Bias-aware Fair Neural Ranking for Addressing Stereotypical Gender Biases

Shirin SeyedSalehi
shirin.seyedsalehi@ryerson.ca
Ryerson University, Canada

Amin Bigdeli
abigdeli@ryerson.ca
Ryerson University, Canada

Negar Arabzadeh
narabzad@uwaterloo.ca
University of Waterloo, Canada

Bhaskar Mitra
Bhaskar.Mitra@microsoft.com
Microsoft Research

Morteza Zihayat
mzihayat@ryerson.ca
Ryerson University, Canada

Ebrahim Bagheri
bagheri@ryerson.ca
Ryerson University, Canada

ABSTRACT

Research has shown that neural rankers can pick up and intensify gender biases. The expression of stereotypical gender biases in retrieval systems can lead to their reinforcement in users' beliefs. As such, the objective of this paper is to propose a *bias-aware* fair ranker that explicitly incorporates a notion of gender bias and hence controls how bias is expressed in documents that are retrieved. The proposed approach is designed such that it learns the notion of relevance between the document and the query from the relevant sampled documents while incorporating the notion of gender bias by penalizing irrelevant biased sampled documents. We show that unlike the state of the art, our approach reduces bias while maintaining retrieval effectiveness over different query sets.

1 INTRODUCTION

A growing number of studies have shown that information retrieval (IR) systems exhibit stereotypical gender biases as they are mainly trained based on large-scale user data [1, 3, 8, 14]. This can have potentially harmful impact on the users' judgements when exposed to unfair and biased search results, e.g., problems caused by discrimination against minority groups. This is concerning especially given the fact that not only do a large number of search engine users heavily rely on retrieval systems on a daily basis but also due to the fact that search results often constitute a major component of important practical systems such as recommendation systems [2, 6, 11], question answering systems [7, 20], intelligent assistants [5, 24], to name a few. Thus, recent works have focused on controlling stereotypical gender biases in retrieval systems. These include the work by Rekabsaz et al. [23] on the impact of neural ranking methods on amplifying gender biases, Bigdeli et al. [4] on the bias embedded in relevance judgement datasets, and Fabris et al. [10] on gender stereotype reinforcement, to name a few. These works have strongly motivated the need to capture and curtail the impact of stereotypical gender biases in neural retrieval. Recently, Rekabsaz et al. [22] proposed an *adversarial* method for gender bias reduction in BERT-based rankers, namely AdvBert. The authors argue that there is a systematic tradeoff between bias and retrieval effectiveness and hence design an adversarial mini-max game to find a balance point between retrieval effectiveness and bias. The game consists of two components (a) a max marginal loss for learning the relevance associations between the query and relevant and irrelevant documents. (b) a cross entropy loss, which attempts to predict whether the vector representation of the document-query pair contains any bias-related information. In the context of the mini-max game, retrieval effectiveness and bias need to form a pareto front with effectiveness on one axis and bias on the other. The optimal point will be an equilibrium between effectiveness and bias. By design, this leads to a competition between the loss functions, which does not necessarily cooperate to learn an optimal representation for relevance and bias at the same time. Therefore, while leading to reduced bias due to the cross entropy loss, this may come at the cost of suboptimal relevance, leading to decrease in retrieval effectiveness.

In contrast to the work by Rekabsaz et al. [22], we propose a method to systematically reduce gender biases in SOTA neural-based rankers while striving to maintain an effective retrieval performance. We design a bias-aware fair neural ranking method that explicitly considers the degree of measurable gender biases associated with sampled documents. We argue that including a bias term associated with sampled documents in neural rankers ensures that the model learns to avoid representations that are affiliated with gender biases and at the same time it learns accurate relevance relationships. In other words, the model will learn to avoid biased representations through the bias term associated with document samples and will also learn relevance through the associations learnt based on the positive documents. Therefore, such an approach strives to maintain its retrieval effectiveness while reducing gender biases.

