De-biasing Relevance Judgements for Fair Ranking

Amin Bigdeli[1(✉)], Negar Arabzadeh[2], Shirin Seyedsalehi[1], Bhaskar Mitra[3], Morteza Zihayat[1], and Ebrahim Bagheri[1]

[1] Toronto Metropolitan University, Toronto, Canada
{abigdeli,shirin.seyedsalehi,mzihayat,bagheri}@torontomu.ca
[2] University of Waterloo, Waterloo, Canada
narabzad@uwaterloo.ca
[3] Microsoft Research, Montreal, Canada
bmitra@microsoft.com

Abstract. The objective of this paper is to show that it is possible to significantly reduce stereotypical gender biases in neural rankers without modifying the ranking loss function, which is the current approach in the literature. We systematically de-bias gold standard relevance judgement datasets with a set of balanced and well-matched query pairs. Such a de-biasing process will expose neural rankers to comparable queries from across gender identities that have associated relevant documents with compatible degrees of gender bias. Therefore, neural rankers will learn not to associate varying degrees of bias to queries from certain gender identities. Our experiments show that our approach is able to (1) systematically reduces gender biases associated with different gender identities, and (2) at the same time maintain the same level of retrieval effectiveness.

1 Introduction

There have been both qualitative and quantitative studies that have effectively shown that stereotypical biases are prevalent in various natural language processing and Information Retrieval (IR) techniques, models and datasets [1,2,6,8,9,11,17,23,24]. Given these tools are often deployed at scale, such biases have the potential to directly impact the lives of many people. More specifically within the context of IR, biased retrieval methods can exacerbate biases by exposing users to a set of biased documents in response to user queries. In order to systematically address such biases, various researchers have proposed methods that can help measure and/or mitigate systematic biases, such as gender biases, in IR methods [7,20,21]. For instance, Rekabsaz et al. [21] compared different neural ranking models and found that the ranked list of documents returned by neural ranking models are more inclined towards the male gender compared to traditional retrieval methods such as BM25. Building on this work, Rekabsaz et al. [20] later proposed the AdvBert model, which is a BERT re-ranker, which leverages adversarial training to de-bias the output encoder of the BERT model from gender inclination. The authors reported that AdvBert increases the fairness level of the ranked list of documents.

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13981, pp. 350–358, 2023.
https://doi.org/10.1007/978-3-031-28238-6_24

However, this comes at the cost of reduced retrieval effectiveness. Recently, some studies have investigated reducing stereotypical biases while maintaining retrieval effectiveness [3,4,22]. For instance, the authors in [22] propose a neural ranking model that incorporates a notion of gender bias in the loss function to reduce gender bias exposure in the retrieved documents. Furthermore, in [3], the authors propose a bias-aware negative sampling training strategy that represents those documents that are not only irrelevant but also biased towards a particular gender as negative samples to the model. As a result, the model learns the concept of relevance and avoids gender biases.

Our work in this paper builds on the foundations of the earlier work [3,4,11,20,22] and attempts to address biases exposed by IR methods while maintaining their effectiveness. While existing studies have shown that it is possible to reduce gender biases among the retrieved list of documents associated with gender neutral queries, none of them investigated psychological biases that exist among the retrieved documents of gender affiliated queries. Our work is inspired by the observations made by [5] that shows gold standard relevance judgement collections such as MS MARCO may include systematic psychological biases. We propose a methodical approach for augmenting relevance judgement datasets with automatically generated pairs of query-documents that can systematically reduce biases when used to train neural rankers. We show that neural rankers trained on our proposed de-biased relevance judgement datasets exhibit significantly lower biases while maintaining comparable levels of retrieval effectiveness.

In summary, our work delivers the following main contributions. First, we propose a systematic approach for automatically building query-document pairs that can be used for training neural rankers. Second, we show how combining our proposed query-document pairs with existing gold standard relevance judgement datasets can lead to the training of less biased neural rankers that have competitive effectiveness. We conduct our experiments on the MS MARCO passage collection and use three widely adopted psychological and stereotypical gender bias measurement methods to show that decrease in bias happens effectively regardless of how gender biases are measured. We also report the effectiveness of our approach on the MS MARCO passage retrieval task.

