F2VAE: A Framework for Mitigating User Unfairness in Recommendation Systems

Rodrigo Borges
rodrigo.borges@tuni.fi
Tampere University
Tampere, Finland

Kostas Stefanidis
konstantinos.stefanidis@tuni.fi
Tampere University
Tampere, Finland

ABSTRACT

Recommendation algorithms are widely used nowadays, especially in scenarios of information overload (i.e., when users have too many options to choose from), due to their ability to suggest potentially relevant items to users in a personalized fashion. Users, nevertheless, might be considered as separated in groups according to sensitive attributes, such as age, gender or nationality, and the recommendation process might be biased towards one of these groups. If observed, this bias has to be mitigated actively, or it can propagate and be amplified over time. Here, we consider a relevant difference of recommendation quality among groups as unfair, and we argue that this difference should be maintained as low as possible. We propose a framework named F2VAE for mitigating user-oriented unfairness in recommender systems. The framework is based on Variational Autoencoders (VAE) and it introduces two extra terms in VAE's standard loss function, one associated to fair representation and another one associated to fair recommendation. The conflicting objectives associated to these terms are discussed in details in a series of experiments considering the bias associated to the users' nationality in a music consumption dataset. We recall recent works proposed for generating fair representations in the context of classification, and we adapt one of these methods to the recommendation task. F2VAE was able to increase the precision by approximately 1% while reducing the unfairness by 21% when compared to standard VAE.

KEYWORDS

Recommender Systems; Bias; Fair Representation; User Fairness; Fairness

1 INTRODUCTION

Recommender systems are personalized systems that were proposed for helping users navigating in large collections of items hosted typically online. Different from search engines, which retrieve the best possible results for a given query, recommenders are trained with historical user/item interactions information, and produce a list of potential relevant items for each user separately. These personalized systems are widespread in the internet nowadays and can been seen in several contexts, from job seeking to music streaming platforms.

The most popular approach applied by recommender systems, named *Collaborative Filtering* (CF), associates each user with a consumption profile, containing all items with which they interacted in the past. Similar profiles are assumed as indicating similarity of preferences, and are used as a resource for calculating suggestions to a target user. It can happen that user profiles are biased towards one specific user attribute considered sensitive, e.g., gender, age, nationality. For instance, we can imagine users from a certain age range being interested in a specific category of products, or users from one specific country listening mostly to certain artists. In situations like these, user profiles from one specific group can be biased for containing specific consumption patterns, and this can naturally happen in any recommendation scenario.

A problematic situation, however, would be the one in which worse recommendation results were systematically delivered to a specific group of users. We can imagine, as an example, a situation in which users from all countries are predominantly satisfied with the suggestions provided by a recommendation algorithm, except from one country, whose users are never satisfied. In this work, we assume a difference in recommendation quality that is frequently observed in a given scenario as systematic, and we consider a situation like this as *unfair*. More specifically, we argue that, ideally, the quality of recommendation results should vary as less as possible among groups of users.

The topic of fairness in algorithmic systems have been extensively discussed in classification [6, 8], ranking [1, 3, 22] and recommendation [17, 21, 24, 25] domains. Among the ones dedicated to measuring and mitigating unfairness from recommendation results, some consider the perspective from the items being recommended [24], some consider the perspective from the users [11, 16], and some consider both [5, 19]. In this work, we are interested in recommendation situations where users are considered as belonging to groups, like in [16], but instead of considering a binary attribute, we expand the notion of unfairness to any attribute that can be separated in categories.

In situations where user profiles[1] contain or can be associated to any sensitive attribute, an auxiliary and neutral representation (i.e., a *fair representation*) might be needed [13, 15, 20, 23]. The main idea here is to calculate a representation for a user profile containing the maximum of information associated with attributes considered non sensitive, while suppressing any potential proxy to any attribute considered sensitive. We adapt one fair representations technique to the task of recommendation, and we discuss to what extent fair representations imply fairness in the recommendation results.

We propose a framework for mitigating user-oriented unfairness based on Variational Autoencoders (VAE) [10]. VAEs were demonstrated as powerful methods for large scale recommendations [12], that apply an encoder/decoder neural network architecture: user/item interaction data is presented in the input, converted

[1] *User profile* and *user attributes* are used interchangeably here. The reader should notice, however, that in many applications users are considered as a list of their attributes (e.g., age, name, zip code, etc.), and in a recommendation scenario users are usually considered as binary lists indicating the items with which they interacted in the past.

to low-dimensional embedding (latent factors), and expanded back to its original dimension. The output is compared with the input, an error is measured, and parameters are adjusted using *backpropagation* [18]. Latent factors are considered as representations of the input data, to which fair representation techniques are applied.

Our contributions are:

- We adapt a method proposed in the domain of fair classification to the task of recommendation. The original method was designed for generating fair representations, and we analyze to which extent both objectives, representation and recommendation, are correlated.
- We propose a new loss function that is responsible for mitigating user bias, here understood as unfairness, from latent factor recommendation models.
- We propose a framework that combines fair representation and fair recommendation results in a single loss function, and we analyze how these objectives are conflicting during the training process.

