A Light-weight Strategy for Restraining Gender Biases in Neural Rankers

Amin Bigdeli[1], Negar Arabzadeh[2], Shirin Seyedsalehi[1], Morteza Zihayat[1], and Ebrahim Bagheri[1]

[1] Ryerson University, Toronto, Canada
{abigdeli,shirin.seyedsalehi,mzihayat,bagheri}@ryerson.ca
[2] University of Waterloo, Waterloo, Canada, {narabzad}@uwaterloo.ca

Abstract. In light of recent studies that show neural retrieval methods may intensify gender biases during retrieval, the objective of this paper is to propose a simple yet effective sampling strategy for training neural rankers that would allow the rankers to maintain their retrieval effectiveness while reducing gender biases. Our work proposes to consider the degrees of gender bias when sampling documents to be used for training neural rankers. We report our findings on the MS MARCO collection and based on different query datasets released for this purpose in the literature. Our results show that the proposed light-weight strategy can show competitive (or even better) performance compared to the state-of-the-art neural architectures specifically designed to reduce gender biases.

1 Introduction

With the growing body of literature on the prevalence of stereotypical gender biases in Information Retrieval (IR) and Natural Language Processing (NLP) techniques [1, 2, 9, 25, 5, 12, 19], researchers have specifically started to investigate how the existence of such biases within the algorithmic and representational aspects of a retrieval system can impact retrieval outcomes and as a result the end users [23, 12, 5, 4, 3]. For instance, Bigdeli et al. [4] investigated the existence of-stereotypical gender biases within relevance judgment datasets such as MS MARCO [17]. The authors showed that stereotypical gender biases are observable in relevance judgements. In another study, Rekabsaz et al. [23] showed that gender biases can be intensified by neural ranking models and the inclination of bias is towards the male gender. In line with the work of Rekabsaz et al., Fabris et al. [8] proposed a gender stereotype reinforcement metric to measure gender inclination within a ranked list of documents. The results of their study revealed that neural retrieval methods reinforce gender stereotypical biases.

To address such gender biases in neural rankers, Rekabsaz et al. [22] have been the first to focus on de-biasing neural rankers by removing gender-related information encoded in the vector representation of the query-document pairs specifically in the `BERT` reranker model. However, effectively de-biasing neural rankers is still a challenging problem due to the following major challenges: (1) The model by Rekabsaz et al., referred to as AdvBert, can reduce bias but this comes at the cost of reduction in retrieval effectiveness. (2) AdvBert introduces new adversarial components within the `BERT` reranker loss function,

requiring structural changes in the architecture of the model. Such changes may not be generalizable to other neural rankers that have alternative loss functions.

In this paper, we address these challenges by proposing a novel training strategy, which 1) can decrease the level of gender biases in neural ranking models, while maintaining a comparable level of retrieval effectiveness, and 2) does not require any changes to the architecture of SOTA neural rankers. Our work is inspired by the findings of researchers such as Qu et al. [21] and Karpukhin et. al. [15] who effectively argue that the performance of neural rankers are quite sensitive to the adopted negative sampling strategy where in some cases retrieval effectiveness of the same neural ranker can be increased by as much as 17% on MRR@10 by changing the negative sampling strategy. On this basis, we hypothesize that it would be possible to control gender biases in neural rankers by adopting an effective negative sampling strategy. We propose a systematic negative sampling strategy, which would expose the neural ranker to representations of gender bias that need to be avoided when retrieving documents. We then empirically show that SOTA neural rankers are able to identify and avoid stereotypical biases based on our proposed negative sampling strategy to a greater extent compared to models such as AdvBert. At the same time, our work exhibits competitive retrieval effectiveness to strong SOTA neural rankers.

