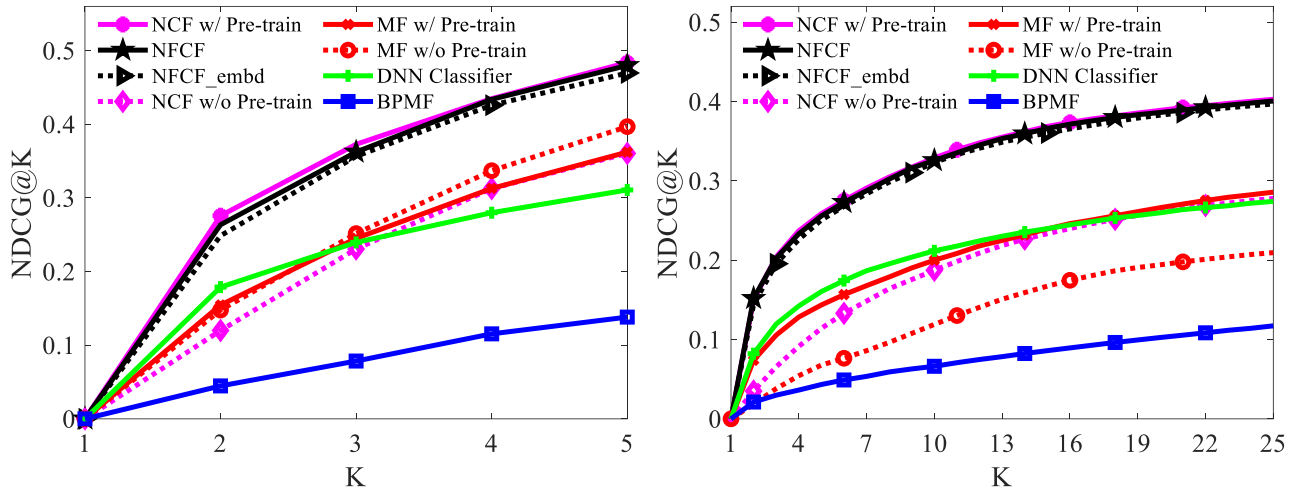
We also use the following fair baseline models:

- **MF**-U_{abs}. The objective of the *MF w/o Pre-train* model is augmented with a smoothed variation of U_{abs} [54] using the Huber loss [28], weighted by a tuning parameter λ.
- **Resampling for Balance**. This method [19] involves pre-processing by resampling the u-i data to produce a gender-balanced version of the training data. First, we extract u-i data for users with known gender and randomly sample same number of male and female users without replacement where i includes both i_n and i_s. Finally, *NCF* and *MF* are trained on the gender-balanced u-i interactions for i_s recommendations.