This work is structured as follows. We start with a review of previous works on the topics of fairness-aware recommendation algorithms, and fair representations, in Section 2. In Section 3, a short description of VAE's assumptions and loss function terms are presented. User-oriented fairness in a recommendation task is formalized in Section 4. The framework for mitigating user unfairness is presented in Section 5, together with some intuition on how different objectives can conflict with each other. Experiments preparation and results are presented, respectively, in Sections 6 and 7. We make some final remarks and conclusions in Section 8.

2 RELATED WORK

2.1 Fairness in Recommendation

Many metrics and methods were already proposed for measuring and mitigating unfairness in the context of recommendation systems. Some works consider the perspective from the items being recommended [24], some consider the perspective from the users [11, 16], and some consider both [5, 19].

A tensor factorization method is proposed in [24], that is capable of isolating and suppressing sensitive user attributes during the recommendation process. *MADr (Mean Average Difference - rating)* is the absolute difference between mean ratings of different groups, assuming users separated in two groups. Our notion of fairness is similar to the one proposed here, except that we are interested in attributes that can assume several values, instead of just two.

It is worth mentioning that items, as well as users, might be separated in groups, according to some of specific features. Uneven distributions of groups of items among groups of users (e.g. people for a certain age range who watch mostly scifi movies) can be also considered unfair [19]. *BS (Dataset Bias)*, *BR (Recommendation Bias)* and *BD (Bias Disparity)* were suggested for measuring this differences, assuming categories of items and protected users groups.

Users might be separated in groups based on the their level of interaction with the recommender system, and the suggestions offered to these groups might be biased, according to [11]. Uneven recommendation results might be also offered to users considered as male or female [16]. The authors submitted several recommendation algorithms to the task of recommending music to users, and they evaluated the fairness in the results provided by these methods considering users as male or female. An unfair recommendation result is considered as the one that varies among user groups, and they demonstrate how biased these algorithms can be. The dataset applied here is also presented as contribution in that same work.

2.2 Fair Representation

Learning Fair Representations (LFR) [23] discusses fair representation in the context of classification, proposes a loss function combining demographic parity, the error measured in the reconstruction of the input data, and classification accuracy.

In [15] the authors were inspired by three fairness metrics proposed previously in the literature: demographic parity, equalized odds, and equal opportunity. The main aim is to mitigate unfair prediction results by learning fair representation of the input data. An adversarial objective is provided for each metric and the model is optimized according to these objectives.

Considering also the situation of fair classification, [7] presents an approach for learning flexible representations that minimize the capability of an adversarial critic. The authors propose Adversarial Learned Fair Representations (ALFR), and test their solution in situations of making decisions free from discrimination by removing private information from images. An stochastic gradient alternate min-max optimizer is proposed to ensure that little or no information about the sensitive variable is present in the representation. The results obtained reflect the method's ability to provide discriminant free representations for standard classification problems.

A Graph-Based approach for Fair Representation (FairGO) is presented in [20]. The method is informed with a sensitive feature set, it takes the user and item embeddings from any recommendation models as input, and it learns a filter space to obfuscate any sensitive information in the sensitive attribute set. The bias is reduced and the recommendation accuracy is maintained, considered as an important goal in fair representation learning.

3 BACKGROUND

VAE. Variational Autoencoders (VAE) are derived from inference models [10] and can be considered as a dimensionality reduction technique, due to its encoder/decoder architecture. VAEs usually assume a Gaussian latent space $\mathbf{z}$ ($p(\mathbf{z}) = \mathcal{N}(\mu, \sigma)$), and the encoded version of input $\mathbf{x}$ is expressed as $p(\mathbf{z}|\mathbf{x})$. The latent variables are decoded back, trying to approximate as much as possible the input $\mathbf{x}$, expressed as $p(\mathbf{x}|\mathbf{z})$.

VAE's loss function combines two terms, one corresponding to the accuracy of input reconstruction, and another for the proximity between the encoded and a Gaussian prior distributions. The accuracy in the specific objective of reconstruction the input is measured as:

$$E_\theta = \mathbb{E}[\log p_\theta(\mathbf{x}|\mathbf{z})] \tag{1}$$

The encoder parameters, which will be optimized during the training process, are denoted as θ. E_θ generates negative values when $p_\theta(\mathbf{x}|\mathbf{z})$ is lower than 1, and in many cases E_θ is presented with a negative sign and submitted to a minimization process. We

maintain a positive sign for the sake of presentation, as the target of a maximization process.

