Debiasing Neural Retrieval via In-batch Balancing Regularization

Yuantong Li[1]*, **Xiaokai Wei**[2†], **Zijian Wang**[2†], **Shen Wang**[2†],
Parminder Bhatia[2], **Xiaofei Ma**[2], **Andrew Arnold**[2]
[1]UCLA
[2]AWS AI Labs
yuantongli@ucla.edu; xiaokaiw, zijwan, shenwa, parmib,
xiaofeim, anarnld@amazon.com

Abstract

People frequently interact with information retrieval (IR) systems, however, IR models exhibit biases and discrimination towards various demographics. The in-processing fair ranking methods provide a trade-offs between accuracy and fairness through adding a fairness-related regularization term in the loss function. However, there haven't been intuitive objective functions that depend on the click probability and user engagement to directly optimize towards this. In this work, we propose the **I**n-**B**atch **B**alancing **R**egularization (IBBR) to mitigate the ranking disparity among subgroups. In particular, we develop a differentiable *normed Pairwise Ranking Fairness* (nPRF) and leverage the T-statistics on top of nPRF over subgroups as a regularization to improve fairness. Empirical results with the BERT-based neural rankers on the MS MARCO Passage Retrieval dataset with the human-annotated non-gendered queries benchmark (Rekabsaz and Schedl, 2020) show that our IBBR method with nPRF achieves significantly less bias with minimal degradation in ranking performance compared with the baseline.

1 Introduction

Recent advancements in Natural Language Processing and Information Retrieval (Palangi et al., 2016; Devlin et al., 2019; Zhao et al., 2020; Karpukhin et al., 2020) have led to great progress in search performances. However, search engines easily expose various biases (e.g., (Biega et al., 2018; Baeza-Yates, 2018; Rekabsaz and Schedl, 2020; Rekabsaz et al., 2021)), which sabotage the trust of human beings from day to day. Many methods have been proposed recently to reduce the bias of the retrievers. Existing fairness-aware ranking methods can be categorized into pre-processing methods, in-processing methods, and post-processing methods (Mehrabi et al., 2021; Zehlike et al., 2021). Pre-processing methods typically focus on mitigating bias in data before training the model. Lahoti et al. (2019) discussed the individual fairness pre-processing method to learn the fair representation of data. However, the representation-based method will undermine the value of the features determined by domain experts (Zehlike et al., 2021). The in-processing methods usually transform the fairness in ranking task into an optimization problem consisting of an accuracy objective and a fairness objective. These methods learn the best balance between these two objectives (Kamishima et al., 2011; Berk et al., 2017; Bellamy et al., 2018; Konstantinov and Lampert, 2021). Zehlike and Castillo (2020) handles different types of bias without knowing the exact bias form; Post-processing algorithms (Singh and Joachims; Zehlike et al., 2017, 2020; Cui et al., 2021) are model agnostic without requiring access to the training process, but these methods re-order the ranking at the expense of accuracy (Menon and Williamson, 2018).

Among recent works on fair neural retrieval, Beutel et al. (2019) introduce the pairwise ranking fairness (PRF) metric for ranking predictions. This pairwise fairness metric evaluates whether there is a difference in accuracy between two groups. Rekabsaz et al. (2021) (AdvBert) mitigates the bias magnitude from the concatenation of query and passage text rather than treating the bias magnitude from query and passage separately through an adversarial neural network.

In this paper, we propose the In-Batch Balancing Regularization (IBBR) method combined with the neural retrieval model. IBBR is an in-processing debiasing method that balances the ranking disparity among different demographic groups by adding an in-batch balancing regularization term to the objective function. We design two batch-level regularization terms, *Pairwise Difference* (PD) and *T-statistics* (TS) that measure biases within demographic groups. In addition, we introduce normed Pairwise Ranking Fairness (nPRF), a relaxed ver-

*Work done during an internship at AWS.
†Equal contribution.