4.3 Experimental Settings

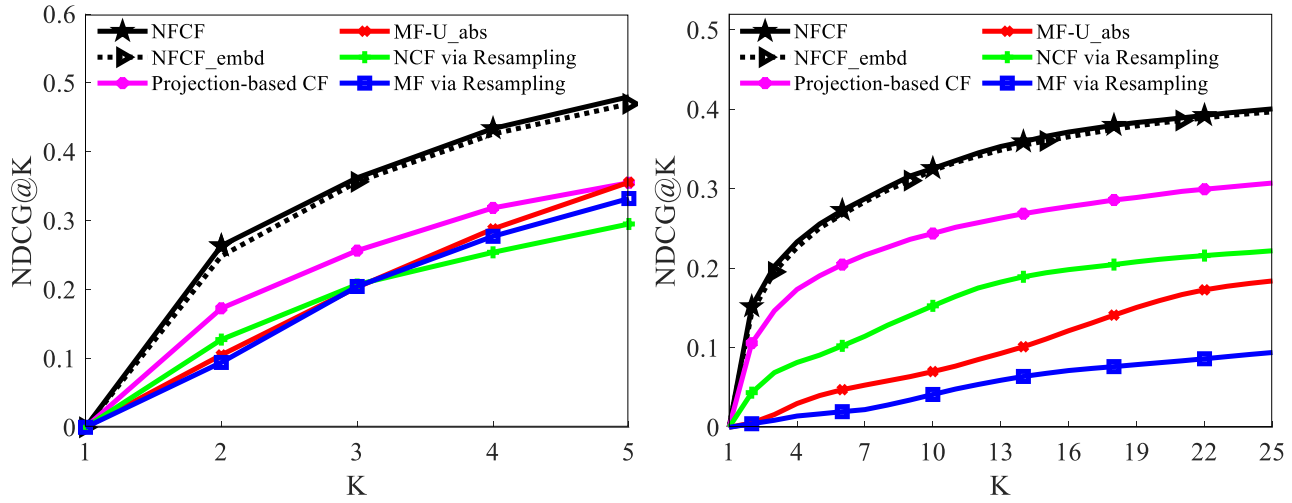
All the models were trained via adaptive gradient descent optimization (Adam) with learning rate = 0.001 using pyTorch where we sampled 5 negative instances per positive instance. The mini-batch size for all models was set to 2048 and 256 for user-page and user-career data, respectively, while the embedding size for users and items was set to 128. The configuration of the *DNN* under *NFCF* and *NFCF_embd* was 4 hidden layers with 256, 64, 32, 16 neurons in each successive layer, "relu" and "sigmoid" activations for the hidden and output layers, respectively. We used the same *DNN* architecture for the *NCF* and *DNN Classifier* models.

For the Facebook dataset, we held-out 1% and 40% from the user-page and user-college major data, respectively, as the test set, using the remainder for training. Since there are fewer users in the *Movielens* dataset, we held-out 1% and 30% from the user-movie and user-career data, respectively, as the test set, using the remainder for training. We further held-out 1% and 20% from the training u-i_n and u-i_s data, respectively, as the development set for each dataset. The fairness penalty was computed for each mini-batch during training. Note that the tuning parameter λ needs to be chosen as a trade-off between accuracy and fairness [22]. We chose λ as 0.1 for *NFCF* and *MF*-U_{abs} via a grid search on the development set according to similar criteria to [22], i.e. optimizing fairness while allowing up to 2% degradation in accuracy (i.e. NDCG) from the corresponding typical model (*NCF w/ Pre-train* and *MF w/o Pre-train*, respectively).



(a) *MovieLens* Dataset (b) *Facebook* Dataset

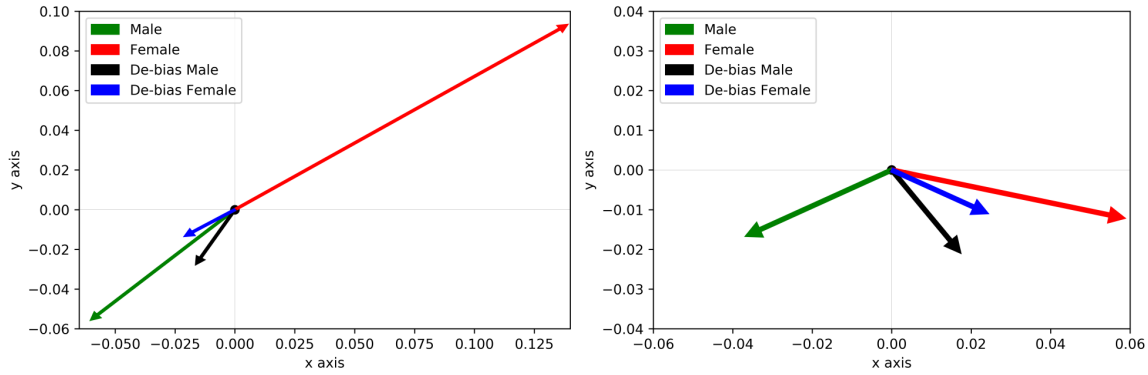
Figure 3: Comparison of proposed models with "typical" baselines that do not consider fairness. Evaluation of Top-K career and college major recommendations on the (a) *MovieLens* (among 17 unique careers) and (b) *Facebook* (among 169 unique majors) datasets, where K ranges from 1 to 5 and 1 to 25, respectively. NCF w/ Pre-train outperforms all the baselines; NFCF performs similarly.