A small divergence between $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$ encourages latent variables to be learnd as Gaussian distributions. The proximity between both distributions act as a regularization factor and can measured with Kullback-Leibner divergence:

$$D_\phi = D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{2}$$

The decoder parameters will also be optimized during the training process, and are denoted as ϕ. When encouraged to learn the first term (E_θ), the model basically adjusts its predictions to the ground truth data, through $p(\mathbf{x}|\mathbf{z})$; and when to the second term (D_ϕ) is given more weight, the model approximates a Gaussian distribution to the one observed in the data, using the Kullback-Leibler divergence.

That said, the parameters are learned according to:

$$\max_\theta \min_\phi L_{\theta,\phi} = E_\theta + D_\phi \tag{3}$$

The performance of VAE models depend on the expressiveness of the inference model, and a good balance between both objectives, E_θ and D_ϕ, will determine its capacity of approximating the decoded version of the latent variables to the original input.

β-VAE. A variation of VAE was proposed in the context of *disentangled representations*, assuming that the data generated according to $p_\theta(\mathbf{x}|\mathbf{z})$ is generated from a fixed number of independent factors. Disentangled representations are defined in [2] as: a representation where a change in one dimension corresponds to a change in one factor of variation, while being relatively invariant to changes in other factors.

The main idea in β-VAE is to encourage $\mathbf{z}$ components to be independent, and one possible strategy for obtaining this is upweighting the KL-divergence term in VAE's loss function. We assume a parameter $\beta > 1$ multiplying D_ϕ to a certain extent [9], expressed as:

$$\max_\theta \min_\phi L_{\theta,\phi} = E_\theta + \beta \cdot D_\phi \tag{4}$$

β-VAEs were already explored in the context of recommendation [12], but instead of encouraging the independence of $\mathbf{z}$ factors, the authors are proposing to set β values lower than 1, in order to reduce regularization in the latent space. The authors reported that setting β equals to 1 would interfere in the learning process of the recommender, lowering its overall performance.

In this same work, the latent representation of a single user ($\mathbf{z}_u$) is transformed by a nonlinear function $f(\cdot) \in \mathbb{R}^I$, to produce a probability distribution $\pi(\mathbf{z}_u)$ over N items, and the log-likelihood is given by:

$$\log p(\mathbf{x}_u|\mathbf{z}_u) = \sum_{i=0}^{N} x_{ui} \log \pi_i(\mathbf{z}_u) \tag{5}$$

4 USER-ORIENTED FAIRNESS IN RECOMMENDATION

In the Collaborative Filtering (CF) scenario, user preferences are typically represented by a *rating matrix* containing information of how many times each user interacted with each item. Formally, assume a dataset of items, $i \in I$, available to users, $u \in U$, and a matrix $\mathbf{R} \in \mathbb{N}^{|U|\times|I|}$ containing a numerical value (feedback) for each *(user-item)* pair, where *user* interacted at some point with *item*. User profiles are usually assumed as the rows of the rating matrix ($\mathbf{x}_u$).

We assume that users are associated to sensitive attributes. S contains potential values for a sensitive attribute s, for example $S = \{male, female\}$, and U_s refers to all users associated to sensitive attribute s. We also assume $\mathcal{F}(u)$ as a function for measuring the quality in the recommendation results for user u, i.e., the score given by user u to the items suggested to him/her.

We assume θ and ϕ as the parameters of a recommendation model, as mentioned in Section 3, and we want to maximize:

$$\max_{\theta,\phi} \sum_{u\in U} \mathcal{F}(u) \tag{6}$$

The recommendation results, however, will vary among groups of users, and we want to ensure that the difference between recommendation results has the lowest difference possible. Thus, we want to minimize:

$$\min_{\theta,\phi} \sum_{s_i\in S} \sum_{s_j\in S} |\mathcal{F}(U_{s_i}) - \mathcal{F}(U_{s_j})| \tag{7}$$

More specifically, we assume a recommender implemented as two consecutive neural networks, $f_\theta(\cdot)$ and $f_\phi(\cdot)$, the first one is responsible for mapping the input ($\mathbf{x}$) to a lower-dimensional latent space ($\mathbf{z}$), and the second one is responsible for decoding this variable back to its original dimension ($\hat{\mathbf{x}}$). Parameters ϕ and θ are adjusted in order to approximate as much as possible the output and the input. The recommendation process can be summarized as encoding the input with $f_\theta(\mathbf{x}) \rightarrow \mathbf{z}$, and decoding $\mathbf{z}$ back with $f_\phi(\mathbf{z}) \rightarrow \hat{\mathbf{x}}$.

5 FAIR REPRESENTATION AND FAIR RECOMMENDATION VARIATIONAL AUTOENCODERS

We now present *ALFR-VAE*, *FaiRVAE*, and *F2VAE*, dedicated respectively to fair representation, fair recommendation and a combinations of both. ALFR-VAE is inspired in Adversarial Learned Fair Representations (ALFR) [7], originally proposed for reducing discrimination associated to a binary user attribute in the context of fair classification. We have adapted the idea to categorical sensitive attributes, and to the context of recommender systems. FaiRVAE aggregates a new term for measuring the bias in the recommendation results, and it is focused on mitigating systematic differences between recommendation results offered to different user groups. F2VAE combines the two terms associated to the two objectives, reducing unfairness in representation and in recommendation results according to one single loss function. This method was designed according to an intuition that fair representations do not necessarily imply in fair recommendation results. User profiles are being neutralized from the perspective of a certain sensitive attribute, and the differences between accuracy observed for groups of users are being reduced at the same time.