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 58 - 66

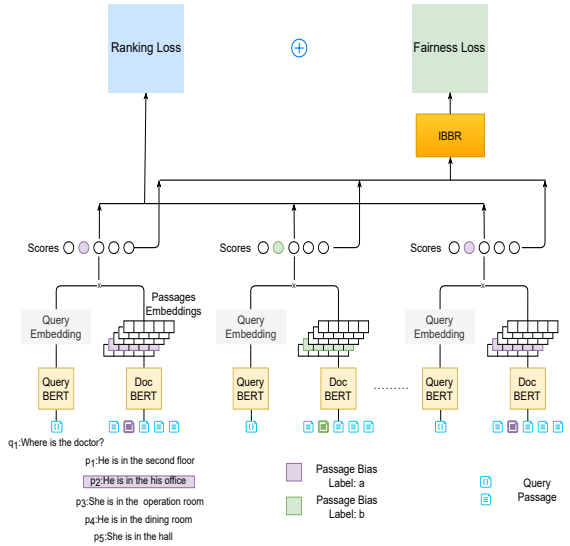

Figure 1: An example of In-Batch Balancing Regularization method. For each query, we calculate the typical ranking loss and the fairness loss from IBBR on top K retrieved passages. We jointly optimize the ranking loss and the fairness loss. There are two ways of computing the IBBR, pairwise difference loss and T-statistics Loss.

sion of the PRF (Beutel et al., 2019) that is differentiable, thus could be directly optimized. We apply IBBR to MS MARCO passage re-ranking task (Nguyen et al., 2016) on gender bias using pre-trained BERT$_{L_2}$ and BERT$_{L_4}$ models (Turc et al., 2019). Empirical results show that our model could achieve significantly less bias with minor ranking performance degradation, striking a good balance between accuracy and fairness. Our contributions can be summarized as follows:

- We introduce IBBR, an in-processing debiasing method based on pairwise difference and T-statistics.
- We introduce normed PRF, a relaxed version of the pairwise ranking fairness (PRF) metric (Beutel et al., 2019). The normed PRF solves the non-differentiable issue and could be directly optimized during training.
- We perform experiments on the MS MARCO passage re-ranking task with IBBR and normed PRF. Empirical results show that IBBR and normed PRF could achieve a statistically significant improvement in fairness while maintaining good ranking performance.

2 PROBLEM DEFINITION

We first introduce notations in the ranking task in §2.1. §2.2 provides the definition of the bias of the passage. In §2.3, we propose the definition of the group fairness in the ranking task.

2.1 Notations in the Ranking Task

Formally, we define the task of *Gender Debiased Neural Retrieval* (GDNR) as: given a query q and top K passages retrieved by the neural retrieval system, we adapt the ranking to mitigate bias in the retrieval result. We first define the whole query set as $Q = \{q_1, q_2, ..., q_N\}$. For each query q_i, we denote $P_i = \{p_{i,1}, p_{i,2}, \ldots, p_{i,j}, \ldots, p_{i,K}\}$ as the corresponding retrieved passages' set for query q_i. With query q_i and corresponding retrieved passages P_i, $s_i = \{q_i, p^+_{i,1}, p^-_{i,2}, \ldots, p^-_{i,K}\}$ is defined as one data pair. Here $p^+_{i,1}$ is the ground truth passage (clicked passage) and $p^-_{i,j}$ is the non-clicked passage, $\forall j \in \{2, 3, ..., K\}$. We use $Y_{i,j} = 1$ to label the passage p_j as a clicked passage, otherwise, $Y_{i,j} = 0$. Finally, the whole dataset is defined as $D = \{s_1, s_2, ..., s_N\}$. For notation simplicity, we use $[1 : K]$ to represent $\{1, 2, ..., K\}$.

2.2 Bias Label of Passage

We first provide the definition of the bias label of one passage, and consider the gender bias as a running example. Rekabsaz and Schedl (2020) use the degree of gender magnitude in the passage to define the bias value, where the gender concept is defined via using a set of highly representative gender definitional words. Such a gender definitional set usually contains words such as *she, woman, grandma* for female definitional words (G_f), and *he, man, grandpa* for males definitional words (G_m).

