arXiv:2006.16742v2 [cs.IR] 15 Apr 2021

FairRec: Fairness-aware News Recommendation with Decomposed Adversarial Learning

Chuhan Wu[1], **Fangzhao Wu**[2], **Xiting Wang**[2], **Yongfeng Huang**[1], **Xing Xie**[2]
[1]Department of Electronic Engineering & BNRist, Tsinghua University, Beijing 100084, China
[2]Microsoft Research Asia, Beijing 100080, China
{wuchuhan15, wufangzhao}@gmail.com, {xitwan, xing.xie}@microsoft.com, yfhuang@tsinghua.edu.cn

Abstract

News recommendation is important for online news services. Existing news recommendation models are usually learned from users' news click behaviors. Usually the behaviors of users with the same sensitive attributes (e.g., genders) have similar patterns and news recommendation models can easily capture these patterns. It may lead to some biases related to sensitive user attributes in the recommendation results, e.g., always recommending sports news to male users, which is unfair since users may not receive diverse news information. In this paper, we propose a fairness-aware news recommendation approach with decomposed adversarial learning and orthogonality regularization, which can alleviate unfairness in news recommendation brought by the biases of sensitive user attributes. In our approach, we propose to decompose the user interest model into two components. One component aims to learn a bias-aware user embedding that captures the bias information on sensitive user attributes, and the other aims to learn a bias-free user embedding that only encodes attribute-independent user interest information for fairness-aware news recommendation. In addition, we propose to apply an attribute prediction task to the bias-aware user embedding to enhance its ability on bias modeling, and we apply adversarial learning to the bias-free user embedding to remove the bias information from it. Moreover, we propose an orthogonality regularization method to encourage the bias-free user embeddings to be orthogonal to the bias-aware one to better distinguish the bias-free user embedding from the bias-aware one. For fairness-aware news ranking, we only use the bias-free user embedding. Extensive experiments on benchmark dataset show that our approach can effectively improve fairness in news recommendation with minor performance loss.

Introduction

Personalized news recommendation techniques are critical for news websites to help users find their interested news and improve their reading experience (Wu et al. 2019d). Many existing methods for news recommendation rely on the news click behaviors of users to learn user interest models (Okura et al. 2017; Wu et al. 2019b). For example, Okura et al. (2017) proposed to learn user representations from the representations of clicked news articles with a GRU network. Wu et al. (2019b) proposed to use personalized attention networks to learn user representations from the representations of clicked news by using the embedding of user ID as attention query. Usually, users with the same sensitive attributes (e.g., genders) may have similar patterns in their news click behaviors. Taking user genders as an example, in Fig. 1 the female users may prefer fashion news while male users may prefer sports news. However, user interest models can easily capture the these patterns and lead to some biases (e.g., gender bias) in the news recommendation results. For example, as shown in Fig. 1, since fashion news may be clicked by more female users while NBA news may be preferred more by male users, the model tends to only recommend fashion news to female users and NBA news to male users. In this scenario, the recommendation results are heavily influenced by the biases brought by sensitive user attributes, and the users interested in both fashion and NBA cannot receive diverse news information, which is unfair and may be harmful for user experience.

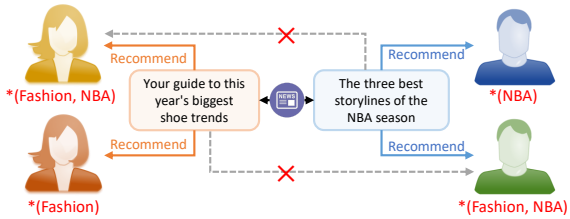


Figure 1: An example of gender bias in news recommendation. *Keywords under users represent their interest.

In this paper, we propose a **fair**ness-aware news **rec**ommendation (FairRec) approach with decomposed adversarial learning and orthogonality regularization, which can effectively alleviate the unfairness in news recommendation brought by the biases related to sensitive user attributes like genders. We propose to decompose the user interest model into two components, where the first one aims to learn a bias-aware user embedding that captures biases related to sensitive user attributes from user behaviors, and the second one aims to learn a bias-free user embedding that mainly encodes attribute-independent user interest information for making fairness-aware news recommendation. In addition, we apply a sensitive user attribute prediction task to the bias-aware user embedding to push it to convey

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

more bias information, and we apply adversarial learning techniques to the bias-free user embedding to eliminate its information on sensitive user attributes. Moreover, we propose an orthogonality regularization method to encourage the bias-free user embedding to be orthogonal to the bias-aware one, which can further remove the information related to sensitive attributes from the bias-free user embedding. To achieve fairness-aware news recommendation, we only use the bias-free user embedding for personalized news ranking. We conduct experiments on a benchmark news recommendation dataset, and the results show that our approach can effectively improve news recommendation fairness with acceptable performance sacrifice.

The major contributions of this paper include:

- This is the first work that explores to improve fairness in news recommendation by proposing a fairness-aware news recommendation framework.
- We propose a decomposed adversarial learning method with orthogonality regularization to learn bias-free user embeddings for fairness-aware news ranking.
- Extensive experiments on real-world dataset demonstrate that our approach can effectively improve fairness in news recommendation.

