vivaFemme: Mitigating Gender Bias in Neural Team Recommendation via Female-Advocate Loss Regularization

Abstract. Neural team recommendation has brought state-of-the-art efficacy while enhancing efficiency at forming teams of experts whose success in completing complex tasks is almost surely guaranteed. Yet proposed methods overlook diversity; that is, predicted teams are male-dominated and female participation is scarce. To this end, pre- and post-processing debiasing techniques have been initially proposed, mainly for being model-agnostic with little to no modification to the model's architecture. However, their limited mitigation performance has proven futile, especially in the presence of extreme bias, urging further development of *in*-process debiasing techniques. In this paper, we are the first to propose an in-process gender debiasing method in neural team recommendation via a novel modification to models' conventional cross-entropy loss function. Specifically, (1) we dramatically penalize the model (i.e., an increase to the loss) for false negative female experts, and meanwhile, (2) we randomly sample from female experts and reinforce the likelihood of female participation in the predicted teams, even at the cost of increasing false positive females. Our experiments on a benchmark dataset withholding extreme gender bias demonstrate our method's competence in mitigating gender bias in feed-forward neural models while maintaining accuracy. On the contrary, our method falls short of addressing bias in Bayesian models, urging further research on debiasing techniques for variational neural models. Our codebase to reproduce our experiments is available at `https://anonymous.4open.science/r/VivaFemme-8F03/`.

1 Introduction

As modern tasks have been surpassing the capacity of individuals, collaborative teams of experts have become vital in today's diverse landscape across academia, industry, law, freelancing, and healthcare. The team recommendation problem, also known as team allocation, team selection, team composition, and team configuration, seeks to automate the assembly of experts in a team whose combined skills solve challenging tasks. Team recommendation can be seen as social information retrieval (Social IR), where the right group of experts, rather than relevant information, is required to accomplish the task at hand.

Traditionally, even now in many scientific and industrial sectors, teams have been formed manually by relying on human experience and instinct, a process that is tedious, error-prone, and suboptimal due to hidden personal and societal biases, a multitude of criteria to optimize, and an overwhelming number of candidates, among other reasons. Notably, the team formation has been heavily influenced by the individuals' subjective opinions which inherit hidden and unfair societal biases, largely ignoring the diversity in recommending expert members

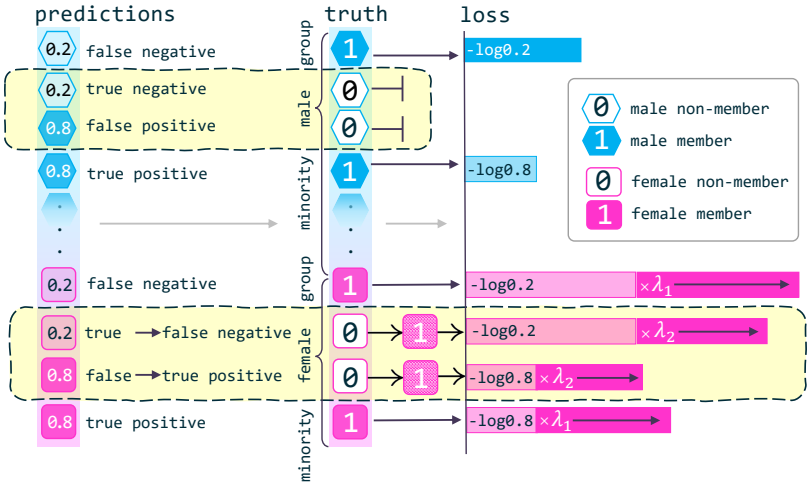

Fig. 1: Female-advocate loss regularization.

of a team, resulting in discrimination and reduced visibility for already disadvantaged *female* experts [9], disproportionate selection of *male* experts, and gender disparities [13].

Within the vast space of expert combinations, machine learning, neural models in particular, have enabled the analysis of massive collections of experts from diverse fields through learning relationships between experts and their skills based on past teams' successes and failures [18,5,6]. However, machine learning-based methods of team recommendation focus solely on maximizing the teams' success rates, overlooking fairness and diversity among team members. Such recommended teams are often successful yet unfair and biased toward the predominant distribution of male experts in the training datasets, with females heavily underrepresented. For example, in the `imdb` dataset for movies, `87.7`% of the cast and crew are male compared to `12.3`% female. Therefore, male cast and crew receive more attention and are more frequently recommended by a machine learning model, leading to discrimination against already disadvantaged female experts. To the best of our knowledge, there is no fairness-aware approach in the neural team recommendation method except that of Loghmani et al. [15], who showed *post*-processing greedy rerankers falls short of maintaining the accuracy of recommended teams when mitigating the popularity bias, urging further tandem integration of *in*-process debiasing techniques. In this paper, we propose viva$\mathscr{F}$emme, an *in*-process debiasing method with respect to the demographic (statistical) parity [16] where the model's recommendation of experts are *in*dependent of their genders. As opposed to pre-processing-based methods, which modify data or its labels before model training via over/under sampling, or post-processing techniques, which seek to improve the fairness of models after training via reranking experts in the final ranked list of recommendations, we focus on balancing model accuracy with fairness considerations during training by adjusting the model's loss in favour of minority female experts. Specifically,

