On the Characteristics of Ranking-based Gender Bias Measures

Anja Klasnja
Ryerson University, Canada
Canada
aklasnja@ryerson.ca

Negar Arabzadeh
University of Waterloo
Canada
narabzad@uwaterloo.ca

Mahbod Mehrvarz
University College London
UK
Mahbod.mehrvarz.20@ucl.ac.uk

Ebrahim Bagheri
Ryerson University
Canada
bagheri@ryerson.ca

ABSTRACT

With increased recent awareness on the possible impact of retrieval techniques on intensifying gender biases, researchers have embarked on defining quantifiable gender bias metrics that can provide the means to concretely measure such biases in practice. While successful in allowing for identifying possible sources of gender bias, there has been little work that systematically explores the characteristics of these metrics. This paper argues that effective future works on gender biases in information retrieval require a careful understanding of the bias metrics in terms of their consistency, robustness, sensitivity and also their relation with psychological characteristics and what they actually measure. Through our experiments, we show that more rigorous work on gender bias metrics need to be pursued as existing metrics may not necessarily be consistent and robust and often capture differing psychological characteristics.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Evaluation of retrieval results**.

KEYWORDS

Gender Bias, Fairness, Gender stereotypes , Bias in Information Retrieval

ACM Reference Format:
Anja Klasnja, Negar Arabzadeh, Mahbod Mehrvarz, and Ebrahim Bagheri. 2022. On the Characteristics of Ranking-based Gender Bias Measures. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3501247.3531540

WebSci '22, June 26–29, 2022, Barcelona, Spain
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9191-7/22/06…$15.00
https://doi.org/10.1145/3501247.3531540

1 INTRODUCTION

Stereotypical gender biases, often in the form of unconscious bias or implicit bias, can lead to discrimination when preconception about attributes or characteristics of a certain gender are imposed. Researchers have already reported that recent advances in natural language processing and Information Retrieval (IR) have the tendency to capture, intensify and exhibit such gender biases at scale [2, 5, 6, 8–10, 12–15, 17, 22, 24, 25]. More specifically within the IR community, a potentially biased retrieval system can expose a large number of users to information that can shape their thoughts, decision making and choices.There have been enlightening recent studies that show information retrieval systems may potentially be exacerbating gender biases [3, 15, 19, 20].

As expected, in order for researchers to be able to possibly explore the extent of gender biases within information retrieval techniques, it seemed imperative to define quantifiable metrics that would indicate the level of stereotypical gender biases exposed by IR techniques. This strategy is aligned with the popular quote *'If you can't measure it, you can't improve it'* attributed to Lord Kelvin and widely practiced in measurement theory. To this end, the pioneering work by Rekabsaz et al. was among the first to quantify gender biases in search retrieval results [20]. They proposed a framework to quantify gender biases in search results by measuring the gender magnitude of retrieved documents. The basic idea behind their approach is to check the existence and/or frequency of a set of predefined gendered terms such as she/woman/queen for female and he/man/king for male within a set of documents. Their proposed metrics, i.e., Term-Frequency based Average Rank Bias (TF-ARaB) and Boolean-based Average Rank Bias (Boolean-ARaB), would check to see if gendered terms associated with a certain gender are more prevalent in the list of documents than terms associated with other genders, and if so, those documents would be considered to be biased towards that specific gender. The idea of measuring the difference between gender affiliated terms, i.e., terms that are inclined towards any of the genders, such as male and female, has also been explored in neural embeddings [4, 7].

Later, inspired by [7, 11, 23], a new gender bias metric was proposed in which the ranking properties of the retrieved list of documents were more taken into consideration. In other words, this proposed gender bias metric, known as the Normalized Fairness of Retrieval Results (NFaiRR), focuses more on a ranked list of documents and compares it to the ideal fair rank that the ranker could possibly retrieve from within the collection [19]. Taking the document's rank into account is an important consideration because the rank of a document is an indicator of the likelihood of the user being exposed to that document. Thus, a document in a lower rank has a lower chance of being exposed to the user. Therefore, gender biases exhibited by lower-ranked documents may not be as impactful as those exposed by higher-ranked documents. Similar to TF-ARaB and Boolean-ARaB, NFaiRR is also based on a predefined list of gendered terms associated with each gender identity.

