Unlearning Protected User Attributes in Recommendations with Adversarial Training

Christian Ganhör
christian.ganhoer@jku.at
Johannes Kepler University Linz
Austria

David Penz
david.penz@jku.at
Johannes Kepler University Linz and
TU Wien
Austria

Navid Rekabsaz
navid.rekabsaz@jku.at
Johannes Kepler University Linz and
Linz Institute of Technology
Austria

Oleg Lesota
oleg.lesota@jku.at
Johannes Kepler University Linz and
Linz Institute of Technology
Austria

Markus Schedl
markus.schedl@jku.at
Johannes Kepler University Linz and
Linz Institute of Technology
Austria

ABSTRACT

Collaborative filtering algorithms capture underlying consumption patterns, including the ones specific to particular demographics or protected information of users, e. g., gender, race, and location. These encoded biases can influence the decision of a recommendation system (RS) towards further separation of the contents provided to various demographic subgroups, and raise privacy concerns regarding the disclosure of users' protected attributes. In this work, we investigate the possibility and challenges of removing specific protected information of users from the learned interaction representations of a RS algorithm, while maintaining its effectiveness. Specifically, we incorporate adversarial training into the state-of-the-art MultVAE architecture, resulting in a novel model, *Adversarial Variational Auto-Encoder with Multinomial Likelihood* (Adv-MultVAE), which aims at removing the implicit information of protected attributes while preserving recommendation performance. We conduct experiments on the MovieLens-1M and LFM-2b-DemoBias datasets, and evaluate the effectiveness of the bias mitigation method based on the inability of external attackers in revealing the users' gender information from the model. Comparing with baseline MultVAE, the results show that Adv-MultVAE, with marginal deterioration in performance (w. r. t. NDCG and recall), largely mitigates inherent biases in the model on both datasets.

CCS CONCEPTS

• **Information systems** → **Collaborative filtering**; • **Computing methodologies** → **Neural networks**.

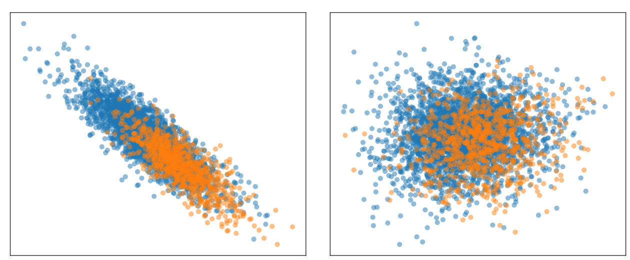
KEYWORDS

recommendation, adversarial training, gender bias, bias mitigation

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531820

ACM Reference Format:
Christian Ganhör, David Penz, Navid Rekabsaz, Oleg Lesota, and Markus Schedl. 2022. Unlearning Protected User Attributes in Recommendations with Adversarial Training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531820



(a) MultVAE **(b) Adv-MultVAE**

Figure 1: Output of an attacker network aiming to infer users' genders from the latent embeddings of the MultVAE and Adv-MultVAE models trained on LFM-2b-DemoBias [21] dataset. The blue and orange markers correspond to male and female users, respectively.

1 INTRODUCTION

In recommender systems (RSs), collaborative filtering algorithms provide recommendations for users (consumers), primarily based on the collected user-item interactions, e. g., through listening to music tracks or watching movies. Among these algorithms, MultVAE [19] learns to recommend items through decoding the variational encoding of user interaction vectors and has shown top results among a variety of deep neural network approaches [5]. While the interaction data does not explicitly contain information about protected user attributes such as gender, race, or age, a model may still encode sensitive information in its latent embeddings. This is depicted in Figure 1a, as the points regarding male and female users in a trained MultVAE model form fairly separated clusters of users according to their genders. These encoded biases in models can lead to strengthening "filter bubbles" based on the demographics of users [1, 8, 9, 28], and to intensifying the existing societal biases in data, thereby increasing unfairness of the RS [11, 21, 23, 24]. They can also raise privacy concerns regarding the disclosure of sensitive information from the recommendations or model parameters [2, 3, 29].