(a) *MovieLens* Dataset (b) *Facebook* Dataset

Figure 4: Comparison of proposed models with fair baselines. Evaluation of Top-K career and college major recommendations on the (a) *MovieLens* (among 17 unique careers) and (b) *Facebook* (among 169 unique majors) datasets, where K ranges from 1 to 5 and 1 to 25, respectively. NFCF outperforms all the baselines; NFCF_embd performs similarly.

To evaluate the performance of item recommendation on the test data, since it is too time-consuming to rank all items for every user during evaluation [27], we followed a common strategy in the literature [20]. For non-sensitive items, we randomly sampled 100 items which are not interacted by the user for each test instance, and ranked the test instance among the 100 items. For sensitive item recommendations, in the case of *Facebook* data we similarly randomly sampled 100 college majors. For the *MovieLens* data, there are only 17 unique careers, so we used the remaining 16 careers when ranking the test instance. The performance of a ranked list is measured by the average Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) [26]. The HR measures whether the test item is present in the top-K list, while the NDCG accounts for the position of the hit by assigning higher scores to hits at top ranks. We calculated both metrics for each test user-item pair and reported the average score. Finally, we computed ϵ_{mean} and U_{abs} on the test data for user-sensitive item to measure the fairness of the models in career and college major recommendations.



(a) *MovieLens* Dataset (b) *Facebook* Dataset

Figure 5: De-biasing pre-trained user embeddings by removing the component along the bias direction v_B (PCA projection) for the (a) *MovieLens* and (b) *Facebook* datasets. PCA was performed based on all embeddings.

MovieLens Dataset				
Ablation study	HR@5 ↑	NDCG@5 ↑	ϵ_{mean} ↓	U_{abs} ↓
NFCF	**0.670**	0.480	**0.083**	**0.009**
w/o pre-train	0.493	0.323	0.112	0.017
w/o de-biasing embeddings	0.665	**0.481**	0.120	0.010
w/o fairness penalty	0.667	0.480	0.097	0.013
replace NCF w/ MF	0.514	0.350	0.122	0.021

Facebook Dataset				
Ablation study	HR@10 ↑	NDCG@10 ↑	ϵ_{mean} ↓	U_{abs} ↓
NFCF	0.551	0.326	**0.302**	**0.024**
w/o pre-train	0.339	0.127	0.613	0.038
w/o de-biasing embeddings	0.556	**0.328**	0.314	**0.024**
w/o fairness penalty	**0.557**	0.327	0.363	0.026
replace NCF w/ MF	0.297	0.112	0.880	0.071

Table 4: Ablation study of NFCF for i_s recommendations on the *MovieLens* and *Facebook* datasets. Higher is better for HR and NDCG; lower is better for ϵ_{mean} and U_{abs}. Removing each model component harms performance and/or fairness.

4.4 Validation of NFCF Model Design

Before comparing to fair recommendation baseline models, we systematically validate our modeling choices for *NFCF*.

Pre-training Task Performance: We first study the performance for *NCF* and *MF* models at the pre-training task, *Facebook page* and *movie* recommendations (Table 3). *NCF* had substantially and consistently better performance compared to *MF* on the larger *Facebook* dataset, and similar overall performance on *MovieLens* (better in 2 of 4 metrics).