We perform extensive empirical experiments based on different query collections and answer three main research questions (RQs): (RQ1) Would the proposed bias-aware fair ranker show an improved balance between bias and retrieval effectiveness compared to the state of the art baseline? (RQ2) Does the proposed bias-aware fair ranker operate consistently regardless of the initial contextual embedding model used in the training? and, (RQ3) Is the performance of the proposed bias-aware fair ranker, in terms of retrieval effectiveness and gender bias, consistent across a range of different datasets? We comparatively evaluate the performance of our proposed approach on two datasets that consist of 1,765 neutral queries by [23], and 215 queries publicly shared in [22]. We find that our proposed bias-aware fair ranker consistently reduces stereotypical gender biases while maintaining a comparable retrieval effectiveness across both datasets.

2 PROBLEM DEFINITION

Assume that Π and $\hat{\Pi}$ as a state of the art ranker and a fair ranker, respectively. Given neutral query set Q, where the set of retrieved documents for each query is expected to have equitable representation of all genders, a *fair ranker* would ideally need to satisfy the following conditions:

Maintain Performance: $\hat{\Pi}$ would need to show comparable retrieval effectiveness to be a practically viable alternative to Π:

$$\mathbb{U}(\hat{\Pi}, Q) \sim \mathbb{U}(\Pi, Q) \quad (1)$$

where $\mathbb{U}(\Pi, Q)$ shows the effectiveness of Π on Q.

Reduce Bias: $\hat{\Pi}$ should exhibit reduced degrees of gender bias on neutral queries to be a more desirable ranker through exposing less stereotypical biases:

$$\beta(\hat{\Pi}, Q) < \beta(\Pi, Q) \tag{2}$$

where $\beta(\Pi, Q)$ is a quantifiable measure of bias as explained in [10, 23]. Our objective is to propose a *fair neural ranker* that exhibits the desirable characteristics outlined in Equations 1 and 2.

3 PROPOSED APPROACH

Bias-aware Fair Ranker. While neural rankers have primarily focused on optimizing $\mathbb{U}(\hat{\Pi}, Q)$, the objective of our work is to also minimize $\beta(\hat{\Pi}, Q)$. We hypothesize that it would be possible to reduce the bias exhibited by a neural ranker if the ranker is explicitly introduced to documents with higher degrees of gender bias. As such, we suggest that a fair ranker should incorporate two considerations: (**C1**) it should capture the associations between a query and its relevant documents in order to satisfy Equation 1, and, (**C2**) it should penalize documents with higher degrees of bias such that they are ranked lower compared to less biased documents, especially, when there is low probability for the document being relevant to the query; hence, satisfying Equation 2.

Given query q, and N^+ and N^- relevant and irrelevant documents to q, one of the widely adopted ranking loss functions in neural rankers [12, 25] is defined as:

$$L = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} max(0, m - \Phi(q, {d_i}^+) + \Phi(q, {d_j}^-) \tag{3}$$

where d^+ and d^- are relevant and irrelevant documents to the input query, respectively, m is a margin, and $\Phi(q, d)$ is the relevance score of document d w.r.t. query q. Irrespective of the degree of bias, this function has already shown to be able to effectively capture C1 by ensuring the $\Phi(q, d^+)$ is maximized and $\Phi(q, d^-)$ is minimized [12, 13, 25]. To meet C2, we propose to incorporate each document's degree of bias in the loss function such that the neural ranker learns to rank documents with higher degrees of bias at lower ranks. One possible approach is to penalize the relevance of each document to the query based on the degree of bias exposed by that document. That is, we incorporate the degree of bias of each document as a penalty term in $\Phi(q, {d_i}^+)$ and $\Phi(q, {d_j}^-)$. This is likely to reduce bias of the retrieved documents since the loss function allows the model to distance highly biased documents from the query.