in *neural* team recommendation models, where the team recommendation has been addressed as a *multilabel* Boolean classification task by and large, and each expert is mapped to a label and would be recommended should the expert's class prediction probability be close to one, (1) we dramatically penalize the model by a substantial increase to the loss if a false negative happens for a female expert, i.e., a female expert who is a member of an optimum team has been overlooked by the model, compared to when the false negative happens for a male expert. For example, from Figure 1, a true female member (red square `1`), which has not been recommended for low probability `0.2`, generates an increased loss of $-\lambda_1 \times \texttt{log(0.2)}$ compared to its male peer (blue hexagon `1`). Additionally, given a *fixed* number of false positive experts by a model, i.e., incorrect recommendation of experts who are *not* members of an optimum team, skewing false positives toward *female* experts neither helps nor hurts the accuracy more, yet improves the participation of minority female experts in the recommended teams. In view of this presumption, (2) we randomly sample from female experts who are not part of an optimum team as if they should have been in the team and reinforce the likelihood of their participation in the recommended teams by adding their loss values to the original training loss at the cost of increasing false positives but in favour of female experts. For example, from Figure 1, there are two female and two male experts who are not members of an optimum team (white squares `0` and white hexagons `0`, respectively). While these male experts have contributed no loss even if the model recommended one of them for the high probability `0.8` (false positive), the female experts have been *virtually* considered members of the optimum team and contributed increased loss values by the factor of λ_2. To illustrate the effectiveness of our proposed female-advocate loss, we perform experiments on a benchmark dataset (`imdb`). Our results show that our proposed loss substantially mitigates gender bias while maintaining the accuracy of the recommended teams in feed-forward neural architectures. However, it falls short of mitigating bias when plugged into variational Bayesian neural models, urging further research on in-process loss-based bias mitigation techniques for such architectures.

2 Related Works

The works related to this paper are largely centered around two areas: 1) neural team recommendation and 2) fairness-aware recommendation.

2.1 Neural Team Recommendation

Thus far, proposed machine learning-based solutions to the team recommendation problem are based on neural models that can be categorized based on their model architecture, including feed-forward, variational Bayesian networks [5], and graph neural networks, and training strategies including negative sampling [5] heuristics and streaming (temporal) training [7]. Sapienza et al. [19] were the first to use graph representations learning in the form of an autoencoder for the team recommendation problem in online multiplayer games. They propose to learn dense vector representations for game players (experts) via random

walks on the co-play network upon which pairwise top-k closest vector representation of experts yields the optimum subset of experts as a team. Sapienza et al.'s method, however, builds autoencoder models independently for each game, resulting in player-specific vector representations for each game and overlooking the joint combination of proficiencies. Other researchers have explored alternative neural architectures. Rad et al. [18] proposed variational Bayesian neural model incorporating uncertainty via probabilistic weights to address overfitting. They employ a multi-layer feed-forward neural network to map the required subset of skills to an optimum subset of experts as the recommended team. Later, Rad et al. [17] used a graph neural network to learn dense vector representation of skills at the input layer of variational Bayesian neural model and showed performance improvements. In the neural team recommendation literature, (non-variational) feed-forward neural networks have also been employed as a baseline to provide a reference level of comparison [5]. Reinforcement learning via neural policy estimators has also been proposed to emulate team formation processes in multi-agent environments, where autonomous agents learn to negotiate and decide team composition for tasks that individual agents cannot complete alone.

Despite the rich body of research, existing neural team recommendation models overlook fairness. Meanwhile, accounting for fairness has gained significant importance in other disciplines such as healthcare [4], information retrieval [3], computer vision [12], ranking and recommendations. In this paper, we are among the first to undertake an empirical investigation addressing the fairness gap in neural team formation by assessing the impact of loss regularization in favor of recommending more female experts while controlling the accuracy of the recommended teams.

2.2 Fairness-aware Recommendation

Fairness in machine learning has been addressed at an individual level, where consistency in treatment toward an individual is expected, and at a group level, where equitable treatment is sought between a disadvantaged (protected) group and the advantaged group as a whole, e.g., female vs. male experts. Fairness-aware approaches aim to identify and assess unfair biases or mitigate them through debiasing algorithms at individual or group levels.