Fine-Tuning Performance: We fine-tuned these models on the interaction of users with the sensitive items for *career* and *college major* recommendations on *MovieLens* and *Facebook* dataset, respectively. Figure 3 shows top-K recommendations from 17 and 169 unique *careers* and *college majors* using several "typical" baseline models that do not involve any fairness constraints, where K ranges from 1 to 5 and 1 to 25 for *MovieLens* and *Facebook* dataset, respectively. *NCF w/ Pre-train* had the best performance in NDCG versus other baselines while our proposed *NFCF* and *NFCF_embd* performed approximately similarly to *NCF w/ Pre-train* for both datasets. Of the typical baselines, *MF w/o Pre-train* and *NCF w/o Pre-train* performed the second best for *MovieLens* and *Facebook* dataset, respectively. For the *MovieLens* dataset, *MF w/o Pre-train* performed better than *MF w/ Pre-train*, presumably due to the relatively small dataset and having relatively few parameters to fine-tune, unlike for the *DNN*-based *NCF* model. *BPMF* performed poorly despite using Bayesian inference for scarce data, perhaps due to [48]'s initialization via older *MF* methods.

Visualization of Embedding De-biasing: We visualized the *PCA* projections of the male and female vectors (Equation 8) before and after the linear projection-based de-biasing embeddings method, where *PCA* was performed based on all the embeddings. Figure 5 shows that the male and female vectors have very different directions and magnitudes. After de-biasing, the male and female vectors had a more similar direction and magnitude to each other.

Ablation Study: Finally, we conducted an ablation study in which the components of the method were removed one at a time. As shown in Table 4, there was a large degradation of the performance of *NFCF* when pre-training was removed (the de-biasing embeddings step was also removed, since there was no pre-trained user vector), or when *NCF* was replaced by *MF*. Removing the de-biased embedding method lead to better HR and NDCG scores, but with an increase in gender bias metrics. Similarly, learning without the fairness penalty led to similar performance in HR and NDCG, but greatly increased gender bias. Therefore, ***both** of the bias correction methods in the* NFCF *model are necessary to achieve the best level of fairness while maintaining a high recommendation accuracy.*

4.5 Performance for Mitigating Gender Bias in Sensitive Item Recommendations

We evaluated performance for career and college major recommendations in terms of accuracy (HR and NDCG) and fairness (ϵ_{mean} and U_{abs}). In Figure 4, we show that our proposed *NFCF* and *NFCF_embd* models clearly outperformed all the fair baseline models in terms of NDCG, regardless of the cut-off K. Another

MovieLens Dataset							
	Models	HR@5 ↑	NDCG@5 ↑	HR@7 ↑	NDCG@7 ↑	ϵ_{mean} ↓	U_{abs} ↓
Proposed Models	NFCF	**0.670**	0.480	0.822	0.536	**0.083**	**0.009**
	NFCF_embd	0.661	0.470	**0.825**	0.531	0.091	0.016
	Projection-based CF	0.514	0.355	0.655	0.408	0.229	0.012
Typical Baselines	NCF w Pre-train	0.667	**0.484**	**0.825**	**0.542**	0.188	0.022
	NCF w/o Pre-train	0.570	0.360	0.762	0.432	0.244	0.026
	MF w Pre-train	0.548	0.362	0.747	0.436	0.285	0.060
	MF w/o Pre-train	0.622	0.397	0.820	0.471	0.130	0.020
	DNN Classifier	0.428	0.311	0.546	0.355	0.453	0.035
	BPMF	0.225	0.138	0.338	0.180	0.852	0.063
Fair Baselines	MF-U_{abs} [54]	0.588	0.356	0.776	0.426	0.096	0.017
	NCF via Resampling [19]	0.443	0.295	0.622	0.362	0.144	0.022
	MF via Resampling [19]	0.542	0.332	0.759	0.413	0.103	0.029

