关于推荐结果公平性的新闻推荐系统的研究。模型方法以及一些概念没有弄懂。

Fairness-aware News Recommendation with Decomposed Adversarial Learning

ABSTRACT

News recommendation is important for online news services. Most news recommendation methods model users' interests from their news click behaviors. Usually the behaviors of users with the same sensitive attributes have similar patterns, and existing news recommendation models can inherit these biases and encode them into news ranking results. Thus, their recommendation results may be heavily influenced by the biases related to sensitive user attributes, which is unfair since users cannot receive sufficient news information that they are interested in. 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. For model training, we propose to learn a bias-aware user embedding that captures the bias information on user attributes from click behaviors, and 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 further purify the bias-free user embedding. 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 acceptable performance loss.

KEYWORDS

Fairness, News recommendation, Decomposed adversarial learning

ACM Reference Format:

. 2020. Fairness-aware News Recommendation with Decomposed Adversarial Learning. In *Proceedings of The Web Conference (WWW 2021)*, Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.475/123_4

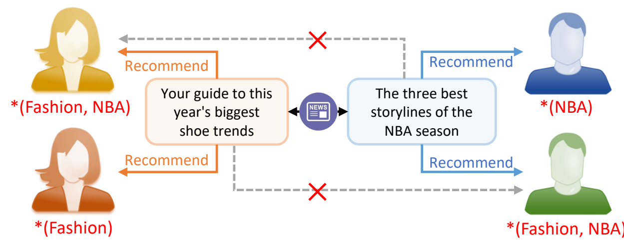


Figure 1: An example of unfairness in news recommendation. *Keywords under users represent their interest.

1 INTRODUCTION

Personalized news recommendation techniques are critical for news websites to help users find their interested news and improve their reading experience [22]. Many existing methods for news recommendation rely on the news click behaviors of users to model their interest [14, 20]. For example, Okura et al. [14] proposed to learn user representations from the representations of clicked news articles with a GRU network. Wu et al. [20] 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 may have similar patterns in their news click behaviors. Taking the gender attribute as an example, in Fig. 1 the female users may prefer fashion news while male users may prefer sports news. However, news recommendation models can easily capture the biases related to sensitive user attributes and encode them into 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 may tend 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 sufficient news information they are interested in as other users, which is unfair and may harm user experience.

In this paper, we propose a Fairness-Aware News recommendation (FAN) 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 via learning bias-free user embeddings. For model training, we propose to learn a bias-aware user embedding that captures biases related to sensitive user attributes from user behaviors, and learn and a bias-free user embedding that only encodes attribute-independent user interest information for making fairness-aware news recommendation. In addition, we apply a user attribute prediction task to the bias-aware user embedding to enhance its ability on capturing 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 push the bias-free user embedding to be orthogonal to the bias-aware one, which can reduce the information related to sensitive attributes in the bias-free user embedding. To achieve fairness-aware news recommendation, we only use the bias-free user embedding for

WWW 2021, April 2021, Ljubljana, Slovenia

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

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:

- We first explore 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 debiased embeddings of users.
- Extensive experiments on real-world dataset demonstrate that our approach can effectively improve fairness in news recommendation with an acceptable performance loss.

2 RELATED WORK

2.1 News Recommendation

News recommendation is an essential technique for online news platform to provide personalized news services. Accurately modeling of user interest is a critical step in news recommendation [20]. In many existing news recommendation methods, the interest of users is modeled by their news click behaviors [7, 18–21, 26]. For example, Okura et al. [14] proposed to use a GRU network to learn user representations from the representations of clicked news. Wang et al. [18] proposed to learn user representations based on the relevance between the representations of clicked and candidate news. Wu et al. [21] proposed to learn user representations from news representations via a combination of multi-head self-attention and additive attention networks. However, these existing methods mainly focus on recommending news that is likely to be clicked and can easily grasp and inherit the biases in the click behaviors of users. Thus, they may make biased news recommendations to users and further cause the problem of filter bubble, which is harmful to user experience. Different from these methods, our approach uses a decomposed adversarial learning approach with orthogonality regularization to learn debiased user embeddings, which can substantially improve news recommendation fairness with minor performance sacrifice.