Debiasing algorithms can be categorized based on their integration into the machine learning pipeline: (1) pre-processing methods modify data or label through re-sampling heuristics before model training, (2) in-processing techniques adjust the optimization process of models to balance accuracy and fairness, and (3) post-processing methods modify model outputs during inference, which may involve altering thresholds, scoring rules, or reranking the recommended list of items. Since preprocessing methods entail significant alterations to datasets, particularly when confronted with extreme biases, and re-sampling adjustments may result in deviation from real-world scenarios, researchers mostly opt for in-processing and post-processing methods. For instance, Zehlike et al. proposed an in-process method by extending the loss function of ListNet, a list-wise learning-to-rank model, with fairness objective based on reducing the discrepancy in average group visibility (exposure) between a protected group and a

non-protected group in the ranking results. From post-processing methods, Feng et al. [8] studied gender bias associated with professional occupations in search results across major search engines and developed a re-ranking algorithm that prioritizes gender parity in top positions while maintaining relevance.

While there are few studies in fairness-aware team recommendation literature such as the work by Barnabò et al [2] that proposed *search*-based greedy approximation for a fair team through the weighted set cover problem, there is no fairness-aware method that mitigates unfair gender bias in machine learning-based team recommendation except that of the work by Loghmani et al. [15] that showed existing greedy post-processing reranking methods fall short of mitigating neural team recommendation models' bias toward popularity experts. To fill the research gap, we present a unique contribution yet pioneering in-process method for machine learning-based team recommendation.

3 Female-Advocate Neural Team Recommendation

In this section, we foremost provide a formal problem statement for neural team recommendation, on the one hand, and demographic parity notions of fairness, on the other hand. Then, we formalize our proposed in-process loss-based method for neural team recommendation to recommend an optimum yet fair team of experts with respect to gender bias in view of demographic parity.

3.1 Neural Team Recommendation

Given a set of skills $\mathcal{S} = \{i\}$ and a set of experts $\mathcal{E} = \{j\}$, an optimum team of experts $\mathbf{e} \subseteq \mathcal{E}$; $\mathbf{e} \neq \emptyset$ that collectively cover the skill set $\mathbf{s} \subseteq \mathcal{S}$; $\mathbf{s} \neq \emptyset$ is shown by $(\mathbf{s}, \mathbf{e})$. Further, $\mathcal{T} = \{(\mathbf{s}, \mathbf{e}) | \mathbf{s} \subseteq \mathcal{S}, \mathbf{e} \subseteq \mathcal{E}\}$ indexes all optimum teams. For a given subset of desired skills $\mathbf{s}$, the goal of the team recommendation problem is to recommend an optimal subset of experts $\mathbf{e}$ that their collaboration as a team leads to success, i.e., $(\mathbf{s}, \mathbf{e}) \in \mathcal{T}$, while avoiding a potentially unsuccessful subset of experts $\mathbf{e}'$, i.e., $(\mathbf{s}, \mathbf{e}') \notin \mathcal{T}$. More concretely, the team recommendation problem is to find a mapping function $\mathcal{F}$ of parameters θ from the power set of skills to the powerset of experts such that $\mathcal{F}_\theta : \mathcal{P}(\mathcal{S}) \rightarrow \mathcal{P}(\mathcal{E}), \mathcal{F}_\theta(\mathbf{s}) = \mathbf{e}$.

Neural team recommendation estimates $\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{s})$ using a multilayer neural network that learns, from $\mathcal{T}$, to map a vector representation of subset of skills $v_{\mathbf{s}}$, to a vector representation of subset of experts $v_{\mathbf{e}}$ by maximizing the posterior probability of $\boldsymbol{\theta}$ in $\mathcal{F}_{\boldsymbol{\theta}}$ over $\mathcal{T}$, that is, $\texttt{argmax}_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}|\mathcal{T})$. For $v_{\mathbf{s}}$, neural team recommendation methods adopt either *i*) the *occurrence* vector representation for $\mathbf{s}$ or *ii*) a dense low d-dimensional vector representation of $\mathbf{s}$, $d << |\mathcal{S}|$, pretrained by e.g., a graph neural network [10]. In the output layer for vector representation of the subset of experts $v_{\mathbf{e}}$, neural team recommendation methods frame the problem as a *multilabel* Boolean classification task and used occurrence vector representation for $\mathbf{e}$, that is, $v_{\mathbf{e}} \in \{0, 1\}^{|\mathcal{E}|}$ where $v_{\mathbf{e}}[j] = 1$ if $j \in \mathbf{e}$, and 0 otherwise, as seen in Figure 1. Using a neural model of one hidden layer $\mathbf{h}$ of size d, without loss of generality to multiple hidden layers, with the input layer $v_{\mathbf{s}}$ and