Building on the proposed quantifiable gender bias metrics, there have been attempts to reduce such biases during retrieval. For instance, ADVBERT is a gender bias mitigation method in the context of deep neural ranking models, which extends BERT-based rankers with an adversarial training mechanism [19]. This method couples the ranker's main objective with bias reduction for maintaining model performance as well as protecting gender attributes of interest. The authors evaluated the effectiveness of their proposed ranker in terms of gender bias reduction using the NFaiRR Metric. Similarly, Bigdeli et al. [1] proposed a bias-aware query expansion framework that reformulates an input query such that it maintains retrieval effectiveness while reducing observed gender biases in the retrieved list of documents based on the ARaB metrics.

While these metrics have allowed researchers to quantify gender biases, they are fundamentally dependent on a set of predefined gendered terms and quantify bias as a function of the presence or frequency of these terms in documents. This dependence on a predefined list of terms, while convenient, could raise questions with regards to the reliability of the defined metrics and the generalizability of the findings that are reported based on these metrics. As such, the objective of our work in this paper is to investigate the characteristics of these metrics and explore whether the metrics can reliably be used to make conclusions about the existence and degree of gender biases in retrieval methods. To this end, we study three main research questions (RQ) as follows:

- **RQ1 (Consistency).** Do existing gender bias metrics exhibit similar behavioral patterns and show comparable degrees of bias over the same datasets?
- **RQ2 (Sensitivity).** How sensitive are the existing gender bias metrics to the pre-defined list of gendered terms that are used to quantify bias?
- **RQ3 (Psychological characteristics).** Are gender bias metrics capturing specific psychological characteristics of gender bias? And if so, what psychological characteristics within each gender are the bias metric capturing?

The broader impact of our work is that it will allow the community to understand the behavior of the existing gender bias metrics and understand the semantics of what these metrics are measuring, to have a better picture of what is captured when used to quantify gender bias in retrieval. We will also advocate for a more comprehensive treatment of stereotypical biases beyond the reporting of quantifiable metrics, which are likely not able to capture the full spectrum of issues associated with biases.

2 RESEARCH FRAMEWORK

The objective of our work is to provide insight into the behavioral characteristics of gender bias metrics and investigate them from the perspectives of consistency, sensitivity, and psychological characteristics. Given the delicacy of the topic of gender biases, and stereotypical biases in general, our hope is to raise awareness into the need to design measurement frameworks for biases that go beyond mere quantification of bias by capturing the psychological characteristics of the biases and contextualizing it within the literature.

2.1 Preliminaries and Datasets

Here we formally introduce the gender bias metrics and the query and document collections that were employed in this paper.

Gender Bias Metrics. We consider the two variations of the ARaB metric, i.e., TF-ARaB and Boolean-ARaB [20] as well as the NFaiRR metric [19] in this paper. Based on the definition in [20], the ARaB metrics are defined for a query q and its top-t retrieved documents D where d_i^q refers to the i^{th} document in D as follows:

$$AraB(q) = \frac{1}{t}\sum_{x=1}^{t}\frac{1}{x}\sum_{i=1}^{x} mag^f(d_i^q) - mag^m(d_i^q) \tag{1}$$

where $mag^{m/f}(d)$ for TF-ARaB and Boolean-ARaB are defined in Equation 2 and 3, respectively:

$$TF : mag^{m/f}(d) = \sum_{w\in V_{f/m}} log(tf(w,d)) \tag{2}$$

$$Boolean : mag^{m/f}(d) = \begin{cases} 1 & if \sum_{w\in V_{f/m}} tf(w,d) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $tf(t,d)$ represents the frequency of term t in document d and $V_{d/f}$ represent the vocabulary list of gendered terms (female and male related terms such as she/her/girl and he/him/boy) as proposed in [19][1].