Facebook Dataset							
	Models	HR@10 ↑	NDCG@10 ↑	HR@25 ↑	NDCG@25 ↑	ϵ_{mean} ↓	U_{abs} ↓
Proposed Models	NFCF	0.551	0.326	0.848	0.401	**0.302**	0.024
	NFCF_embd	0.557	**0.333**	0.850	0.397	0.359	**0.022**
	Projection-based CF	0.419	0.244	0.674	0.307	0.407	0.030
Typical Baselines	NCF w Pre-train	**0.559**	0.329	**0.851**	**0.403**	0.376	0.027
	NCF w/o Pre-train	0.402	0.187	0.762	0.278	0.785	0.039
	MF w Pre-train	0.372	0.200	0.717	0.286	0.875	0.077
	MF w/o Pre-train	0.267	0.119	0.625	0.210	0.661	0.029
	DNN Classifier	0.379	0.212	0.630	0.274	0.633	0.070
	BPMF	0.131	0.066	0.339	0.117	1.173	0.084
Fair Baselines	MF-U_{abs} [54]	0.163	0.007	0.627	0.184	0.629	0.026
	NCF via Resampling [19]	0.315	0.153	0.586	0.222	0.442	0.025
	MF via Resampling [19]	0.103	0.041	0.314	0.094	0.756	0.039

Table 5: Comparison of proposed models with the baselines in career and college major recommendations on *MovieLens* (17 *careers*) and *Facebook* (169 *majors*). Higher is better for HR and NDCG; lower is better for ϵ_{mean} and U_{abs}. NFCF greatly improves fairness metrics and beats all baselines at recommendation except for NCF w Pre-train, a variant of NFCF without its fairness correction.

variant of our proposed method, *Projection-based CF*, performed the second best on both datasets out of all of the fair models.

In Table 5, we show detailed results for the top 5 and top 7 recommendations on *MovieLens* and for the top 10 and top 25 recommendations on the *Facebook* dataset. Our proposed *NFCF* model was the most fair career and college major recommender in terms of ϵ_{mean}, while our *NFCF_embd* was the most fair in terms of U_{abs} on the *Facebook* dataset. In the case of the *MovieLens* dataset, our *NFCF* model was the most fair recommender model in terms of both fairness metrics. *NCF w/ Pre-train* performed best in the HR and NDCG metrics on both datasets. *NFCF* and *NFCF_embd* had nearly as good HR and NDCG performance as *NCF w/ Pre-train*, while also mitigating gender bias. We also found that our proposed fair models *NFCF* and *NFCF_embd* sometimes outperformed *NCF w/ Pre-train* in terms of HR and NDCG. For example, HR@5 and NDCG@10 for *NFCF* on the *MovieLens* and *NFCF_embd* on the *Facebook* data, respectively. This counter-intuitive result is presumably due to the regularization behavior of the fairness penalty on the objective which can sometimes lead fair models to reduce overfitting to some extent compared to the typical model, a phenomenon which has previously been observed by [45].

As expected, we also found that the pre-training and fine-tuning approach reduced overfitting for *NCF w/ Pre-train*, and thus improved the fairness metrics by reducing bias amplification. *NCF w/ Pre-train* outperforms most of the fair baselines in terms of both fairness and accuracy-based measures which validates effectiveness of pre-training and fine-tuning neural method for career recommendations. This was not the case for *MF w/ Pre-train*, presumably due to the limited number of pre-trained parameters to fine-tune. *Projection-based CF* and *MF-U_{abs}* also showed relatively good performance in mitigating bias in terms of U_{abs} compared to the typical models, but with a huge sacrifice in the accuracy. Similarly, *NCF via Resampling* and *MF via Resampling* had poor performance in accuracy, but improved fairness to some extent over their corresponding

MovieLens Dataset			
NFCF		NCF w/o Pre-train	
Male	Female	Male	Female
college/grad student	college/grad student	sales/marketing	customer service
executive/managerial	executive/managerial	academic/educator	academic/educator
academic/educator	technician/engineer	executive/managerial	artist
technician/engineer	academic/educator	doctor/health care	writer
programmer	programmer	college/grad student	college/grad student