We approach this issue by proposing Adv-MultVAE, a novel bias-aware recommendation model which enhances MultVAE with adversarial training to reduce encoded biases. The Adv-MultVAE model, while learning to provide effective recommendations, simultaneously forces its latent embeddings to become invariant with respect to a given protected attribute of the consumers. This results in reducing the distinguishability of the sub-populations in the model (as shown in Figure 1b), hence making the recommendation "blind" to the protected attribute while maintaining the model's recommendation performance. We particularly adopt MultVAE, as it achieved top results among a variety of different deep neural network based approaches [5].

To assess the merits of our approach regarding both bias mitigation and recommendation performance, we conduct a set of experiments on the MovieLens-1m [14] and LFM-2b-DemoBias [21, 26] datasets covering the domains of movies and music, respectively. We focus on gender as the protected attribute and evaluate the accuracy and balanced accuracy of an attacker network to quantify the effect of bias mitigation. Moreover, we assess the models' recommendation performance via NDCG and recall. Adv-MultVAE successfully reduces inherent gender bias, whilst marginally decreasing performance mainly caused by the challenges imposed during model selection.

Brief Review of Related Work. As surveyed by Deldjoo et al. [6], adversarial training in combination with latent factor recommendation algorithms is investigated for various purposes by a few recent studies. In particular, Beigi et al. [2] propose a novel model based on Bayesian Personalized Ranking (BPR), which uses attacker networks to increase the model's privacy. In this model, the attacker networks aim to infer sensitive user information by looking at the output recommendations of the network, and the whole model is optimized such that no sensitive information can be inferred from the recommendations. Similarly, Zhang et al. [34] intend to mitigate biases of classifiers by utilizing adversarial networks, resulting in reducing the leakage of sensitive user attributes into the model predictions. In contrast to these studies, our proposed model aims to remove implicitly encoded sensitive information from its latent space rather than the output space. Moreover, unlike some approaches [18, 31] that apply filtering layers on top of their user embeddings to drop unwanted information, Adv-MultVAE is trained with the objective that the information of the protected attributes is removed from the model in the first place. Concerning bias mitigation in RSs, Zhu et al. [35] introduce the debiased personalized ranking model, in which the adversarial training aims to identify which item group, such as movie genre in the movie domain, the recommendation belongs to. This information is subsequently removed to mitigate *item popularity bias*. In contrast to this work, we study bias mitigation from the consumer side. More recently, based on adversarial training, Wu et al. [30] explore the mitigation of consumer bias in news recommendation, and several recent studies[16, 17, 23, 33] approach fairness in the representation of gender-related documents in information retrieval. Our work extends these studies by introducing a novel bias-aware recommendation model based on variational autoencoders.

The paper is structured as follows: We introduce Adv-MultVAE in Section 2. In Section 3, we present our experimental setup and the datasets and metrics we use to evaluate our approach. Section 4 provides an analysis of our results, which we extend by open challenges and limitations in Section 5. Finally, we conclude this work in Section 6. Our code together with all resources is available at **https://github.com/CPJKU/adv-multvae**

2 ADVERSARIAL MULTVAE

In this section, we describe the architecture of our *Adversarial Variational Auto-Encoder with Multinomial Likelihood* (Adv-MultVAE) model. We first provide an overview of the baseline MultVAE, followed by explaining our adversarial extension. We finally describe the procedure of adversarial attacking used to assess the effectiveness of bias mitigation. Figure 2 depicts the outline of the proposed Adv-MultVAE model.