2 Problem Definition

Let $D_{q_i}^{\mathcal{M}} = [d_1^{q_i}, d_2^{q_i}, ..., d_m^{q_i}]$ be a list of initial retrieved documents for a query q_i by a first-stage retrieval method $\mathcal{M}$. Also, let us define $\mathcal{R}$ as a neural ranking model that accepts q_i and its initial retrieved list of documents $D_{q_i}^{\mathcal{M}}$ and generates $\Psi_{q_i}^{\mathcal{R}}$, which is the re-ranked version of $D_{q_i}^{\mathcal{M}}$ based on the neural ranker $\mathcal{R}$. The objective of this paper is to train a neural ranker $\mathcal{R}'$ through a bias-aware negative sampling strategy in a way that the following conditions are met:

$$\frac{1}{|Q|}\sum_{q\in Q} Bias(\Psi_q^{\mathcal{R}'}) < \frac{1}{|Q|}\sum_{q\in Q} Bias(\Psi_q^{\mathcal{R}}), \tag{1}$$

$$\frac{1}{|Q|}\sum_{q\in Q} Utility(\Psi_q^{\mathcal{R}'}) \simeq \frac{1}{|Q|}\sum_{q\in Q} Utility(\Psi_q^{\mathcal{R}}) \tag{2}$$

Where $Bias(\Psi_q^{\mathcal{R}})$ is a level of bias of the top-k retrieved list of documents for query q_i by Ranker $\mathcal{R}$, as defined in [23], and $Utility(\Psi_q^{\mathcal{R}})$ is the retrieval effectiveness of ranker $\mathcal{R}$ based on metrics such as MRR. We are interested in training $\mathcal{R}'$ based on the same neural architecture used by $\mathcal{R}$, only differing in the negative sampling strategy such that the retrieval effectiveness of $\mathcal{R}$ and $\mathcal{R}'$ remain comparable while the level of bias in $\mathcal{R}'$ is significantly reduced.

3 Proposed Approach

The majority of SOTA neural rankers utilize the set of top-k documents retrieved by a fast and reliable ranker such as BM25 as a weakly-supervised strategy for negative sampling [21, 15, 14, 16, 11, 13, 10, 20]. In this paper, instead of randomly sampling $N \leq m$ negative samples from top-k retrieved documents by BM25, we systematically select N negative samples such that the neural ranker is exposed to stereotypical gender biases that need to be avoided when ranking documents.

Let β be a non-negative continuous function that measures the genderedness of any given document. An implementation of β function can be obtained from [23]. Given β and N, the retrieved document set $D_q^{\mathcal{M}}$ is sorted in descending order based on $\beta(d_i)$ ($d_i \in D_q^{\mathcal{M}}$) and the top-$N$ documents form a non-increasing list S_q^{β} such that $\{\forall i \in [1, \ldots, N], \beta(s_i) \geq \beta(s_{i+1})\}$.

As the documents in S_q^{β} exhibit the highest degree of gender bias compared to the rest of the documents in $D_q^{\mathcal{M}}$, we suggest that S_q^{β} can be served as the negative sample set due to two reasons: (1) S_q^{β} is a subset of $D_q^{\mathcal{M}}$, as such, when using the random negative sampling strategy, S_q^{β} may have been chosen as the negative sample set; therefore, it is unlikely that the choice of S_q^{β} as the negative sample set results in decreased retrieval effectiveness; and (2) S_q^{β} consists of documents with the highest degree of gender bias and hence, the neural ranker would not only have a chance to learn that these documents are not relevant but also to learn to avoid biased gender affiliated content within these documents and hence avoid retrieving gender-biased documents at retrieval time.