output layer $v_{\mathbf{e}}$, a neural team recommendation method can be formalized as,

$$\mathbf{h} = \pi(\boldsymbol{\theta}_1 v_{\mathbf{s}} + \mathbf{b}_1) \tag{1}$$

$$logits \rightarrow \mathbf{z} = \boldsymbol{\theta}_2 \mathbf{h} + \mathbf{b}_2 \tag{2}$$

$$v_{\mathbf{e}} = \sigma(\mathbf{z}) \tag{3}$$

where π is a nonlinear activation function, $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \cup \boldsymbol{\theta}_2 \cup \mathbf{b}_1 \cup \mathbf{b}_2$ are learnable parameters for $\mathcal{F}$, and σ is the `sigmoid` function to interpret the model's output as each expert's predicted probability of membership in $\mathbf{e}$. During training, given a team $(\mathbf{s}, \mathbf{e})$, neural models tune the parameters $\boldsymbol{\theta}$ by maximizing the posterior probability of $\boldsymbol{\theta}$ in $\mathcal{F}_{\boldsymbol{\theta}}$ over $\mathcal{T}$. By Bayes theorem:

$$\texttt{argmax}_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}|\mathcal{T}) \propto p(\mathcal{T}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{(\mathbf{s},\mathbf{e})\in\mathcal{T}} p(\mathbf{e}|\mathbf{s}, \boldsymbol{\theta}) \tag{4}$$

where $p(\mathcal{T})$ is independent of $\boldsymbol{\theta}$, $p(\mathcal{T}|\boldsymbol{\theta})$ is the likelihood and can be estimated via average *binary* cross-entropy over $\boldsymbol{e}$:

$$p(\mathbf{e}|\mathbf{s}, \boldsymbol{\theta}) = \prod_{j\in\mathbf{e}} \sigma(\mathbf{z}[j]) \propto \sum_{j\in\mathbf{e}} \texttt{log}\ \sigma(\mathbf{z}[j]) \tag{5}$$

and $p(\boldsymbol{\theta})$ is the prior joint probability of weights, which is unknown. The *true* prior probability of weights $p(\boldsymbol{\theta})$ cannot be calculated analytically or efficiently sampled, and as such, existing works either assume the *uniform* probability distribution over all possible real-values of $\boldsymbol{\theta}$ and proceed with maximum likelihood estimation $p(\mathcal{T}|\boldsymbol{\theta})$ [5] using (non-variational) feed-forward neural model, or estimate $p(\boldsymbol{\theta})$ by Gaussian distribution and calculate the maximum a posterior via a variational Bayesian neural model [11,10].

Nonetheless, existing loss-based optimizations are prone to overfitting male experts when training data suffers from a dominant distribution of male experts in successful teams, and the female experts have participated sparingly, reinforcing the unfair gender disparity.

3.2 Demographic Parity

To eschew varied interpretations and to provide actionable criteria to design and evaluate fairness-aware methods, fairness has been mathematically formalized, with a level of abstraction from an underlying real-world scenario, based on well-known notions of justice and equity at a group level, like females vs. males, including demographic parity, equality of odds, and equality of opportunity. In this paper, we focus on the *demographic parity* notion of fairness and defer the exploration of other notions to future work.

Given the protected attribute *gender* of values $\{f : \text{female}, m : \text{male}\}$, we divide experts into the protected groups of $\mathcal{G}_f = \{j_f\} \subseteq \mathcal{E}$ and $\mathcal{G}_m = \{j_m\} \subseteq \mathcal{E}$ for female experts vs. male ones. Given D the set of decisions, demographic parity requires a decision $d \in D$ for members of protected groups to be *in*dependent of the value of the protected attribute [16]. Formally,

$$\forall d \in D : p(\hat{d}|j_f) = p(\hat{d}|j_m) \tag{6}$$

Table 1: Statistics of the imdb.

	raw	filtered
#teams	507,034	32,059
#unique experts	876,981	2,011
#unique female experts	**-**	**248**
#unique male experts	-	1,763
#unique skills	28	23
#team w/ single expert	**322,918**	**0**
#team w/ single skill	315,503	15,180
avg #expert per team	1.88	3.98
avg **#female expert per team**	**-**	**0.01**
avg #male expert per team	-	3.91
avg #skill per team	1.54	1.76
avg #team per expert	1.09	62.45
avg #skill per expert	1.59	10.85

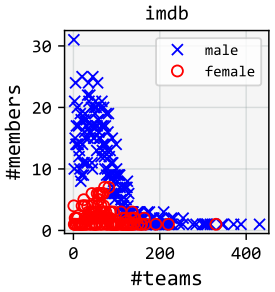

Fig. 2: Sparse vs. dominant distribution of female and male experts.