2.2 Fairness-aware Recommendation

The problem of fairness in recommendation has attracted much attention in recent years [1]. Some studies explore the problem of provider-side fairness, e.g., items from different providers have a fair chance of being recommended [9, 11, 12]. There are also several methods that address the problem of customer-side fairness, e.g., provide similar rankings to users with different sensitive attributes [2, 6, 23, 27], which is also the problem studied in this paper. Many methods study customer-side fairness on e-commerce scenarios by using ratings to indicate fairness [24]. For example, Yao and Huang [24] 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. [5] proposed to use probabilistic soft logic (PSL) rules to balance the ratings for both users in different groups by un-biasing the ratings for each item. Besides e-commerce platforms, Geyik et al. [6] also 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. However, news recommendation aims to rank news rather than users, and there is no explicit user rating to indicate fairness. Different from these methods, our approach is based on adversarial learning techniques, which aim to learn debiased user embeddings for fairness-aware news recommendation.

3 METHODOLOGY

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

3.1 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 $\mathcal{D}^c = \{D^c_1, D^c_2, ..., D^c_M\}$, 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 $\mathcal{D}^r = \{D^c_{i_1}, D^c_{i_2}, ..., D^c_{i_K}\}$. 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.

3.2 Fairness-Aware News Recommendation Framework

In this section, we introduce the framework of the proposed **f**airness-**a**ware **n**ews recommendation (FAN) method, as shown in Fig. 2. Its major function is to compute a fairness-aware news ranking score for each candidate news of a user, which is further used to generate fairness-aware news recommendation results for her. Concretely, our *FAN* framework uses a news model to learn the embeddings of candidate news, a 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 the neural news and user models in the NRMS [21] 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

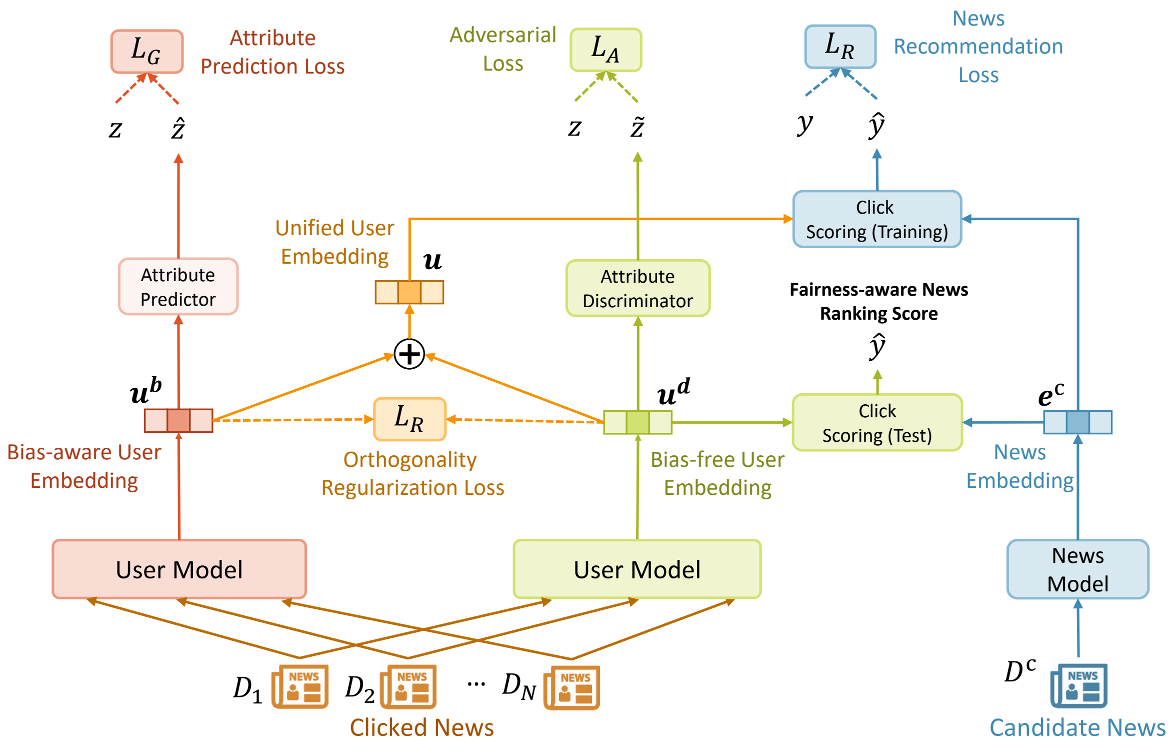


Figure 2: The architecture of our *FAN* approach.