Considering the highest gender-biased documents as the negative sample set may be a strict requirement and is not desirable as it might cause the neural ranker forgets the need of learning document relevance and only focuses on learning to avoid gender-biased documents. In order to avoid interpreting all gendered-biased documents as irrelevant during the training process, we relax the negative sampling strategy through a free-parameter λ. According to λ, a subset of negative documents is selected from S_q^{β} (NS_{Biased}) and the rest of the negative document set is randomly selected from the original pool $D_q^{\mathcal{M}}$ (NS_{Rnd}).

$$NS_{Biased} = \{d_i \in S_q^{\beta} | i \leq \lambda \times N\}$$

$$NS_{Rnd} = \{Rand(d \in D_q^{\mathcal{M}}) | d \notin NS_{Biased}\},$$

such that $|NS_{Rnd}| + |NS_{Biased}| = N$ and the final set of negative samples would be $NS = NS_{Rnd} \cup NS_{Biased}$.

4 Experiments

Document Collection. We adopt the MS MARCO collection consisting of over 8.8M passages and over 500k queries with at least one relevant document.
Query Sets. We adopt two query sets in our experiments that have been proposed in the literature for evaluating gender bias: (1) The first query set (QS1) includes 1,765 neutral queries [23]. (2) The second query set (QS2) includes 215 fairness-sensitive queries that are considered as socially problematic topics [22].
Bias Metrics. We adopt two bias measurement metrics from the literature to calculate the level of biases within the retrieved list of documents: (1) The first metric is introduced by Rekabsaz et al. [23] and measures the level of bias in a document based on the presence and frequency of gendered terms in a document, referred to as Boolean and TF ARaB metrics. (2) The Second metric is NFaiRR which calculates the level of fairness within the retrieved list of documents by calculating each document's neutrality score proposed in [22]. We note that less ARaB and higher NFaiRR metric values are desirable.

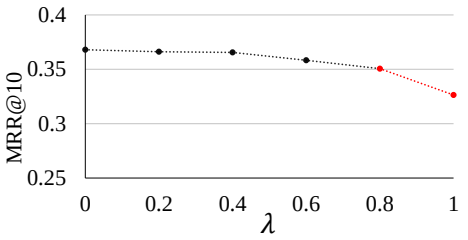

Fig. 1. Impact of λ on neural ranker performance on MS MARCO Dev Set. The red points indicate statistically significant drop in performance.

Neural Ranker	Training Schema		Change
	Original	Ours	
`BERT (base)`	0.3688	0.3583	-2.84%
`DistilRoBERTa (base)`	0.3598	0.3475	-3.42%
`ELECTRA (base)`	0.3332	0.3351	+0.57%

Table 1. Comparison between the performance (MRR@10) of the base ranker and the ranker trained based on our proposed negative sampling strategy when $\lambda = 0.6$ on MS MARCO Dev Set.

Neural Rankers. To train the neural rerankers, we adopted the cross-encoder architecture as suggested in SOTA ranking literature [18, 21] in which two sequences (Query and candidate document) are passed to the transformer and the relevance score is predicted. As suggested by the Sentence Transformer Library [3], we fine tuned different pre-trained transformer models, namely, `BERT-base-uncased` [7], `DistilRoBERTa-base` [24], `ELECTRA-base` [6], `BERT-Mini` [26], and `BERT-Tiny` [26]. For every query in the MS MARCO training set, we considered 20 negative documents (N=20) from the top-1000 unjudged documents retrieved by BM25 [17] (based on random sampling and our proposed sampling strategy).