MultVAE. The MultVAE model consists of two parts: First, the encoder network $f(\cdot)$ receives the input vector $\boldsymbol{x}$ containing the interaction data of a user and infers a low-dimensional latent distribution. Considering a standard Gaussian distribution as the prior ($\mathcal{N}(0, I)$) and using the *reparameterization trick* [15], this distribution is characterized by $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ learnable vectors, from which the latent vector $\boldsymbol{z}$ is sampled. [1] The second part is the decoder network $g(\cdot)$, which aims to reconstruct the original input $\boldsymbol{x}$ from the latent vector $\boldsymbol{z}$ by predicting $\boldsymbol{x'}$. We refer to the loss function of MultVAE as $\mathcal{L}^{\text{rec}}(\boldsymbol{x})$, defined below:

$$\mathcal{L}^{\text{rec}}(\boldsymbol{x}) = \mathcal{L}^{\text{MULT}}(g(z), \boldsymbol{x}) - \beta \mathcal{L}^{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}), \mathcal{N}(0, I)) \quad (1)$$

where $\mathcal{L}^{\text{MULT}}$ is the input reconstruction loss, and $\mathcal{L}^{\text{KL}}$ is the regularization loss aiming to keep the latent distribution of the encoder close to the prior, whose influence is adjusted by the hyperparameter β. We refer to Liang et al. [19] for more details.

Adv-MultVAE. Our proposed model extends MultVAE with an adversarial network, referred to as $h(\cdot)$. The adversarial network is added as an extra head over the latent vector and aims to predict from latent vector $\boldsymbol{z}$ a specific protected attribute of the user. $h(\cdot)$ is typically a feedforward network, which is optimized with respect to the $\boldsymbol{y}$ vector containing the user's protected attribute as classification labels. The training process of Adv-MultVAE aims to simultaneously remove the information of the protected attribute from $\boldsymbol{z}$, and maintain recommendation performance. To this end, the loss of the model is defined as the following min-max problem:

$$\arg\min_{f,g} \arg\max_{h} \mathcal{L}^{\text{rec}}(\boldsymbol{x}) - \mathcal{L}^{\text{adv}}(\boldsymbol{x}, \boldsymbol{y}) \quad (2)$$

where the loss of the adversarial network $\mathcal{L}^{\text{adv}}$ is defined as the cross-entropy loss ($\mathcal{L}^{\text{CE}}$) between the predicted and actual value of the protected attribute: $\mathcal{L}^{\text{adv}}(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{L}^{\text{CE}}(h(z), \boldsymbol{y})$. In fact, the

[1] In short, reparametrization trick allows sampling the random variable $\boldsymbol{z}$ by reparameterizing the sampling process with an auxiliary stochastic variable, thereby maintaining the ability to perform back-propagation on $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$.

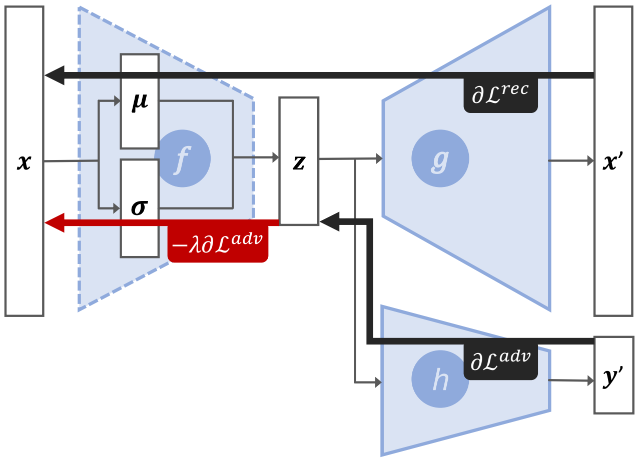


Figure 2: Outline of Adv-MultVAE. The bold arrows show the flow of gradients during backward pass, where the red color indicates the reversed gradient for learning latent embeddings (z) invariant to the protected attribute (y).

loss defined in Eq. 2 aims to maximize the prediction ability of $h(\cdot)$ to discover all sensitive information when $\boldsymbol{z}$ is given, while it minimizes the encoded information in $\boldsymbol{z}$ concerning the protected attribute.

