Addressing Gender-related Performance Disparities in Neural Rankers

Shirin Seyedsalehi
Ryerson University
Toronto, Canada
shirin.seyedsalehi@ryerson.ca

Amin Bigdeli
Ryerson University
Toronto, Canada
abigdeli@ryerson.ca

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

Morteza Zihayat
Ryerson University
Toronto, Canada
mzihayat@ryerson.ca

Ebrahim Bagheri
Ryerson University
Toronto, Canada
bagheri@ryerson.ca

ABSTRACT

While neural rankers continue to show notable performance improvements over a wide variety of information retrieval tasks, there have been recent studies that show such rankers may intensify certain stereotypical biases. In this paper, we investigate whether neural rankers introduce retrieval effectiveness (performance) disparities over queries related to different genders. We specifically study whether there are significant performance differences between male and female queries when retrieved by neural rankers. Through our empirical study over the MS MARCO collection, we find that such performance disparities are notable and that the performance disparities may be due to the difference between how queries and their relevant judgements are collected and distributed for different gendered queries. More specifically, we observe that male queries are more closely associated with their relevant documents compared to female queries and hence neural rankers are able to more easily learn associations between male queries and their relevant documents. We show that it is possible to systematically balance relevance judgment collections in order to reduce performance disparity between different gendered queries without negatively compromising overall model performance.

CCS CONCEPTS

• **Information systems** → **Learning to rank**.

KEYWORDS

Information Retrieval, Neural Rankers, Gender Bias, Responsible-AI

ACM Reference Format:
Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2022. Addressing Gender-related Performance Disparities in Neural Rankers. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3531888

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531888

1 INTRODUCTION

With the increased real-world adoption of information retrieval and natural language processing techniques, concerns about the intensification of stereotypical biases, such as gender biases, by these techniques have increased. Specifically, within the information retrieval literature, there have been documented work that point to systematic biases exhibited by retrieval methods [1, 6]. For instance, Rekabsaz et al. [13] were among the first to report that neural rankers have the potential to intensify gender biases in their retrieved list of documents even when dealing with non-gendered queries. In another similar study, Fabris et al. [4] showed that neural retrieval methods may reinforce stereotypical gender biases within the final retrieved list of documents shown to the users. While researchers such as Rekabsaz et al. and Fabris et al. have explored the role of neural rankers, Bigdeli et al. have explored this issue from a complementary view and studied whether stereotypical gender biases could be observed in gold standard relevance judgment collections [3]. They explored the relevance judgements of the MS MARCO collection [7] and found that relevance judgements do exhibit biases towards/against different genders.

Given the existence of such biases in retrieval methods, there have been attempts to develop methods that can contain or reduce the levels of stereotypical biases in the retrieved results [12]. As a prominent example, Rekabsaz et al. propose an adversarial loss function for neural ranking models that encodes the notion of bias and reduces gender biases within the ranked list of documents shown to the users. In another similar study, Seyedsalehi et al. [14] focused on including an explicit term in the loss function of neural rankers to control for stereotypical gender biases. These methods have shown that it is possible to create a trade-off between retrieval effectiveness and observed degrees of bias[2].

While impactful, we note that existing works primarily focus on measuring stereotypical gender biases in neural rankers based on the affiliation of the retrieved documents with certain genders. For instance, metrics such as Boolean-RaB [13], TF-ARaB [13], and FaiRR [12] have been introduced that often rely on some measure of the frequency of gender-affiliated words within a document to ascribed preference towards that gender to a document. For instance, if a document consists of a larger number of male-affiliated words

compared to female-affiliated words, then this document is considered to be more biased towards the male gender. On this basis, the major assumption of existing works has been that the expectation from a neural ranker is not to show any preference towards any gender when given non-gendered queries (these queries are also sometimes known as neutral queries). However, to the best of our knowledge, there are no earlier works that have systematically analyzed the performance of neural rankers with regards to gendered queries. As such, our work in this paper is among the first to understand, regardless of the bias induced by neural rankers, whether neural rankers provide a similar level of retrieval effectiveness for different gender-affiliated queries or not.

