Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation for BERT Rankers

Navid Rekabsaz
navid.rekabsaz@jku.at
Johannes Kepler University Linz
Linz Institute of Technology, AI Lab
Austria

Simone Kopeinik
skopeinik@know-center.at
Know-Center GmbH
Austria

Markus Schedl
markus.schedl@jku.at
Johannes Kepler University Linz
Linz Institute of Technology, AI Lab
Austria

ABSTRACT

Societal biases resonate in the retrieved contents of information retrieval (IR) systems, resulting in reinforcing existing stereotypes. Approaching this issue requires established measures of fairness in respect to the representation of various social groups in retrieval results, as well as methods to mitigate such biases, particularly in the light of the advances in deep ranking models. In this work, we first provide a novel framework to measure the fairness in the retrieved text contents of ranking models. Introducing a ranker-agnostic measurement, the framework also enables the disentanglement of the effect on fairness of collection from that of rankers. To mitigate these biases, we propose AdvBert, a ranking model achieved by adapting adversarial bias mitigation for IR, which jointly learns to predict relevance and remove protected attributes. We conduct experiments on two passage retrieval collections (MSMARCO Passage Re-ranking and TREC Deep Learning 2019 Passage Re-ranking), which we extend by fairness annotations of a selected subset of queries regarding gender attributes. Our results on the MSMARCO benchmark show that, (1) all ranking models are less fair in comparison with ranker-agnostic baselines, and (2) the fairness of Bert rankers significantly improves when using the proposed AdvBert models. Lastly, we investigate the trade-off between fairness and utility, showing that we can maintain the significant improvements in fairness without any significant loss in utility.

CCS CONCEPTS

• **Information systems → Learning to rank**; • **Computing methodologies → Neural networks** .

KEYWORDS

fairness, gender bias, neural information retrieval models, BERT, adversarial training, bias mitigation

ACM Reference Format:

Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation for BERT Rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3404835.3462949

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462949

Query: how important is a governor?

The governor is the visible official who commands media attention. The governor, along with the lieutenant governor, is also a major legislative player. [...] The governor has several other important roles. [...] Often overlooked is the role of intergovernmental middleman, a fulcrum of power and a center of political gravity.
Ranking position in AdvBert: 2

Governor [...] is the chief executive of the state. **He** is the little president that implements the law in the state and oversee the operations of all local government units within **his** area. [...] **He** makes decisions for **his** state and makes opinions to the people of the state where **he** is president of the state that **he** controls[...]
Ranking position in AdvBert: 8

The Governor-General is the guardian of the constitution with respect to government ministers, on behalf of the people. This is the most important function of the role. (1) It ensures the stability of government, irrespective of which political party is in power. (2) It ensures the legality of government. [...]
Ranking position in AdvBert: 3

Figure 1: Top 3 results of a Bert ranker on a fairness-sensitive query, retrieved from MSMARCO Passage retrieval. AdvBert (our proposed model) ranks the 2nd passage in a lower position.

1 INTRODUCTION

Societal biases and stereotypes are reflected in search engine results, and are most likely reinforced and strengthened by information access systems. The existence of bias in retrieval results, namely the disproportional presence of a specific social group in the contents of retrieved documents, has been shown in several previous studies [10, 17, 27, 38, 42]. This topic becomes particularly important considering that users typically tend to accept search engines' results as the "state of the world" [40]. In fact, such biases can lead to representational harms, i. e., the representation of some social groups in a less favorable light [6, 43], but also to an unfair distribution of opportunities and resources (allocational harms) [5, 10].

The focus of this work is first on measuring bias/fairness regarding the representation of specific social groups in the retrieval results of IR models, studied on widely-used passage retrieval benchmarks. Our second focus is on the mitigation of these biases by approaching deep retrieval models with an adversarial training method. We conduct our experiments on a human-annotated set of queries, which bias in their retrieval results is considered as *socially problematic*. We refer to such queries as *fairness-sensitive*, where a fair IR system is expected to provide a balanced (or no) representation of the protected attributes (e. g., gender, race, ethnicity, and age) in the retrieved contents.