Considering the well-known complexities of optimizing min-max loss function [13], following previous work [10, 23, 32], we add a gradient reversal layer $grl(\cdot)$ [12] between $\boldsymbol{z}$ and the adversarial network $h(\cdot)$. During training, $grl(\cdot)$ acts as the identity function in the forward pass, while it scales the calculated gradient by $-\lambda$ in the backward pass. The $grl(\cdot)$ network does not have any effect on the model at inference time. We refer to the parameter λ as *gradient reversal scaling*. By employing $grl(\cdot)$ in the model, the overall loss in Eq. 2 can now be reformulated to a standard risk minimization setting:

$$\begin{aligned} &\arg\min_{f,g,h} \mathcal{L} = \mathcal{L}^{\text{rec}}(\boldsymbol{x}) + \mathcal{L}^{\text{adv}}(\boldsymbol{x}, \boldsymbol{y}), \\ &\mathcal{L}^{\text{adv}}(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{L}^{\text{CE}}(h(grl(z), \boldsymbol{y}) \end{aligned} \tag{3}$$

This formulation enables optimizing the model through standard gradient-based loss minimization.

Adversarial Attacks. After training the model (whether Mult-VAE or Adv-MultVAE), we examine to which extent the information of the protected attribute remains in the model, i. e., to which degree this information can still be recovered. To this end, once the training is complete, an attacker network $h^{\text{atk}}(\cdot)$ is introduced to the model, which aims to predict the protected attribute $\boldsymbol{y}$ from the latent vector $\boldsymbol{z}$. Similar to $h(\cdot)$, the attacker $h^{\text{atk}}(\cdot)$ is defined as a feedforward network. During training the attacker, all model parameters remain unchanged (are frozen) and only the attacker parameters are updated. The prediction performance of the attacker – relative to a random predictor – is used as a metric to quantify the degree of bias in the underlying model.

Dataset		Users	Items	Interactions
ML-1M	All	6,040	3,416	999,611
	Male	4,331		753,313
	Female	1,709		246,298
LFM2b-DB	All	19,972	99,639	2,829,503
	Male	15,557		2,385,427
	Female	4,415		444,076

Table 1: Statistics of the datasets used in our experiments.

3 EXPERIMENT SETUP

In this section, we describe the setup of our experiments. To ensure reproducibility, our dataset splits, code and hyperparameters are available at **https://github.com/CPJKU/adv-multvae**.

Datasets. We evaluate our approach on two standardized datasets containing user-item interactions as well as partial demographic information of their users: **(1) MovieLens-1M (ML-1M) [14]**[2] contains ratings of users on movies as well as the users' gender, age, and occupation information. We binarize the interactions by setting the values of the rated items to one, and the rest to zero. Finally, we only keep the users that rated at least 5 movies, and the movies with at least 5 user interactions; **(2) LFM2b-DB [21]**[3] is a subset of the LFM-2b dataset, which provides a collection of music listening records of users, for whom partial demographic information (gender, age, country) is available. We follow the same experiment setting as in Melchiorre et al. [21]. In particular, we only keep the user-item interactions with a play count of at least 2, and binarize the interactions. Moreover, for computational reasons, 100,000 tracks are randomly sampled from the data. Finally, we keep only the users with at least 5 track interactions, and the tracks that are listened to at least 5 times. The statistics of the datasets are reported in Table 1. With both datasets, we focus on the users' gender as the protected attribute for our experiments.[4]

Data Splits. Following the exact setting of Melchiorre et al. [21], we apply a user split strategy [22]. In this setting, the users (and their corresponding interactions) are split into 5 folds for cross-validation, where 3 folds make up the training set, and 1 fold each makes up the validation and test set. For the training set, we further perform random upsampling of female users (as the minority group) to achieve an balanced dataset, which supports bias mitigation in models [21]. For validation and testing, the interactions in each set are further split: 80% are used as model input, the remaining 20% for calculating the evaluation metrics.

Evaluation. We use two popular recommendation performance metrics: recall@k, namely the fraction of relevant items in the top k recommended items, and NDCG@k, which weights the relevance of the top k recommended items based on their ranking positions. As common, we set the cut-off threshold k to 10. Additionally, we

[2] https://grouplens.org/datasets/movielens/1m
[3] http://www.cp.jku.at/datasets/LFM-2b
[4] The provided gender information of the users in the datasets are limited to female and male, neglecting the more nuanced definition of genders. Despite this limitation, the introduced model is generic and can be applied to non-binary settings too.

