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

——AAAI 2021

MOTIVATION

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.

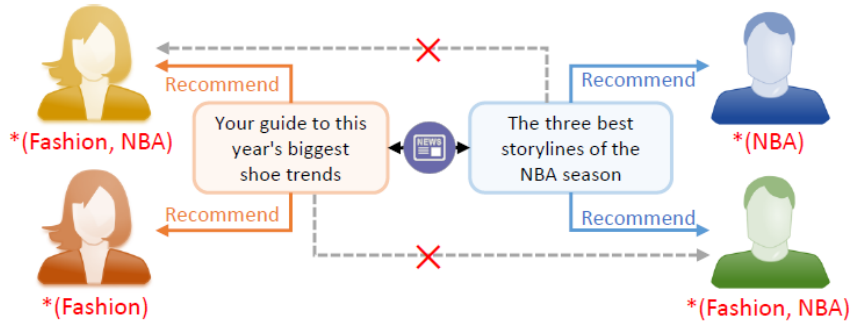


Figure 1: An example of gender bias in news recommendation. *Keywords under users represent their interest.

CONTRIBUTIONS

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.

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 $\mathcal{D}^r = \{D_{i_1}^c, D_{i_2}^c, ..., D_{i_K}^c\}$.

若模型能通过推荐结果的topk准确预测出属性Z，则判定模型为不公平的。

FRAMEWORK

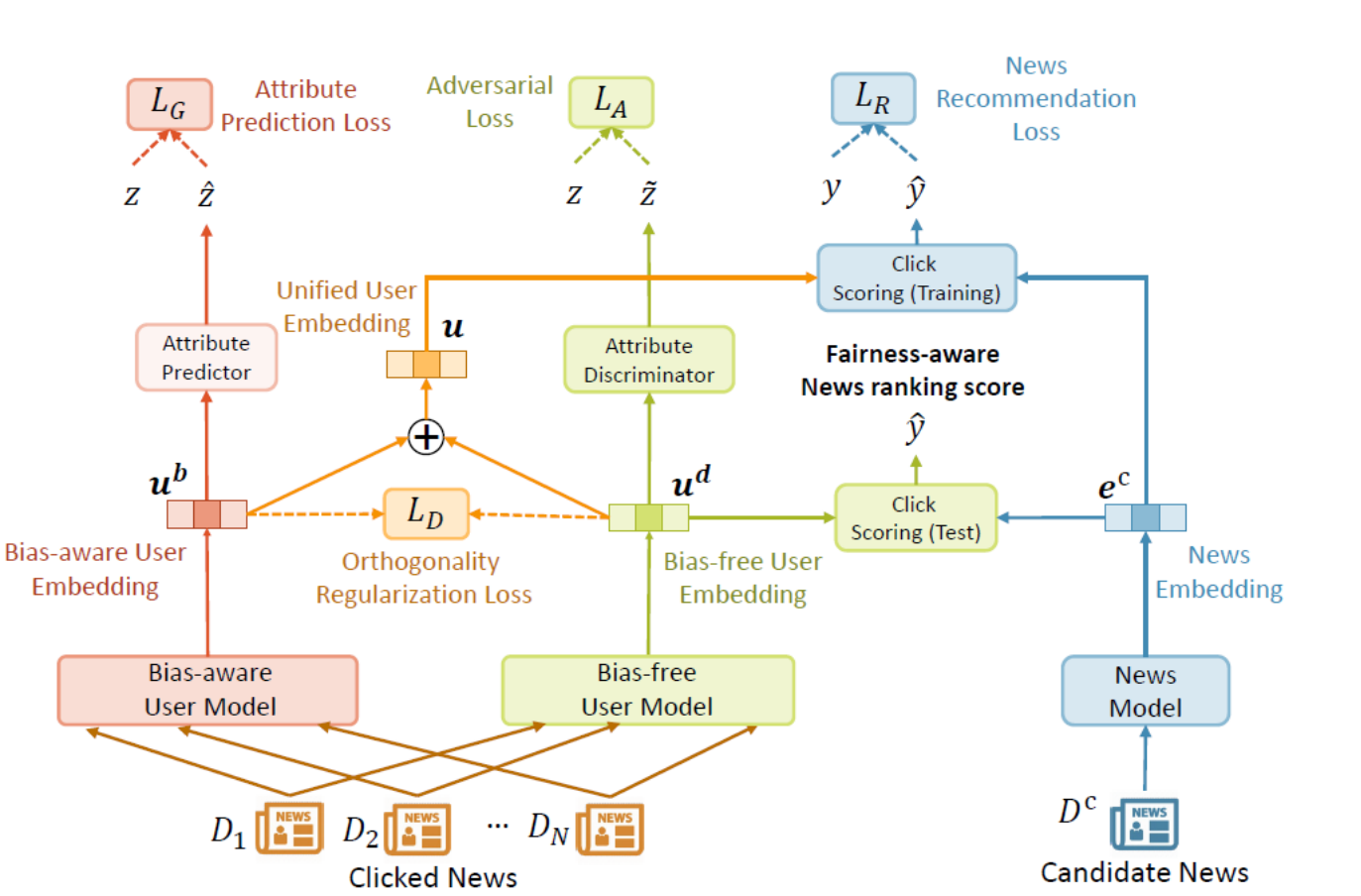


Figure 2: The architecture of our *FairRec* approach.