The other gender bias metric referred to as NFaiRR is defined based on neutrality of the content of a document. A document is considered gender-neutral if it includes either no or a balanced representation of the specific gender. This is formally defined as follows:

$$NFaiRR_q(D) = \sum_{i}^{t} w(d_i)(d_i^q)\frac{1}{log_2(1+i)} \tag{4}$$

where $w(d) =$

$$\begin{cases} 1 & if mag^f(d) + mag^m(d) < \tau \\ 1 - \sum_{g\in[m,f]} |\frac{mag^g(d)}{\sum_{x\in[m,f]} mag^x(d)} - \frac{1}{2}| & \text{otherwise} \end{cases} \tag{5}$$

where $mag^{f/m}(d)$ in NFaiRR is defined as $\sum_{w\in V_{f/m}} tf(w,d)$. Also, τ is a threshold on each document's gender magnitude used to determine each document's gender-neutrality. Furthermore, the NfaiRR metric becomes ranker agnostic through its normalization according to the best possible fairness result that can be achieved from reordering the retrieved documents in D [19].

It should be noted that the NFaiRR metric measures gender bias on the scale of 0 to 1. The fairer the retrieved results are, the closer the NfaiRR metric would be to 1. As NFaiRR gets closer to 0, the higher the gender magnitude of the list of documents would be. On the other hand, the ARaB metric measures gender polarity. Positive

[1] https://github.com/CPJKU/FairnessRetrievalResults/tree/main/resources

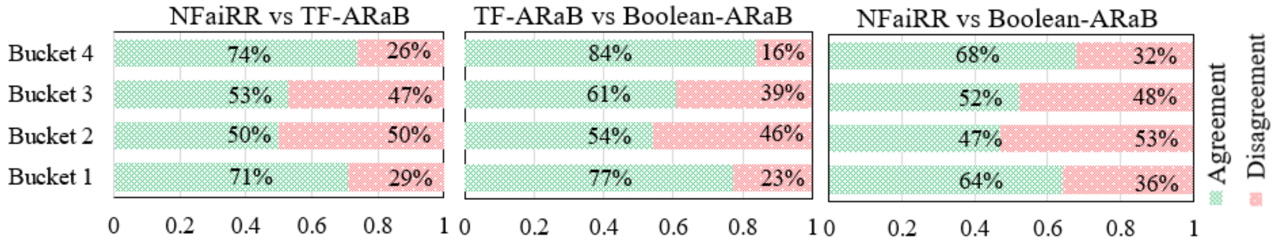

Figure 1: Consistency between queries with different degrees of bias. The green and red bars represent the agreement and disagreement between pairs of metrics respectively.

and negative ARaB values indicate male and female inclination, as in $mag^f(d)$ and $mag^m(d)$, respectively. The greater the absolute value of ARaB for a document is, the higher its likelihood of being bias would be.

Query and Document Collections. We ran our experiments on the well-known MS MARCO [2] passage collection datasets [16]. MS MARCO is a large-scale dataset which includes over 8 million passages and 500K queries for training and developing purposes. However, in order to examine gender biases in the retrieved results of different ranking systems, we additionally adopt a set of *gender-neutral queries* to measure the imposed biases into retrieval methods as released in [20]. This set is a human-annotated query set including gender-neutral queries from the MS MARCO dev set. It includes 1,765 queries that have been labelled as gender-neutral by crowd workers on the Amazon Mechanical Turk platform and has been used by different studies for measuring gender biases in retrieval systems [3, 19, 20]. We retrieve documents for these neutral queries using the BM25 retrieval method, which was also used as a baseline for the official MS MARCO reranking task. We report the bias among top-retrieved documents of BM25 [21, 26] for this query set only at cut-off 10.

2.2 Research Questions

Research Question 1 — Consistency. The purpose of the first research question is to investigate whether the gender bias measures show consistent behavior to each other. The reason this is important is because papers in the field use a different subset of gender bias metrics to show that their debiasing approaches have lead to the reduction of stereotypical gender biases. The expectation would be that if such methods have been able to systematically reduce gender biases, that such a reduction would be observed regardless of the gender bias metric that is used. Understanding the degree of consistency of the gender bias metrics allows the community to get more in-depth insight into the reported results in studies that use different metrics. A lack of consistency between the metrics could indicate that reporting a subset of the metrics would not be sufficient for getting a full picture of the impact on gender bias reduction.