Related Work

News Recommendation

News recommendation is an essential technique for online news platform to provide personalized news services. Accurately modeling of user interest is critical for news recommendation (Wu et al. 2019b). In many existing news recommendation methods, the interest of users is modeled by their news click behaviors (Wang et al. 2018; Wu et al. 2019a; Zhu et al. 2019; An et al. 2019; Wu et al. 2019b,c; Qi et al. 2020; Wang et al. 2020; Hu et al. 2020). For example, Okura et al. (2017) proposed to use a GRU network to learn user representations from the representations of clicked news. Wang et al. (2018) proposed to learn user representations based on the relevance between the representations of clicked and candidate news. Wu et al. (2019c) proposed to learn user representations from clicked news via multi-head self-attention networks. These existing methods usually learn news recommendation models from users' news click behaviors. However, their models can easily grasp the similar patterns in the behaviors of users with the same sensitive attributes and lead to biased news recommendation results. Thus, the users may not receive diverse news information, which is harmful to user experience. Different from these methods, in our approach we propose a decomposed adversarial learning approach with orthogonality regularization to learn bias-free user embeddings for fairness-aware news ranking, which can substantially improve news recommendation fairness with small performance sacrifice.

Fairness-aware Recommendation

The problem of fairness in recommendation has attracted much attention in recent years (Beutel et al. 2019; Ekstrand, Burke, and Diaz 2019; Fu et al. 2020; Patro et al. 2020). Some studies explore the problem of provider-side fairness, e.g., items from different providers have a fair chance of being recommended (Lee et al. 2014; Kamishima et al. 2014; Liu et al. 2019). There are also several methods that address the problem of customer-side fairness, e.g., provide similar recommendations for users with different sensitive attributes (Xiao et al. 2017; Zhu, Hu, and Caverlee 2018; Burke, Sonboli, and Ordonez-Gauger 2018). Many methods study customer-side fairness on e-commerce scenarios by using ratings to indicate fairness (Yao and Huang 2017). For example, Yao and Huang (2017) proposed four different metrics based on the predicted and real ratings of users with different attributes to measure unfairness. They proposed to regularize collaborative filtering models with one of the unfairness metrics to explore the model performance in minimizing each form of unfairness. Farnadi et al. (2018) proposed to use probabilistic soft logic (PSL) rules to balance the ratings for both users in different groups by unbiasing the ratings for each item. These methods mainly aim to balance the recommendation performance for users with different sensitive attributes. Geyik, Ambler, and Kenthapadi (2019) explored several re-ranking rules to provide fair rankings of LinkedIn users based on their ranking scores and the desired proportions over different user attributes. This method aims to provide fair rankings of users with different attributes. Different from these methods, our approach focuses on the fairness of news recommendation results rather than accuracy, and we need to rank news rather than users. We propose a decomposed adversarial learning method to learn bias-free user embeddings, which is used to generate fairness-aware news recommendation results.

Methodology

In this section, we first present the problem definitions of this paper, then introduce the details of our fairness-aware news recommendation framework with decomposed adversarial learning and orthogonality regularization.

Problem Definition

For a target user u with the sensitive attribute z, we assume that she has clicked N news articles, which are denoted as $\mathcal{D} = \{D_1, D_2, ..., D_N\}$. We denote the candidate news set for this user as $\mathcal{D}^c = \{D_1^c, D_2^c, ..., D_M^c\}$, where M is the number of candidate news. The gold click labels of the target user u clicking these candidate news are denoted as $[y_1, y_2, ..., y_M]$. The click labels predicted by the news recommendation model are denoted as $[\hat{y}_1, \hat{y}_2, ..., \hat{y}_M]$. Candidate news are sorted by these predicted click labels, and the top K ranked candidate news set (regarded as the recommendation result) is denoted as $\mathcal{D}^r = \{D_{i_1}^c, D_{i_2}^c, ..., D_{i_K}^c\}$. The unfairness of the recommendation result $\mathcal{D}^r$ is defined as how discriminative it is for inferring the sensitive user attribute z. If z can be predicted from $\mathcal{D}^r$ more accurately, the recommendation result is more unfair since it is more heavily influenced by the sensitive user attribute.

Framework of FairRec

First, we introduce the framework of the proposed **fair**ness-aware news **rec**ommendation (FairRec) method, as shown