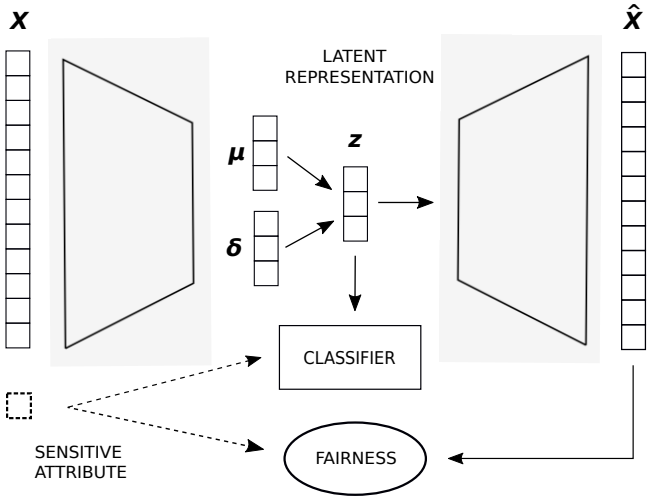

Figure 1: The general scheme of a F2VAE recommender network.

ALFR-VAE. is inspired in Adversarial Learned Fair Representations (ALFR) [7], originally proposed in the context of fair classification. ALFR quantifies how discriminative the representation is considering a binary sensitive attribute, and proposes a binary log-loss for measuring the bias. We are also interested in how discriminative $\mathbf{z}$ is for the prediction task, but using the categorical log-loss of a classifier network trained to predict s from $\mathbf{z}$. The classifier has its parameters adjusted together with the recommender's, and its accuracy is calculated with:

$$C_\theta(\mathbf{z}, s) = -\mathbb{E}\Big[\sum_{i=1}^{N} s \cdot \log(Pred_i(\mathbf{z}))\Big] \tag{8}$$

We consider the classifier parameters also as θ, N classes of a sensitive attribute, and $Pred_i$ as the Softmax probability for i^{th} class. The new fairness-aware recommendation loss function is given by:

$$\max_\theta \min_\phi L_{\theta,\phi} = E_\theta + D_\phi + C_\theta \tag{9}$$

The main idea here is to increase the classification loss, in order to reduce as much information as possible of a specific sensitive attribute.

FaiRVAE. A systematic difference between accuracy in the recommendation results delivered to different groups of users is considered here as unfair. We recall that users are separated in groups according to attributes considered sensitives (i.e., gender, nationality, age) and we want to ensure that, at each training step, big differences of reconstruction accuracy are neutralized.

In Fair Recommendation VAE (FaiRVAE), we aggregate a new term for mitigating user bias in the recommendation results, expressed as:

$$F_\phi(u) = \Big|\frac{1}{|U_s|}\sum_{c\in U_s}\mathbb{E}[\log p_\theta(\mathbf{x}_c|\mathbf{z}_c)] - \frac{1}{|U|}\sum_{u\in U}\mathbb{E}[\log p_\theta(\mathbf{x}_u|\mathbf{z}_u)]\Big| \tag{10}$$

where s is the sensitive attribute to which user u is associated to. The final loss function can be expressed as:

$$\max_\theta \min_\phi L_{\theta,\phi} = E_\theta + D_\phi + F_\phi \tag{11}$$

F2VAE. In Fair Representation and Recommendation via Variational Autoencoders (F2VAE), we aggregate all terms in a single loss function. The potential issue in this process is that if any of these objectives is conflicting with another, then the optimization process might not converge.

The final loss function is expressed as:

$$\max_\theta \min_\phi L_{\theta,\phi} = E_\theta + \beta \cdot D_\phi + \gamma \cdot C_\theta + \tau \cdot F_\phi \tag{12}$$

where E_θ is the proximity between input and output, that needs to be maximized; D_ϕ is the divergence between the inner representation and the approximate distribution, that need to be minimized; C_ϕ is the prediction error measured for the classifier that tries to guess the sensitive attribute, that needs to be maximized; and F_ϕ is the user unfairness, that needs to be minimized. A general scheme is presented in Figure 1.

We included two new hyperparameters, γ and τ, for controlling the strength of each new term, associated to fair representations and fair recommendation results respectively. The main intuition that led to this loss function is that neither latent space regularization leads necessarily to fair representation, nor fair representation leads necessarily to fair recommendations results. Moreover, gathering all these terms together would increase the chances that the model weights will be adjusted to achieve one single objective: mitigate unfairness associated to users' sensitive attributes.