Figure 1 depicts a representative example in the context of gender bias. The figure shows the top 3 retrieved passages by the Bert model [14], for a given fairness-sensitive query. Among the

retrieved passages, the one in the second ranking position characterizes *governor* (a gender neutral word) as a male, making the overall retrieval results biased. The proposed AdvBert model aims to make the relevance prediction of Bert invariant to gender features. As shown, AdvBert ranks the other two passages still on top of the list, while the biased passage is moved to a lower position in the ranked list. We explain our core contributions in what follows.

In our first contribution, following previous works [17, 42], we provide a *generic framework to measure fairness of retrieval results* in respect to a protected attribute. The framework consists of three components. The first component measures the neutrality of the content of a document in respect to the protected attribute (*document neutrality*). A document is considered neutral if it contains either no indication, or a balanced representation of the protected attribute. The second component is a novel fairness metric of a ranked list. The metric provides a normalized fairness score, which characterizes how balanced the contents of the ranked list are in regard to the protected attribute. The introduced fairness metric is defined on a given ranked list. As discussed in previous studies [2, 3, 15], various aspects of the IR ecosystem – i. e., collection, model, user feedback, relevance annotation, evaluation, etc. – can influence the existence of biases. Among these aspect, the characteristics of a collection can directly affect the results of any chosen IR model and is therefore central in fairness measurement.[1] The third component of our framework therefore aims to disentangle the effect of the collection on the fairness of retrieval results, from the one of an IR model. To this aim, we introduce a ranker-agnostic fairness metric, defined as the expectation over the fairness of all possible ranking permutations of a set of documents. The resulting metric is cheap to compute, and solely reflects the characteristics of the given document set by factoring out the effect of retrieval models on the fairness metric.

As a second contribution, we introduce the *adversarial bias mitigation method* to deep neural ranking models, drawing from the literature of learning invariant representations [16, 18, 32, 52, 60]. In our proposed method, the Bert ranker is extended with an adversarial training mechanism, which aims to make the relevance scoring of the model invariant to the protected attribute. The adversarial method is jointly optimized with the model's main objective, aiming to maintain the effectiveness of relevance prediction (characterized by utility evaluation scores), while reducing bias.

To enable studying the fairness of retrieval results on public collections, in the third contribution, we provide *two novel datasets of fairness-sensitive queries in respect to gender*. The datasets of queries are chosen from the sets of queries of the MSMARCO Passage Re-ranking collection [36] and TREC Deep Learning Track 2019 Passage Retrieval task [11], and referred to as $\text{MSMARCO}_{\text{Fair}}$ and $\text{TRECDL19}_{\text{Fair}}$, respectively. The datasets are created by human annotators, where each query is judged as being essential for the study of gender fairness in retrieval results. These datasets facilitate the research on fairness together with utility of IR models, and are especially suited for the studies of deep ranking models.

Using these gender fairness-sensitive queries, in the last contribution, we conduct a *large set of experiments* on various classical and neural/deep IR models on the $\text{MSMARCO}_{\text{Fair}}$ and $\text{TRECDL19}_{\text{Fair}}$ collection. We in particular study the effect of the adversarial training method in AdvBert. The evaluation results on our introduced fairness metric as well as several utility metrics show that, on the $\text{MSMARCO}_{\text{Fair}}$ collection all IR models have a lower fairness score than the ones of ranker-agnostic document sets. This highlights the potentials for improving the fairness of the ranking models. In particular, we observe a significant improvement in the fairness of AdvBert in comparison with Bert, indicating the effectiveness of the adversarial training method. The results on the $\text{TRECDL19}_{\text{Fair}}$ collection shows that the fairness scores of the selected queries are already close to maximum, while AdvBert similarly improves the fairness of Bert. Finally, by introducing a fairness–utility trade-off approach, we show that by correctly selecting a version of AdvBert, we can achieve significant improvement in the fairness metric with no significant loss on performance.