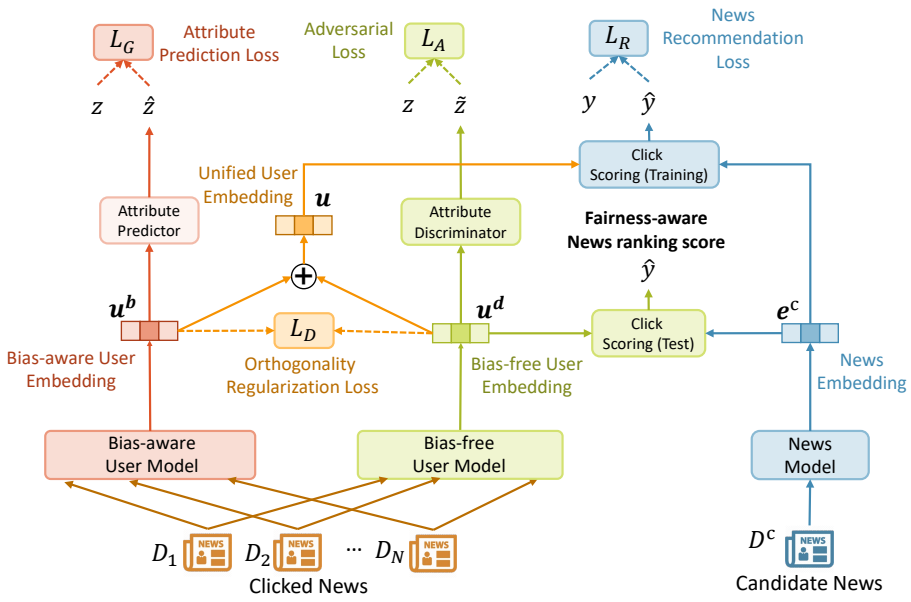


Figure 2: The architecture of our *FairRec* approach.

in Fig. 2. It mainly aims to compute a fairness-aware news ranking score for each candidate news of a user, which is further used to rank candidate news and generate fairness-aware news recommendation results for this user. More specifically, our *FairRec* framework uses a news model to learn the embeddings of candidate news, a bias-free user model to learn the bias-free embeddings of users which minimally contain the bias information on the sensitive user attribute, and a click scoring model to compute the fairness-aware news ranking scores based on the bias-free user embedding and candidate news embeddings. We briefly introduce these components as follows.

The news and user models in our approach are based on those in the NRMS (Wu et al. 2019c) method. The news model learns news representations from news titles. It first uses a multi-head self-attention network to capture the contexts of words within a news title, and then uses an attentive pooling network to learn news representations by modeling the importance of different words. We denote the representation of the candidate news D^c learned by the news model as $\mathbf{e}^c$. The user model learns the representation of a target user u from her clicked news $[D_1, D_2, ..., D_N]$. It first uses a news model to learn the representations of these clicked news, then uses a combination of multi-head self-attention network and attentive pooling network to obtain the unified user representations. We denote the bias-free user embedding learned by this user model as $\mathbf{u}^d$. Finally, the click scoring module computes the fairness-aware ranking score $\hat{y}$ based on the bias-free user embedding $\mathbf{u}^d$ and the candidate news embedding $\mathbf{e}^c$. Following many previous methods (Okura et al. 2017; Wu et al. 2019b), we use the dot product function to compute the fairness-aware ranking score by evaluating the relevance between the bias-free user embedding and candidate news embedding, i.e., $\hat{y} = \mathbf{u}^d \cdot \mathbf{e}^c$. The ranking scores of candidate news are further used for personalized news ranking and display.

Decomposed Adversarial Learning with Orthogonality Regularization

Then, we introduce the details of the proposed decomposed adversarial learning and orthogonality regularization method for learning bias-free user embeddings. In our fairness-aware recommendation framework, a core problem is how to learn the bias-free user embedding $\mathbf{u}^d$ from users' news click behaviors. However, since the users with the same sensitive attribute usually have some similar patterns in their news click behaviors, the user model can easily capture these patterns from users' news click behaviors and generate biased user embeddings. Thus, it is non-trivial to learn bias-free user embeddings from the biased user behaviors.

Adversarial learning is a technique that can be used to learn bias-free deep representations from biased data (Madras et al. 2018; Elazar and Goldberg 2018). Its mission is to enforce the deep representations to be maximally informative for predicting the labels of the main task, and meanwhile to be minimally discriminative for predicting sensitive attributes (Du et al. 2019). Thus, adversarial learning has the potential to learn bias-free user embeddings by removing the bias information about sensitive user attributes. A straightforward way is to apply an attribute discriminator to the user embeddings learned by the user model to infer the sensitive user attribute, and penalize the model according to the negative gradients from the adversarial loss that indicates the informativeness of user embeddings for sensitive user attribute prediction. At the same time, the user embeddings are also used to evaluate the relevance between the user and candidate news for news recommendation model training. Unfortunately, users' sensitive attributes

may be informative for the main news recommendation task, and the bias information related to the sensitive user attribute may be encoded into the user embeddings, making it difficult to be removed by adversarial learning. As an alternate, we propose to decompose the user interest model into two components, i.e., a bias-aware one that mainly aims to learn bias-aware user embeddings that capture the bias information on sensitive user attributes, and a bias-free one that only encodes the attribute-independent information of user interest into bias-free user embeddings. To push the bias-aware user embedding to be more attribute-discriminative, we propose to apply a sensitive attribute prediction task to the bias-aware user embedding. The user attribute z is predicted by an attribute predictor as follows[1]:

$$\hat{z} = \text{softmax}(\mathbf{W}^b \mathbf{u}^b + \mathbf{b}^b), \tag{1}$$

where $\mathbf{W}^b$ and $\mathbf{b}^b$ are parameters, $\hat{z}$ is the predicted probability vector. The loss function for attribute prediction is crossentropy, which is formulated as:

$$\mathcal{L}_G = -\frac{1}{U} \sum_{j=1}^{U} \sum_{i=1}^{C} z_i^j \log(\hat{z}_i^j), \tag{2}$$

where z_i^j and $\hat{z}_i^j$ respectively stand for the gold and predicted probability of the j-th user's attribute in the i-th class, and U is the number of users.