where $\hat{d}$ is the predicted decision for the correct decision d. In fair team recommendation, decisions are about the Boolean membership status of experts in the recommended subset of experts $\mathbf{e}$, i.e., $j \in \mathbf{e}$ or $j \notin \mathbf{e}$. Hence, e.q. 6 becomes:

$$\forall j_f \in \mathcal{G}_f, j_m \in \mathcal{G}_m; \overbrace{[p(j_f \in \mathbf{e}) = p(j_m \in \mathbf{e})]}^{\text{true positives}} \wedge \overbrace{[p(j_f \notin \mathbf{e}) = p(j_m \notin \mathbf{e})]}^{\text{true negatives}} \quad (7)$$

Intuitively, demographic parity enforces the membership in a team to be independent of values of the protected attribute for team members, i.e., no regard to their gender or any other protected characteristics.

From e.q. 5, neural team recommendation methods calculate the loss values only for those who should be members of the optimum team, i.e., $j \in \mathbf{e}$ or true positives, primarily for training efficiency, overlooking $j \notin \mathbf{e}$ or true negatives in e.q. 7; The number of experts in an optimum team $|\mathbf{e}|$, is significantly less than the number of unique experts $|\mathcal{E}|$, i.e., $|\mathbf{e}| \ll |\mathcal{E}|$ as in imdb where $3.98 \ll 2{,}011$, hence, calculating the loss for experts who should *not* be in the optimum team, i.e., $j \notin \mathbf{e}$ or true negatives, for every training samples would be computationally prohibitive. Secondly, even for the true positives in e.q. 7, their loss function (e.q. 5) is oblivious to the prior distribution of female and male experts in the training datasets. Therefore, the dominant distribution of male experts in training instances of teams results in more loss values and further frequent updates to the neural model's parameters but for male experts, leaving little to no update for female experts due to their sparse distribution in teams.

3.3 vivaFemme: Female-Advocate Loss Regularization

The overarching theme of our paper is to propose a gender-aware loss function that advocates the minority protected group, herein, the female experts $\mathcal{G}_f$, to dampen the effect of the majority protected group, i.e., the male experts $\mathcal{G}_m$. We propose vivaFemme, which modifies the original loss in e.q. 5 via two regularization components with respect to e.q. 7, as explained hereafter.

Foremost, we make a distinction for the loss values of female and male expert members of an optimum team, i.e., $\mathbf{e} = \{j_f \in \mathbf{e}\} \cup \{j_m \in \mathbf{e}\}$, or female true positives vs. male ones. For female experts $j_f \in \mathbf{e}$, given their very sparse distribution in teams in a biased dataset, thereby very low probability of their

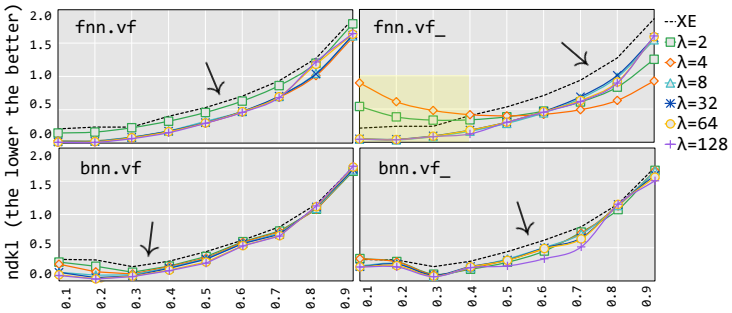

Fig. 3: Comparative fairness results for an increasing ratio of female experts.

encounter during a model training, we drastically penalize the model by amplifying the loss values via a punitive coefficient λ_1 should it miss the correct recommendations for such scarce and invaluable opportunities. In contrast, the model expects frequent encounters with male experts during training, whose accumulative loss values over many teams would provide sufficient guides toward their correct recommendations. In doing so, we strictly avoid the model's tendency to under-represent female experts in the optimum teams and provide a balance in the true positive part of e.q. 7. Formally,

$$p(\mathbf{e}|\mathbf{s},\boldsymbol{\theta}) \propto \overbrace{\sum_{j_m \in \mathbf{e}} \log \sigma(\mathbf{z}[j_m])}^{\text{male true positives}} + \overbrace{\sum_{j_f \in \mathbf{e}} \lambda_1 \times \log \sigma(\mathbf{z}[j_f])}^{\text{female true positives}} \tag{8}$$