2 Proposed Approach

In this work, our hypothesis is that a neural ranker trained on a balanced relevance judgment dataset has a lower likelihood of exhibiting biased retrieval performance. We aim to augment relevance judgement datasets with query and document pairs that have controlled and matched degrees of bias and hence allow the retrieval method to learn a balanced measure of relevance without being inclined towards certain stereotypical biases. We hypothesize that the augmentation of an existing relevance judgement dataset with our proposed dataset leads to a consistent reduction in bias while maintaining comparable effectiveness. Developing such a balanced dataset cannot be accomplished through crowdsourcing due to several reasons: (1) the collection of a large number of judged queries by human participants is very expensive; and (2) given gender biases may be unconsciously embedded in the labelers' beliefs, it is still possible that the newly collected data still suffer from such biases. Therefore, we propose an automated method to generate pairs of query and relevant documents that have controlled degrees of bias.

2.1 Query-Relevant Document Pairs Generation

Given our objective, we need to first automatically generate a set of queries and their associated relevant documents that would be then further filtered to ensure a balanced representation of bias. To this end, we adopt a translation approach that translates a document into a query representation. We train a transformer model based on existing query-relevant document pairs that are already available in the relevance judgement dataset. Thus, the transformer learns to generate queries for an input document. The details of the transformer is provided in the experimental setup section. With the transformer, we are able to generate queries for each document in any given document corpus; producing a set of query and relevant document pairs.

Given the generated pairs of query and relevant documents, we need to selectively choose comparable queries from different genders. We recognize that gender identities go beyond a binary framework, and in practice requires a careful treatment of a spectrum of gender identities, however given available datasets consist predominantly of binary gender queries, we build two classes of queries affiliated with the male and female genders. The idea is that there should always be a corresponding query in one gender that matches a query in the other where the documents associated with these queries show comparable degrees of stereotypical bias. Such an approach would develop a relevance judgement dataset that controls for bias across different gender-affiliated queries. We determine the gender of each query using the proposed model in [5] which is fine-tuned over the manually classified gendered queries released by [21]. We assume that the gender of the document associated with each query to be the same as the gender of the query. As such, we produce a large number of query and relevant document pairs that have been predictively labeled with gender affiliation information.

2.2 Balancing Biases on Query-Document Pairs

With the generated query-document pairs, we aim to perform a controlled matching process with balanced representation of queries and documents from each gender affiliation such that the matched queries exhibit the same degrees of bias regardless of the gender of the query or document. To this end, we assume, as suggested in the literature [12,18], that each document can be characterized through a set of psychological processes such as affective, cognitive, and perceptual processes, to name a few. Such psychological characteristics of a document can be captured through the widely-adopted Linguistic Inquiry and Word Count (LIWC) toolkit [19]. Let us assume that each document can be characterized by a set of n different psychological characteristics, namely $P_1, P_2, ...P_n$. Let $\phi(d)$ be the document psychological characteristic representation for document d, based on its psychological characteristics as $\phi(d) = [P_1(d), P_2(d), ..., P_n(d)]$

where $\phi(d)$ is an n-dimensional vector whose individual elements quantify the different psychological characteristics observed in document d. We benefit from this document representation to perform the matching process between queries affiliated with different gender identities. Consider D_f and D_m to be the set of relevant documents affiliated with female and male queries, respectively. The degree of similarity of a document in $d_i \in D_f$ and another $d_j \in D_m$ is computed as the cosine similarity of their representations $\phi(d_i)$ and $\phi(d_j)$.

In order to build comparable pairs of queries from across gender identities, for each query in one gender identity, we identify a matching query from the other gender identity such that their associated relevant documents' psychological characteristics are most similar to one another. This will produce a collection of pairs of queries from different gender identities that are associated with relevant documents that have similar psychological characteristics. The benefit of this is that given the queries from each gender identity are paired through a matching process, the degree of bias exposed to each gender-affiliated query is no different than the other and hence bias is controlled across the two classes. We propose that the augmentation of existing relevance judgment datasets such as MS MARCO with our proposed matched query-document pairs has the potential to systematically control the stereotypical gender biases.

3 Experiments

Passage Collection. We employed the MS MARCO passage collection dataset that consists of 8,841,822 passages [14].