Results and Findings. The objective of our work is to show that a selective negative sampling strategy can systematically reduce gender bias while maintaining retrieval effectiveness. As such, we first investigate how our proposed negative sampling strategy affects retrieval effectiveness. We note statistical significance is measured based on paired t-test with α=0.05. In Figure 1, we demonstrate the performance of a SOTA `BERT-base-uncased` neural ranker trained with our proposed negative strategies when changing λ from [0,1] with 0.2 increments. Basically, when $\lambda = 0$ the model is trained with all randomly negative samples from BM25 retrieved documents (baseline) and when $\lambda = 1$, the N negative samples are the most gendered documents in $D_q^{\mathcal{M}}$. Based on Figure 1, we observe that gradual increase in λ will come at the cost of retrieval effectiveness. However, the decrease is only statistically significant when $\lambda > 0.6$. Thus we find that when up to 60% of negative samples are selected based on our proposed negative sampling strategy, the retrieval effectiveness remains comparable to the base ranker. As mentioned earlier, this drop in performance is due to the fact that the model would learn the concept of avoiding gender-biased documents and not the concept of relevance due to the large number of gender-biased negative samples. We further illustrate the performance of adopting our proposed negative sampling strategy with $\lambda = 0.6$ on other pre-trained language models including `ELECTRA` and `DistilRoBERTa` in Table 1. For these pre-trained language models, similar to `BERT`, we observe that retrieval effectiveness remains statistically comparable to the base ranker and no statistically significant changes occur in terms of performance. Thus we conclude that it is possible to adopt our proposed bias-aware negative sampling strategy (e.g. at $\lambda = 0.6$) and maintain comparable retrieval effectiveness. We note all our code and the run files are publicly available[4].

[3] `https://www.sbert.net/`

[4] `https://github.com/aminbigdeli/bias_aware_neural_ranking`

Table 2. Retrieval effectiveness and the level of fairness and bias across three neural ranking models trained on query sets QS1 and QS2 when $\lambda = 0.6$ at cut-off 10.

Query Set	Neural Ranker	Training Schema	MRR@10	NFaiRR		ARaB			
				Value	Improvement	TF	Reduction	Boolean	Reduction
QS1	BERT (base)	Original	0.3494	0.7764	-	0.1281	-	0.0956	-
		Ours	0.3266	0.8673	11.71%	0.0967	24.51%	0.0864	9.62%
	DistilRoBERTa (base)	Original	0.3382	0.7805	-	0.1178	-	0.0914	-
		Ours	0.3152	0.8806	12.83%	0.0856	27.33%	0.0813	11.05%
	ELECTRA (base)	Original	0.3265	0.7808	-	0.1273	-	0.0961	-
		Ours	0.3018	0.8767	12.28%	0.0949	25.45%	0.0855	11.03%
QS2	BERT (base)	Original	0.2229	0.8779	-	0.0275	-	0.0157	-
		Ours	0.2265	0.9549	8.77%	0.0250	9.09%	0.0156	0.64%
	DistilRoBERTa (base)	Original	0.2198	0.8799	-	0.0338	-	0.0262	-
		Ours	0.2135	0.9581	8.89%	0.0221	34.62%	0.0190	27.48%
	ELECTRA (base)	Original	0.2296	0.8857	-	0.0492	-	0.0353	-
		Ours	0.2081	0.9572	8.07%	0.0279	43.29%	0.0254	28.05%

Table 3. Comparing AdvBert training strategy and our approach at cut-off 10.

Neural Ranker	Training Schema	MRR@10	NFaiRR		ARaB			
			Value	Improvement	TF	Reduction	Boolean	Reduction
BERT-Tiny	Original	0.1750	0.8688	-	0.0356	-	0.0296	-
	AdvBert	0.1361	0.9257	6.55%	0.0245	31.18%	0.0236	20.27%
	Ours	0.1497	**0.9752**	**12.25%**	0.0099	**72.19%**	0.0115	**61.15%**
BERT-Mini	Original	0.2053	0.8742	-	0.0300	-	0.0251	-
	AdvBert	0.1515	0.9410	7.64%	0.0081	**73.00%**	0.0032	**87.26%**
	Ours	0.2000	**0.9683**	**10.76%**	0.0145	51.67%	0.0113	54.98%

We now investigate the impact of our proposed negative sampling strategy on reducing gender biases. To this end, using each of the SOTA rankers, we re-rank the queries in each two query sets (QS1 and QS2) and report their retrieval effectiveness as well as their bias metrics (ARaB and NFaiRR) measurements in Table 2. As shown, we observe that for both of the query sets the level of fairness (NFaiRR) increases, while the level of bias (ARaB) decreases across all of the three neural ranking models. In addition and more importantly, the decrease in gender biases does not come at the cost of significant reduction in retrieval effectiveness. In summary, our proposed negative sampling strategy is able to maintain retrieval effectiveness while reducing bias and increasing fairness.