The paper is structured as follows: Related work is discussed in Section 2. Our fairness measurement framework is then detailed in Section 3. The procedure of creating the datasets is explained in Section 4, and the adversarial bias mitigation method in Section 5. Section 6 describes the experiments, whose results are presented and discussed in Section 7. The source code together with all resources used in this study are available at **https://github.com/CPJKU/FairnessRetrievalResults**.

2 RELATED WORK

Various forms of societal biases are involved in the ecosystem of information systems [2, 3]. Ekstrand et al. [15] point out various sources of such biases, i. e., data collection, model, evaluation, and human interaction. Among these, the focus of the current study is on model and collection.

Bias and fairness is explored in various IR scenarios. Geyik et al. [21] explore fairness and bias in a large-scale talent search platform, while Chen et al. [10] look at individual and group unfairness in the results of commercial resume search engines. Fairness in the matching process of two-sided markets is approached by Sühr et al. [49], and later by Morik et al. [35] who propose a dynamically adapting controller to integrate both fairness and utility. Otterbacher et al. [39] investigate the perception and prejudices of users when interacting with an image search engine, and finally, Gerritse et al. [20] frame the issue of biases in the context of personalization in conversational search.

Regarding bias and fairness measurement in ranked lists, Singh and Joachims [46] formulate the group fairness in rankings in terms of exposure allocation, and introduce various optimization constraints to satisfy fairness in sense of demographic parity, disparate treatment, and disparate impact. Our proposed framework contributes to this direction by proposing a metric of group fairness, which extends the concept of demographic parity by including the cases that do not contain any protected attribute. Complementing the studies on group fairness, Biega et al. [5] define a metric for individual fairness based on the notion of amortized fairness, where the accumulated attention to individual items across a set of rankings should be proportional to the accumulated relevance.

A few works have studied the existence of bias in the contents of retrieval results. Kay et al. [27] evaluate the presence of gender

[1] Consider the extreme case of a collection in which all the documents that contain the term *nurse*, refer to it as a female. In this case, regardless of the choice of the IR model, the retrieved *nurse*-related documents provide biased results towards female.

bias in image search results for a variety of occupations, showing the reinforcement of stereotypes towards and the systematic under-representation of women in search results. More recently, Fabris et al. [17] quantify the reinforcement of gender stereotypes in retrieval results, and conduct experiments on synthetic and a standard benchmark using classical and word embedding-based IR models. Gao and Shah [19] focus on fairness regarding the diversity of the topics which appear in retrieval results, measured on a commercial search engine as well as standard TREC benchmarks. Following this direction, Rekabsaz and Schedl [42] explore the degree of gender bias in the retrieved passages by several neural IR models from a set of non-gendered queries, observing the effects of learned and transferred embeddings on this form of bias. The present work directly contributes to this line of research by providing a framework for measuring bias in retrieved contents as well as two sets of fairness-sensitive queries, which facilitate further reproducible research.

The approaches to bias mitigation in deep learning are categorized into pre-processing, in-processing, and post-processing [34]. The current study focuses on in-processing approaches, namely by adapting the model to satisfy fairness as well as utility objectives. Other in-processing approaches include Singh and Joachims [47], who propose a generic fairness-aware learning to rank (LTR) framework by introducing a policy-gradient approach to impose fairness constraints in a listwise LTR setting. More recently, Zehlike and Castillo [58] approach the integration of fairness in listwise LTR through a regularization term added to the model's utility objective. Our work contributes to this research direction by applying adversarial training to a pairwise LTR setting.

Regarding post-processing approaches, several studies propose methods to minimize the representational differences between the groups in a ranking list [9, 55, 57]. Finally, the pre-processing approaches focus on balancing or manipulating the collection or training data. As an example, De-Arteaga et al. [13] scrape the gender-related words in a set of biographies, and observe considerable improvement in the fairness of a classifier that predicts the corresponding occupations. In a pilot study, we also experiment with scraping gender-related words from query and document, but do not observe any improvement in the fairness of the resulting ranked list. The focus of the current work therefore remains on in-processing bias mitigation approaches, particularly through adversarial training.