Query Sets. For the purpose of measuring psychological characteristics, we use the set of gendered queries introduced in [5]. We also employ two different query sets that consist of neutral queries. The first query set is a human-annotated dataset, which consists of 1,765 neutral queries [21]. The other dataset [20] consists of 215 queries in which the queries are neutral in nature, but the retrieved documents exhibit biases.

Bias Measurement. We adopt two strategies to measure gender biases and refer to these as *proxy measures of bias* because while they have been used in the most recent papers on gender bias in IR, they have not yet been empirically or theoretically shown to be the best or at least reliable measures of bias. The **first approach** relies on measuring differences observed across pairs of gender-affiliated queries. We measure the degree of bias based on the metric proposed in [5] which measures bias as the degree to which male and female affiliations are observed within a document based on psychometric properties offered in LIWC [19]. We measure the difference between the psychological characteristics of queries affiliated with different gender identities as a sign of bias towards a certain gender identity. The **second approach** is based on the bias measurement strategy proposed in [21]. The authors propose two metrics based on (1) presence (Boolean) and (2) term frequency of gendered terms for measuring gender bias within a document. They further expand their proposed metrics over the retrieved list of documents for the queries in the dataset by proposing the

Average Ranking Bias (ARaB) metric, which calculates the degree to which a ranked list of documents are biased towards a specific gender.

3.1 Experimental Setup

To generate query and relevant document pairs, we fine-tuned a T5 transformer model based on the query-document pairs of the MSMARCO training set as suggested in [16]. Using this transformer, we generate queries for each document. Furthermore, to estimate the gender affiliation of queries, we adopt the BERT model released in [5].

As a result of generating queries based on the T5 transformer and estimating query gender affiliations using the fine-tuned BERT, we produce 298,389 female, 460,776 male, and 8,056,297 neutral queries. Each of these queries are associated with one relevant judgment document used to generate the query. Furthermore, for each document in this collection, we produce $\phi(d)$ based on LIWC psychological characteristics, namely affective processes, cognitive processes, drives, and personal concerns, and their subprocesses, which constitute a total of 22 subprocesses. Inspired by [10], we augment the small training set of MS MARCO with data from our generated query-document pairs using different ratios with 10% increments.

Based on the de-biased datasets, we leverage the BERT transformer model for passage ranking introduced by Nogueira et al. [15] and train `BERT-base-uncased` on the original dataset, i.e., the small training set of MS MARCO, as well as the newly developed de-biased datasets. We use OpenMatch [13] to fine-tuning for the ranking task with batch size of 64, learning rate of 2e-5, and epoch of 1. We also set the max document length and max query length to 150 and 20, respectively. We publicly release our code, models, results and datasets for general use.[1] We note that while the query-document pairs are included in our dataset, the predicted gender affiliation of the queries are hidden as these were solely predicted based on a fine-tuned BERT model and may not be reflective of true gender affiliations.

Table 1. MRR on original and de-biased datasets. * indicates statistically significant decrease in effectiveness. (two-tailed paired t-test 95% confidence).

Training set	Ratio	MRR@10	Reduction (%)
Original	–	0.3080	–
De-biased	0.05	0.3100	0.65%
	0.15	0.3039	−1.33%
	0.25	0.3002	−2.53%
	0.35	0.2905	−5.68%*

[1] https://github.com/aminbigdeli/balanced-relevance-judgment-collection.

Table 2. Impact of training on de-biased dataset on the difference in psychological characteristics of gender-affiliated queries.

Training dataset		Affective processes	Cognitive processes	Drives	Personal concerns
Original dataset	Female queries	0.0315	0.0725	0.0545	0.0600
	Male Queries	0.0290	0.0521	0.0641	0.0829
	Difference	0.0025	0.0204	0.0095	0.0229
De-biased dataset	Female queries	0.0304	0.0730	0.0536	0.0546
	Male queries	0.0288	0.0563	0.0624	0.0747
	Difference	0.0016	0.0167	0.0088	0.0201
Reduction (%)		36.00%	18.13%	7.37%	12.22%

3.2 Results and Findings

Impact on Retrieval Effectiveness. The objective of our work has been to reduce proxy measures of bias while maintaining retrieval effectiveness. As such, we investigate how the same model [15] performs when trained on different training datasets including the MS MARCO small training set and the de-biased datasets with different ratios. We measure retrieval effectiveness based on the 6,980 queries of the small dev set of MS MARCO collection based on the standard leaderboard metric, i.e., MRR@10.