Facebook Dataset			
NFCF		NCF w/o Pre-train	
Male	Female	Male	Female
psychology	psychology	philosophy	psychology
english literature	english literature	psychology	nursing
graphic design	music	computer science	sociology
music	theatre	biochemistry	graphic design
nursing	nursing	business admin.	business marketing
liberal arts	history	political science	elementary education
business admin.	sociology	business management	cosmetology
biology	liberal arts	medicine	accounting
history	business admin.	law	physical therapy
criminal justice	biology	finance	music

Table 6: Top 5 (among 17 unique careers) and 10 (among 169 unique majors) most frequent career and college major recommendations on the *MovieLens* and *Facebook* datasets, respectively, to the overall male and female users using NFCF and NCF w/o Pre-train models.

"typical" models, *NCF w/o Pre-train* and *MF w/o Pre-train*, respectively. Although it is intuitive to use a classification-based method to recommend sensitive items which typically occur only once such as careers, the results show that our *NFCF* and *NFCF_embd* methods comprehensively outperformed *Projection-based CF* and *DNN Classifier* in terms of all measures on both datasets.

As a further qualitative experiment, we recommended the top-1 career and college major to each test male and female user via the *NFCF* and *NCF w/o Pre-train* models. In Table 6, we show top 5 and 10 most frequent recommendations to the overall male and female users among the 17 and 169 unique careers and majors for *MovieLens* and *Facebook* dataset, respectively. *NFCF* was found to recommend similar careers to both male and female users on average for both datasets, while *NCF w/o Pre-train* encoded societal stereotypes in its recommendations. For example, *NCF w/o Pre-train* recommends *computer science* to male users and *nursing* to female users on the *Facebook* dataset while it recommends *executive/managerial* to male users and *customer service* to female users on the *MovieLens* dataset.

5 DISCUSSION AND FUTURE WORK

In this paper, we investigated gender bias in recommender systems trained on social media data for suggesting sensitive items (e.g. college majors or career paths). For social media data, we typically have abundant implicit feedback for user preferences of various *non-sensitive* items in which gender disparities are acceptable, or even desirable (e.g. "liked" Facebook pages, movies or music), but limited data on the *sensitive* items (e.g., users typically have only one or two college majors or occupations). User embeddings learned from the non-sensitive data can help recommend the sparse sensitive items, but may encode harmful stereotypes, as has been observed for word embeddings [8]. Furthermore, the distribution of sensitive items typically introduces further unwanted bias due to societal disparities in academic concentrations and career paths, e.g. from the "leaky pipeline" in STEM education [5].

We developed a practical solution for *gender* de-biased career recommendations while resolving the above challenges. Although we generally aimed to predict discrete class labels such as college majors, we intentionally framed the fair career recommendation task as a *recommender system problem* rather than a *classification problem*. We use this approach because as our results showed in Table 5, the personalized predictions in this task benefited from collaborative filtering, which outperformed classification baselines. Furthermore, the components of our proposed method such as debiasing embeddings, pre-training, fine-tuning, and fairness interventions via penalty term can potentially be transferred to other models, e.g. neural graph collaborative filtering [52], and applied directly to mitigate other demographic biases, e.g. *race*, *age*, and *nationality*.

In general, the disparate behavior of typical recommendation systems (e.g. see *NCF w/o Pre-train* in Table 6) may partly reflect legitimately differing real-world preferences in career choices by women and men. However, according to a report by the US Department of Commerce [5], gender disparity in STEM jobs can also be attributed to factors such as strong gender stereotypes and a lack of female role models, and reducing this gender disparity is an untapped opportunity to improve the economic competitiveness and innovative capacity of the USA, and to decrease the gender wage gap. The equitable predictions produced by *NFCF* are one step in this direction.

The main limitation of our approach is that it is designed and evaluated for a single protected attribute, i.e. *gender*. For multiple protected attributes, although it is straightforward to measure differential fairness-based penalty term [22], it is not clear how the bias direction can be computed accurately in the intermediate step between pre-training and fine-tuning. In future work, inspired by [43, 51, 56], we plan to address this limitation with an adversarial network included in the fine-tuning step which aims to make the user embeddings independent from multiple protected attributes simultaneously.