6 EXPERIMENTAL SETUP

6.1 Data Preparation

We elected LFM-2b dataset [16] to be used in our experiments. The dataset contains music listening habits extracted from LastFM[2] platform, and is, to our knowledge, the only dataset available for recommendation experiments that is substantially big and that contains demographic (gender, nationality and age) information associated to users. Not all users provided their demographic information, and then, we started by filtering the ones who provided gender, nationality and age information. We ended with 40,374 users in total.

We removed duplicated records, i.e. user/track events happening more than once, for obtaining binary associations between users and tracks. We removed also users with less than 10 interactions, and tracks with less than 100 interactions, considered as not having enough interaction data. We ended with 40,374 users, 767,304 tracks and 171,381,362 listening events.

[2]https://www.last.fm/

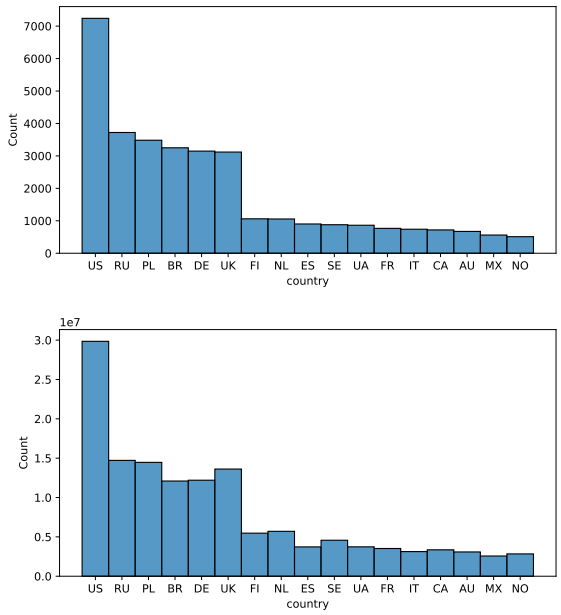

Figure 2: (Top) The number of users associated to each country in a LFM-2b subset. (Bottom) The number of interactions associated to each country in the same subset.

We did not want to restrict our experiments to attributes containing binary or continuous values, and we decided to user nationality as a sensitive attribute. We maintained the labels of countries associated with more than 500 users (shown in the top of Figure 2), and the remaining countries were grouped together in a single category. In the end, we had 18 possible values for the nationality. We measured also the number of interactions associated to each country for checking any correlation between recommendation accuracy and number of interactions.

The recommenders were implemented using Pytorch[3], and the network architecture was inspired in [12], but with layers with higher dimensions: $[767, 303 \rightarrow 2000 \rightarrow 200 \rightarrow 2000 \rightarrow 767, 303]$. In the case of ALFR-VAE, a linear layer is added to the network architecture, with dimensions $[200 \rightarrow 18]$, corresponding to the classifier indicated in Figure 1, named *adversarial classifier*. The latent variable calculated for each user profile is submitted to the classifier, that is trained to predict the corresponding user nationality. But in fact, the classification loss is being maximized, and the recommender will then be encouraged to build latent representations that are independent of this specific attribute value[4].

All models were trained for 100 epochs, and performance and fairness metrics were measured at every epoch applied to the data split separated for validation. The batch size was set to 500, within which F_ϕ (Equation 10) was calculated. The learning rate was set to 1e-4, and it decreases by a factor of 0.1 in epochs 50 and 75. The KL divergence factor was gradually introduce during the training, as proposed in [12], with the help of an annealing factor. The idea is to start with a low β value so the model has time to learn its weights according to its accuracy term, before applying regularization.

After all models were already trained, a new classifier (*auxiliary classifier*) was designed for classifying latent variables in their original category, i.e. user nationality. The intuition in here is that unbiased latent representations would lead to low classification accuracy, because of not containing information about that category. A Multi Layer Perceptron (MLP) was implemented with dimensions $[18 \rightarrow 500 \rightarrow 18]$.

We calculated a latent representation for all users in the dataset. 80% of them are separated for training the classifier, and 20% for testing. In order to make an unbiased classifier, we selected the country with the smallest number of users, which in this case was Norway with 500 users, and we elected 500 random users from each of the 18 countries. The MLP was trained to predict the country of a certain user given its embedding as an input. All classifiers were trained for 100 epoch, and for 5 consecutive times. The average accuracy is presented together with its variance as final results.

6.2 Formalization

Users U are split into train/validation/test subsets. Each user is represented as a profile containing the set of tracks they listened to, and a corresponding sensitive attribute, in this case their nationality. The profiles are also split into query and ground truth subsets. When requested with a query, the recommender retrieves a set of suggested tracks $N = \{n_1, n_2, \ldots\}$, ordered by relevance, which are compared with a set of target tracks T (the ground truth); these are used in the computation of the performance metrics defined below. Within this process, an inner representation $\mathbf{z}$ is generated for each user, as the result of an encoding process.