More specifically, we are interested in exploring the following objectives in this paper:

- To quantify any possible differences between the retrieval effectiveness shown by neural rankers between queries from different genders;
- To offer insight into the reasons as to why neural rankers may exhibit different retrieval effectiveness for queries from different genders;
- To offer a systematic approach to alleviate any differences between the retrieval effectiveness of gendered queries such that the overall performance of the neural rankers are maintained and the performance of gendered queries reach a comparable level.

We perform our experiments on the queries and documents of the widely-adopted MS MARCO passage retrieval collection[1] and benefit from earlier work from Rekabsaz et al. [13] who have collected a set of queries affiliated with different gender types from human judges. In summary, we find that there is a noticeable performance difference between queries affiliated with different genders where queries related to the male gender show a consistent higher performance compared to queries related to the female gender. We find that this can, in part, be due to how relevance judgment documents were distributed in the collection, i.e., relevance judgment documents for male queries were much more similar to their query compared to the relevance judgment documents for the female queries. We further show that it is possible to reduce such performance disparities across queries affiliated with different genders by controlling how relevance judgment documents are distributed over different query gender types.

We note that all of our the data, code and results are made publicly accessible[2].

2 PRELIMINARIES

It is important to note that while we recognize the importance of different gender identities, the work in this paper follows existing works in the literature that analyze the impact of neural rankers on male and female related queries [3]. In this context, query q is considered to be a gendered query if it aims to inquire about information related to a specific gender. A gendered query may explicitly include gender-related words such as mother, woman, and girl or may implicitly target a specific gender such as a query about pregnancy. In contrast, a query is considered to be a neutral query if it inquires about general non-gendered information. Table 1 presents some examples of male, female, and neutral queries adopted from the dataset released by Rekabsaz et al. [13].

Table 1: Sample gendered and neutral queries from Rekabsaz et al. [13].

Gender Affiliation	Query
5*Female	how old is elle mckinnon?
	effects of being pregnant.
	who played carol brady?
	what is mammogram screening?
	when did selena quintanilla die?
5*Male	how old is tom selleck?
	what is tony hawk's stan lee net worth?
	how tall is jon jones?
	how many rings does michael jordan has?
	when did nelson davis die?
5*Neutral	in what year had youtube started?
	sick of a palsy biblical definition.
	before how long should i go for passport renewal?
	what is the purpose of dna replication?
	how much does a lab cost per month?

Given an initial dataset of gendered and neutral queries such as the one released in [13], Bigdeli et al. [3] have shown that it is possible for a function Ψ to be trained to represent a mapping from $q \in Q$ to each particular gender affiliation:

$$g = (\Psi; \eta) \tag{1}$$

Where g represents female, male, and neutral genders and η is the set of parameters of the mapping function Ψ. Given Ψ, any query set Q can be divided into female Q^f, male Q^m and neutral query sets. A fair neural ranker is expected to show similar retrieval effectiveness for both female and male queries such that:

$$U(Q^f) \sim U(Q^m) \tag{2}$$

Where $U(Q)$ measures the average performance, e.g., mean average precision, of the neural ranker R on the query set Q. We define *gender-related retrieval effectiveness disparity* when $U(Q^m) > U(Q^f)$ (or vice versa) and this observed difference is statistically significant through a statistical significance test such as a paired t-test.

3 METHODOLOGY

In order to explore gender-related retrieval effectiveness disparities over male and female queries, we benefit from the widely adopted MS MARCO dataset [7], which consists of 8,841,822 passages in its collection set, and 500K pairs of query and relevant documents. We further adopt the BERT-based query classifier proposed in [3] to act as Ψ and classify the training set queries of MS MARCO and obtain 14,000 male and 14,000 female queries. In addition, we classify the queries in the test set of MS MARCO and obtain 1,405 male and 1,405 female queries. Furthermore, and in order to explore the performance disparities by neural rankers, we adopt the well-known first stage retriever known as Sentence-BERT[11]. Sentence-BERT has shown to have promising and stable performance as well as being computationally inexpensive compared to other neural

[1] https://microsoft.github.io/msmarco/
[2] https://github.com/shirinssalehi/Addressing-Gender-related-Performance-Disparities-in-Neural-Rankers

models. We report performances for Sentence-BERT when different contextualized embeddings are used for it.