Usually, the supervision of the main recommendation task may also encode the bias information about sensitive user attribute into the bias-free user embedding. Thus, in order to eliminate the bias information, we propose to apply adversarial learning to the bias-free user embedding. More specifically, we use a attribute discriminator to predict user attributes according to the bias-free user embedding as follows:

$$\tilde{z} = \text{softmax}(\mathbf{W}^d \mathbf{u}^d + \mathbf{b}^d), \tag{3}$$

where $\mathbf{W}^d$ and $\mathbf{b}^d$ are parameters. The adversarial loss function of the discriminator is similar to the attribute predictor, which is formulated as follows:

$$\mathcal{L}_A = -\frac{1}{U} \sum_{j=1}^{U} \sum_{i=1}^{C} z_i^j \log(\tilde{z}_i^j). \tag{4}$$

To avoid the discriminator from inferring user attributes from the bias-free user embedding, we use the negative gradients of the discriminator to penalize the model.

Unfortunately, the bias-free user embedding may still contain some information related to the sensitive user attribute. This is because the discriminator usually cannot perfectly infer the sensitive user attribute, and there are shifts between the decision boundary of the discriminator and the real distribution of the sensitive user attribute. Since the bias-free user embedding generated by the user model only needs to cheat the discriminator, it does not necessarily fully remove the information of sensitive user attributes. To solve this problem, we propose an orthogonality regularization method to further purify the bias-free user embedding. Concretely, it regularizes the bias-aware user embedding and bias-free user embedding by encouraging them to be orthogonal to each other. The regularization loss function is formulated as follows:

$$\mathcal{L}_D = \frac{1}{U} \sum_{i=1}^{U} \left| \frac{\mathbf{u}_i^b \cdot \mathbf{u}_i^d}{||\mathbf{u}_i^b|| \cdot ||\mathbf{u}_i^d||} \right|, \tag{5}$$

where $\mathbf{u}_i^b$ and $\mathbf{u}_i^d$ are respectively the bias-aware and bias-free embeddings of the i-th user.

Model Training

Finally, we introduce how to train the models in our approach. In our *FairRec* framework, the bias-aware user embedding mainly contains the information on sensitive user attribute, and the bias-free user embedding mainly encodes attribute-independent user interest information. The information in both embeddings is correlated with the main recommendation task.[2] Thus, we add both user embeddings together to form a unified one for training the recommendation model, i.e., $\mathbf{u} = \mathbf{u}^b + \mathbf{u}^d$. We denote the representation of the candidate news D^c as $\mathbf{e}^c$, which is encoded by the news model. The probability of a user u clicking news D^c is predicted by $\hat{y} = \mathbf{u} \cdot \mathbf{e}^c$. Following (Huang et al. 2013; Wu et al. 2019c), we use negative sampling techniques to construct labeled samples for news recommendation model training. For each candidate news clicked by a user, we randomly sample T negative news in the same session which are not clicked. The loss function for news recommendation is the negative log-likelihood of the posterior click probability of clicked news, which is formulated as follows:

$$\mathcal{L}_R = -\frac{1}{N_c} \sum_{i=1}^{N_c} \log\left[\frac{\exp(\hat{y}_i)}{\exp(\hat{y}_i) + \sum_{j=1}^{T} \exp(\hat{y}_{i,j})}\right], \tag{6}$$

where $\hat{y}_i$ and $\hat{y}_{i,j}$ are the click scores of the i-th clicked candidate news and its associated j-th negative news, respectively. N_c is the number of clicked candidate news for training. The entire framework is trained collaboratively, and the final loss function for the recommendation model (except the discriminator) is a weighted summation of the news recommendation, attribute prediction, orthogonality regularization and adversarial loss functions, which is formulated as follows:

$$\mathcal{L} = \mathcal{L}_R + \lambda_G \mathcal{L}_G + \lambda_D \mathcal{L}_D - \lambda_A \mathcal{L}_A, \tag{7}$$

where λ_G, λ_D and λ_A are coefficients that control the importance of their corresponding losses.

Experiments

Dataset and Experimental Settings

In our experiments, we focus on gender parity in validating the effectiveness of our fairness-aware news recommendation approach. The dataset used in our experiments is provided by (Wu et al. 2019d), which contains the news impression logs of users and their gender labels (if available).

[1] We assume the attribute is a categorical variable here.

[2] In fact, bias-independent user interest information may also exist in both kinds of user embeddings. We will explore how to push the bias-independent user interests to be maximally captured by the bias-free user embedding in our future work.

#users	10,000	avg. #words per news title	11.29
#news	42,255	#clicked news logs	503,698
#impressions	360,428	#non-clicked news logs	9,970,795

Table 1: Statistics of the dataset.

It contains 10,000 users and their news browsing behaviors (from Dec. 13, 2018 to Jan. 12, 2019), and 4,228 users provide their gender label (2,484 male users and 1,744 female users). For the users without gender labels, the attribute prediction and adversarial losses are deactivated. The logs in the last week are used for test, and the rest are used for model training. In addition, we randomly sample 10% of training logs for validation. The statistics of this dataset are summarized in Table 1.

In our experiments, pre-trained Glove (Pennington, Socher, and Manning 2014) embeddings are used to initialize the word embeddings. Adam (Kingma and Ba 2015) is used as the model optimizer, and the learning rate is 0.001. The dropout (Srivastava et al. 2014) ratio is 0.2. The loss coefficients in Eq. (7) are all set to 0.5. These hyperparameters are tuned on the validation set.