The loss function in Equation 3 aims at maximizing the relevance score of relevant documents to the query while minimizing the association between negative samples and the query. We penalize the biased relevant documents as $\Phi_B(q, {d_i}^+) = \Phi(q, {d_i}^+) - \Psi({d_i}^+)$ and biased irrelevant documents as $\Phi_B(q, {d_j}^-) = \Phi(q, {d_j}^-) + \Psi({d_j}^-)$ where $\Psi(d_i)$ measures the gender bias of document d_i. We will, later in the experimental setup section, discuss that $\Psi(d_i)$ can be computed based on the gender bias measures proposed by Rekabsaz et al [23] and Bigdeli et al [4]. Now, we propose to rewrite the loss function of a fair neural ranker as follows:

$$L = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} max(0, m - \Phi_B(q, d_i^+) + \Phi_B(q, d_j^-)) \tag{4}$$

Bias-Performance Trade off. It is possible to argue that penalizing both the positive and negative documents with respect to their degrees of bias can come at the expense of ranker effectiveness for two reasons: **(a)** Those relevant documents that are biased will receive lower likelihood of being retrieved and ranked at the top of the retrieved list of documents. We note that while retrieving biased documents is undesirable outcome, it is also quite undesirable to retrieve irrelevant yet unbiased documents. **(b)** Large-scale relevance judgment datasets often include very sparse relevance judgements per query (e.g., MS MARCO [18], has only 1.06 relevant documents per query on average). Under such circumstances, if the only relevant document is penalized based on its degree of bias, the ranker will not have a chance to learn the concept of relevance.

On this basis, we propose to relax the penalty terms in Equation 4 by capturing the concept of document bias over only negative document samples (d^-) in the loss function. By relaxing the penalty term associated with positive document samples, although the proposed loss function would not directly penalize positive documents, it would still capture (1) bias through the penalty term associated with negative samples, and (2) the concept of relevance through the association with positive documents and the unbiased irrelevant document samples. By relaxing the penalty term for relevant documents, we rewrite the loss function in Equation 4 as follows:

$$L = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} max(0, m - \Phi(q, d_i^+) + \Phi_B(q, d_j^-)) \tag{5}$$

From a theoretical point of view, our proposed bias term in the loss function can be seen as a regularizer for the margin. A larger bias value results in a greater margin between the positive and negative relevance scores, which in turn serves as a stricter constraint on the more biased documents. In the case of neural rankers, one could look at the problem in the embedding space. In such space, the training objective is in fact trying to adjust the representation vectors in the embedding space such that for each query representation, the vector representations of its relevant documents are closer to it as compared to the vector representations of the irrelevant documents. In such a scenario, adding bias to the relevance score of the irrelevant document pushes the vector representation of the biased irrelevant document farther away from the vector representation of the query. Therefore, during inference, biased irrelevant documents will gain a smaller score compared to less-biased or even unbiased irrelevant documents. Consequently, the final retrieved list has a higher likelihood of consisting of a lower degree of bias compared to the original model (trained on the original loss function in Equation 3). Furthermore, in our proposed approach, the bias term is only applied to irrelevant documents, and as such, the relevant documents' score is immune from being subject to change due to the added bias term. Hence, the retrieval effectiveness is maintained; hence satisfying both C1 and C2 conditions.

4 EXPERIMENTS

Passage Collection. We conduct our experiments on the MS MARCO [18] passage collection dataset that consists of 8,841,822 passages.

Query Sets. In order to investigate whether our proposed approach that uses a bias-aware loss function can reduce biases in the ranked list of documents for queries, we require a set of gender neutral queries whose retrieved documents are not expected to exhibit any gender biases. Using gender neutral queries would rule out the potential need to have gender biases within the retrieved document and hence has a better chance of revealing gender bias in the retrieval method. To this end, We employ two query

Table 1: Comparison of our approach with Baseline Methods.