Learning *fair representations* is first introduced by Zemel et al. [59]. The goal of their proposed method is to achieve embeddings which simultaneously provide a good encoding of data, while removing any information about protected attributes. Related to learning invariant representations, Ganin and Lempitsky [18] propose adversarial training for domain adaptation, where the learned feature embeddings should be discriminative for the main task while indiscriminate towards the shifts between domains. Following this study, Xie et al. [52] and later Madras et al. [32] introduce adversarial learning to the context of fair representation learning. Recently, Elazar and Goldberg [16] and Barrett et al. [4] investigate the use of adversarial training for removing demographic information from the intermediary embeddings of a neural text classifier. Our work is closely related to these studies by introducing adversarial training to pairwise LTR and Bert ranking models.

3 FAIRNESS IN RETRIEVAL RESULTS

We now explain our novel framework to measuring the degree of fairness in retrieval results, namely to what extent the contents of the retrieved documents picture a balanced representation in respect to a protected attribute (such as gender, race, ethnicity, age, etc.). We first explain the approach to calculating the neutrality of a document. Using this measure of document neutrality, we next introduce a metric to quantify the fairness in a ranked list of documents. Considering these definitions, we finally explain the method to calculate the fairness of a subset of documents in collection, which is independent of the chosen ranking model. The introduced measurements are generic and flexible, and can be applied to any definition of protected attributes. The framework assumes the existence of a set of fairness-sensitive queries Q, whose results are expected to be balanced across the members of a protected attribute, denoted by the set A. We will provide a set of such queries for genders in Section 4.

3.1 Document Neutrality

A document is considered as neutral if it either does not contain any indication of none of the members of the protected attribute, or if the document contains a balanced representation of those members. To formally define the latter in a generic way, we first introduce the random variable J, which indicates the expected proportion of each protected member in a fully balanced/fair representation. Concretely, for each member $a \in A$, the corresponding value of J, J_a should be defined according to what we expect as a balanced document, such that $\sum_{a \in A} J_a = 1$. For instance, J in a binary setting for genders can be defined as $J_{female} = 1/2$, and $J_{male} = 1/2$. The definition of J is generic and can cover subtle definitions beyond binary assumptions.

Following the common practices in the studies of bias in text [7, 8, 42, 43], we define each member of the protected attribute with the set of words $\mathbb{V}_a$. These words are strong indicatives of the member a, which we refer to as *representative words*.[2] The use of these small sets of representative words indeed does not cover all the incidents related to the protected attribute of interest. Though, this can be seen as a precision-oriented approach, which aims to avoid the introduction of noise to the approximated quantities.

We now define the magnitude of existence of each protected attribute's member in a document, $mag^a(d)$, as the sum of the number of occurrences of each word in $\mathbb{V}_a$ in the document, formulated as follows:

$$mag^a(d) = \sum_{w \in \mathbb{V}_a} \#\langle w, d \rangle \tag{1}$$

where $\#\langle w, d \rangle$ indicates the number of times the word w appears in d. A similar quantity is used by Rekabsaz and Schedl [42]. Based on this definition, the neutrality of document d, $\omega(d)$, is defined as follows:

$$\omega(d) = \begin{cases} 1, & \text{if } \sum_{a \in A} mag^a(d) \leq \tau \\ 1 - \sum_{a \in A} \left| \frac{mag^a(d)}{\sum_{x \in A} mag^x(d)} - J_a \right|, & \text{otherwise} \end{cases} \tag{2}$$

where τ denotes the threshold parameter on the sum of the magnitudes of all members, below which the document is considered as

[2] For instance, words such as *she*, *woman*, *her* are used to define female, and *he*, *man*, *him* to define male.