It is important to also compare our work against the most recent neural ranking model designed to increase fairness, namely AdvBert [22]. Unlike our proposed work which retains the same neural architecture of the original ranker and only changes the negative sampling strategy, AdvBert proposes an adversarial neural architecture to handle gender biases. The authors of AdvBert have publicly shared their trained models based on BERT-Tiny and BERT-Mini and only for QS2. For the sake of comparison, we compare our work with AdvBert based on these two models and on QS2. Based on the results reported in Table 3, we make the following observations: (1) For the models based on BERT-Tiny, neither our model nor AdvBert significantly drop retrieval effectiveness; however, the fairness (NFaiRR) and bias (ARaB) measures are notably more favorable for our proposed approach. (2) Similar observations can be made for BERT-Mini as well. In this case, the retrieval effectiveness of our proposed approach is substantially higher than AdvBert and at the same time the reported level of fairness (NFaiRR) is also higher. However, in terms of bias metrics, AdvBert has de-

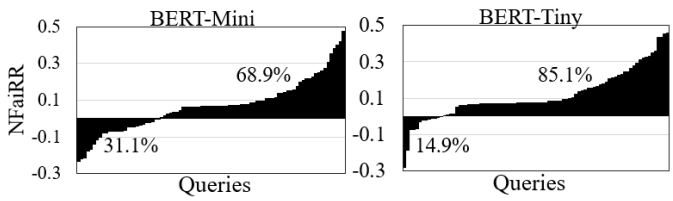

Fig. 2. Comparative analysis between AdvBert and our proposed approach based on NFaiRR at cut-off 10 on a per query basis.

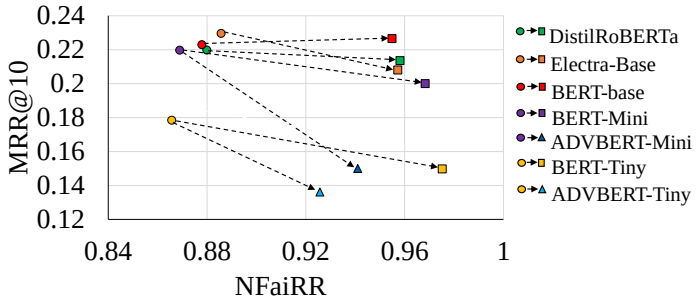

Fig. 3. Comparing the base rankers vs our proposed approach and AdvBert in terms of performance and bias when using different pre-trained language models on QS2.

creased both TF ARaB and Boolean ARaB more than our proposed approach.

We further compare the level of fairness between our proposed approach and AdvBert on a per-query basis. To this end, for `BERT-Tiny` and `BERT-Mini` we calculate the level of fairness within the ranked list of documents returned by our method and AdvBert. Followed by that, we subtract the level of fairness of each query and report the results in Figure 2. As shown, the number of queries that have seen improvement in their fairness metric (NFaiRR) based on our approach compared to AdvBert as well as the degree fairness has been impacted. Positive values show improved fairness by our approach compared to AdvBert while negative values show otherwise. As shown, 69% and 85% of the queries have seen increased fairness based on our proposed approach on BERT Mini and Tiny, respectively. We contextualize this by mentioning that on both models, the retrieval effectiveness of our proposed approach is also higher than AdvBert.

Now, let us illustrate the robustness of our proposed approach across all the neural rankers by showing the level of their effectiveness and fairness on QS2 (and not QS1 since AdvBert is not available on QS1) in Figure 3. As shown, when adopting our proposed approach, the level of fairness increases notably, while retrieval effectiveness remains at a comparable level. We further observe that while AdvBert is able to increase fairness (not to the extent of our proposed approach), it does so at the cost of a notable decrease in retrieval effectiveness.