Since the problem studied in this paper is the fairness of recommendation results rather than accuracy (He, Burghardt, and Lerman 2020), the fairness metrics based on user ratings used in several existing methods (Yao and Huang 2017; Farnadi et al. 2018) may not be suitable. To quantitatively measure the fairness of news recommendation results, we propose to use the prediction performance of sensitive user attribute based on the top K ranked candidate news in each session as the indication of recommendation fairness. The attribute prediction model contains a user model to learn user embeddings and an attribute predictor with a dense layer to infer the attributes. Since the dataset has an imbalanced gender distribution and there are system gender biases in the impression logs brought by news recall and pre-ranking, we build a new dataset from the original dataset to better evaluate recommendation fairness. We down-sample the number of male users to balance user gender, and use the entire news set as the candidate news set $\mathcal{D}^c$ for ranking to avoid impression gender bias. We use 80% of users for training the attribute prediction model, 10% for validation and the rest 10% for test. Following (Wu et al. 2019d), we use accuracy and macro F-score as the metrics to indicate fairness, where lower scores mean better recommendation fairness. To evaluate the performance of news recommendation, we use the average AUC, MRR, nDCG5 and nDCG10 scores of test sessions. We independently repeat each experiment 10 times and report the average results with standard deviations.

Performance Evaluation

In this section, we evaluate the performance of our *FairRec* approach in terms of fairness and news recommendation. We compare *FairRec* with various baseline methods for news recommendation, including: (1) LibFM (Rendle 2012), a popular recommendation tool based on factorization machine; (2) EBNR (Okura et al. 2017), an embedding-based news recommendation method that employs autoencoders to learn news representations and a GRU network to generate user representations; (3) DKN (Wang et al. 2018), using knowledge-aware CNNs to encode news representations and the relevance between representations of clicked news and candidate news to build user representations; (4) DAN (Zhu et al. 2019), using CNN to learn news representations and attentive LSTM to form user representations; (5) NPA (Wu et al. 2019b), using personalized attention networks to learn news and user representations; (6) NRMS (Wu et al. 2019c), using a combination of multi-head self-attention and additive attention to learn news and user representations. In addition, we compare the recommendation fairness of several additional methods, including: (7) MR (Yao and Huang 2017), using an unfairness loss to regularize our recommendation model. We regard the predicted click scores as "ratings"; (8) AL (Wadsworth, Vera, and Piech 2018), applying adversarial learning to the single user embedding; (9) ALGP (Zhang, Lemoine, and Mitchell 2018), using gradients projection in adversarial learning. (10) Random, ranking candidate news randomly, which is used to show the ideal recommendation fairness. The recommendation fairness of different methods under $K = 1, 3, 5$ or 10 and their recommendation performance are respectively shown in Tables 2 and 3. From the results, we have several observations.

First, compared with random ranking, the recommendation results of most methods are biased. This is possibly because users with the same attributes such as demographics usually have similar patterns in their behaviors, and user models may inherit these biases and encode them into the news ranking results. Second, compared with the methods that do not consider the fairness of recommendation (e.g., DAN, NPA and NRMS), fairness-aware methods (MR, AL, ALGP and FairRec) yield better recommendation fairness. Among them, the methods based on adversarial learning techniques perform better than the model regularization (MR) method that uses an unfairness loss to regularize the model. It shows that adversarial learning is more effective in improving the fairness of recommendation results by reducing the bias information in user embeddings. Third, compared with AL and ALGP, our approach achieves better recommendation fairness with a substantial margin. This may be because in AL and ALGP there are shifts between the decision boundaries of their discriminators and the real attribute distributions. Since the bias-free user embeddings only need to deceive the discriminator, they may not be orthogonal to the space of sensitive user attribute, which means that the bias information is not fully removed. Our approach uses a decomposed adversarial learning method with orthogonality regularization, which can learn bias-free user embeddings more effectively. Fourth, our approach can effectively improve recommendation fairness and meanwhile keep good recommendation performance. Compared with random ranking, our approach can almost achieve comparable recommendation fairness under different K. In addition, the recommendation performance of our approach is quite competitive. It outperforms several strong baseline methods like DKN and DAN, and the performance sacrifice is not large compared with its basic model NRMS that does