In order to measure the level of consistency between pairs of metrics, we bin the 1,765 gender neutral queries into four equal-sized bins after sorting the queries based on their level of gender bias as determined by each gender bias metric separately. This produces three sets of four equal-sized bins of queries, each set belongs to one gender bias metric. Now, it would be expected the if the metrics showed a consistent behavior that the queries in each comparable bin across different metrics would consist of the same set of queries. In other words, if query q was determined to be a part of the first bin of queries (most gender biased queries) by TF-ARaB, it is expected that Boolean-ARaB and NfaiRR would also place q in their first bin. In order to measure such consistency across the three gender bias metrics, we measure the percentage of overlap between the bins generated by each pair of methods. Figure 1 illustrates the degree of overlap between each pair of metrics and as such visualizes the consistency between NFaiRR, TF-ARaB, Boolean-ARaB metrics. The higher the number of overlaps (green part of each chart) is, the higher the consistency between the pair of metrics would be. From this Figure, we can interpret that for the queries with the most biased results (bin 1) and least biased results (bin 4), we observe a relatively higher consistency between the metrics compared to the middle bins, i.e., bins 2 and 3. This is an important observation especially when considering the degree of overlap between NFaiRR and the two ARaB metrics on bins 2 and 3 for two reasons: (1) The degree of overlap in such cases is $\sim 50\%$ indicating that the metrics do not show a consistent measurement of gender bias on at least half of the queries in the query set, and (2) the inconsistency is higher for queries that are sitting at the middle of the ranked list of gender biased queries. This shows that while the metrics have higher degrees of consistency for highly biased and highly unbiased queries, they may fail to show similar determination of gender bias for less clear cut queries.

To dig deeper into the inconsistencies between the three metrics, we ranked the queries based on their degree of gender bias as measured by each of the metrics. We computed the absolute difference between the ranking of each query based on pairs of gender bias metrics as reported in Figure 2. It is expected that the gender bias ranks of queries do not differ substantially when ranked based on the different gender bias metrics, if the metrics were to be consistent. However, as illustrated in Figure 2, TF-ARaB and NFaiRR metrics show discernible disagreement on their developed rankings. The level of disagreement is lower when comparing Boolean-ARaB and NFaiRR compared to TF-ARaB and NFaiRR, as well as Boolean-ARaB and TF-ARaB, however, it is still notable.

As a result of RQ1, we find that gender bias metrics are not necessarily consistent with each other and hence the findings of gender debiasing methods that report subsets of these metrics may not be generalizable.

[2] https://microsoft.github.io/msmarco/

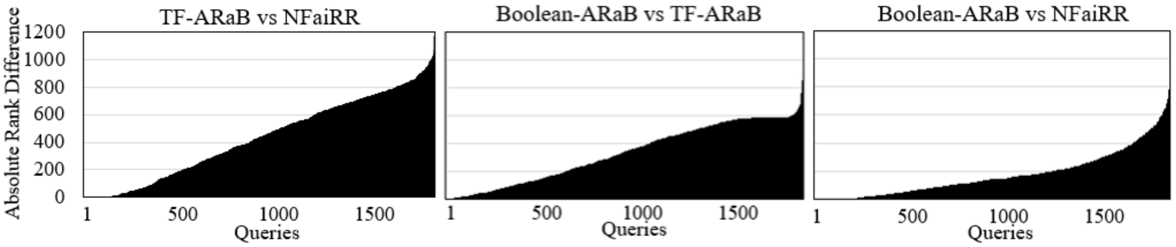

Figure 2: Consistency between the three metrics on a per-query gender bias ranking.

Research Question 2 — Sensitivity. Given all three gender bias metrics are dependent on a predefined list of gendered terms affiliated with each gender identity, the objective of the second research question is to explore how sensitive the gender bias measurements are to the set of terms included in the predefined lists. In order to study the sensitivity of gender bias metrics, we randomly subsample from each list of gender-related terms to observe how the behavior of each metric changes as the list of gendered terms change. We randomly subsample 25%, 50% and 75% of the terms from the list of 400 gender-related terms proposed by [19] and study how the bias measures change according to these perturbations in the source gendered term list.