Next, with regard to the true negatives in e.q. 7, we assume that given a *fixed* number of false recommendations of experts, a model that incorrectly recommends more female experts than male experts is preferable, because, while it yields the same performance drain, it advocates more participation of female experts in teams. Indeed, in a biased dataset with a predominant distribution of male experts, not only does the model become biased toward recommending more male experts under the binary cross-entropy, but also produces more false recommendations of male experts and, hence, unbalancing the true negatives part of the e.q. 7. To counter, we enforce a neural model to recommend more female experts for an optimum team, even incorrectly and knowing the cost of female false positives (which is negligible in view of a male-dominant dataset), by considering random samples of female experts who are *not* members of the optimum team as *virtually* correct members of the team, i.e., $j'_f \notin \mathbf{e} \rightarrow j'_f \in^* \mathbf{e}$ and calculate their losses by a punitive coefficient λ_2. Hence, should a neural model falsely recommend incorrect experts, it favours female experts over male experts. Formally,

$$p(\mathbf{e}|\mathbf{s},\boldsymbol{\theta}) \propto \sum_{j_m \in \mathbf{e}} \log \sigma(\mathbf{z}[j_m]) + \sum_{j_f \in \mathbf{e}} \lambda_1 \times \log \sigma(\mathbf{z}[j_f]) + \overbrace{\sum_{j'_f \sim \mathbb{P}: j'_f \notin \mathbf{e}} \lambda_2 \times \log \sigma(\mathbf{z}[j'_f])}^{\text{female false negatives}} \tag{9}$$

where $\mathbb{P}$ is the probability distribution from which we draw k female experts j_f as *virtually* correct member of the optimum team $(\mathbf{s},\mathbf{e})$ where $j_f \in \mathbf{e}$ but $j'_f \notin \mathbf{e}$.

4 Experiments

We present our experiments to answer the following research questions:
RQ1: Can our proposed loss function mitigate the gender bias in neural team recommendation methods while maintaining models' accuracy?
RQ2: Is the proposed loss's impact consistent across different desired distributions of female vs. male experts in the recommended teams?
RQ3: To what extent does the random sampling of female experts contribute to gender bias mitigation?

4.1 Setup

Dataset. We assess viva$\mathscr{F}$emme's effectiveness on `imdb`, a widely recognized benchmark dataset within the domain of team recommendation [5], where each entry is a moving picture such as a movie or a tv series including the top-10 short-listed cast and crew such as directors, producers, actors and actresses. We consider each movie as an optimum team, which has been successfully produced, the cast and crew as the team members, and the genres as the skills of the team. It's important to note that the utilization of `imdb` in team recommendation literature differs from its applications in movie recommender systems or movie review analysis; our objective here is to assemble teams of cast and crew for movie *production* rather than movie recommendation. For the cast and crew's gender labels, we inferred the gender of some cast and crew by their role identified as actor or actress. For the rest, we utilized `genderize` [1]. We filtered out singleton and sparse teams with less than 3 members, as suggested by [18]. Table 1 reports statistics on the raw and filtered dataset. Also, as seen in Figure 2, male experts are dominating teams while female experts have participated sparingly.

Baselines. We compare the impact of our proposed loss function on mitigating the gender bias of two reference neural architectures: (1) feed-forward non-Bayesian (non-variational) neural network (`fnn`) [5,18] and (2) the-state-of-the-art Bayesian (variational) neural network(`bnn`). Both models include a single hidden layer of size `d=128` and `leaky relu` is the activation function for the hidden layer. For the input layer, we used sparse occurrence vector representations (multi-hot encoded) of skills of size $|\mathcal{S}|$. The output layer is the sparse occurrence vector representations (multi-hot encoded) of experts of size $|\mathcal{E}|$. We trained the neural models using the binary cross-entropy (`xe`) as the biased baseline, viva$\mathscr{F}$emme without random samplings of female experts (`vf_`), and viva$\mathscr{F}$emme with random sampling (`vf`) of $k = |\mathcal{G}_f|$ female experts from $\mathbb{P}$ =`uniform` distribution for increasing punitive coefficients $\lambda_1 = \lambda_2 = \lambda \in \{2,4,8\ldots,128\}$.