Dataset	Model	Bias↓		Performance↑	
		Acc	BAcc	NDCG	Recall
ML-1M	MultVAE$_{\text{Best}}$	0.692	0.707	**0.621**	**0.596**
	MultVAE$_{\text{Last}}$	0.699	0.693	0.591†	0.566†
	Adv-MultVAE	**0.565**	**0.572**	0.593†	0.569†
LFM2b-DB	MultVAE$_{\text{Best}}$	0.703	0.717	**0.211**	**0.192**
	MultVAE$_{\text{Last}}$	0.709	0.717	0.206†	0.189†
	Adv-MultVAE	**0.631**	**0.609**	0.206†	0.189†

Table 2: The results of bias mitigation in terms of accuracy and balanced accuracy of adversarial attackers (lower values indicate less bias), as well as recommendation performance (NDCG and recall). The best results are shown in bold. The sign † indicates a significant decrease in performance metrics in comparison to MultVAE$_{\text{Best}}$.[5]

measure the effectiveness of the models in terms of bias mitigation using the accuracy (Acc) and balanced accuracy (BAcc) of the attacker when predicting the users' gender. We use BAcc as a proper metric in imbalanced classification settings [4]. It reports the average recall per class (female/male) where a value of 0.50 indicates a fully debiased network. We report the performance and the bias mitigation results as the average over all test sets' results across cross-validation folds. We test the statistical significance of the differences of the performance metrics using the Wilcoxon signed-rank test [25] with a confidence level of 95%.

Models and Training. We train the MultVAE and Adv-MultVAE models for 100 and 200 epochs on LFM2b-DB and ML-1M, respectively. For MultVAE, we select the best performing model across the training epochs based on the validation NDCG results. We refer to this model as **MultVAE$_{\text{Best}}$**. For **Adv-MultVAE**, we conduct model selection based on the BAcc results of the adversarial network, mainly resulting in the selection of the model at the very last steps of training. To have a comparable setting between MultVAE and Adv-MultVAE, we also report the results of MultVAE when the selected model is at the last epoch, denoted as **MultVAE$_{\text{Last}}$**. Additionally, we perform a hyperparameter search over embedding size, parameter β, various dropouts, learning rate, weight decay, and gradient reversal scaling λ. In each setting, the best performing model on NDCG on the validation set is chosen and evaluated on the test set. The range of hyperparameters and the ones used for the final models are available together with our published code.

4 RESULTS AND ANALYSES

Table 2 reports the results of the bias mitigation and recommendation performance metrics. We report the best performing results in bold, and indicate the significant decrease in the results of performance metrics in comparison with MultVAE$_{\text{Best}}$ with the † sign.

[5] We omit testing for significance on Acc and BAcc as they are both metrics calculated on the whole test set, which would require running the experiments many times to gather sufficient data. However, a McNemar's test [20], which we use to determine whether the attackers on the different models achieve different performances, signals a significant difference of the results ($p < 0.05$).

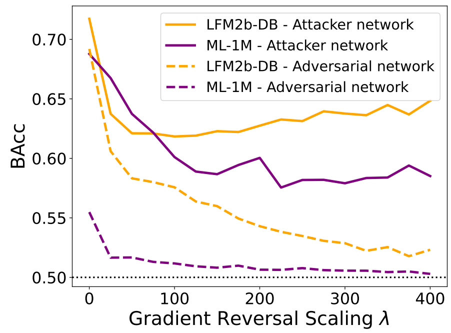
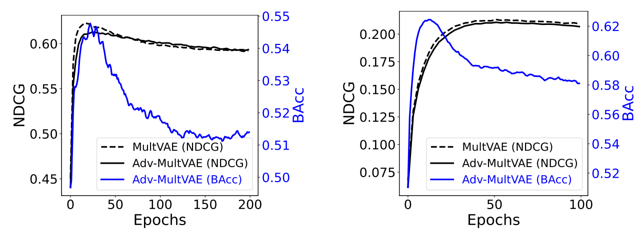


Figure 3: Balanced accuracy of adversarial and attacker networks over a range of gradient reversal scalings λ.