3.1 Gender-related Performance Disparity

We explore gender-related performance disparities by comparing the performance of the set of 1,405 male queries to that of the 1,405 female queries. We note that these queries are all identified from the MS MARCO test query set. We report the performance of the neural ranker with different contextualized embeddings for the neural ranker over male and female queries in Table 2. We find that regardless of the adopted contextualized embedding in the neural ranker, there is notable statistically significant disparity between the retrieval effectiveness (MRR@10) of male queries compared to female queries. This difference is at least 14.47% on the MiniLM contextual embedding. This is a significant performance disparity. In the following section, we explore the possible reasons for such a large disparity in the performance of gendered queries.

3.2 Source of Disparity for Neural Rankers

In order to understand the possible reasons for the large disparity between the performance of gendered queries, we consider the core component of neural rankers which is their loss function. The loss function of a neural ranker learns the concept of relevance based on the relevance judgements [5, 8, 10]. The loss function often leverages the relevant and irrelevant documents to the query and compares the ranking results with a gold standard relevance judgment dataset to improve the quality of ranking. Now, given neural rankers often fine-tune an existing contextual language model such as MiniLM, or BERT, they are likely to prefer existing association already observed between terms and documents in the pretraining dataset. As such, the performance of neural rankers would be higher on those queries whose relevance judgment documents are closely related with the query in the initial contextual embedding. In other words, the more immediately related the query and its relevant judgment document are, the more effectively the neural ranker would be able to learn their association. For instance, if the terms of the query appear in the relevance judgment document of that query, the neural ranker would be able to seamlessly learn the association between the query and its relevant document. However, if the terms in the query have never been observed in similar contexts to terms of its relevant document, the neural ranker will have a hard time learning such relevance. As such, we hypothesize that it may be possible that the degree of association between male queries and their relevance judgment documents is much higher than that of the female queries and their associated relevant documents.

In order to empirically explore this hypothesis, we first quantify the degree of association between queries and their relevance judgments and compare it between male and female queries. Let us assume $\gamma(q, d)$ is a deterministic function to calculate a similarity score between a query q and a document d based on the common words between the query and the document. For each female query $q^f \in Q^f$ and male query $q^m \in Q^m$ the score between the query and its relevant document d_q^{rel} can be calculated as:

$$\begin{aligned} S^f &= \gamma(q^f, d_{q^f}^{rel}) \\ S^m &= \gamma(q^m, d_{q^m}^{rel}) \end{aligned} \tag{3}$$

Table 2: Performance of the neural ranker on gendered queries (MRR@10).

	Male	Female	Difference (%)
BERT-tiny	0.4065	0.3397	19.66 %
MiniLM	0.4913	0.4292	14.47 %
BM25	0.2783	0.2284	21.85 %

We define two distributions $P_{S^f}(s)$and $P_{S^m}(s)$ as the histogram of female and male query scores over the female and male queries, respectively. A fair relevance judgment collection is expected to have the same distribution of $\gamma(.)$ for both female and male scores. For each of the female and male query sets, we employ the BM25 score function as γ and calculate S^f, and S^m and then P_{S^f} and P_{S^m} accordingly. Figure 2 shows the results of these two sets of distributions. Based on Figure 2, we find that the distribution of γ for female queries is skewed towards the left while that of male queries is skewed to the right. This indicates that there are a larger number of male queries that have a higher degree of association with their relevant documents compared to female queries.

Based on this observation, we further hypothesize that the performance disparity observed for neural rankers can be, in part, due to the degree of association between queries and their relevant documents. Given male queries have a higher degree of association with their relevant documents, neural rankers are able to learn their association more easily and hence leading to better performance on such queries.

3.3 Controlling Gender-related Performance Disparities

In order to empirically validate our hypothesis with regards to a possible source for the performance disparity between female and male queries, one would expect that such disparity be removed, or at least minimized, if the distribution of male-related and female-related degrees of association between queries and their relevant documents were kept at a comparable rate. In other words, if the distributions observed in Figure 1 were fully overlapping, then the expectation would be that the performance of female and male queries would also be comparable. For this purpose and in order to develop fully overlapping distributions for male and female queries, we define a gender-balanced relevance judgment collection as a set of query-document pairs which have a similar distribution of scores over the queries:

$$\{(q^f, d_{q^f}^{rel}), (q^m, d_{q^m}^{rel}) | P_{S^f}(s) \sim P_{S^m(s)}\} \tag{4}$$