4.2 Evaluation Strategy and Metrics

We randomly select 15% of teams for the test set and perform 5-fold cross-validation on the remaining teams for model training and validation that results in one trained model per fold. Given an optimum team $(\mathbf{s}, \mathbf{e})$ from the test set, we compare the top-k ranked list of experts, recommended by the model of each fold, with the observed subset of experts $\mathbf{e}$ and report the average accuracy of

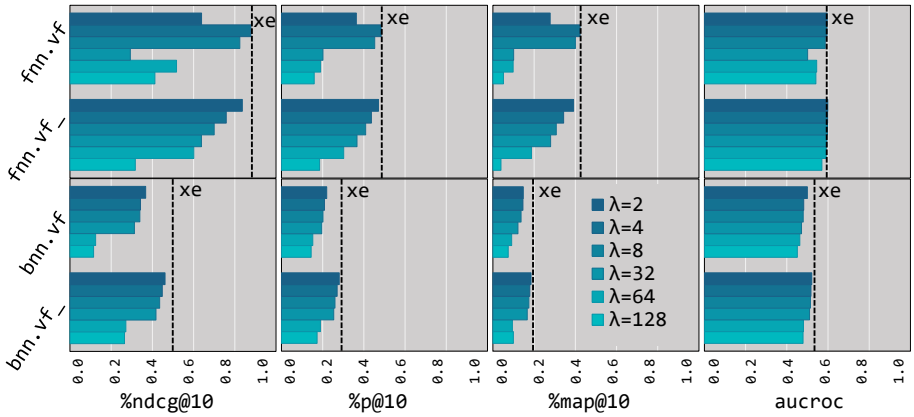

Fig. 4: Comparative accuracy results with/out sampling.

models on all folds in terms of normalized discounted cumulative gain (`ndcg`), mean average precision (`map`), and precision (`p`) at top-`10` as well as area under the receiver operating characteristic (`rocauc`). In all such metrics, the higher value indicates more accurate recommendations.

To evaluate fairness, we used `ndkl`, which measures the *divergence* of the distribution of a protected group in the ranked list of recommendations with the ideal fair distribution using Kullback–Leibler [14], and the lower divergence the better with being `0` in the ideal equal distributions. We measure `ndkl` for increasing ratios of `0.1` to `0.9` for female experts as the ideal fair distributions.

4.3 Result

In response to **RQ1**, from Figure 3 (left), we observe that our proposed loss with random sampling consistently yields less divergence across increasing desired ratios of female experts in the recommended teams, hence, better fairness compared to the binary cross-entropy (`xe`) across neural models and increasing punitive coefficients. For instance, at the ideal ratio of `0.9` where we expect to observe `90%` female experts, not unexpectedly, all loss functions fall short of reaching the ideal distribution, yet our loss function obtains closer distribution across increasing values of λ. For lower ratios, like `0.5` where we expect a balance between female and male experts in the recommended teams, while all loss functions have been more or less effective, viva$\mathscr{F}$emme is the most effective one across different punitive coefficients in both `bnn` and `fnn` models. Further, from Figure 4 and for `fnn`, we can observe that viva$\mathscr{F}$emme with sampling (`fnn.vf`) could obtain the fairest distributions for different ideal ratios of female experts yet maintaining success rate in terms of accuracy metrics for $\lambda = 4$ or with a minor decrease for $\lambda = 8$. Based on our observations from Figure 3 (left) and Figure 4, we can conclude that viva$\mathscr{F}$emme can mitigate the unfair gender bias against female experts in feed-forward neural team recommendation methods while maintaining the model's accuracy. Looking at the Bayesian neural model in Figure 4 (`bnn.*`), we foremost observe its generally poor performance compared to `fnn`, as already shown by Dashti et al. [5], which has been attributed to

the small number of skills (genres of movies) and fair distribution of movies over skills. In contrast with `fnn`, our proposed loss function mitigates `bnn`'s bias with little to no discounts on the model's accuracy when there is *no* sampling in our loss function (`bnn.vf_`). As seen, while increasing the punitive coefficients in our loss function reduces the `bnn`'s accuracy, the negative impact is more pronounced when we apply random sampling (`bnn.vf`). Predictably, our findings are aligned with Dashti et al. [5] where sampling methods have shown futile for `bnn` models. In sum, our results urge further investigation for loss function to mitigate `bnn`'s bias while maintaining accuracy.

In response to **RQ2**, whether viva$\mathscr{F}$emme's impact is consistent across different desired distributions of female experts in the recommended teams, from Figure 3, we observe a consistent trend of lower `ndkl` values for our method across neural models and increasing ratios from `0.1` to `0.9`, compared to the binary cross-entropy (`xe`). Therefore, viva$\mathscr{F}$emme is effective for different desired distributions of female experts in the recommended teams. However, as we mentioned in **RQ1**, `bnn.*` recommends fairer teams but at the cost of slightly lower accuracy.