Cutoff@10

Models		TF		Boolean		LIWC	
		Value	Δ%	Value	Δ%	Value	Δ%
Mini	Original	0.006	-	0.008	-	0.654	-
	Ours	0.005	-22.22%	0	-94.46%	0.595	-9.13%
	ADVBERT	-0.002	-74.72%	-0.007	-21.47%	0.426	-28.42%
Tiny	Original	-0.031	-	-0.03	-	0.664	-
	Ours	-0.023	-26.52%	-0.024	-19.07%	0.619	-6.79%
	ADVBERT	0.009	-71.93%	0.008	-74.12%	0.467	-24.65%

Cutoff@20

Models		TF		Boolean		LIWC	
		Value	Δ%	Value	Δ%	Value	Δ%
Mini	Original	0.004	-	0.007	-	0.541	-
	Ours	-0.001	-68.59%	-0.004	-45.88%	0.497	-8.29%
	ADVBERT	0.006	32.62%	0.002	-68.16%	0.425	-14.5%
Tiny	Original	-0.018	-	-0.015	-	0.548	-
	Ours	-0.013	-28.00%	-0.012	-20.39%	0.477	-12.85%
	ADVBERT	0.010	-43.93%	0.010	-34.10%	0.442	-7.44%

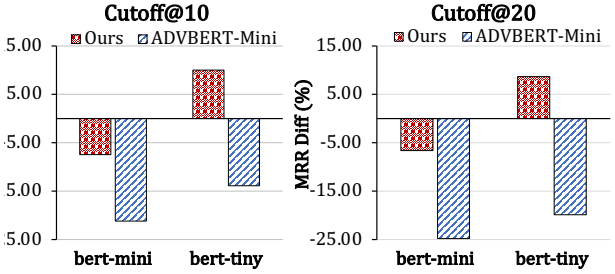

Figure 1: Our approach versus AdvBert based on BERT-mini and BERT-tiny.

sets that consist of gender neutral queries. The first dataset is a human-annotated dataset introduced in [23], which consists of 1,765 queries that were labelled as gender neutral by Amazon Turkers. The second dataset is introduced by Rekabsaz et al. [22] and consists of 215 queries that are neutral in nature and do not have inclination towards a specific gender, but their retrieved list of documents could have social gender role stereotypes.

Neural Rankers. For training neural rankers, queries and documents are first tokenized and further, due to transformer models' limitations, are truncated to at most 64 and 512 tokens, respectively. The special starting token of BERT ($[CLS]$), which has shown to yield a reasonable representation of the whole sequence [21], allows us to obtain an embedding vector e_{CLS} for the input tokens, which is obtained when the query and the document are separated by the special token $[SEP]$ and passed through f, i.e., the BERT model. e_{CLS} is then fed into a single layer neural network to obtain the relevance score $\Phi(q, d)$ as the output. We employ the OpenMatch toolkit [17] for training our models. The network architecture and training hyper-parameters for all of the models are the default settings provided by the OpenMatch.

Quantifying Bias. To assess the validity of our assumption that the proposed loss function is able to systematically decrease gender biases in neural rankers, we calculate the gender bias in the final retrieved list and compare it to the gender bias of the list retrieved by the Original model. As required by $\Psi(d_i)$ in Equations 4 and 5, we adopt two approaches to quantify gender biases: (1) the two metrics proposed by Rekabsaz et al. [23] The first one (Boolean) is based on the presence of biased terms in a document, and the second one considers the term-frequency of biased terms. These metrics are extended to the bias of a retrieved list and are referred to as the ARaB metrics. (2) the method proposed by Bigdeli et al [4], which measures the stereotypical biases present in a document. In their work, the authors employ the Linguistic Inquiry and Word Count (LIWC) toolkit [19] for calculating the female or male inclinations of a document. These approaches allow us to show that even with bias metrics that measure bias using completely different ways, our model effectively reduces gender bias.