The definition of the bias of the passage in our method is different from (Rekabsaz and Schedl, 2020) who assume that one passage has two magnitudes: female magnitude and male magnitude. However, we assume that one passage has only one implication or tendency and use the gender magnitude difference as the bias value. So the bias value of the passage p, $mag(p)$, defined as

$$mag(p) = \sum_{w \in G_m} \log |\langle w, p \rangle| - \sum_{w \in G_f} \log |\langle w, p \rangle|, \quad (1)$$

where $|\langle w, p \rangle|$ refers to the number of occurrences of the word w in passage p, $w \in G_m$ or G_f. Furthermore, we define the bias label for the passage p as $d(p)$, if $mag(p) > 0$, then $d(p) = 1$ (male-biased); if $mag(p) < 0$, then $d(p) = -1$ (female-biased); if $mag(p) = 0$, then $d(p) = 0$ (neutral). So for each retrieved passage $p_{i,j}, j \in$

$[1:K], i \in [1:N]$, it has one corresponding bias label $d(p_{i,j}) \in \{-1, 0, 1\}$.

2.3 Group Fairness in Ranking

In §2.3.1, we introduce one metric of ranking group fairness (pairwise ranking fairness) proposed by (Beutel et al., 2019). In §2.3.2, we provide a more refined definition of pairwise ranking fairness.

2.3.1 Pairwise Ranking Fairness

If $R(p) \in [0, 1]$ is the ranking score of passage p from one retrieval model , $\text{PRF}_m(s_i)$ measures the probability level of a male-biased random passage selected from the male group m higher than all random female-biased passages of data pair s_i

$$\text{PRF}_m(s_i) = \frac{1}{n_1^m(s_i)n_0(s_i)} \sum_{j \in g_1^m(s_i)} \sum_{k \in g_0(s_i)} \mathbb{1}[R(p_{i,j}) \geq R(p_{i,k})], \tag{2}$$

where $g_1^m(s_i) = \{j | d(p_{i,j}) = 1, Y_{i,j} = 1, j \in [1:K]\}$ represents passages clicked ($Y_{i,j} = 1$) as well as belonging to male biased group ($d(p_{i,j}) = 1$). $n_1^m(s_i) = |g_1^m(s_i)|$ represents the number of male-biased clicked passages. $g_0(s_i) = \{j | Y_{i,j} = 0, j \in [1:K]\}$ represents the group of non-clicked passages. $n_0(s_i)$ represents the number of all non-clicked samples in retrieved passages. Beutel et al. (2019) use the probability that a clicked sample is ranked above another non-clicked sample for the sample query as the pairwise accuracy. The pairwise fairness asks whether there is a difference between two groups when considering the pairwise accuracy as the fairness level metric.

However, we find that PRF is not directly applicable as an argument in the regularizer of a loss function that works as a trade-off of accuracy and fairness. Because PRF is a *0-normed objective function*, which is non-convex and non-differentiable. So we propose a modified PRF that can be optimized directly.

2.3.2 Normed-Pairwise Ranking Fairness

We propose a relaxed version of PRF called normed-PRF (nPRF), which measures the degree of group fairness in retrieval results for a given query and considering the ranking performance as well. The detailed definition of nPRF_m is defined over all clicked male-biased passages p_i in a data pair s is

$$\text{nPRF}_m = \left\{ \frac{1}{n_1^m(s_i)n_0(s_i)} \sum_{j \in g_1^m(s_i)} \sum_{k \in g_0(s_i)} |R(p_{i,j})|^2 \mathbb{1}[R(p_{i,j}) \geq R(p_{i,k})] \right\}^{\frac{1}{2}}, \tag{3}$$

where $n_1^m(s_i)$ is the number of all clicked male-biased passages in a data pair, usually $n_1^m(s_i) = 1$ in the ranking system.