6 RELATED WORK

The recommender systems research community has begun to consider issues of fairness in recommendation. A frequently practiced strategy for encouraging fairness is to enforce *demographic parity* among different protected groups. Demographic parity aims to ensure that the set of individuals in each protected group have similar overall distributions over outcomes [55]. Some authors have addressed the unfairness issue in recommender systems by adding a regularization term that enforces demographic parity [7, 31–35]. However, demographic parity is only appropriate when user preferences have no legitimate relationship to the protected attributes. In recommendation systems for typical items such as movies, user preferences are indeed often influenced by protected attributes such as gender, race, and age [13]. Therefore, enforcing demographic parity may significantly damage the quality of recommendations. Fair recommendation systems have also been proposed by penalizing disparate distributions of prediction error [54], by making recommended items independent from protected attributes such as gender, race, or age [30], and by isolating protected attributes in tensor-based recommendations [58]. In addition, [10, 11] 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 different groups. Pareto efficiency-based fairness-aware group recommendation [53] was also proposed, however this method is not effective in personalized fair recommendations. Furthermore, a simple technique using fair tf-idf was recently proposed [17] to mitigate demographic bias in the AI-based resume screening process. Unlike previous methods, we develop neural network method for fair collaborative filtering on social media data that focuses on mitigating bias in career recommendations.

7 CONCLUSION

We investigated gender bias in social-media based collaborative filtering. To address this problem, we introduced Neural Fair Collaborative Filtering (*NFCF*), a pre-training and fine-tuning method which corrects gender bias for recommending sensitive items such as careers or college majors with little loss in performance. On the *MovieLens* and *Facebook* datasets, we achieved better performance and fairness compared to an array of state-of-the-art models.

ACKNOWLEDGMENTS

This work was performed under the following financial assistance award: 60NANB18D227 from U.S. Department of Commerce, National Institute of Standards and Technology. This material is based upon work supported by the National Science Foundation under Grant No.'s IIS1850023; IIS1927486. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Michelle Alexander. 2012. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* The New Press.
[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.
[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.
[4] Mathieu Bastian, Matthew Hayes, William Vaughan, Sam Shah, Peter Skomoroch, Hyungjin Kim, Sal Uryasev, and Christopher Lloyd. 2014. LinkedIn skills: large-scale topic extraction and inference. In *Proceedings of the 8th ACM Conference on Recommender systems.* 1–8.
[5] David N Beede, Tiffany A Julian, David Langdon, George McKittrick, Beethika Khan, and Mark E Doms. 2011. Women in STEM: A gender gap to innovation. *Economics and Statistics Administration Issue Brief* 04-11 (2011).
[6] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.
[7] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 2212–2220.
[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems.* 4349–4357.
[9] Steven Broad and Meredith McGee. 2014. Recruiting Women into Computer Science and Information Systems. *Association Supporting Computer Users in Education* (2014).
[10] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).
[11] Robin Burke, Nasim Sonboli, Masoud Mansoury, and Aldo Ordoñez-Gauger. 2017. Balanced Neighborhoods for Fairness-aware Collaborative Recommendation. In *FATREC Workshop on Responsible Recommendation Proceedings.* 5.
[12] A. Campolo, M. Sanfilippo, M. Whittaker, A. Selbst K. Crawford, and S. Barocas. 2017. *AI Now 2017 Symposium Report.* AI Now.
[13] Olivia Chausson. 2010. Who watches what?: Assessing the impact of gender and personality on film preferences. Paper published online on the MyPersonality project website.
[14] P.H. Collins. 2002 [1990]. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment (2nd ed.).* Routledge.
[15] K. Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *U. Chi. Legal F.* (1989), 139–167.
[16] Angela Y Davis. 2011. *Are prisons obsolete?* Seven Stories Press.
[17] Ketki V Deshpande, Shimei Pan, and James R Foulds. 2020. Mitigating Demographic Bias in AI-based Resume Filtering. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization.* 268–275.
[18] Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word Vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics.* 879–887.