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 [14, 20], 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.

3.3 Decomposed Adversarial Learning with Orthogonality Regularization

Next, 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 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 biases from users' news click behaviors and generate bias-aware user embeddings. Thus, it is a non-trivial task 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 [4, 13]. 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 [3]. 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 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 user embeddings into two components, i.e., a bias-aware one that mainly aims to capture the bias information on sensitive user attributes and a bias-free one that only encodes the attribute-independent information of user interest. 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]:

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

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

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 recommendation 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.

3.4 Model Training

Then, we introduce how to train the models in our approach. In our *FAN* 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. 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 [8, 21], 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.

4 EXPERIMENTS

4.1 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 [22], which contains the news impression logs of users and their gender labels (if available). 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.

Table 1: Statistics of the dataset.

#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

In our experiments, pre-trained Glove [15] embeddings are used to initialize the word embeddings. Adam [10] is used as the model optimizer, and the learning rate is 0.001. The loss coefficients in Eq. (7) are all set to 0.5.[2] These hyperparameters are tuned on the validation set.[3] In the news recommendation scenario there is no explicit user rating, and the impression news sets of different users are usually different. Thus, the fairness metrics used in several existing methods [6, 24] may not be suitable. To quantitatively measure the fairness of news recommendation results, we propose to use

[2]The results of hyperparameter search are included in supplementary materials.
[3]Complete settings are included in supplements.

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

Methods	Top 1		Top 3		Top 5		Top 10	
	Accuracy	Macro-F	Accuracy	Macro-F	Accuracy	Macro-F	Accuracy	Macro-F
LibFM	62.96±0.95	53.73±0.89	65.13±0.81	60.07±0.80	66.99±0.76	61.69±0.78	68.37±0.69	65.41±0.66
EBNR	63.64±0.83	54.21±0.82	65.51±0.76	60.46±0.77	67.49±0.75	62.06±0.74	68.73±0.69	65.75±0.68
DKN	63.66±0.78	54.30±0.80	65.58±0.79	60.52±0.80	67.53±0.73	62.17±0.73	68.99±0.71	65.80±0.72
DAN	63.71±0.81	54.26±0.79	65.59±0.75	60.54±0.74	67.51±0.74	62.19±0.75	69.01±0.70	65.83±0.72
NPA	63.88±0.82	54.34±0.84	65.72±0.77	60.75±0.75	67.59±0.71	62.32±0.73	69.14±0.65	65.89±0.62
NRMS	63.89±0.86	54.40±0.83	65.78±0.75	60.79±0.76	67.64±0.72	62.35±0.70	69.19±0.63	66.01±0.68
MR	62.96±0.91	53.48±0.83	64.57±0.82	58.83±0.81	66.19±0.73	60.82±0.70	68.36±0.65	65.12±0.67
AL	62.55±0.85	52.80±0.83	63.31±0.74	57.62±0.75	65.43±0.68	59.88±0.66	66.86±0.62	63.55±0.61
ALGP	62.48±0.86	52.72±0.82	63.09±0.75	57.31±0.73	65.21±0.66	59.43±0.67	66.16±0.61	63.28±0.63
FAN	**62.10**±0.80	**52.41**±0.76	**62.61**±0.69	**54.36**±0.68	**62.95**±0.62	**55.98**±0.63	**63.39**±0.59	**57.13**±0.58
Random	62.08±0.91	52.39±0.90	62.57±0.79	54.27±0.79	62.86±0.78	55.91±0.76	63.12±0.68	56.97±0.67

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

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
FAN	61.95±0.22	29.01±0.21	31.25±0.18	40.24±0.21

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 to infer the attributes. We use 80% of test sessions for training the attribute prediction model, 10% for validation and the rest 10% for test. Following [22], we use accuracy and macro Fscore as the metrics to indicate fairness, and 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.