Results and Findings.[1] In **RQ1**, we compare our work against the only state of the art method that is focused on bias reduction in neural rankers, namely AdvBert[22], by comparing the retrieval effectiveness as well as bias reduction of the two methods. We additionally compare our work and AdvBert against the original rankers. In order to make a fair comparison between our work and AdvBert, we adopt two pre-trained contextual embeddings, namely BERT-Tiny and BERT-Mini as proposed in the AdvBert paper. We are only able to compare our work against AdvBert based on the 215 neutral queries set introduced earlier since this is the only dataset for which runs are available for AdvBert. The results are shown in Table 1 and Figure 1. Considering the first query set and in terms of retrieval effectiveness, our method consistently shows better performance compared to AdvBert. As seen in Figure 1, on BERT-Mini, our model has only dropped in performance by about $\sim$ 5% on both @10 and @20 whereas the AdvBert method has experienced between $20-25\%$ decline in retrieval effectiveness. More notably, when using BERT-Tiny, our approach increases retrieval effectiveness by $\sim$ 10% while AdvBert shows a decrease in retrieval effectiveness between $19-22\%$. From the perspective of bias reduction on the ARaB and LIWC metrics reported in Table 1, we find that both loss functions are able to reduce bias in most of the cases where AdvBert is showing a greater overall degree of bias reduction. There are several instances where our method show comparable or even higher degrees of bias reduction. This is notable as our bias-aware loss function minimally drops retrieval effectiveness on BERT-Mini and even increases on BERT-Tiny while showing consistent bias reduction. We believe that given the importance of retrieval effectiveness for the usefulness and utility of a neural ranker, it is important for the ranker to maintain retrieval effectiveness while attempting to reduce bias. Based on comparison with AdvBert, we show that the retrieval effectiveness of this baseline has significantly declined in favor of reduced bias. However, our work takes a more balanced approach where effectiveness is maintained or improved while bias is reduced.

To answer **(RQ2)**, we applied our approach on different contextual embeddings, namely ALBERT, ELECTRA and Distil-RoBERTa. ALBERT [15] introduces two parameter reduction

[1]We note that all of our code, models, and results are publicly available online: https://github.com/biasaware1/biasaware.

Table 2: Retrieval Effectiveness & Degree of Bias based on the 215 query dataset [22].

	Cutoff@10								Cutoff@20							
	Effectiveness		Bias						Effectiveness		Bias					
Models	MRR		TF		Boolean		LIWC		MRR		TF		Boolean		LIWC	
	Value	Δ%	Value	Δ%	Value	Δ%	Value	Δ%	Value	Δ%	Value	Δ%	Value	Δ%	Value	Δ%
ALBERT	0.227	-	0.022	-	0.018	-	0.617	-	0.234	-	0.018	-	0.015	-	0.533	-
ALBERT$^{Bias}_{Aware}$	0.232	2.11%	0.020	-8.77%	0.014	-20.20%	0.587	-4.86%	0.238	1.93%	0.016	-6.31%	0.013	-15.46%	0.472	-11.55%
BERT Mini	0.203	-	0.006	-	0.008	-	0.654	-	0.207	-	0.004	-	0.007	-	0.541	-
BERT Mini$^{Bias}_{Aware}$	0.188	-7.44%	0.005	-22.22%	0.000	-94.46%	0.595	-9.13%	0.193	-6.62%	-0.001	-68.59%	-0.004	-45.88%	0.497	-8.29%
BERT Tiny	0.173	-	-0.031	-	-0.030	-	0.664	-	0.180	-	-0.018	-	-0.015	-	0.548	-
BERT Tiny$^{Bias}_{Aware}$	0.190	9.83%	-0.023	-26.52%	-0.024	-19.07%	0.619	-6.79%	0.195	8.69%	-0.013	-28.00%	-0.012	-20.39%	0.477	-12.86%
DistilRoBERTa	0.175	-	0.030	-	0.013	-	0.676	-	0.182	-	0.026	-	0.013	-	0.534	-
DistilRoBERTa$^{Bias}_{Aware}$	0.172	-2.17%	0.016	-45.68%	0.007	-47.18%	0.694	2.59%	0.177	-2.48%	0.011	-55.95%	0.005	-59.43%	0.511	-4.32%
ELECTRA small	0.199	-	-0.009	-	-0.012	-	0.707	-	0.207	-	-0.005	-	-0.005	-	0.569	-
ELECTRA small$^{Bias}_{Aware}$	0.203	2.01%	0.005	-46.15%	-0.002	-83.37%	0.678	-4.03%	0.210	1.25%	0.002	-63.80%	0.000	-91.82%	0.517	-9.09%