Methods	Top 1		Top 3		Top 5		Top 10	
	Accuracy	Macro-F	Accuracy	Macro-F	Accuracy	Macro-F	Accuracy	Macro-F
LibFM	59.78±0.64	59.34±0.62	63.25±0.61	63.04±0.60	64.63±0.59	64.46±0.56	66.42±0.54	66.25±0.51
EBNR	61.65±0.70	61.31±0.67	65.40±0.64	65.12±0.64	66.86±0.61	66.72±0.60	68.65±0.51	68.49±0.50
DKN	61.88±0.74	61.54±0.71	65.84±0.67	65.61±0.66	67.33±0.63	67.19±0.63	69.12±0.56	68.98±0.55
DAN	62.54±0.72	62.29±0.70	66.22±0.70	65.97±0.69	67.96±0.67	67.79±0.66	69.74±0.54	69.57±0.52
NPA	62.67±0.68	62.31±0.67	66.43±0.67	66.13±0.65	68.07±0.64	67.84±0.62	69.85±0.52	69.62±0.49
NRMS	63.13±0.71	62.75±0.70	66.89±0.68	66.54±0.66	68.32±0.67	67.96±0.65	70.12±0.59	69.94±0.56
MR	60.75±0.76	60.55±0.73	63.27±0.67	62.98±0.64	65.45±0.68	65.23±0.65	67.24±0.60	67.01±0.57
AL	58.86±0.75	58.51±0.73	62.67±0.65	62.41±0.63	64.92±0.63	64.61±0.61	66.70±0.54	66.39±0.52
ALGP	57.93±0.71	57.64±0.70	61.84±0.66	61.62±0.65	63.73±0.61	63.52±0.60	65.52±0.51	65.30±0.49
FairRec	**51.11**±0.69	**50.99**±0.66	**52.20**±0.61	**52.06**±0.60	**52.83**±0.54	**52.61**±0.54	**53.40**±0.48	**53.12**±0.46
Random	50.11±0.30	50.09±0.28	50.04±0.21	50.03±0.20	50.06±0.17	50.03±0.16	50.02±0.14	50.01±0.10

Table 2: News recommendation fairness of different methods. Lower scores indicate better fairness. The best results except random ranking are in bold.

Methods	AUC	MRR	nDCG@5	nDCG@10
LibFM	56.83±0.51	24.20±0.53	26.95±0.49	35.64±0.52
EBNR	60.94±0.24	28.22±0.25	30.31±0.23	39.60±0.24
DKN	60.34±0.33	27.51±0.29	29.75±0.31	38.79±0.30
DAN	61.43±0.31	28.62±0.30	30.66±0.32	39.81±0.33
NPA	62.33±0.25	29.46±0.23	31.57±0.22	40.71±0.23
NRMS	62.89±0.22	29.93±0.20	32.19±0.18	41.28±0.18
FairRec	61.95±0.22	29.01±0.21	31.25±0.18	40.24±0.21

Table 3: News recommendation performance of different methods. Higher scores indicate better results.

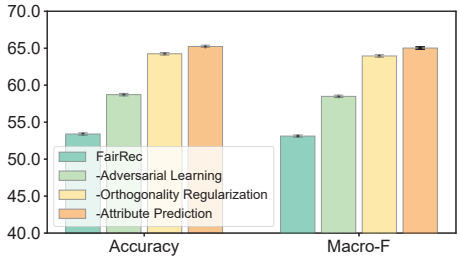


Figure 3: The effectiveness of decomposed adversarial learning. Lower scores represent better fairness.

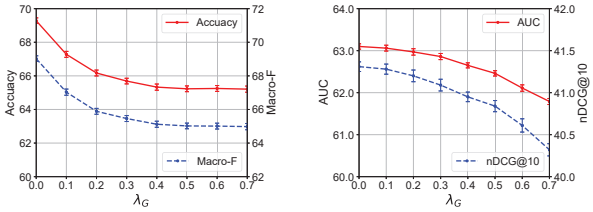
not consider recommendation fairness. These results validate that our approach can effectively improve fairness in news recommendation with minor performance loss.

Effectiveness of Decomposed Adversarial Learning

In this section, we conduct several ablation studies to verify the effectiveness of the core components in our *FairRec* approach, i.e., attribute prediction, adversarial learning and orthogonality regularization. We compare the recommendation fairness (under $K = 10$) of *FairRec* and its variants with one of these components removed, and the results are illustrated in Fig. 3. We have several findings from this plot. First, applying the attribute prediction task to the bias-aware user embedding is very important. This is because the attribute prediction task can greatly enhance the ability of bias-aware user embedding in bias modeling, which can help further remove the bias information from the bias-free user embedding. Second, applying adversarial learning to the bias-free user embedding is helpful for improving the fairness of news recommendation. This is because adversarial learning can encourage the bias-free user embedding to minimize the information for inferring the sensitive user attributes. Third, the orthogonality regularization added to the bias-aware and bias-free user embeddings can also effectively improve the recommendation fairness. It is because that this auxiliary regularization can push the bias-free user embedding to be orthogonal to the bias-aware user embedding and hence contains less bias information on sensitive user attributes.