We propose two strategies for developing gender-balanced relevance judgment collection. In the first strategy, we divide the queries into k bins based on their degree of association with their relevant documents. Once the female and male queries are placed into appropriate bins, we randomly sample male queries and for each sampled male query, we also sample a corresponding female query from the same bin that has a similar degree of association with its relevant document as the male query. We continue the sampling process until there are no more corresponding female or male queries left. We refer to this process as *balancing using query matching*. The balanced histograms developed as a result of this

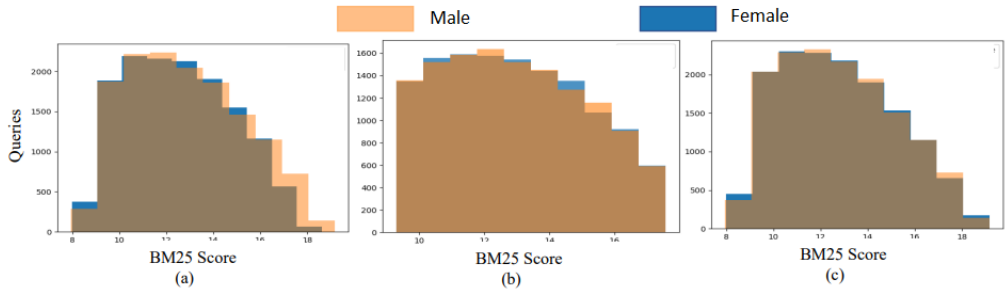

Figure 1: Distribution of the degree of association between queries and their relevance judgments for both male and female queries (a) on the MS MARCO training set, (b) after systematically balancing using query generation., and (c) after systematically balancing using query matching.

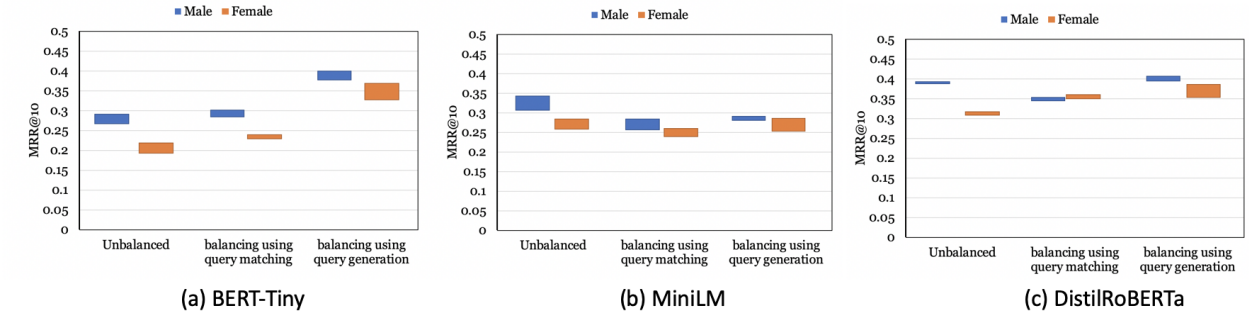

Figure 2: Performance of neural rankers trained on unbalanced and balanced query sets with query matching and query generation.

process is illustrated in Figure 2. Although this strategy results in a balanced distribution over male and female query scores, it comes at the cost of eliminating some of the labeled training data which is not efficient.

In the second strategy, and in order to avoid eliminating any of the queries from either the male or female sets, we employ the Transformer-based query generation method, known as DocT5 [9], to generate queries for the documents in the MS MARCO collection. We generate 5 queries per document for each document in MS MARCO. Each of the generated queries is then labeled for their gender based on $\Psi(.)$. This will provide us with a set of synthetically developed gendered queries for which we can compute their degree of association with their relevant document. The relevant document for each synthetic query is the document from which DocT5 generated the document to begin with. Similar to the previous strategy, we divide the gendered queries into bins based on their degree of association with their relevant documents. Now, given the synthetic gendered queries, we sample queries from the synthetic gendered queries to fill in each bin with synthetic queries that have corresponding scores for that bin until the distribution of male and female queries in that bin is balanced. We repeat this process for all bins. The balanced histograms developed as a result of this process are illustrated in Figure 2.