6.3 Metrics

The LFM-2b dataset provides substantially large user profiles (i.e. users who listened to a big number of tracks), possibly because of the big time span comprehended in the user/track interaction data. Normalizing the number of right predictions by the total number of missing tracks then, also known as recall, provided extremely low values. Precision, on the other hand, measures the proportion of relevant items that were presented by the algorithm normalized by the length of the presented list. Precision seemed a more reasonable metric for this specific dataset.

Precision at K (PREC@K) measures the relative number of correct predictions in the first K ranked suggestions:

$$PREC@K = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}[n_k \in T], \tag{13}$$

where $\mathbb{I}$ is an indicator function, n_k the item ranked in position k.

Unfairness at K (UFAIR@K) is calculated as the summation of the differences between PREC@K measured for all pairs of user groups. The results is normalized by the number of comparisons, according to:

[3]https://pytorch.org/

[4]The source code for reproducing the experiments is available at https://github.com/rcaborges/F2VAE

$$UFAIR@K = \frac{1}{n \times m} \sum_{n=1}^{|C|} \sum_{m=1}^{|C|} \Big| \frac{1}{|C_n|} \sum_{u \in C_n} PREC@K(u) - \frac{1}{|C_m|} \sum_{u \in C_m} PREC@K(u) \Big|, \quad (14)$$

where C is a set containing all countries in the dataset, and C_n is the subset of users associated to nationality $c \in C$.

The amount of information of s contained in $\mathbf{z}$ is measured with cross-entropy (Equation 8). We differentiate the classifiers applied in ALFR-VAE and the one designed for measuring fairness in representations by naming the former one as *adversarial* and the latter *auxiliary*.

7 EXPERIMENTAL RESULTS

We now present the results of the experiments in two steps, first we evaluate the efficiency of the method to generate fair representations of users in the dataset, and then we calculate its capacity of removing nationality bias in recommendation results and the impact in recommendation quality. We report results for K equals to 1, 10 and 20 in order to have a better understanding of the effects of our method within a range of positions in the ranked results.

7.1 VAE x β-VAE

We tested β values higher and lower than 1, the parameter was set respectively to 10, and to 0.5. The higher the value assumed by β, the stronger regularization is applied on the latent variables . Stronger regularization meaning, in this case, that these latent variables are being modeled more accurately as normal distributions.

Latent variables generated by models trained with three different β values were submitted to the task of sensitive attribute prediction. The task was designed to measure to what extent these inner representations can be considered fair. There is a clear correlation between β values and unfairness in the inner representations, reflected in auxiliary classification accuracy results shown in Table 1. In this experiment, higher accuracy values indicate a higher chance of latent variables being associated with users' sensitive attribute.

The impact on recommendation and fairness metrics can be also seen in Table 1. Stronger regularization, i.e. higher β values, lead to higher accuracy in the recommendation results, according to precision measured for three values of K. But it leads also to higher unfairness in the recommendation results. The lowest values for unfairness were observed for standard VAE, with β equals to 1.

7.2 VAE x ALFR-VAE

It seemed reasonable to assume that reducing mutual information between sensitive attributes and the latent variable during the training process, would lead to fairer recommendation results. But our results showed that this is not necessarily the case. Instead, increasing the parameter γ, responsible for encouraging the adversarial classifier to neutralize the latent representation, led to more accurate recommendation results. The hyperparameter γ was set equals to 5, and 10, for encouraging mutual information reduction.

It is worth remembering the intuition behind the adversarial classifier and parameter γ. Latent factor models usually map the input data in a reduced space, latent space, before expanding this representation back to its original dimension. A regularization factor applied by these models, here assumed as D_ϕ, improves learning convergence by approximating the latent space distribution to normal distributions. This regularization has an interesting property of encouraging independence between latent components, to a certain extent when these components would be even associated to semantic properties contained in the input data [14]. But encouraging independence can potentially increase unfairness, for example, if one latent component is closely associated to users' nationality, and then, to this component could be given more emphasis because of its relevance in the recommendation task.

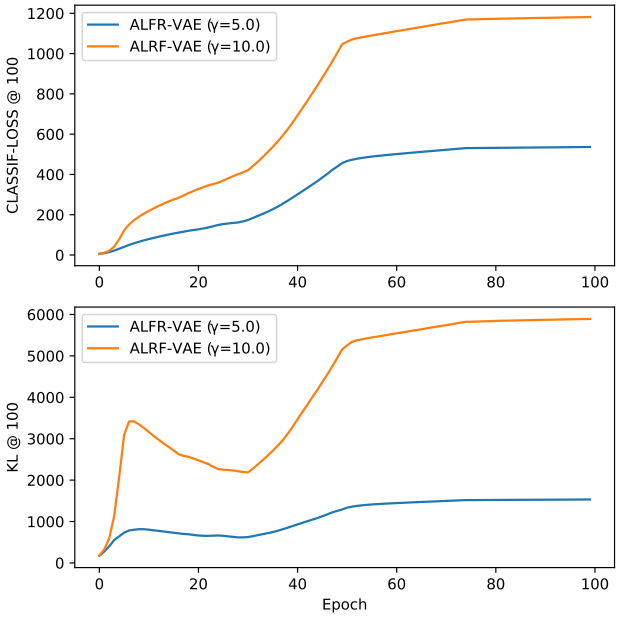

Figure 3: Adversarial loss and KL divergence measured for ALFR-VAE with different values of γ.