In order to avoid the drawback of PRF being non-differentiable, we multiply the square of the ranking score ($|R(p_{i,j})|^2$) of the passage p_j to the indicator function $\mathbb{1}[R(p_{i,j}) \geq R(p_{i,k})]$, which is differentiable. Besides, $\frac{1}{n_0(s_i)} \sum_{j \in g_0(s_i)} |R(p_{i,j})|^2 \mathbb{1}[R(p_{i,j}) \geq R(p_{j,k})]$ measures the average harm of the biased passage $p_{i,j}$. If this value is large, it means that on average, these non-clicked passages are more relevant to the clicked passage $p_{i,j}$. This contributes more harm to the society since people are more willing to accept the ranking result. If this value is small, it means that on average, these non-clicked passages are less irrelevant to the click passage p_i. This contributes less harm to the society since people are less willing to accept the ranking result. Thus, the nPRF not only considers the magnitude of the ranking performance of the retrieval results but also inherits the explainable society impact into the PRF.

3 Algorithms

In this section, we create a regularizer based on the nPRF to mitigate the gender bias. In §3.1, we introduce necessary components for the neural retrieval task. In §3.2, we provide the definition of the ranking loss and two fairness loss functions, *Pairwise Difference Loss* and *T-statistics Loss*, acting as a regularizer, named as *in-batch balancing regularization method* (IBBR).

3.1 Rank Model

Given the data set D, we use the two-tower dense passage retriever (DPR) model (Karpukhin et al., 2020) as our retrieval model. DPR uses two dense encoders E_P, E_Q which map the given text passage and input query to two d-dimensional vectors ($d = 128$) and retrieves K of these vectors which are close to the query vector. We define the ranking score between the query and the passage using the dot product of their vectors produced from DPR as $\text{sim}(q_i, p_{i,j}) = z_{q_i}^\top z_{p_{i,j}}$, where $z_{q_i} = E_Q(q_i)$ and

$z_{p_{i,j}} = E_P(p_{i,j})$ are the corresponding query and passage dense embeddings.

Remarks. Here we use two-tower DPR for two reasons. (I) Computational considerations. Humeau et al. (2019) thoroughly discussed the pros and cons between cross-encoders (Nogueira and Cho, 2019) and bi-encoders such as DPR and stated that cross-encoders are too slow for practical use. (II) Using cross-encoders can cause ill-defined problem such as, if the query's bias label belongs to groups m and the passage's bias label belongs to group f, the concatenation of these two texts' bias label is unclear, based on the definition provided in Eq. (2) from (Rekabsaz et al., 2021). So the two-tower BERT model is applied separately on the query and document to tackle this ill-defined problem. Here we only consider the DPR as our ranking model.

Encoders. In our work, in order to demonstrate the robustness of IBBR, we use two BERT models (Turc et al., 2019), (1) tiny BERT (base, uncased); (2) mini BERT (base, uncased) as our encoders, and take the representation at the [CLS] token as the output.

Inference. For the data pair s_i, the ranking score $R(p_{i,j})$ of passage p_j for query q_i is simply the inner product of $\text{sim}(q_i, p_{i,j})$ produced by DPR encoders.

3.2 Loss Functions

3.2.1 Ranking Loss

The ranking loss is the *negative log-likelihood loss* by computing the inner product of query and passage embeddings to measure the ranking performance for the data pair s_i,

$$\text{L}^{\text{Rank}} = -\log \frac{e^{(\text{sim}(q_i, p^+_{i,1}))}}{e^{(\text{sim}(q_i, p^+_{i,1}))} + \sum_{j=2}^{K} e^{(\text{sim}(q_i, p^-_{i,j}))}}.$$

3.2.2 Fairness Loss

To mitigate the bias for two groups, we use the ranking disparity as a measure to evaluate the fairness level of the neural retrieval system. And this ranking disparity works as a regularization in the loss function. Here we propose two regularization terms as follows.