Figure 3 demonstrates the TF-ARaB, Boolean-ARaB and NFaiRR gender bias metric measurements with the full gendered list as well as the subsampled lists. The box plots represent the mean (×), Median, first quartile and third quartile as well as the minimum and maximum bias values for each metric. As shown in this Figure, a slight change in the list of gendered terms can cause a noticeable difference between the measured biases. We observe that the range of all three bias metrics varies considerably when the number of terms in the original list changes. As the number of gender-related terms decreases, the range of gender bias metrics decreases as well. A smaller range of bias metric values indicates that the metric is not able to discern between the degree of gender bias of each query and is considering all queries to have similar degrees of gender bias; hence, the queries would be indistinguishable from each other from a gender bias perspective. In addition, when comparing the two ARaB metrics with NFaiRR, we observe that ARaB metrics show less sensitivity to the initial list of gendered terms compared to NFaiRR; however, they are more prone to reducing the range of bias metric measurements and hence potentially leading to ineffective measurements of bias over queries.

Research Question 3 — Psychological characteristics. In the third research question, we are particularly interested in understanding the psychological characteristics behind each gender bias metric. Understanding the psychological characteristics affiliated with the gender bias metrics would allow the community to go beyond the mere reporting of numerical quantifiers of gender bias and understand the underlying concepts that are driving the observed quantifiers of gender bias. To do so, we measure the Pearson ρ correlation between each gender bias metric and psychological attributes from the Linguistic Inquiry and Word Count (LIWC) toolkit [18] across the top-10 retrieved documents. LIWC is a well established toolkit that is capable of quantifying more than 70 different psychological characteristics of any textual content. Based on the correlation between metric values and LIWC psychological characteristics, we rank-order the psychological characteristics and report the top-5 in Table 1. It should be noted that all the psychological attributes mentioned in Table 1 shows statistically significant Pearson correlations with $\alpha = 0.05$.

Table 1: Top-5 most correlated LIWC attributes with each metric.

Rank	NFaiRR	TF-ARaB	Boolean-ARaB
1	Male Ref.	Male Ref.	Male Ref.
2	Past Focus	Female Ref.	Female Ref.
3	Present Focus	Past Focus	Past Focus
4	Tentative	Present Focus	Present Focus
5	Causation	Tentative	Tentative

Based on the ranking of psychological characteristics in Table 1, we make several observations: (1) when considering the three gender bias metrics (first three columns of the table), all three metrics are highly correlated with the male references characteristic from LIWC. In other words, the gender bias metrics are in principle inclined towards the male gender themselves. (2) In the NFaiRR metric, the female references characteristic is not even included in the top-5 characteristics, which again points to the metric being negatively biased against female references. (3) From a psychological characteristics point of view, it seems that the three metrics do capture similar semantic concepts related to past and present focus, and tentativeness. However, this also highlights the fact that these measures are making judgements about the degree of gender bias in retrieval systems based only on a limited number of psychological characteristics. One would expect that the treatment of the concept gender bias to receive a broader range of concepts from different angles, and (4) we find that the psychological characteristics of the gender bias metrics are dependent on the choice of gendered terms in the predefined list. Therefore, the psychological characteristics of the metrics can change depending on the adopted gendered terms and hence can make the interpretation of the findings based on these metrics difficult, if different sets of gendered terms are adopted in different studies.

In summary and based on the three research questions, we find: (1) the three bias metrics show a reasonably consistent behavior when identifying high biases or high unbiased queries, however, they are not consistent on others, (2) the gender bias metrics are sensitive to the choice of the gendered terms used for quantifying gender bias, and as such, changes to the list can result in unexpected changes in the measurements, and (3) the psychological characteristics of the gender bias metrics is essentially dependent on the