An adversarial classifier is added to this model to encourage that the minimum information from users' sensitive attribute, in this case their nationality, is propagated to the decoder. The mutual information between nationality and latent representations is being minimized, through cross entropy loss, and it can happen that the two objectives, mutual information and distribution divergence minimization, are conflicting with each other.

In Figure 3 the reader can see the evolution of adversarial classification loss and KL divergence measured for the validation data split at each consecutive training epoch. Parameter β is maintained fixed equals to 1, and parameter γ is set to 5, and 10. It is evident how both objectives are conflicting, and that for higher values of γ, the KL divergence measured for the latent variables increases significantly.

ALFR-VAE was responsible for the best precision results, and the worst bias measurements (Table1). The auxiliary classification loss, proposed for measuring how predictive latent representations are in the task of predicting the corresponding users' nationality,

also increased for higher γ values. User representations got more biased when more strength was given to the adversarial classifier loss term.

7.3 VAE x FaiRVAE

FaiRVAE methods, on the other hand, were responsible for the best results measured for unfairness, i.e. the highest fairness. The same methods however, were also responsible for the lowest precision results, as seen in Table1.

Setting hyperparameter τ with values higher then 1 led to divergence in the optimization process, and then we tested setting it with values lower than 1. Higher values of τ generated fairer representations, whereas lowest values generated fairer precision results. In the case when precision is higher, for $K = 1$, user unfairness was substantially reduced, by around 0.26% if compared to standard VAE, and by around 35% if compared with ALFR-VAE. In this same configuration, the precision result was reduced by approximately 12% taking the standard VAE as a reference.

7.4 F2VAE

Our main aim in this study is to provide users with relevant and fair recommendation results. We assume a fair situation as the one where users are being treated similarly, regardless of their attributes considered sensitive. In this study we considered nationality as the attribute according to which users were gathered in 18 groups.

Moreover, providing users with fair suggestions is not enough: the overall quality of results needs to be maintained as high as possible. In other words, a fairness-aware recommender would ideally provide users with the best recommendation results while ensuring that no one is being discriminated because of any attribute considered sensitive.

With F2VAE we were able to increase the recommendation quality while reducing unfairness. We compared the results with the original VAE model, and, in the case where $K = 1$, F2VAE was able to increase the precision in approximately 1% while reducing the unfairness by 21%. As one can see in Table 1.

Figure 4 shows the average PREC@1 calculated for each country, for methods ALFR-VAE and F2VAE. One can notice some adjustments made by the F2VAE to provide fairer recommendation results. AU had its average precision increased while SE, which had the highest precision, was penalized and got closer to the average among countries.

One common phenomenon observed in recommendation systems, usually named *popularity bias* [4], refers to the fact that more active users generate more interaction data and end up being offered with more accurate results. According to this idea, one could argue that more accurate recommendation results are expected in the case of users originating from countries associated to higher number of interactions (Figure 2). But the results calculated for each country show that higher accuracy results were actually obtained for countries associated to low number of interactions, such as IT and SE; and that low accuracy results were calculated for countries associated to a big number of interactions, such as US and UK.

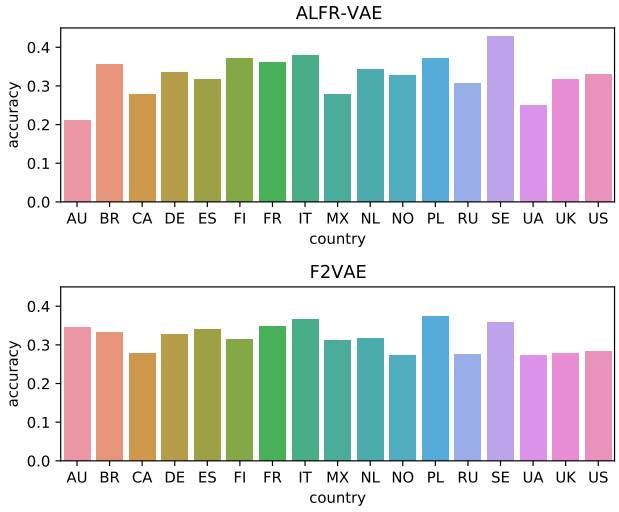

Figure 4: The average PREC@1 measured for each country. In the first row applying ALFR-VAE and in the second row applying F2VAE.

8 CONCLUSIONS

We have proposed a method for mitigating bias in recommendation systems associated to users' nationality, considered here as a categorical sensitive attribute. The method can be also applied to any other kind of attribute, for example gender or age, as long as these attributes can be expressed in categories.