(I) Pairwise Difference Loss. The pairwise difference (PD) loss $\text{L}_{\text{P}}^{\text{Fair}}$ measures the average ranking disparity between two groups m and f over a batch size (B) of data pairs,

$$\text{L}_{\text{P}}^{\text{Fair}} = \frac{1}{n_m n_f} \sum_{c \in p^m_{[1:B]}} \sum_{d \in p^f_{[1:B]}} (\text{nPRF}_m(s_c) - \text{nPRF}_m(s_d))^2, \tag{4}$$

where $P^m_{[1:B]} = \{i | p_{i,j} \in g^m_1(s_i), i \in [1:B], j \in [1:K]\}$ is the set that the clicked passage belongs to group m over batch size B data. $P^f_{[1:B]} = \{i | p_{i,j} \in g^f_1(s_i), i \in [1:B], j \in [1:K]\}$ is the set that the clicked passage belongs to group f over batch size B data, and $n_m = |P^m_{[1:B]}|$ and $n_f = |P^f_{[1:B]}|$.

Remarks. If there are many $\text{nPRF}_m(s_x)$ which are different from other $\text{nPRF}_f(s_y)$, this means that group m and group f have different fairness level over this batch data and will introduce more loss. However, this PD loss does not consider distribution information over this batch data, and imbalanced-data issue when group m and group f samples are imbalanced. Thus we propose the T-statistics loss to overcome this.

(II) T-statistics Loss. The design of T-statistics (TS) loss is also based on the ranking disparity but considers the second order information (variance effect) of each group for each batch data. We use the square of T-statistics as the ranking disparity measure and defined as,

$$\text{L}_T^{\text{Fair}} = \left\{ (\hat{\mu}_m - \hat{\mu}_f)^2 / \sqrt{\hat{\text{var}}_m / n_m + \hat{\text{var}}_f / n_f} \right\}^2,$$

where $\hat{\mu}_m = \frac{1}{n_m} \sum_{j \in P^m_{[1:B]}} \text{nPRF}_m(j)$ is the mean of the male group's nPRF, and $\hat{\text{var}}_m = \frac{1}{n_m} \sum_{j \in P^m_{[1:B]}} (\text{nPRF}_m(j) - \hat{\mu}_m)^2$ is the variance of the male group's nPRF. Besides, $\hat{\mu}_f, \hat{\text{var}}_f$ can be defined similarly.

Remarks. This TS loss can provide a robust measure for the ranking disparity especially when the batch data pair is imbalanced. The square of the T-statistics, i.e., χ^2 distribution, provides the theoretical guarantee and power to reject the similarity between group m and group f.

Total Loss. The total loss will be the sum of the ranking loss and fairness loss, represented as $L^{\text{total}}_{[1:B]} = L^{\text{rank}}_{[1:B]} + \lambda L^{\text{fair}}_{[1:B]}$, where L^{fair} can be the PD loss or TS Loss. λ is a hyperparameter to control the balance of the fairness loss and ranking loss. In the experiment, we try manually and automatically to tune λ_{fair}. The details of our method can be found in Figure 1.

4 Experiments

In this section, we describe data resources in §4.1, experiment setup in §4.2, evaluation metrics in Section 4.3, baseline models in Section 4.4, and corresponding result analysis in Section 4.5.

4.1 Dataset

We experimented on the passages re-ranking task from MS MARCO (Nguyen et al., 2016). This collection includes 8.8 million passages and 0.5 million queries, comprised of question-style queries from Bing's search logs, accompanied by human-annotated clicked/non-clicked passages. Additionally, data bias labels over this dataset are available from (Rekabsaz and Schedl, 2020).

Data For DPR. The whole dataset is composed of total 537,585 queries and $K * 537,585$ retrieved passages where $K = 200$, for the baseline DPR model. Each query has top K passages including one ground truth and 199 negative samples. The details of splitting the dataset used for training, development, and test (7:2:1) for the DPR model can be found in Appendix A Table 3. There are 126 queries used for the final evaluation.

Data For Fair Model. The fairness dataset (Rekabsaz and Schedl, 2020) is also created upon this MS MARCO dataset. These queries were annotated into one of four categories: non-gendered (1765), female (742), male (1,202), other or multiple genders (41). Here we only use the non-gendered queries, and assume the query is unbiased given it does not have any gender definitional terms. There are 1,252 unique queries in total. Examples of non-gendered queries are: *what is a synonym for beautiful?*, *what is the meaning of resurrect?*, etc.