Given the newly balanced relevance judgments based on the *balancing using query matching* and *balancing using query generation* strategies, we train each neural ranker again and separately once for each of the balanced relevance judgments. The results of the performance of the rankers are reported in Figure 2 for the different contextual embeddings. We make the following observations based on the reported figures:

- When applying either of the two balancing strategies, we observe a consistent and statistically significant reduction in the retrieval effectiveness disparity between female and male queries. This is most clearly observable on the DistilRoBERTa contextualized embedding where the performance of male and female queries are consistent while the average retrieval effectiveness over all queries is maintained at the same rate as when the unbalanced relevance judgment was used to train the neural ranker. This shows that it is possible to effectively balance the relevance judgment datasets in order to reduce retrieval effectiveness disparity between different gendered queries while at the same time maintaining similar retrieval effectiveness over all queries.

- When comparing between the two balancing strategies, we observe that the *balancing using query generation* leads to a lower disparity between female and male queries but at

the same time also shows comparable overall performance (in cases even better performance) compared to the model trained on the unbalanced dataset. We believe that this can be due to the fact that the query generation strategy is actually introducing new queries to the training set, which both increases the number and diversity of samples that the neural ranker is exposed to. For this reason, this strategy leads to better retrieval effectiveness as well as a lower disparity between the female and male-gendered queries.

We believe that the observations from our empirical studies confirm our hypothesis about one of the potential sources of retrieval effectiveness disparity between gendered queries. Our experiments show that the degree of association between a male query and its relevant document is much higher than that of the association between a female query and its relevant document. When such differences are controlled for and removed using either query matching or query generation strategies, the retrieval effectiveness disparity is reduced. As such, we conclude that one of the sources for retrieval effectiveness disparity for gendered queries is related to how relevant relevance judgment documents are collected for gendered queries. These relevant judgements systematically favor male over female queries due to the nature of the relevant judgment documents and their association with male and female queries.

4 CONCLUDING REMARKS

While earlier works have extensively explored how neural rankers may intensify gender biases even for queries that are considered to be neutral, i.e., they are not associated with or related to any gendered information, to the best of our knowledge, there are no existing works that explore the impact of query gender on retrieval effectiveness. In this paper, we empirically show that there is a systematic retrieval effectiveness disparity between male and female queries where neural rankers perform consistently better over male compared to female queries. We show that such performance disparity can be due to the characteristics of the relevance judgment collections where male queries have a much higher degree of association with their relevant documents compared to female queries. This can allow the neural ranker to more easily learn the relevance between male queries and their relevant documents compared to female queries. Furthermore, we show that a disciplined approach towards balancing relevance judgment collections can lead to a lower disparity between gendered query performances while being able to maintain comparable retrieval effectiveness over all queries.

REFERENCES

[1] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. 2022. A Light-Weight Strategy for Restraining Gender Biases in Neural Rankers. In *European Conference on Information Retrieval*. Springer, 47–55.

[2] Amin Bigdeli, Negar Arabzadeh, Shirin Seyersalehi, Morteza Zihayat, and Ebrahim Bagheri. 2021. 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*.

[3] 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*.

[4] 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.

[5] Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540* (2021).

[6] Anja Klasnja, Negar Arabzadeh, Mahbod Mehrvarz, and Ebrahim Bagheri. 2022. On the Characteristics of Ranking-based Gender Bias Measures. In *WebSci'22* (2022-03-30) *(The 14th International ACM Conference on Web Science in 2022 (WebSci'22), 26 – 29, June, 2022, Universitat Pompeu Fabra, Barcelona, Spain)*.

[7] 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*.

[8] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[9] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docTTTTTquery. *Online preprint* (2019).

[10] Harshith Padigela, Hamed Zamani, and W Bruce Croft. 2019. Investigating the successes and failures of BERT for passage re-ranking. *arXiv preprint arXiv:1905.01758* (2019).

[11] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR* abs/1908.10084 (2019). arXiv:1908.10084 http://arxiv.org/abs/1908.10084

[12] 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).

[13] 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.

[14] Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Bhaskar Mitra, Morteza Zihayat, and Ebrahim Bagheri. 2022. Bias-aware Fair Neural Ranking for Addressing Stereotypical Gender Biases. In *EDBT/ICDT 2022* (2022-01-01).