(a) ML-1M

(b) LFM2b-DB

Figure 4: Performance evaluation results on the validation sets during the training of models for both Adv-MultVAE and MultVAE (NDCG on the left y-axes). The BAcc of the adversarial network in Adv-MultVAE is shown based on the scale of the right y-axes.

Considering the balanced accuracy results of the standard MultVAE model in Table 2, we observe that the gender of the users can be identified from their consumption patterns with considerably high values, namely with ∼ 0.71 BAcc in ML-1M and ∼ 0.72 in LFM2b-DB, in comparison to 0.5 of a bias-free model. This confirms the existence of gender bias in MultVAE and consequently in its provided recommendations, as also similarly reported in previous studies [7, 21, 27]. Looking at the results of Adv-MultVAE, we observe a significant decrease in the balanced accuracy of attackers, indicating the effectiveness of Adv-MultVAE for mitigating encoded societal biases. Despite the large decreases, we should note that the predictions are not yet fully random (BAcc > 0.5), indicating that the model still contains considerable biases.

Let us have a closer look at the impact of gradient reversal scaling λ on the bias mitigation method. Figure 3 shows the balanced accuracy of the attackers as well as adversarial networks for both datasets over a range of λ values. As shown, by increasing λ, the BAcc of the adversarial networks generally decrease and saturate at some point. In particular, the adversarial network of the model trained on ML-1M much earlier reaches the BAcc of close to 0.5. However, the attackers' BAcc are consistently higher than the ones of adversaries, indicating that the adversarial network could not discover and remove all sensitive information.

Impact of adversarial training on performance. Comparing the performance evaluation results of Adv-MultVAE to that of $\text{MultVAE}_{\text{Best}}$ in Table 2, we observe a significant drop in NDCG and recall on both datasets. At first, this drop might be seen as the *cost* of applying adversarial bias mitigation. However, we should consider that the model selection of Adv-MultVAE is done based on the lowest value of BAcc rather than the highest NDCG (as in MultVAE). This might not be an ideal model selection criterion as it ignores a potential trade-off between BAcc and NDCG, and under-emphasizes recommendation performance of Adv-MultVAE.

To better understand this, Figure 4 shows the NDCG results on the validation sets of the two datasets for both, Adv-MultVAE and MultVAE models, over the training epochs. The figure also depicts BAcc of the adversarial network of Adv-MultVAE. As shown, both models achieve higher values of NDCG in early epochs, and by continuing the training, the performances slightly decrease (due to overfitting). However, BAcc in Adv-MultVAE considerably decreases in later epochs, where NDCG has already decreased due to the slight overfitting.

In order to examine the performances while factoring out the effect of model selection and thereby enabling a fairer comparison, we compare the recommendation results of $\text{MultVAE}_{\text{Last}}$ with Adv-MultVAE. As reported in Table 2, we observe no significant differences in NDCG and recall between these two models, for both datasets.

5 OPEN CHALLENGES AND LIMITATIONS

The analysis in the previous section highlights the inherent challenge in training adversarial networks for bias mitigation. Since this method aims to simultaneously satisfy two objectives (increasing recommendation performance and decreasing bias), it is challenging to train and select the right model that optimizes both aspects.

We should also shed light on two limitations of our adversarial bias mitigation method for RSs. First, our adversarial approach aims to reduce the correlations in the model to the protected attribute based on the observed data. This approach, like other data-oriented bias mitigation methods, might lack effective generalization, particularly when the model is evaluated on other domains or out-of-distribution data. The second limitation of adversarial bias mitigation for RSs is that, while the method aims to make the recommendations agnostic to protected attributes, it does not directly account for the perception of users regarding the bias in the recommended results.