4.2 Performance Evaluation

In this section, we evaluate the performance of our *FAN* approach in terms of fairness and news recommendation. We compare *FAN* with several baseline methods for news recommendation, including: (1) LibFM [16], a popular recommendation tool based on factorization machine; (2) EBNR [14], an embedding-based news recommendation method that employs autoencoders to learn news representations and a GRU network to generate user representations; (3) DKN [18], 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 [26], using CNN to learn news representations and attentive LSTM to form user representations; (5) NPA [20], using personalized attention networks to learn news and user representations; (6) NRMS [21], 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 [24], using an unfairness loss to regularize our recommendation model. We regard the predicted click scores as "ratings"; (8) AL [17], applying adversarial learning to the single user embedding; (9) ALGP [25], 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 FAN) 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 at 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

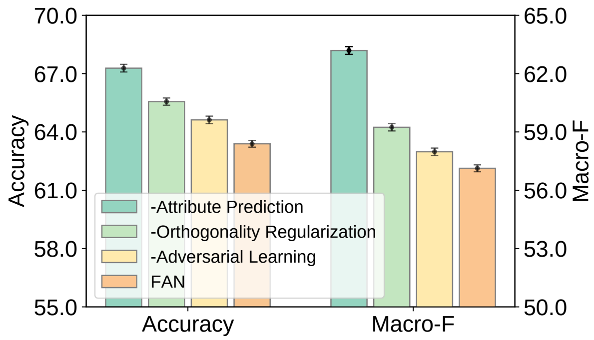


Figure 3: The effectiveness of decomposed adversarial learning. Lower scores represent better fairness.

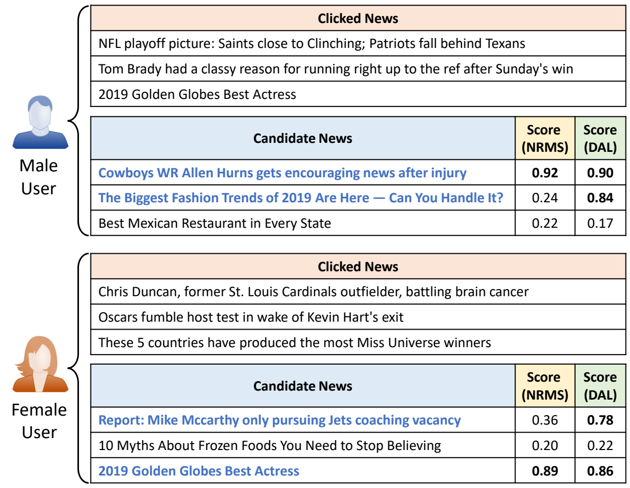
ranking, our approach consistently achieves comparable recommendation fairness under different values of K. In addition, the recommendation performance of our approach is comparable with many existing methods, and the performance sacrifice is acceptable compared with the basic model NRMS that does not consider recommendation fairness. These results validate that our approach can effectively improve fairness in news recommendation with acceptable performance loss.

4.3 Effectiveness of Decomposed Adversarial Learning

In this section, we conduct several ablation studies to verify the effectiveness of the core components in our *FAN* approach, i.e., attribute prediction, adversarial learning and orthogonality regularization. We compare the recommendation fairness (under $K = 10$) of FAN 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 is beneficial for further removing 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 discriminating 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.

4.4 Case Study

We conduct several case studies to verify that our approach can improve the fairness of news recommendation. 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 *FAN*. The results are illustrated in Fig. 4. From the top table in Fig. 4, we can



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 (DAL)
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**
Best Mexican Restaurant in Every State	0.22	0.17

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 (DAL)
Report: Mike Mccarthy only pursuing Jets coaching vacancy	0.36	**0.78**
10 Myths About Frozen Foods You Need to Stop Believing	0.20	0.22
2019 Golden Globes Best Actress	**0.89**	**0.86**

Figure 4: Clicked news and the ranking scores of candidate news from a male user and a female user. The clicked candidate news are in blue.

infer that this male user may be interested in football and Golden Globes. However, the NRMS method that does not consider recommendation fairness predicts a high ranking score for the candidate news about sports (Cowboys WR...) while assigns the candidate news about fashion (The Biggest...) a low score, 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 may be interested in baseball games, and she may also have some interests in football. However, the news about football is assigned relatively low scores, since football news may be preferred more by male users. These results reflect the unfairness in news recommendation. Fortunately, Fig. 4 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.

5 CONCLUSION

In this paper, we propose a fairness-aware news recommendation approach with decomposed adversarial learning and orthogonality regularization. We propose to use two parallel user models to learn a bias-aware user embedding to capture 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 acceptable performance sacrifice.

REFERENCES

[1] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *KDD*. 2212–2220.

[2] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *FAT**. 202–214.

[3] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2019. Fairness in Deep Learning: A Computational Perspective. *arXiv preprint arXiv:1908.08843* (2019).

[4] Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *EMNLP*. 11–21.