4.2 Experiment Setup

The maximum length of query and passage are set to 100. Batch size B is 150 optimized over $\{100, 120, 150\}$. Learning rate is $3e^{-5}$ optimized over $\{3e^{-6}, 3e^{-5}, 3e^{-4}\}$. A warmup ratio of 10% with linear scheduler and a weight decay of 0.01 are set. In addition, we searched the fairness penalty parameter $\lambda = [0.1, 0.5, 1, 5, 10]$ (Best). We also experimented setting the λ_{fair} as a trainable parameter (Auto). All experiments are conducted ten times and we reported the average.

4.3 Evaluation Metrics

Ranking metrics. We use Recall@10, MRR, and NDCG@10 to evaluate the ranking performance.

Fairness metrics. We use RaB@5, RaB@10, and ranking disparity |ΔA-PRF| to evaluate the fairness magnitude.

RaB$_t$. RaB$_t$ is a measurement of ranking bias, which is based on the average of the gender magnitude of passages at top t ranking list (Rekabsaz and Schedl, 2020). To measure the retrieval bias, RaB calculates the mean of the gender magnitudes of the top t (5 or 10) retrieved documents for the data pair s_i, for females, $\text{qRaB}_t^f(s_i) = \frac{1}{t}\sum_{j=1}^{t} mag_f(p_{i,j})$. Using these values, the RaB metric of the query q, $\text{RaB}_t(s_i) = \text{qRaB}_t^m(s_i) - \text{qRaB}_t^f(s_i)$, and the RaB metric of the retrieval model over all the queries, $\text{RaB}_t = \frac{1}{N}\sum_{s_i \in D} \text{RaB}_t(s_i)$. The smaller the absolute value of RaB$_t$, the less the ranking disparity is.

|ΔA-PRF|. |ΔA-PRF| measures the ranking disparity over two groups, which is the difference over two averaged PRF, $|\Delta\text{A-PRF}| = |\frac{1}{|T_m|}\sum_{i \in T_m} \text{PRF}_i - \frac{1}{|T_f|}\sum_{i \in T_f} \text{PRF}_i|$, where T_m is the dataset that the clicked passage belongs to group m, $T_m = \{i | Y_{i,j} = 1, d_{i,j} = 1, \forall i \in [1 : N]\}$. With the running example, we denote $|T_m|$ as the number of male-biased clicked pairs and similar definitions are for T_f and $|T_f|$. The smaller the |ΔA-PRF| is, the smaller the ranking disparity is. If |ΔA-PRF| is close to zero, it means that the retrieved results are relatively fair since the two groups' PRF are close to each other. To avoid selection bias, |ΔA-PRF| measures the whole dataset's fairness level rather than the subset's result such as top 5 and top 10.

4.4 Baseline Models

The baseline methods contain the classical IR models, BM25, and RM3 PRF, and neural based models: Match Pyramid (MP), Kernel-based Neural Ranking Model (KNRM), Convolutional KNRM (C-KNRM), Transformer-Kernel (TK), and the fine-tuned BERT Model. These results are available in in Appendix Section A. For the BERT rankers, we use BERT-Tiny (BERT$_{L_2}$) and BERT-Mini (BERT$_{L_4}$).

4.5 Results Analysis

Ranking Performance. In Table 1, we present the result of original BERT$_{L_2}$ and BERT$_{L_4}$ and BERT$_{L_2}$ and BERT$_{L_4}$ with IBBR (PD and TS). We found that in BERT$_{L_2}$, after adding IBBR, the ranking performance decreases 2.2% in Recall@10 and the bias level decreases 80% when applying the TS. Overall, TS outperforms PD on average