6 CONCLUSION AND FUTURE DIRECTIONS

This work addresses the challenge of mitigating societal biases in RSs from the user perspective. To this end, we extend the widely-used MultVAE model with an adversarial component, and propose the novel Adv-MultVAE architecture. Our approach aims to decrease the model bias in terms of latent information about the protected user attribute in the model and consequently also in the provided recommendations. We conduct experiments on two datasets (ML-1M and LFM2b-DB) and evaluate the amount of recoverable sensitive user information (gender in our experiments) from the models, as well as the models' recommendation accuracy. Our results show that the introduced Adv-MultVAE model, despite yielding a marginal performance decrease which is mainly caused by the dilemma in model selection, provides a substantial reduction in the amount of encoded protected information, offering a bias- and privacy-aware alternative.

We envision addressing the limitation discussed in the previous section as future directions of this work, particularly by exploring the generalization aspects of the method, as well as the perception of end-users regarding the biases in recommendations. We would also like to gain further insights into which user groups specifically are affected by our approach. Moreover, finding a balance between BAcc and NDCG for optimization and model selection might be a possibility to mitigate the slight performance loss of Adv-MultVAE.

ACKNOWLEDGMENTS

This work received financial support by the Austrian Science Fund (FWF): P33526 and DFH-23; and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research, through grant LIT-2020-9-SEE-113 and LIT-2021-YOU-215.

REFERENCES

[1] Christine Bauer. 2019. Allowing for equal opportunities for artists in music recommendation. *CoRR* abs/1911.05395 (2019). arXiv:1911.05395 http://arxiv.org/abs/1911.05395

[2] Ghazaleh Beigi, Ahmadreza Mosallanezhad, Ruocheng Guo, Hamidreza Alvari, Alexander Nou, and Huan Liu. 2020. Privacy-Aware Recommendation with Private-Attribute Protection using Adversarial Learning. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 34–42. https://doi.org/10.1145/3336191.3371832

[3] Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. 2019. Privacy Preserving Text Representation Learning. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media, HT 2019, Hof, Germany, September 17-20, 2019*, Claus Atzenbeck, Jessica Rubart, and David E. Millard (Eds.). ACM, 275–276. https://doi.org/10.1145/3342220.3344925

[4] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The Balanced Accuracy and Its Posterior Distribution. In *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*. IEEE Computer Society, 3121–3124. https://doi.org/10.1109/ICPR.2010.764

[5] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 101–109. https://doi.org/10.1145/3298689.3347058

[6] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2021. A Survey on Adversarial Recommender Systems: From Attack/Defense Strategies to Generative Adversarial Networks. *ACM Comput. Surv.* 54, 2 (2021), 35:1–35:38. https://doi.org/10.1145/3439729

[7] 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, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 172–186. http://proceedings.mlr.press/v81/ekstrand18b.html

[8] Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 242–250. https://doi.org/10.1145/3240323.3240373

[9] Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Erik Knudsen, Helle Sjøvaag, Kristian Tolonen, Øyvind Holmstad, Igor Pipkin, Eivind Throndsen, Agnes Stenbom, Eivind Fiskerud, Adrian Oesch, Loek Vredenberg, and Christoph Trattner. 2021. Towards responsible media recommendation. *AI and Ethics* (02 Nov 2021). https://doi.org/10.1007/s43681-021-00107-7

[10] Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical*

Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 11–21. https://doi.org/10.18653/v1/d18-1002

[11] Lisette Espín-Noboa, Claudia Wagner, Markus Strohmaier, and Fariba Karimi. 2021. Inequality and Inequity in Network-based Ranking and Recommendation Algorithms. *CoRR* abs/2110.00072 (2021). arXiv:2110.00072 https://arxiv.org/abs/2110.00072

[12] Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (JMLR Workshop and Conference Proceedings, Vol. 37)*, Francis R. Bach and David M. Blei (Eds.). JMLR.org, 1180–1189. http://proceedings.mlr.press/v37/ganin15.html

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 2672–2680. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[14] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (dec 2015), 19 pages. https://doi.org/10.1145/2827872

[15] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.6114

[16] Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2022. Grep-BiasIR: A Dataset for Investigating Gender Representation-Bias in Information Retrieval Results. *arXiv preprint arXiv:2201.07754* (2022).