[19] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on Fairness, Accountability and Transparency*. 172–186.

[20] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*. 278–288.

[21] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. Bayesian Modeling of Intersectional Fairness: The Variance of Bias. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 424–432.

[22] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1918–1921.

[23] Snorre S Frid-Nielsen. 2019. Find my next job: labor market recommendations using administrative big data. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 408–412.

[24] Francisco Gutiérrez, Sven Charleer, Robin De Croon, Nyi Nyi Htun, Gerd Goetschalckx, and Katrien Verbert. 2019. Explaining and exploring job recommendations: a user-driven approach for interacting with knowledge-based job recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 60–68.

[25] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *ACM transactions on interactive intelligent systems (TIIS)* 5, 4 (2015), 1–19.

[26] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 1661–1670.

[27] 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*. International World Wide Web Conferences Steering Committee, 173–182.

[28] Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in Statistics*. Springer, 492–518.

[29] Rashidul Islam, Kamrun Naher Keya, Shimei Pan, and James Foulds. 2019. Mitigating demographic biases in social media-based recommender systems. *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Social Impact Track)* (2019).

[30] Toshihiro Kamishima and Shotaro Akaho. 2017. Considerations on Recommendation Independence for a Find-Good-Items Task. In *FATREC Workshop on Responsible Recommendation Proceedings*.

[31] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Enhancement of the Neutrality in Recommendation. In *Decisions@ RecSys*. 8–14.

[32] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2013. Efficiency Improvement of Neutrality-Enhanced Recommendation. In *Decisions@ RecSys*. Citeseer, 1–8.

[33] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2014. Correcting Popularity Bias by Enhancing Recommendation Neutrality. In *RecSys Posters*.

[34] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Issei Sato. 2016. Model-based approaches for independence-enhanced recommendation. In *2016 IEEE 16th International Conference on Data Mining Workshops*. IEEE, 860–867.

[35] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.

[36] Krishnaram Kenthapadi, Benjamin Le, and Ganesh Venkataraman. 2017. Personalized job recommendation system at LinkedIn: Practical challenges and lessons learned. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 346–347.

[37] Os Keyes, Jevan Hutson, and Meredith Durbin. 2019. A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[38] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 426–434.

[39] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.

[40] 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.

[41] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019. iFair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering*. IEEE, 1334–1345.

[42] A. Lorde. 1984. Age, race, class, and sex: Women redefining difference. In *Sister Outsider*. Ten Speed Press, 114–124.

[43] Gilles Louppe, Michael Kagan, and Kyle Cranmer. 2017. Learning to pivot with adversarial networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 982–991.

[44] Stella Lowry and Gordon Macpherson. 1988. A blot on the profession. *British medical journal (Clinical research ed.)* 296, 6623 (1988), 657.

[45] Kamrun Naher Keya, Rashidul Islam, Shimei Pan, Ian Stockwell, and James R Foulds. 2020. Equitable Allocation of Healthcare Resources with Fair Cox Models. In *AAAI Fall Symposium on AI in Government and Public Sector*. AAAI FSS.

[46] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

[47] Philip Oreopoulos. 2011. Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy* 3, 4 (2011), 148–71.

[48] Ruslan Salakhutdinov and Andriy Mnih. 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning*. 880–887.

[49] S. Truth. 1851. Ain't I a Woman? Speech delivered at Women's Rights Convention, Akron, Ohio.

[50] Cia Verschelden. 2017. *Bandwidth recovery: Helping students reclaim cognitive resources lost to poverty, racism, and social marginalization*. Stylus Publishing, LLC.

[51] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).

[52] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*. 165–174.

[53] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-aware group recommendation with pareto-efficiency. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 107–115.

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

[55] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[56] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[57] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2979–2989.

[58] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1153–1162.