nPRF	IBBR	λ_{fair}	Ranking Metric			Fairness Metric		
			Recall@10 ↑	**MRR ↑**	**NDCG ↑**	**|ΔA-PRF| ↓**	**RaB@5 ↓**	**RaB@10 ↓**
DPR BERT(L2)			0.357	0.164	0.196	0.005	0.091	0.079
	PD	Best	0.238 (-33.3%)	0.112 (-31.7%)	0.124 (-36.7%)	0.034 (+580%)	0.094 (+3.3%)	0.083 (+5.1%)
		Auto	0.270 (-24.3%)	0.126 (-23.2%)	0.143 (-27.0%)	0.033 (+560%)	0.098 (+7.7%)	0.083 (+5.1%)
	TS	Best	**0.349 (-2.2%)**	**0.170 (+3.6%)**	**0.198 (+1.0%)**	**0.001**‡ **(-80%)**	**0.091 (0%)**	**0.075**‡ **(-5.1%)**
		Auto	0.333 (-6.7%)	0.160 (-2.4%)	0.185 (-5.6%)	0.006 (+20%)	0.109 (+19.7%)	0.077 (-2.5%)
DPR BERT(L4)			0.429	0.205	0.243	0.043	0.016	0.011
	PD	Best	0.381 (-11.1%)	0.213 (+3.9%)	**0.236 (-2.8%)**	0.034‡ (-20.9%)	0.033 (+106%)	0.025 (+127%)
		Auto	0.373 (13.1%)	**0.214 (+4.4%)**	0.234 (-3.7%)	0.030‡ (-30.2%)	0.033 (+106%)	0.021 (+90.9%)
	TS	Best	0.365 (-14.9%)	0.193 (-5.9%)	0.217 (-10.7%)	**0.000**‡ **(-100%)**	**0.003**‡ **(-81.3%)**	**0.012 (+9.1%)**
		Auto	**0.389 (-9.3%)**	0.205 (0%)	0.234 (-3.7%)	0.022‡ (-48.8%)	0.004‡ **(-75.0%)**	0.017 (+54.5%)

Table 1: The ranking and fairness results of two IBBR methods, pairwise difference and T-statistics, combined with nPRF in BERT_{L_2} and BERT_{L_4} models. We compare IBBR with baseline models DPR L2, L4 in the re-ranking tasks and experimenting with different fairness hyperparameter λ_{fair} tuning methods. The bold value in each column shows the best result in that metric. ↑ and ↓ indicate larger/smaller is better in corresponding definition of metrics. ‡ indicates statistically significant improvement (p-value< 0.05) over the DPR baseline in fairness metrics.

when considering the ranking metrics because it downgrades the ranking metric less, which can be found in the ranking metric columns. This phenomenon exists both in hand-tuned or auto-tuned hyperparameter λ_{fair} and BERT_{L_2} and BERT_{L_4}.

Fairness Performance |ΔA-PRF|. BERT_{L_2} + TS can achieve 80% reduction in mitigating |ΔA-PRF| bias. The |ΔA-PRF| fairness metric in BERT_{L_4}+TS can achieve 100% reduction in mitigating bias compared with the original BERT_{L_4}. Besides, PD performs unsatisfied in the fairness metric compared with TS in BERT_{L_2} and BERT_{L_4}, we found that the variance of nPRF and the imbalance affects the performance of PD, which is usually found in the training phase (#male-biased > #female-biased). Overall nPRF + TS can achieve the best performance in mitigating the |ΔA-PRF| ranking disparity, which achieves our goal in mitigating the ranking disparity.

Fairness Performance RaB. As for RaB, we hope to use another fairness metric to demonstrate our regularization's robustness. We realize RaB is focusing on the top-ranking result and |ΔA-PRF| is focusing on the overall ranking result by definition. We present the RaB result in the last column. In the last two columns, the TS method is still better than the PD method on average. For RaB@5, the TS method's performance is similar to the PD method in BERT_{L_2} (3.3% vs 0%); The TS method's performance is better than the PD method in BERT_{L_4} (106% vs -81.3%). For RaB@10, in BERT_{L_2}, the TS method is similar to the PD method (5.1% vs -2.5%); In BERT_{L_4}, the TS method is better than the PD method (90.9% vs 9.1%). After evaluating the the fairness level on BERT_{L_4} and BERT_{L_2}, we found that the more complicated the model is, the more bias it is, which is also demonstrated in (Rekabsaz and Schedl, 2020). We find that the RaB performance not consistent with the |ΔA-PRF| is mainly because |ΔA-PRF| is focusing more on the lower-ranked passages and RaB is focusing the higher-ranked passages. This makes these two fairness metrics are relatively exclusive. However, when the ranking system performs well (rank the clicked passage high), the |ΔA-PRF| will finally consider the overall ranking result.