[5] Golnoosh Farnadi, Pigi Kouki, Spencer K Thompson, Sriram Srinivasan, and Lise Getoor. 2018. A fairness-aware hybrid recommender system. In *FATREC@ RecSys*.

[6] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In *KDD*. 2221–2231.

[7] Linmei Hu, Chen Li, Chuan Shi, Cheng Yang, and Chao Shao. 2020. Graph neural news recommendation with long-term and short-term interest modeling. *Information Processing & Management* 57, 2 (2020), 102142.

[8] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*. 2333–2338.

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

[10] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[11] Eric L Lee, Jing-Kai Lou, Wei-Ming Chen, Yen-Chi Chen, Shou-De Lin, Yen-Sheng Chiang, and Kuan-Ta Chen. 2014. Fairness-aware loan recommendation for microfinance services. In *SocialCom*. 1–4.

[12] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized fairness-aware re-ranking for microlending. In *RecSys*. 467–471.

[13] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *ICML*. 3384–3393.

[14] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *KDD*. 1933–1942.

[15] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[16] Steffen Rendle. 2012. Factorization machines with libfm. *TIST* 3, 3 (2012), 57.

[17] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. In *FAT/ML*.

[18] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep knowledge-aware network for news recommendation. In *WWW*. 1835–1844.

[19] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with attentive multi-view learning. In *IJCAI*. AAAI Press, 3863–3869.

[20] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Npa: Neural news recommendation with personalized attention. In *KDD*. 2576–2584.

[21] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Multi-Head Self-Attention. In *EMNLP-IJCNLP*. 6390–6395.

[22] Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural Gender Prediction from News Browsing Data. In *CCL*. Springer, 664–676.

[23] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-aware group recommendation with pareto-efficiency. In *RecSys*. 107–115.

[24] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *NIPS*. 2921–2930.

[25] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *AIES*. 335–340.

[26] Qiannan Zhu, Xiaofei Zhou, Zeliang Song, Jianlong Tan, and Li Guo. 2019. Dan: Deep attention neural network for news recommendation. In *AAAI*, Vol. 33. 5973–5980.

[27] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *CIKM*. 1153–1162.

SUPPLEMENTARY MATERIALS

Hyperparameter Settings

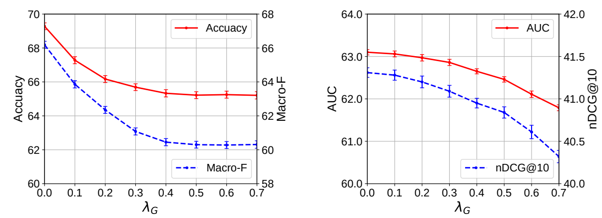
The settings of hyperparameters used in our approach are summarized in Table 4.

Table 4: Detailed hyperparameter settings.

Hyperparameters	Value
word embedding dimension	300
# heads in Transformer	16
output dim of each head	16
negative sampling ratio T	4
dropout	0.2
λ_G	0.5
λ_D	0.5
λ_A	0.5
optimizer	Adam
learning rate	1e-3
batch size	30

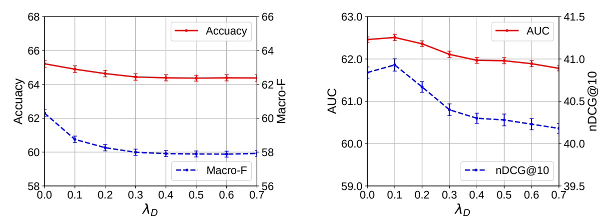
5.1 Hyperparameter Analysis

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. 5(a) and 5(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 better 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. 6(a) and 6(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 balance 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. 7(a) and 7(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 efforts on hyperparameter searching.



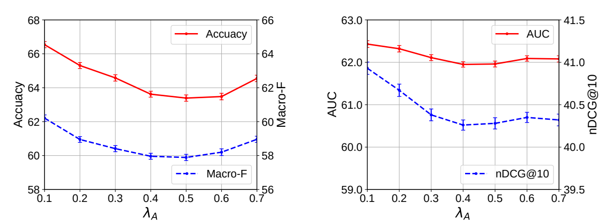
(a) Fairness. (b) Performance.

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



(a) Fairness. (b) Performance.

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



(a) Fairness. (b) Performance.

Figure 7: The news recommendation fairness and performance w.r.t. different λ_A.