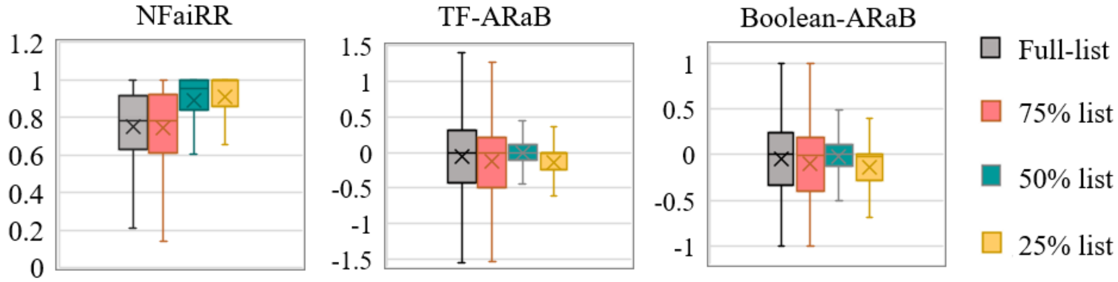

Figure 3: Sensitivity of the three gender bias metrics w.r.t to the source gendered lists. In each sub-plot y-axis indicates the gender bias value based on the mentioned metric.

list of gendered terms that are included in the predefined list. Furthermore, NFaiRR is by nature inclined towards the male affiliation psychological characteristic and hence maybe biased itself.

3 CONCLUDING REMARKS

We believe that understanding gender biases and being able to effectively measure and quantify them is an essential step towards addressing such biases; however, we suggest that attempts to quantify bias need to be done through a careful theory-driven approach that offers consistent metrics with well-understood semantics allowing for reproducible, repeatable and generalizable findings. Our experiments show that existing gender bias metrics may not necessarily consistent with each other, may be sensitive to small perturbations of their gendered list of terms, and might not have clear alignment with psychological characteristics beyond the choice of terms that are included in their gendered list. We suggest that future attempts at defining gender bias metrics should provide an opportunity to access repeatable and robust measurements of gender bias regardless of any predefined terms, and offer more in-depth description of the psychological characteristics of the gender biases. This would allow the community to contextualize findings within the relevant psychological and/or sociological literature and provide the means to understand the sources and reasons for the observed biases, which could then be used to systematically addressing gender bias at scale.

REFERENCES

[1] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. 2021. On the Orthogonality of Bias and Utility in Ad hoc Retrieval. In *SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1748–1752. https://doi.org/10.1145/3404835.3463110

[2] 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.

[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, ECIR 2021 (Lecture Notes in Computer Science)*, Vol. 12657. Springer, 216–224.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. (2016), 4349–4357.

[5] Shikha Bordia and Samuel R. Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. *CoRR* abs/1904.03035 (2019).

[6] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[7] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Inf. Process. Manag.* 57, 6 (2020), 102377.

[8] Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. *CoRR* abs/1901.03116 (2019). arXiv:1901.03116

[9] Emma J Gerritse, Faegheh Hasibi, and Arjen P de Vries. 2020. 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*. 133–136.

[10] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. (2019).

[11] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 417–432.

[12] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does Gender Matter? Towards Fairness in Dialogue Systems. *CoRR* abs/1910.10486 (2019). arXiv:1910.10486

[13] Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Mitigating Gender Bias for Neural Dialogue Generation with Adversarial Learning. *CoRR* abs/2009.13028 (2020). arXiv:2009.13028

[14] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*. Springer, 189–202.

[15] Kaiji Lu, Piotr Mardziel, Fangjing Wu, and Anupam Datta. 2020. Gender Bias in Neural Natural Language Processing. 12300 (2020), 189–202.

[16] 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, Vol. 1773. CEUR-WS.org.

[17] Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, Michael D Ekstrand, Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, et al. 2021. FACTS-IR: fairness, accountability, confidentiality, transparency, and safety in information retrieval. In *ACM SIGIR Forum*, Vol. 53. ACM New York, NY, USA, 20–43.

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

[19] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. (2021), 306–316. https://doi.org/10.1145/3404835.3462949

[20] 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.

[21] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. 500-225 (1994), 109–126.

[22] Shirin SeyedSalehi, Amin Bigdeli, Negar Arabzadeh, Bhaskar Mitra, Morteza Zihayat, and Ebrahim Bagheri. 2022. Bias-aware Fair Neural Ranking for Addressing Stereotypical Gender Biases. (2022).

[23] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*. ACM.

[24] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. *CoRR* abs/1906.00591 (2019). arXiv:1906.00591

[25] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019. Getting Gender Right in Neural Machine Translation. *CoRR* abs/1909.05088 (2019). arXiv:1909.05088

[26] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.