Table 1 shows the results of the model when trained based on different augmentation ratios. We increased the augmentation ratio until the retrieval effectiveness of the model dropped significantly below the performance of the model that was trained on MS MARCO dataset without augmentation.

Table 3. The impact of training `BERT-base-uncased` on the de-biased dataset on proxy measures of gender bias based on different neutral query sets. Reduction (%) values are computed based on actual metric values, while the metric values are rounded to three decimal points.

Query set	Training set	TF ARaB		Boolean ARaB		LIWC	
		Value	Reduction	Value	Reduction	Value	Reduction
QS1	Original	0.072	–	0.059	–	0.011	–
	De-biased	0.059	18.05%	0.049	16.95%	0.011	5.98%
QS2	Original	0.029	–	0.017	–	0.006	–
	De-biased	0.019	34.48%	0.011	35.29%	0.005	16.67%

It is expected that as the number of synthetically generated data pairs used to augment the original dataset increases, the effectiveness of the model drops gradually. This is because the query-document pairs are included in our synthetic dataset such that they would balance the degrees of bias and since they are synthetic query-document pairs, they are not as effective for training the model to learn query-document relevance. On the other hand, the expectation would be that a larger ratio of synthetic data would lead to drop in bias. As such, the preference would be to include as much synthetically generated data as possible to reduce bias. As shown in Table 1, we increase the ratio until the time when the decrease in performance becomes, statistically speaking, significantly lower than the model trained on the original set. This happens when the ratio is set to 35%. Therefore, we employ the ratio of 25% in the rest of our experiment.

Impact on Proxy Measures of Bias. We investigate the impact of our approach on the reduction of the proxy measures of bias.
Bias Observed on Gender Affiliated Queries: We adopt the gendered queries released in [5] and calculate the psychological characteristics observed in the ranked list of each of the models trained on original MS MARCO in comparison to the one trained on the 25% de-biased dataset.

As shown in Table 2, the model trained on the de-biased dataset substantially reduces the differences found on the expression of psychological characteristics between the queries in the different gender identities. A higher reduction in the differences between the gender identities, in Table 2, is a positive indication of the success to bridge the gap between the representation of psychological characteristics in documents retrieved in relation to the gendered queries.

Bias Observed on Neutral Queries: We adopt two different query sets: (**QS1**) 1,765 neutral queries from [21], and (**QS2**) 215 queries from [20]. We investigate if the `BERT-base-uncased` model trained on the de-biased dataset is able to reduce the gender biases among the ranked documents for neutral queries. We report the level of gender bias among the top-10 ranked list of documents for neutral queries using both classes of ARaB metric proposed in [21] in Table 3. As shown in the table, gender inclination among the ranked list of documents for neutral queries decreases significantly when they are retrieved by the model trained on the de-biased dataset in terms of both Boolean and Term Frequency ARaB measures. To validate our findings, we also report the difference between the male and female affiliation for the top-10 ranked list of documents to measure the degree of gender inclination in Table 3. As shown, the reduction of bias associated with gender affiliation computed by LIWC is consistent with both of the ARaB measures and can be observed over all of the datasets. According

Table 4. BERT-Tiny trained on our de-biased dataset vs AdvBert-Tiny. Reduction (%) values are computed based on actual metric values, while the metric values are rounded to three decimal points and reported in this table.

Query set	Training set	Utility	TF ARaB		Boolean ARaB		LIWC	
		MRR@10	Value	Reduction (%)	Value	Reduction (%)	Value	Reduction (%)
QS1 (Rekabsaz et. al 2020)	Original	0.219	0.076	–	0.063	–	0.012	–
	De-biased	0.199	0.047	38.15%	0.042	32.06%	0.010	10.74%
	AdvBert	0.189	0.064	15.78%	0.058	7.30%	0.009	24.79%
QS2 (Rekabsaz et. al 2021)	Original	0.175	0.005	–	0.006	–	0.005	–
	De-biased	0.163	0.001	79.19%	0.000	97.01%	0.005	11.11%
	AdvBert	0.149	0.009	−85.98%	0.007	−16.67%	0.005	14.81%

to Table 3, the proposed de-biased dataset for training neural ranking models can reduce gender inclination among the retrieved list of documents for neutral queries.