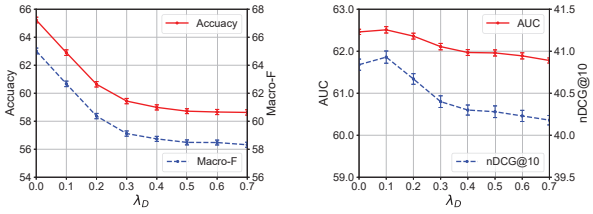
Hyperparameter Analysis

In this section, we explore the influence of several critical hyperparameters, i.e., the loss coefficients λ_G, λ_D and λ_A in Eq. (7) on the fairness and performance of news recommendation. Since there are three hyperparameters, their influence is evaluated independently. Firstly, we vary the value of λ_G without the decomposition loss and adversarial learning, and plot the fairness results under $K = 10$ in Figs. 4(a) and 4(b). We see the attribute prediction task can help improve the recommendation fairness, and the improvement increases when λ_G grows from 0. However, the improvement is marginal when it is larger than 0.5, and the performance declines more rapidly. Thus, a moderate value for λ_G (e.g., 0.5) may be preferable to achieve good fairness without too heavy performance loss. Then, we vary the value of λ_D under $\lambda_G = 0.5$ and adversarial learning deactivated. The results are shown in Figs. 5(a) and 5(b). From these results, we also find that the recommendation fairness improves with the increasing of λ_D, and the performance may decline when λ_D is too large. Thus, a proper range of λ_D (0.3-0.6) can achieve a good tradeoff between recommendation fairness and performance. For convenience, we choose the same value for λ_D as λ_G, i.e., 0.5. Finally, we activate the adversarial discriminator and vary λ_A under $\lambda_G = \lambda_D = 0.5$. The results are shown in Figs. 6(a) and 6(b). We find that if λ_A is too small or too large, the recommendation results are less fair. This may be because the adversaries cannot achieve an appropriate equilibrium and the attribute label is leaked to the bias-free user embedding. Thus, a moderate value of λ_A is also necessary, and for convenience of hyperparameter selection, we choose $\lambda_A = \lambda_G = \lambda_D = 0.5$ to avoid too heavy effort on hyperparameter searching.



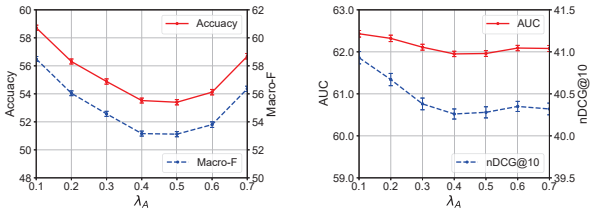
(a) Fairness. (b) Performance.

Figure 4: The news recommendation fairness and performance w.r.t. different λ_G.



(a) Fairness. (b) Performance.

Figure 5: The news recommendation fairness and performance w.r.t. different λ_D.



(a) Fairness. (b) Performance.

Figure 6: The news recommendation fairness and performance w.r.t. different λ_A.

Case Study

We conduct several case studies to show that our approach can improve the fairness of news recommendation results. We randomly select a male user and a female user, and predict the ranking scores of candidate news based on their clicked news using *NRMS* and *FairRec*. The results are illustrated in Fig. 7. From the top table in Fig. 7, we can infer that this male user may be interested in football and Golden Globes. However, the NRMS method that does not consider recommendation fairness provides a top rank for the candidate news about sports (Cowboys WR...) while assigns candidate news about fashion (The Biggest...) a low rank, which may be because fashion news is more likely to be preferred by female users. However, this user may also be interested in this news because it in fact has some inherent relatedness with the clicked news "2019 Golden Globes Best Actress". Similar phenomenon also exists in the ranking results of the female user. We can infer that this user

Male User

Clicked News
NFL playoff picture: Saints close to Clinching; Patriots fall behind Texans
Tom Brady had a classy reason for running right up to the ref after Sunday's win
2019 Golden Globes Best Actress

Candidate News	Score (NRMS)	Score (FairRec)
Cowboys WR Allen Hurns gets encouraging news after injury	**0.92**	**0.90**
The Biggest Fashion Trends of 2019 Are Here — Can You Handle It?	0.24	**0.84**
8 things making the rich even richer	0.36	0.23
Chefs reveal the 20 items they never make from scratch	0.30	0.19
Best Mexican Restaurant in Every State	0.22	0.17

Female User

Clicked News
Chris Duncan, former St. Louis Cardinals outfielder, battling brain cancer
Oscars fumble host test in wake of Kevin Hart's exit
These 5 countries have produced the most Miss Universe winners

Candidate News	Score (NRMS)	Score (FairRec)
2019 Golden Globes Best Actress	**0.87**	**0.90**
Report: Mike Mccarthy only pursuing Jets coaching vacancy	0.24	**0.81**
9 Ravens who could be potential salary cap casualties this offseason	0.20	**0.75**
10 Myths About Frozen Foods You Need to Stop Believing	0.30	0.22
Here's Why Saunas Are So Good For You	0.22	0.11

Figure 7: Comparison between the recommendation results of *NRMS* and *FairRec* for a male and a female user. The clicked candidate news are in blue.

may be interested in baseball games, and she may also have some interests in football. However, the news about football is assigned relatively low ranks, since football news may be preferred more by male users. These results reflect the unfairness in news recommendation. Fortunately, Fig. 7 shows that our approach can recommend the fashion news to male users and NFL news to female users for better satisfying their interest. It indicates that our approach can effectively improve fairness in news recommendation.

Conclusion

In this paper, we propose a fairness-aware news recommendation approach with decomposed adversarial learning and orthogonality regularization. We propose to decompose the user interest model into two parallel ones to respectively learn a bias-aware user embedding that captures bias information and a bias-free user embedding for fairness-aware news ranking. In addition, we apply an attribute prediction task to the bias-aware user embedding to enhance its ability on bias modeling, and apply adversarial learning techniques to the bias-free user embedding to eliminate its bias information on user attributes. Besides, we propose an orthogonality regularization method that pushes both user embeddings to be orthogonal to each other, which can better remove user attribute information from the bias-free user embedding. Extensive experiments show that our approach can substantially improve news recommendation fairness with minor performance sacrifice.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant numbers U1936208, U1936216 and 61862002.