The combination of standard VAE framework with two new terms, one dedicated to fair representation and another one dedicated to fair recommendation, led to positive results. The proposed method was able to increase the precision in approximately 1% while reducing the unfairness by 21% when compared to standard VAE.

In our experiments, fair representations did not lead necessarily to fairness in the recommendation results. Instead, reducing the variance among groups of users within training batches turned out to reduce the unfairness in the recommendation results. Moreover, reducing the mutual information between latent variables and sensitive attributes led to an increase of recommendation accuracy and unfair latent representations.

As future works, we are planning to apply the same method to another sensitive attributes, i.e. gender and age, and we are also planning to consider combinations of two or more attributes, known as *subgroups*.

REFERENCES

[1] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing Fair Ranking Schemes. In *Proceedings of the 2019 International Conference on Management of Data (SIGMOD '19)*. Association for Computing Machinery, New York, NY, USA, 1259–1276. https://doi.org/10.1145/3299869.3300079

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[3] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International*

Table 1: Average Results for LFM-2b dataset. The best (bold) and worst (underline) values are indicated in the table.

	CLASS. ACCURACY	PREC@1	PREC@10	PREC@20	UFAIR@1	UFAIR@10	UFAIR@20
VAE	**0.507 (0.011)**	0.306	0.231	0.203	0.0479	0.0264	0.0209
β-VAE (β = 0.5)	0.515 (0.018)	0.303	0.226	0.198	0.0482	0.0267	0.0213
β-VAE (β = 10.0)	0.510 (0.012)	0.320	0.238	0.208	0.0501	0.0270	0.0223
ALFR-VAE (γ = 5.0)	0.557 (0.017)	0.321	0.238	0.207	0.0481	0.0295	0.0252
ALFR-VAE (γ = 10.0)	0.584 (0.016)	**0.337**	**0.249**	**0.215**	0.0545	0.0275	0.0243
FaiRVAE (τ = 0.2)	0.524 (0.014)	0.307	0.228	0.200	**0.0352**	0.0271	0.0210
FaiRVAE (τ = 0.5)	0.515 (0.006)	0.302	0.223	0.196	0.0517	**0.0218**	**0.0190**
F2VAE (γ = 5.0, τ = 0.5)	0.537 (0.012)	0.310	0.233	0.201	0.0377	0.0234	0.0208

ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR. 405–414.

[4] Rodrigo Borges and Kostas Stefanidis. 2020. On Measuring Popularity Bias in Collaborative Filtering Data. In *Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, March 30, 2020*, Vol. 2578.

[5] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogin, and Tommaso Di Noia. 2021. A Flexible Framework for Evaluating User and Item Fairness in Recommender Systems. *User Modeling and User-Adapted Interaction (UMUAI)* (jan 2021). https://doi.org/10.1007/s11257-020-09285-1.

[6] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. 214–226.

[7] Harrison Edwards and Amos J. Storkey. 2016. Censoring Representations with an Adversary. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

[8] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *CoRR* (2016).

[9] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

[10] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR*.

[11] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-Oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021 (WWW '21)*. 624–632. https://doi.org/10.1145/3442381.3449866

[12] 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*. 689–698.

[13] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. 2016. The Variational Fair Autoencoder. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

[14] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning Disentangled Representations for Recommendation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. 5712–5723. https://proceedings.neurips.cc/paper/2019/hash/a2186aa7c086b46ad4e8bf81e2a3a19b-Abstract.html

[15] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research)*, Jennifer G. Dy and Andreas Krause (Eds.), Vol. 80. PMLR, 3381–3390. http://proceedings.mlr.press/v80/madras18a.html

[16] 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. *Information Processing & Management* 58, 5 (2021), 102666. https://doi.org/10.1016/j.ipm.2021.102666

[17] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in rankings and recommendations: an overview. *The VLDB Journal* (Oct 2021). https://doi.org/10.1007/s00778-021-00697-y.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. *Learning Internal Representations by Error Propagation*. MIT Press, Cambridge, MA, USA, 318–362.

[19] Virginia Tsintzou, Evaggelia Pitoura, and Panayiotis Tsaparas. 2019. Bias Disparity in Recommendation Systems. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark, September 20, 2019*, Vol. 2440. http://ceur-ws.org/Vol-2440/short4.pdf

[20] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. 2021. *Learning Fair Representations for Recommendation: A Graph-Based Perspective*. Association for Computing Machinery, New York, NY, USA, 2198–2208. https://doi.org/10.1145/3442381.3450015

[21] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 2921–2930.

[22] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *CIKM*. 1569–1578.

[23] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 (ICML'13)*. III–325–III–333.

[24] 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 (CIKM '18)*. 1153–1162. https://doi.org/10.1145/3269206.3271795

[25] 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 '20)*. 449–458. https://doi.org/10.1145/3397271.3401177