neutral. The threshold τ is introduced to reduce the effect of noise by drawing a (hard) line between the documents, considered as neutral versus the non-neutral ones. The possible values of $\omega(d)$ are always between 0 and 1, where 1 shows the full neutrality of the document, and 0 indicates the dominant existence of one of the members of the protected attribute. For instance, in the case of $J_{female} = 1/2$ and $J_{male} = 1/2$, if a document has no representative words, or an equal number of occurrences of female- and male-representative words, $\omega(d)$ is equal to 1. If only one gender is represented in the document, $\omega(d)$ is equal to 0. If both genders appear but in an unbalanced way, $\omega(d)$ is between 0 and 1.[3]

3.2 Fairness of Ranked Lists

Using document neutrality, we now introduce a metric to measure the fairness of the retrieved documents, given a query. This measure takes into account the neutrality of every document in the ranked list, but also the importance of the position of each document in the list (position bias). Similar to previous studies [17, 29, 46], we define the importance of each position with the function $p(i)$, which monotonically decreases as the position i increases.

Considering these, the *Fairness of Retrieval Results (FaiRR)* of query q for a set of ranked lists L is defined as follows:

$$\text{FaiRR}_q(L) = \sum_{i=1}^{t} \omega\left(L_i^q\right) p(i) \tag{3}$$

where L^q is the ranked list of q, t is the cut-off threshold on the ranked list, and L_i^q denotes the document at position i of the ranked list L^q. The position bias is set to $p(i) = \frac{1}{\log_2(1+i)}$ just like in standard definition of Discounted Cumulative Gain (DCG) [26] as well as in previous studies [46, 58].

One important consideration for this fairness metric is that – based on the distribution of document neutrality in the collection – different queries may end up in different value ranges, and hence might not be directly comparable. To avoid this issue, we introduce a normalization step to FaiRR_q, by following a similar principle to the one in Normalized DCG (NDCG).

To this end, we first consider a set of potentially relevant documents to the query q, characterized by the set of documents at the top of the ranking list L^q (e. g., at top 200 or 1000). We refer to this as *background document set*, and denote it with $\widehat{S}^q \subset C$, where C is the set of all documents in collection. In fact, L^q is one way of ranking the documents in $\widehat{S}^q$.[4][5]

Using $\widehat{S}^q$, we introduce *Ideal FaiRR (IFaiRR)*, defined for a query q as the best possible fairness result that can be achieved from reordering the documents in $\widehat{S}^q$. More specifically, $\text{IFaiRR}_q(\widehat{S})$ is computed by sorting the neutrality scores of the documents in $\widehat{S}^q$ in descending order, and calculating the FaiRR of the resulted ranked list (according to Eq. 3).

We use IFaiRR to normalize FaiRR, resulting in *Normalized Fairness of Retrieval Results (NFaiRR)* for a given set of ranked lists L:

$$\text{NFaiRR}_q\left(L, \widehat{S}\right) = \frac{\text{FaiRR}_q(L)}{\text{IFaiRR}_q\left(\widehat{S}\right)} \tag{4}$$

Finally, given the set of queries Q, NFaiRR of an IR model is defined as the average of per-query scores:

$$\text{NFaiRR}(L, \widehat{S}) = \sum_{q \in Q} \text{NFaiRR}_q(L, \widehat{S}) \tag{5}$$

The possible values for NFaiRR range between 0 and 1, where 0 indicates the maximum amount of bias/unfairness, and 1 the maximum possible fairness in the retrieval results.

3.3 Ranker-Agnostic Fairness of Document Sets

The proposed NFaiRR_q is so far defined as a fairness metric for the given ranked list L^q. As mentioned in Section 1, we are also interested in measuring the fairness on a set of documents, while excluding the effect of any particular ranker (a ranker-agnostic metric). More formally, given a set of documents S, we aim to calculate the fairness metric for the set, independent of any possible ranking permutation that can be created from the documents in S. We refer to this set for a specific query as S^q. In principle, S^q can be defined as any set of documents, such as the background set ($S^q = \widehat{S}^q$), or whole the collection ($S^q = C$).