Table 3: Retrieval Effectiveness & Degree of Bias based on the 1,765 query dataset [23].

	Cutoff@10								Cutoff@20							
	Effectiveness		Bias						Effectiveness		Bias					
Models	MRR		TF		Boolean		LIWC		MRR		TF		Boolean		LIWC	
	Value	Δ%	Value	Δ%	Value	Δ%	Value	Δ%	Value	Δ%	Value	Δ%	Value	Δ%	Value	Δ%
ALBERT	0.329	-	0.078	-	0.058	-	1.308	-	0.335	-	0.074	-	0.058	-	1.126	-
ALBERT$^{Bias}_{Aware}$	0.332	0.91%	0.072	-7.69%%	0.054	-6.77%	1.266	-3.17%	0.337	0.57%	0.069	-6.46%	0.055	-6.14%	1.067	-5.22%
BERT Mini	0.288	-	0.065	-	0.051	-	1.329	-	0.294	-	0.065	-	0.052	-	1.143	-
BERT Mini$^{Bias}_{Aware}$	0.276	-4.10%	0.065	0.24%	0.049	-3.03%	1.220	-8.18%	0.282	-4.08%	0.060	-8.12%	0.047	-10.68%	1.021	-10.70%
BERT Tiny	0.262	-	0.072	-	0.057	-	1.307	-	0.269	-	0.071	-	0.056	-	1.119	-
BERT Tiny$^{Bias}_{Aware}$	0.269	2.67%	0.073	0.32%	0.056	-1.10%	1.233	-5.67%	0.276	2.49%	0.070	-2.17%	0.055	-3.13%	1.026	-8.32%
DistilRoBERTa	0.238	-	0.086	-	0.066	-	1.240	-	0.245	-	0.082	-	0.066	-	1.047	-
DistilRoBERTa$^{Bias}_{Aware}$	0.229	-3.87%	0.074	-13.70%	0.057	-13.59%	1.176	-5.20%	0.236	-3.80%	0.069	-16.65%	0.055	-16.97%	0.998	-4.68%
ELECTRA small	0.301	-	0.072	-	0.056	-	1.336	-	0.306	-	0.070	-	0.055	-	1.128	-
ELECTRA small$^{Bias}_{Aware}$	0.299	-0.53%	0.067	-6.65%	0.054	-4.44%	1.244	-6.89%	0.305	-0.29%	0.062	-11.65%	0.050	-9.39%	1.048	-7.08%

techniques that are capable of reducing the memory usage of BERT, which in effect also increases training speed. ELECTRA [9] proposes a more sample-efficient pre-training task called Replaced Token Detection (RTD) that leverages GAN's for the task of pre-training language models and shows superior performance over BERT. DistilRoBERTa is the distilled version of the RoBERTa pre-trained model [16], which is built on the BERT language masking strategy and aims to improve performance by optimizing BERT's hyperparameters . We report our findings in Tables 2 and 3 for each of the two different query sets, respectively. We observe that our proposed approach improves retrieval effectiveness on three of the contextual embeddings and remain competitive on the other two contextual embeddings (between $2-7\%$ drop of effectiveness on two of the contextual embeddings). On the other hand, however, we report that our proposed approach consistently reduces bias in the majority of cases for both variations of the ARaB metric as well as the LIWC metric. As such, we conclude that the utility of our proposed approach is not limited to a certain contextual embedding.