[17] Klara Krieg, Emilia Parada-Cabaleiro, Markus Schedl, and Navid Rekabsaz. 2022. Do Perceived Gender Biases in Retrieval Results Affect Relevance Judgements?. In *Proceedings of the Workshop on Algorithmic Bias in Search and Recommendation at the European Conference on Information Retrieval (ECIR-BIAS 2022).*

[18] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards Personalized Fairness Based on Causal Notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, Virtual Event Canada, 1054–1063. https://doi.org/10.1145/3404835.3462966

[19] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 689–698. https://doi.org/10.1145/3178876.3186150

[20] Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 2 (01 Jun 1947), 153–157. https://doi.org/10.1007/BF02295996

[21] Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. *Inf. Process. Manag.* 58, 5 (2021), 102666. https://doi.org/10.1016/j.ipm.2021.102666

[22] Zaiqiao Meng, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2020. Exploring Data Splitting Strategies for the Evaluation of Recommendation Models. In *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, Rodrygo L. T. Santos, Leandro Balby Marinho, Elizabeth M. Daly, Li Chen, Kim Falk, Noam Koenigstein, and Edleno Silva de Moura (Eds.). ACM, 681–686. https://doi.org/10.1145/3383313.3418479

[23] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 306–316. https://doi.org/10.1145/3404835.3462949

[24] Navid Rekabsaz and Markus Schedl. 2020. Do Neural Ranking Models Intensify Gender Bias?. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 2065–2068. https://doi.org/10.1145/3397271.3401280

[25] Denise Rey and Markus Neuhäuser. 2011. Wilcoxon-Signed-Rank Test. In *International Encyclopedia of Statistical Science*, Miodrag Lovric (Ed.). Springer, 1658–1659. https://doi.org/10.1007/978-3-642-04898-2_616

[26] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2b: A Dataset of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis. In *ACM SIGIR Conference on Human Information Interaction and Retrieval.* 337–341.

[27] Markus Schedl, David Hauger, Katayoun Farrahi, and Marko Tkalcic. 2015. On the Influence of User Characteristics on Music Recommendation Algorithms. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings (Lecture Notes in Computer Science, Vol. 9022)*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.). 339–345. https://doi.org/10.1007/978-3-319-16354-3_37

[28] Ningxia Wang and Li Chen. 2021. User Bias in Beyond-Accuracy Measurement of Recommendation Algorithms. In *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Huurnink, and Even Oldridge (Eds.). ACM, 133–142. https://doi.org/10.1145/3460231.3474244

[29] Udi Weinsberg, Smriti Bhagat, Stratis Ioannidis, and Nina Taft. 2012. BlurMe: inferring and obfuscating user gender based on ratings. In *Sixth ACM Conference on Recommender Systems, RecSys '12, Dublin, Ireland, September 9-13, 2012*, Padraig Cunningham, Neil J. Hurley, Ido Guy, and Sarabjot Singh Anand (Eds.). ACM, 195–202. https://doi.org/10.1145/2365952.2365989

[30] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairness-aware News Recommendation with Decomposed Adversarial Learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021.* AAAI Press, 4462–4469. https://ojs.aaai.org/index.php/AAAI/article/view/16573

[31] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. 2021. Learning Fair Representations for Recommendation: A Graph-based Perspective. In *Proceedings of the Web Conference 2021.* ACM, Ljubljana Slovenia, 2198–2208. https://doi.org/10.1145/3442381.3450015

[32] Qizhe Xie, Zihang Dai, Yulun Du, Eduard H. Hovy, and Graham Neubig. 2017. Controllable Invariance through Adversarial Feature Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 585–596. https://proceedings.neurips.cc/paper/2017/hash/8cb22bdd0b7ba1ab13d742e22eed8da2-Abstract.html

[33] George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022. Mitigating bias in search results through set-based document reranking and neutrality regularization. In *Proceedings of the 45th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2022.* ACM.

[34] 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.* ACM, New Orleans LA USA, 335–340. https://doi.org/10.1145/3278721.3278779

[35] Ziwei Zhu, Jianling Wang, and James Caverlee. 2020. Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 449–458. https://doi.org/10.1145/3397271.3401177