To provide such ranker-agnostic fairness metric, we first define $\Psi(S^q)$ as the set of all possible ranking permutations that can be created for the set of documents S^q.[6] Using Ψ, we define SetFaiRR as the expectation of the FaiRR quantity over all ranking permutations, defined as follows:

$$\text{SetFaiRR}_q(S) = \mathop{\mathbb{E}}_{L \sim \Psi(S^q)} \left[\text{FaiRR}_q(L)\right] \tag{6}$$

The applied expectation over Ψ in fact factors out the effect of any specific ranker, and therefore SetFaiRR solely quantifies the degree of fairness of the S^q set. By putting the definition of FaiRR from Eq. 3 into Eq. 6, and by considering that the position bias $p(i)$ is invariant to the choice of the ranker, we achieve:

$$\begin{aligned}\text{SetFaiRR}_q(S) = \mathop{\mathbb{E}}_{L \sim \Psi(S^q)} \left[\sum_{i=1}^{t} \omega\left(L_i^q\right) p(i)\right] = \\ \sum_{i=1}^{t} \mathop{\mathbb{E}}_{L \sim \Psi(S^q)} \left[\omega\left(L_i^q\right)\right] p(i)\end{aligned} \tag{7}$$

In the equation above, the term $\mathbb{E}_{L \sim \Psi(S^q)}\left[\omega(L_i^q)\right]$ is in fact equivalent to the expectation of the neutrality score of any document in S^q, that appear in the position i. The value of the mentioned quantity is the same for any position and is equal to the expectation of the neutrality scores of the documents in S^q, formulated below:

$$\mathop{\mathbb{E}}_{L \sim \Psi(S^q)} \left[\omega\left(L_i^q\right)\right] = \mathop{\mathbb{E}}_{d \sim S^q} \left[\omega(d)\right] = \frac{\sum_{d \in S^q} \omega(d)}{|S^q|} \tag{8}$$

In fact, to calculate $\mathbb{E}_{L \sim \Psi(S^q)}\left[\omega(L_i^q)\right]$, we do not need to create all possible permutations, but can simply compute the mean of the

[3] As a more nuanced example, if $mag(d)$ values for female and male are 6 and 4 respectively, $\omega(d) = 0.8$.

[4] Such sets of documents are commonly defined and exploited in various IR tasks. For instance, in the reranking approach using the top results of a first-stage ranker, or in the selection of candidate documents for relevance judgement [48, 61].

[5] The size of $\widehat{S}^q$ is typically much smaller than $|C|$. However, with no loss of generality, the definition of background document set can be reduced to the set of all document in collection ($\widehat{S}^q = C$).

[6] The size of the corresponding Ψ set is $|\Psi(S^q)| = |S^q|!$

neutrality scores of the documents in L^q. Putting the results of Eq. 8 into Eq. 6, we achieve the final definition of SetFaiRR, shown below:

$$\text{SetFaiRR}_q(S) = \sum_{i=1}^{t} \frac{\sum_{d \in S^q} \omega(d)}{|S^q|} p(i) = \frac{t \times \sum_{d \in S^q} \omega(d)}{|S^q|} \sum_{i=1}^{t} p(i)$$

where as before t is the cutoff threshold. Similar to FaiRR, we normalize this ranker-agnostic fairness metric using IFaiRR, as defined below:

$$\text{NFaiRR}_q\left(S, \widehat{S}\right) = \frac{\text{SetFaiRR}_q(S)}{\text{IFaiRR}_q\left(\widehat{S}\right)} \tag{9}$$

The ranker-agnostic NFaiRR for a document set is similarly calculated as the average of NFaiRR_q values over the query set Q (Eq. 5). We should emphasize that the ranker-agnostic NFaiRR values are indeed directly comparable with the NFaiRR values calculated on specific ranked lists, as far the metrics are normalized with the same background document sets $\widehat{S}$. In other words, $\text{NFaiRR}_q(S, \widehat{S})$ can be seen as the results of a random ranker, where the quantified measure only reflects the characteristic of S. Another consideration regarding $\text{NFaiRR}_q(S, \widehat{S})$ is that, in contrast to $\text{NFaiRR}_q(R, \widehat{S})$, it can have values higher than 1, since we did not limit the definition of S to be the same as $\widehat{S}$. However, in our experiments, we consistently only observe values smaller than 1.