References

An, M.; Wu, F.; Wu, C.; Zhang, K.; Liu, Z.; and Xie, X. 2019. Neural news recommendation with long-and short-term user representations. In *ACL*, 336–345.

Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Wei, L.; Wu, Y.; Heldt, L.; Zhao, Z.; Hong, L.; Chi, E. H.; et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *KDD*, 2212–2220.

Burke, R.; Sonboli, N.; and Ordonez-Gauger, A. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *FAT**, 202–214.

Du, M.; Yang, F.; Zou, N.; and Hu, X. 2019. Fairness in Deep Learning: A Computational Perspective. *arXiv preprint arXiv:1908.08843* .

Ekstrand, M. D.; Burke, R.; and Diaz, F. 2019. Fairness and discrimination in retrieval and recommendation. In *SIGIR*, 1403–1404.

Elazar, Y.; and Goldberg, Y. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *EMNLP*, 11–21.

Farnadi, G.; Kouki, P.; Thompson, S. K.; Srinivasan, S.; and Getoor, L. 2018. A fairness-aware hybrid recommender system. In *FATREC@ RecSys*.

Fu, Z.; Xian, Y.; Gao, R.; Zhao, J.; Huang, Q.; Ge, Y.; Xu, S.; Geng, S.; Shah, C.; Zhang, Y.; et al. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *SIGIR*, 69–78.

Geyik, S. C.; Ambler, S.; and Kenthapadi, K. 2019. Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In *KDD*, 2221–2231.

He, Y.; Burghardt, K.; and Lerman, K. 2020. A Geometric Solution to Fair Representations. In *AIES*, 279–285.

Hu, L.; Li, C.; Shi, C.; Yang, C.; and Shao, C. 2020. Graph neural news recommendation with long-term and short-term interest modeling. *Information Processing & Management* 57(2): 102142.

Huang, P.-S.; He, X.; Gao, J.; Deng, L.; Acero, A.; and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*, 2333–2338.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2014. Correcting Popularity Bias by Enhancing Recommendation Neutrality. In *RecSys Posters*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Lee, E. L.; Lou, J.-K.; Chen, W.-M.; Chen, Y.-C.; Lin, S.-D.; Chiang, Y.-S.; and Chen, K.-T. 2014. Fairness-aware loan recommendation for microfinance services. In *SocialCom*, 1–4.

Liu, W.; Guo, J.; Sonboli, N.; Burke, R.; and Zhang, S. 2019. Personalized fairness-aware re-ranking for microlending. In *RecSys*, 467–471.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning Adversarially Fair and Transferable Representations. In *ICML*, 3384–3393.

Okura, S.; Tagami, Y.; Ono, S.; and Tajima, A. 2017. Embedding-based news recommendation for millions of users. In *KDD*, 1933–1942.

Patro, G. K.; Biswas, A.; Ganguly, N.; Gummadi, K. P.; and Chakraborty, A. 2020. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *WWW*, 1194–1204.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Qi, T.; Wu, F.; Wu, C.; Huang, Y.; and Xie, X. 2020. Privacy-Preserving News Recommendation Model Learning. In *EMNLP: Findings*, 1423–1432.

Rendle, S. 2012. Factorization machines with libfm. *TIST* 3(3): 57.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR* 15(1): 1929–1958.

Wadsworth, C.; Vera, F.; and Piech, C. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. In *FAT/ML*.

Wang, H.; Wu, F.; Liu, Z.; and Xie, X. 2020. Fine-grained Interest Matching for Neural News Recommendation. In *ACL*, 836–845.

Wang, H.; Zhang, F.; Xie, X.; and Guo, M. 2018. DKN: Deep knowledge-aware network for news recommendation. In *WWW*, 1835–1844.

Wu, C.; Wu, F.; An, M.; Huang, J.; Huang, Y.; and Xie, X. 2019a. Neural news recommendation with attentive multi-view learning. In *IJCAI*, 3863–3869. AAAI Press.

Wu, C.; Wu, F.; An, M.; Huang, J.; Huang, Y.; and Xie, X. 2019b. Npa: Neural news recommendation with personalized attention. In *KDD*, 2576–2584.

Wu, C.; Wu, F.; Ge, S.; Qi, T.; Huang, Y.; and Xie, X. 2019c. Neural News Recommendation with Multi-Head Self-Attention. In *EMNLP-IJCNLP*, 6390–6395.

Wu, C.; Wu, F.; Qi, T.; Huang, Y.; and Xie, X. 2019d. Neural Gender Prediction from News Browsing Data. In *CCL*, 664–676. Springer.

Xiao, L.; Min, Z.; Yongfeng, Z.; Zhaoquan, G.; Yiqun, L.; and Shaoping, M. 2017. Fairness-aware group recommendation with pareto-efficiency. In *RecSys*, 107–115.

Yao, S.; and Huang, B. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *NIPS*, 2921–2930.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *AIES*, 335–340.

Zhu, Q.; Zhou, X.; Song, Z.; Tan, J.; and Guo, L. 2019. Dan: Deep attention neural network for news recommendation. In *AAAI*, volume 33, 5973–5980.

Zhu, Z.; Hu, X.; and Caverlee, J. 2018. Fairness-aware tensor-based recommendation. In *CIKM*, 1153–1162.