- a news model（ to learn the embeddings of candidate news）
- a bias-free user model
- a click scoring model (to compute the fairnessaware news ranking scores based on the bias-free user embedding and candidate news embeddings.)

DECOMPOSED ADVERSARIAL LEARNING WITH ORTHOGONALITY REGULARIZATION

- decompose the user interest model into two components
  - A bias-aware one to learn bias-aware user embeddings

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

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

  - A bias-free one to learn bias-free user embeddings.

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

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

use the negative gradients of the discriminator to penalize the model.

DECOMPOSED ADVERSARIAL LEARNING WITH ORTHOGONALITY REGULARIZATION

- an orthogonality regularization method to further purify the bias-free user embedding.

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

DECOMPOSED ADVERSARIAL LEARNING WITH ORTHOGONALITY REGULARIZATION

- Add both user embeddings together for training the recommendation model $\mathbf{u} = \mathbf{u}^b + \mathbf{u}^d$.
- The probability of a user u clicking news Dc is predicted by $\hat{y} = \mathbf{u} \cdot \mathbf{e}^c$
- For each news clicked , randomly sample T negative news in the same session which are not clicked

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

- final loss is a weighted summation of the news recommendation, attribute prediction, orthogonality regularization and adversarial loss functions

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

RESULT

#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.

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.

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.

RESULT

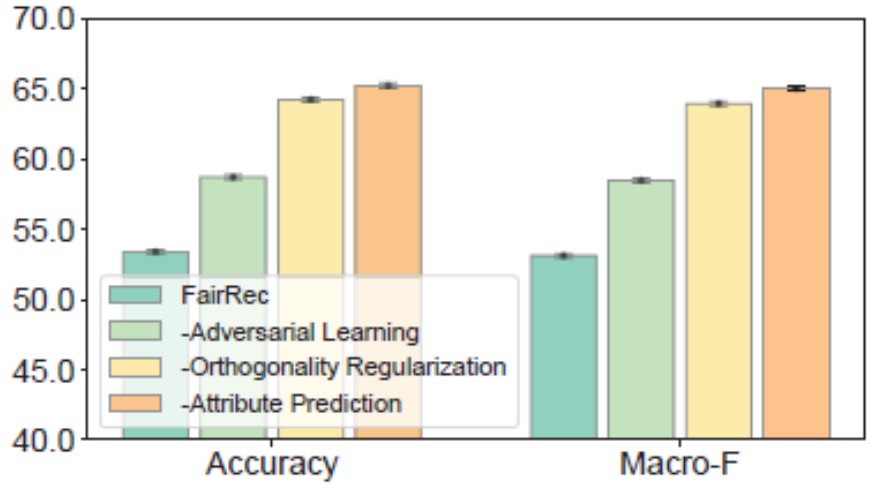
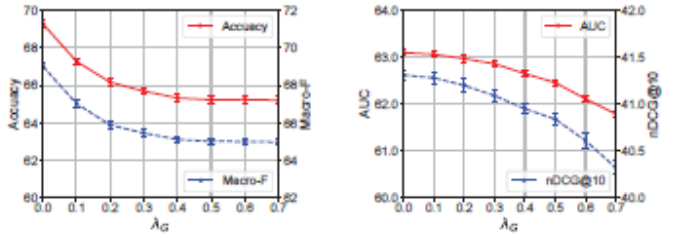
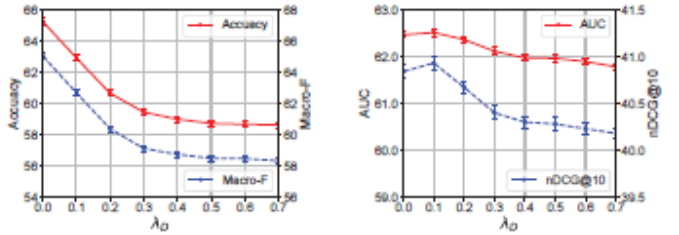


Figure 3: The effectiveness of decomposed adversarial learning. Lower scores represent better fairness.



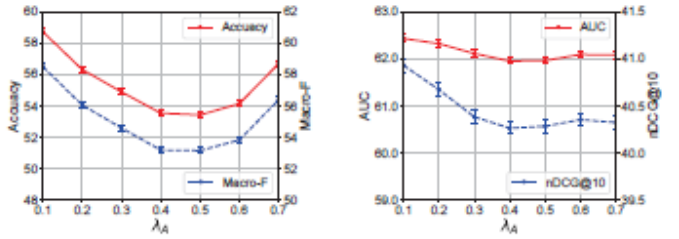
(a) Fairness. (b) Performance.

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.

RESULT

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.

RECOMMEND FAIRNESS

- Population imbalance
- Multiside fairness
- Position bias
- Exposure bias