Comparative Analysis. We compare our work with a recent method proposed by [20], known as AdvBert. As suggested in their paper, we adopt the `BERT-Tiny` model and train it based on the method proposed by the authors over the original MS MARCO dataset. We additionally, train the same `BERT-Tiny` model without adversarial training on the original MS MARCO dataset as well as our proposed de-biased dataset. We report the performance of these models on the two sets of neutral queries. Table 4 shows the results in terms of ARaB and LIWC for the three models and across two query sets. Our approach shows superior retrieval effectiveness compared to AdvBert. This speaks to the objective of our work to maintain effectiveness while addressing bias. In terms of the proxy measures of bias and specifically when considering the ARaB metrics, our proposed approach shows consistent superior performance over AdvBert in both query sets and variations of the ARaB metric. On the other hand, when comparing the degree of bias reduction based on the LIWC-based gender affiliation, we find that AdvBert has a higher degree of bias reduction but this has come at the cost of effectiveness.

4 Concluding Remarks

We proposed an approach to generate matched query-document pairs across gender identities for systematically reducing stereotypical biases that are learnt by neural rankers. Our approach distinguishes itself from existing methods in that (1) it systematically reduces gender biases, and also (2) maintains comparable levels of retrieval effectiveness.

References

1. Baeza-Yates, R.: Bias on the web. Commun. ACM **61**(6), 54–61 (2018)
2. Baeza-Yates, R.: Bias in search and recommender systems. In: Fourteenth ACM Conference on Recommender Systems, p. 2 (2020)
3. Bigdeli, A., Arabzadeh, N., Seyedsalehi, S., Zihayat, M., Bagheri, E.: A light-weight strategy for restraining gender biases in neural rankers. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 47–55. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_6
4. 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)
5. Bigdeli, A., Arabzadeh, N., Zihayat, M., Bagheri, E.: Exploring gender biases in information retrieval relevance judgement datasets. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12657, pp. 216–224. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72240-1_18

6. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
7. Fabris, A., Purpura, A., Silvello, G., Susto, G.A.: Gender stereotype reinforcement: measuring the gender bias conveyed by ranking algorithms. Inf. Process. Manag. **57**(6), 102377 (2020)
8. Font, J.E., Costa-Jussa, M.R.: Equalizing gender biases in neural machine translation with word embeddings techniques. arXiv preprint arXiv:1901.03116 (2019)
9. 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)
10. Ju, J.H., Yang, J.H., Wang, C.J.: Text-to-text multi-view learning for passage re-ranking. arXiv preprint arXiv:2104.14133 (2021)
11. Klasnja, A., Arabzadeh, N., Mehrvarz, M., Bagheri, E.: On the characteristics of ranking-based gender bias measures. In: 14th ACM Web Science Conference 2022, pp. 245–249 (2022)
12. Li, J., Ott, M., Cardie, C., Hovy, E.: Towards a general rule for identifying deceptive opinion spam. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1566–1576 (2014)
13. Liu, Z., Zhang, K., Xiong, C., Liu, Z.: OpenMatch: an open-source package for information retrieval. arXiv e-prints pp. arXiv-2102 (2021)
14. Nguyen, T., et al.: MS MARCO: a human generated machine reading comprehension dataset. In: CoCo@ NIPS (2016)
15. Nogueira, R., Cho, K.: Passage re-ranking with BERT. arXiv preprint arXiv:1901.04085 (2019)
16. Nogueira, R., Lin, J., Epistemic, A.: From doc2query to docTTTTTquery. Online preprint (2019)
17. Olteanu, 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 (2021)
18. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557 (2011)
19. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC 2001. **71**(2001) (2001)
20. 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)
21. 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)
22. Seyedsalehi, S., Bigdeli, A., Arabzadeh, N., Mitra, B., Zihayat, M., Bagheri, E.: Bias-aware fair neural ranking for addressing stereotypical gender biases. In: EDBT, pp. 2–435 (2022)
23. Seyedsalehi, S., Bigdeli, A., Arabzadeh, N., Zihayat, M., Bagheri, E.: Addressing gender-related performance disparities in neural rankers. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2484–2488 (2022)
24. Sun, T., et al.: Mitigating gender bias in natural language processing: literature review. arXiv preprint arXiv:1906.08976 (2019)