Finally, to answer **RQ3**, we compare across Tables 2 and 3 to investigate if retrieval effectiveness is consistent when comparing the two datasets and if reduction in bias can be consistently seen in both datasets and across all bias measurement metrics. We find that the worst case reduction in retrieval effectiveness was seen on BERT-Mini at cut-off=10 on the first query set, which is 7.44% and the largest improvement is observed on BERT-Tiny on the first query set equal to 9.83%. Otherwise, the degree of retrieval effectiveness remains comparable to the base ranker. We also observe that bias has consistently reduced across both query sets and for all bias metrics. As such, we conclude that the proposed approach is robust to different query sets.

5 CONCLUDING REMARKS

In summary, we find that: (1) our proposed approach shows a more balanced approach to dealing with gender bias compared to the state of the art method, AdvBert. While it does not reduce bias as much as AdvBert, it does in fact effectively reduce bias but not at the cost of a significant drop in effectiveness; and, (2) Our proposed approach showed consistent balanced performance on maintaining retrieval effectiveness and reduced bias regardless of the initial contextual embedding that it is trained on and/or the query set that it is tested on; therefore, showing robust behavior.

REFERENCES

[1] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[2] Krisztian Balog, Filip Radlinski, and Alexandros Karatzoglou. 2021. On Interpretation and Measurement of Soft Attributes for Recommendation. *arXiv preprint arXiv:2105.09179* (2021).

[3] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*. 405–414.

[4] Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2021. Exploring Gender Biases in Information Retrieval Relevance Judgement Datasets. In *Advances in Information Retrieval - 43rd European Conference on IR Research*.

[5] Giovanni Campagna, Silei Xu, Mehrad Moradshahi, Richard Socher, and Monica S Lam. 2019. Genie: A generator of natural language semantic parsers for virtual assistant commands. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*. 394–410.

[6] Chong Chen, Min Zhang, Weizhi Ma, Yiqun Liu, and Shaoping Ma. 2020. Jointly non-sampling learning for knowledge graph enhanced recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 189–198.

[7] Long Chen, Ziyu Guan, Wei Zhao, Wanqing Zhao, Xiaopeng Wang, Zhou Zhao, and Huan Sun. 2019. Answer identification from product reviews for user questions by multi-task attentive networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 45–52.

[8] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[9] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).

[10] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management* 57, 6 (2020), 102377.

[11] Shanshan Feng, Lucas Vinh Tran, Gao Cong, Lisi Chen, Jing Li, and Fan Li. 2020. HME: A Hyperbolic Metric Embedding Approach for Next-POI Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1429–1438.

[12] Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2020. Complementing lexical retrieval with semantic residual embedding. *arXiv preprint arXiv:2004.13969* (2020).

[13] Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. 2020. Learning-to-Rank with BERT in TF-Ranking. *arXiv preprint arXiv:2004.08476* (2020).

[14] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. 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*. 3819–3828.

[15] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[17] Zhenghao Liu, Kaitao Zhang, Chenyan Xiong, Zhiyuan Liu, and Maosong Sun. 2021. OpenMatch: An Open Source Library for Neu-IR Research. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). https://doi.org/10.1145/3404835.3462789

[18] 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*.

[19] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[20] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR*.

[21] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. 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*. 5835–5847.

[22] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation for BERT Rankers. *arXiv preprint arXiv:2104.13640* (2021).

[23] Navid Rekabsaz and Markus Schedl. 2020. Do Neural Ranking Models Intensify Gender Bias?. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2065–2068.

[24] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W White. 2019. VERSE: Bridging screen readers and voice assistants for enhanced eyes-free web search. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 414–426.

[25] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. RepBERT: Contextualized Text Embeddings for First-Stage Retrieval. *arXiv preprint arXiv:2006.15498* (2020).