Regarding **RQ3**, that is, to what extend each of the viva$\mathscr{F}$emme's component contribute to its effectiveness, from Figure 3 (right), we observe that viva$\mathscr{F}$emme without random sampling for female experts (`fnn.vf_`) follows similar trend as in viva$\mathscr{F}$emme with random sampling for female experts (`fnn.vf`) from ideal ratios above `0.4` across the increasing range of punitive coefficients (λ) for feed-forward model (`fnn.vf_`). However, for lower ratios like `0.3` to `0.1`, viva$\mathscr{F}$emme without random sampling for female experts (`fnn.vf_`) yield higher values of `ndkl` and falls short of providing the desired distribution in `fnn`. Moreover, from Figure 4, `fnn.vf_` fails to maintain the model's performance across accuracy metrics and increasing values of punitive coefficients, except that of $\lambda = 2$. Therefore, while the effort by the viva$\mathscr{F}$emme without positive sampling is noteworthy, it is insufficient, and the random sampling component is critically required for viva$\mathscr{F}$emme's consistent mitigation of unfair gender bias.

For the Bayesian model (`bnn.*`), our loss function, with and without random sampling, has lower `ndkl`, hence, better fairness across all coefficients. However, its positive impact on fairness is marginal compared to the binary cross-entropy baseline (`xe`), indicating further research on loss functions for variational models.

5 Conclusion

In this paper, we addressed gender bias in neural team recommendation models via an in-process loss-based method. Wherein we substantially penalize the model for failing to recommend true female expert members while promoting more female inclusion by favoring them over male experts in case of incorrect recommendations. Our experiments on the `imdb` dataset demonstrate the effectiveness of our approach in debiasing gender imbalances in feed-forward neural models while maintaining accuracy. For variational Bayesian neural models, while our proposed loss function reduces gender bias, it comes at a slight accuracy reduction. Future work involves developing an in-process method to mitigate gender bias in variational models as well as studying our method under

other notions of fairness, e.g., equality of opportunity, on datasets from diverse domains like patents and `github`.

References

1. `https://genderize.io/`, [Online; accessed 16-June-2023]
2. Barnabò, G., Fazzone, A., Leonardi, S., Schwiegelshohn, C.: Algorithms for fair team formation in online labour marketplaces10033. In: WWW 2019
3. Bigdeli, A., Arabzadeh, N., SeyedSalehi, S., Zihayat, M., Bagheri, E.: Gender fairness in information retrieval systems. SIGIR '22
4. Chen, R.J., Wang, J.J., Williamson, D.F., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., Mahmood, F.: Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature biomedical engineering (2023)
5. Dashti, A., Samet, S., Fani, H.: Effective neural team formation via negative samples. In: CIKM (2022)
6. Dashti, A., Saxena, K., Patel, D., Fani, H.: Opentf: A benchmark library for neural team formation. In: CIKM (2022)
7. Fani, H., Barzegar, R., Dashti, A., Saeedi, M.: A training strategy for future collaborative team prediction. In: ECIR (2024)
8. Feng, Y., Shah, C.: Has ceo gender bias really been fixed? adversarial attacking and improving gender fairness in image search. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022)
9. Hajian, S., Bonchi, F., Castillo, C.: Algorithmic bias: From discrimination discovery to fairness-aware data mining. In: SIGKDD. pp. 2125–2126 (2016)
10. Hamidi Rad, R., Bagheri, E., Kargar, M., Srivastava, D., Szlichta, J.: Retrieving skill-based teams from collaboration networks. SIGIR '21, Association for Computing Machinery (2021)
11. Hamidi Rad, R., Fani, H., Bagheri, E., Kargar, M., Srivastava, D., Szlichta, J.: A variational neural architecture for skill-based team formation. ACM Trans. Inf. Syst. (2023)
12. Jalal, A., Karmalkar, S., Hoffmann, J., Dimakis, A., Price, E.: Fairness for image generation with uncertain sensitive attributes. In: International Conference on Machine Learning. PMLR (2021)
13. Kay, M., Matuszek, C., Munson, S.A.: Unequal representation and gender stereotypes in image search results for occupations. In: Proceedings of the 33rd annual acm conference on human factors in computing systems (2015)
14. Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics (1951)
15. Loghmani, H., Fani, H.: Bootless application of greedy re-ranking algorithms in fair neural team formation. In: Boratto, L., Faralli, S., Marras, M., Stilo, G. (eds.) Advances in Bias and Fairness in Information Retrieval (2023)
16. Pitoura, E., Stefanidis, K., Koutrika, G.: Fairness in rankings and recommendations: an overview. VLDB J. (2022)
17. Rad, R.H., Bagheri, E., Kargar, M., Srivastava, D., Szlichta, J.: Subgraph representation learning for team mining. In: WebSci (2022)
18. Rad, R.H., Fani, H., Kargar, M., Szlichta, J., Bagheri, E.: Learning to form skill-based teams of experts. In: CIKM '20 (2020)
19. Sapienza, A., Goyal, P., Ferrara, E.: Deep neural networks for optimal team composition. Frontiers Big Data (2019)