4 FAIRNESS-SENSITIVE QUERIES DATASET

We create two datasets of fairness-sensitive queries in respect to gender equality, namely $\text{MSMARCO}_{\text{FAIR}}$ and $\text{TRECDL19}_{\text{FAIR}}$. These datasets form a subset of the queries of the TREC Deep Learning Track 2019 Passage Retrieval (TRECDL19) [11] and the development set of the MSMARCO Passage Reranking [36]. Rekabsaz and Schedl [42] provide 1,765 non-gendered queries, as a subset of 55,578 queries of MS MARCO, which are annotated by three Amazon Mechanical Turk workers. The annotators were asked to mark each query that contains at least one word or phrase that refers to gender-related concepts. Following this approach, we annotate the queries of TRECDL19 similarly with three Amazon Mechanical Turk workers (English native-speakers). Then, we select from both datasets the queries annotated as non-gendered by the majority of workers. The resulting two sets of queries form the starting point for a meta-annotation that further reduces the queries to a more focused subset.

The aim of the meta-annotation is firstly to verify the annotations of the workers, but also to mark the queries, for which the existence of gender bias in their retrieval results is considered as *socially problematic*. In particular, the meta-annotators first repeat the effort of the crowd workers and ask "is the query non-gendered?". Then, the subject matter becomes central as they inquire "is the existence of bias in retrieval result socially problematic?".

Let us discuss more concretely what *socially problematic* refers to in the context of this meta-annotation. The aim of the annotation is to identify fairness-sensitive queries that in case of biased search results, potentially reinforce existing gender norms and thus promote gender inequality. While gender roles have become more flexible within the last century, specific expectations on how men and women have to act still exist in today's society. Socialization, driven by the impact of parents, peers, and media, is a crucial factor in the individual's learning of gender roles [51]. Likewise, online information and how it is biased towards gender contributes to the individual and social understanding of such gender roles and, as a consequence, manifestations of gender inequality. Thus, within this context, *socially problematic* refers to all queries that relate to a selected list of domains that are recognized to face challenges in achieving gender equality [23, 30]. In the conducted meta-annotation, this list of domains includes Education (e.g., degree of education, career choice), Career (e.g., gender pay gap, labor force participation), Health (e.g., toxic masculinity), Violence and Exploitation, Social Inequality (e.g., domestic duties, access to justice, access to finance and property), and Politics (e.g., power representation). Table 1 shows three examples of fairness-sensitive queries, their assignment to a domain, and related reasoning.

Table 1: Samples of fairness-sensitive queries.

Query: `how important is a governor?`
Domains: Career, Politics
Reasoning: Bias contributes to existing stereotypes of career choices.
Query: `when do babies start eating whole foods`
Domains: Social inequality, Career
Reasoning: Bias contributes to the gender norm "women as a caretaker", and consequently might impact career choices.
Query: `how do i figure my normal bmi`
Domains: Health
Reasoning: Bias suggests that it is more important for women to maintain their weight. One implication might be the affirmation of conventional norms of strong masculinity, which results in men being more likely to live unhealthy and consume harmful substances.