5 Conclusion

In this paper, we present a novel in-processing in-batch balancing regularization method to mitigate ranking disparity and retain ranking performance. We also overcome the non-differentiable and non-convex properties of the 0-normed PRF and propose the nPRF. We conduct experiments on the MS MARCO dataset and find that the nPRF with T-statistics regularization method outperforms other methods in terms of fairness metrics and ranking metrics. In future work, we will consider generalizing our method to multiple protected variables such as age, income, etc, and also addressing bias in the query by employing adversarial networks.

Bias Statement

In this paper, we study gender bias in the neural retrieval system. If a ranking system allocates resources or opportunities unfairly to specific gender groups (e.g., less favorable to females), this creates allocation harm by exhibiting more and more male-dominated passages, which also forms a more biased dataset in turn. When such a ranking system is used in reality, there is an additional risk of unequal performance across genders. Our work is to explore the bias level of the dense passage retrieval model

with BERT_{L_2} and BERT_{L_4} on the MS MARCO passage reranking task. Thus, the community can use these benchmarks with a clearer understanding of the bias level, and can work towards developing a fairer model.

References

Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM*, 61(6):54–61.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in recommendation ranking through pairwise comparisons. In *KDD*, pages 2212–2220. ACM.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *SIGIR*, pages 405–414. ACM.

Sen Cui, Weishen Pan, Changshui Zhang, and Fei Wang. 2021. Towards model-agnostic post-hoc adjustment for balancing ranking fairness and algorithm utility. In *KDD*, pages 207–217.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, Minneapolis, Minnesota.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781, Online.

Nikola Konstantinov and Christoph H Lampert. 2021. Fairness through regularization for learning to rank. *arXiv preprint arXiv:2102.05996*.

Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*, pages 1334–1345. IEEE.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):694–707.

Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal biases in retrieved contents: Measurement framework and adversarial mitigation for bert rankers. *ArXiv preprint*, abs/2104.13640.

Navid Rekabsaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias? In *SIGIR*, pages 2065–2068. ACM.

Ashudeep Singh and Thorsten Joachims.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *ArXiv preprint*, abs/1908.08962.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1569–1578. ACM.

Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2849–2855. ACM / IW3C2.

Meike Zehlike, Philipp Hacker, and Emil Wiedemann. 2020. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery*, 34(1):163–200.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in ranking: A survey. *ArXiv preprint*, abs/2103.14000.

Sendong Zhao, Yong Huang, Chang Su, Yuantong Li, and Fei Wang. 2020. Interactive attention networks for semantic text matching. In *ICDM*, pages 861–870. IEEE.

A Appendix

In this section, we provide the baseline model performance in Table 2. We also provide the training, development, and test of the origin dataset and the fairness dataset (with fairness label) in Table 3.

Model	Ranking Metric			Fairness Metric		
	Recall@10	**MRR**	**NDCG**	**D-PRF**	**RaB@5**	**RaB@10**
BM25	0.230	0.107	0.125	-	-	-
RM3 PRF	0.209	0.085	0.104	-	-	-
MP	0.295	0.141	0.191	-	-	-
KNRM	0.297	0.169	0.167	-	-	-
C-KNRM	0.325	0.170	0.197	-	-	-
TK	0.360	0.212	0.231	-	-	-
DPR(L2)	0.357†	0.164†	0.196†	0.005	0.091	0.079
DPR(L4)	0.429†	0.205†	0.243†	-0.043	0.016	0.011

Table 2: IR Model and DPR, † indicates significant improvement over BM25.

Data	**Train**	**Dev**	**Test**	**Total**
DPR	510,586	26,873	126	537,585
Fairness[1]	876	250	126	1,252

Table 3: The number of training, development and testing examples for the DPR model and fairness model