5 Concluding Remarks

We have shown that it is possible to adopt a simple yet effective sampling strategy for training neural rankers such that gender biases are reduced while retrieval effectiveness is maintained. Through our experiments, we show that a light-weight strategy is able to show competitive (or even better) tradeoff between bias reduction and retrieval effectiveness compared to adversarial neural rankers that are specifically designed for restraining gender biases.

References

1. Baeza-Yates, R.: Bias on the web. Communications of the ACM **61**(6), 54–61 (2018)
2. Baeza-Yates, R.: Bias in search and recommender systems. In: Fourteenth ACM Conference on Recommender Systems, pp. 2–2 (2020)
3. Bigdeli, A., Arabzadeh, N., Seyersalehi, S., Zihayat, M., Bagheri, E.: On the orthogonality of bias and utility in ad hoc retrieval. In: Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2021)
4. Bigdeli, A., Arabzadeh, N., Zihayat, M., Bagheri, E.: Exploring gender biases in information retrieval relevance judgement datasets. In: European Conference on Information Retrieval, pp. 216–224. Springer (2021)
5. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
6. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Fabris, A., Purpura, A., Silvello, G., Susto, G.A.: Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. Information Processing & Management **57**(6), 102,377 (2020)
9. Font, J.E., Costa-Jussa, M.R.: Equalizing gender biases in neural machine translation with word embeddings techniques. arXiv preprint arXiv:1901.03116 (2019)
10. Gao, L., Callan, J.: Unsupervised corpus aware language model pre-training for dense passage retrieval. arXiv preprint arXiv:2108.05540 (2021)
11. Gao, L., Dai, Z., Callan, J.: Coil: Revisit exact lexical match in information retrieval with contextualized inverted list. arXiv preprint arXiv:2104.07186 (2021)
12. Gerritse, E.J., Hasibi, F., de Vries, A.P.: Bias in conversational search: The double-edged sword of the personalized knowledge graph. In: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, pp. 133–136 (2020)
13. Han, S., Wang, X., Bendersky, M., Najork, M.: Learning-to-rank with bert in tf-ranking. arXiv preprint arXiv:2004.08476 (2020)
14. Izacard, G., Grave, E.: Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282 (2020)
15. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906 (2020)
16. Macdonald, C., Tonellotto, N.: On approximate nearest neighbour selection for multi-stage dense retrieval. arXiv preprint arXiv:2108.11480 (2021)
17. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: Ms marco: A human generated machine reading comprehension dataset. In: CoCo@ NIPS (2016)
18. Nogueira, R., Cho, K.: Passage re-ranking with bert. arXiv preprint arXiv:1901.04085 (2019)
19. Olteanu, A., Garcia-Gathright, J., de Rijke, M., Ekstrand, M.D., Roegiest, A., Lipani, A., Beutel, A., Olteanu, A., Lucic, A., Stoica, A.A., et al.: Facts-ir: fairness,

accountability, confidentiality, transparency, and safety in information retrieval. In: ACM SIGIR Forum, vol. 53, pp. 20–43. ACM New York, NY, USA (2021)
20. Pradeep, R., Nogueira, R., Lin, J.: The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. arXiv preprint arXiv:2101.05667 (2021)
21. Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W.X., Dong, D., Wu, H., Wang, H.: Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5835–5847 (2021)
22. Rekabsaz, N., Kopeinik, S., Schedl, M.: Societal biases in retrieved contents: Measurement framework and adversarial mitigation for bert rankers. arXiv preprint arXiv:2104.13640 (2021)
23. Rekabsaz, N., Schedl, M.: Do neural ranking models intensify gender bias? In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2065–2068 (2020)
24. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
25. Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.W., Wang, W.Y.: Mitigating gender bias in natural language processing: Literature review. arXiv preprint arXiv:1906.08976 (2019)
26. Turc, I., Chang, M.W., Lee, K., Toutanova, K.: Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962 (2019)