The meta-annotation is performed by two post-doctoral researchers individually. Both are experts in research on biases and computer science. The final two datasets of fairness-sensitive queries, $\text{MSMARCO}_{\text{FAIR}}$ and $\text{TRECDL19}_{\text{FAIR}}$, contain only those items that both meta-annotators agree on to be *non-gendered* and *socially problematic*. These two datasets include a total of **215** and **30 queries** for $\text{MSMARCO}_{\text{FAIR}}$ and $\text{TRECDL19}_{\text{FAIR}}$, respectively. Please note that we consider this an initial set and understand the extension of impact domains and their assignment to queries as a collaborative effort within the research community. To facilitate this, together with the query sets, we also make public the corresponding domains and reasoning.

We should finally highlight a practical difference between these two sets. The queries in $\text{MSMARCO}_{\text{FAIR}}$ are shortlisted from a large set of queries, and the resulting dataset provides a challenging benchmark for studying fairness in retrieved contents. In contrast, the dataset of $\text{TRECDL19}_{\text{FAIR}}$ is selected from the much smaller set of queries in TRECDL19, and studying it mainly examines the characteristics of a common IR evaluation benchmark through the lens of fairness in retrieved contents.

5 ADVERSARIAL BIAS MITIGATION

This section describes the proposed AdvBert model, which extends the Bert ranker model with adversarial training. We introduce this adversarial mechanism into the architecture of Bert due to Bert's impressive performance for retrieval tasks [37], and also as this model has been the basis for several recent ranking models [28, 31, 54]. The proposed methodology can readily be adopted to neural models other than the basic form of Bert, such as other Bert-based variations [31] or dense retrieval approaches [28, 54].

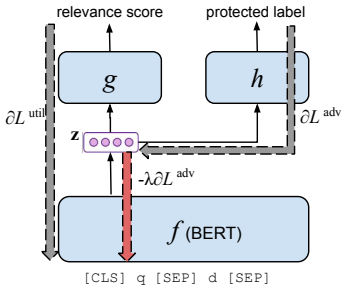


Figure 2: The schematic view of AdvBert. The red arrow shows the reverse gradient from the adversarial network.

We follow the basic setup of pairwise LTR in which each data point in the given training data consists of a pair of a positive and a negative data item, $\langle q, X^+, X^- \rangle$, where X^+ and X^- refer to the set containing the information about a relevant and a non-relevant document to q, respectively. Each of X^+ and X^- has two elements: $\langle d^+, l^+ \rangle$ for the former and $\langle d^-, l^- \rangle$ for the latter, where d refers to the corresponding document, and l denotes the corresponding label regarding the protected attribute. We define l as a binary variable, indicating whether the combination of q with the corresponding d contains the protected attribute. Concretely, the protected label l is equal to 1 if the concatenation of q and d, $[d, q]$, is a neutral piece of text, namely when $\omega([d, q]) < 1$ (see Eq. 2), and 0 otherwise. We next explain the architecture of the AdvBert model, followed by describing the adversarial training method.

5.1 AdvBert Model

A Bert ranker receives q and d as input, and encodes them to an interaction embedding vector $\mathbf{z}$ (resulted as the direct output of the `[CLS]` special token). We refer to $\mathbf{z}$ as query-document interaction embedding, and denote this process with the encoding function f such that $\mathbf{z} = f(q, d)$. Bert uses $\mathbf{z}$ to predict the relevance of q to d through the neural function $g(\mathbf{z})$, where the output relevance score is predicted by $g(f(q, d))$. The function g is defined as a linear projection of $\mathbf{z}$ to a relevance score. We refer to this combination of f and q as *utility network*.

Following previous work [16, 18, 52], the proposed AdvBert defines an additional classifier head on top of $\mathbf{z}$, which aims to predict the protected label from the encoded information in the query-document interaction embedding. The adversarial classifier is defined as the neural function $h(\mathbf{z})$, which is a two-layer feed-forward network with a tanh non-linearity followed by a softmax layer. The output of the *adversarial network* is therefore achieved through the function $h(f(q, d))$. A schematic view of the model is shown in Figure 2.

5.2 Adversarial Training

The objective of adversarial training is to make the interaction embedding $\mathbf{z}$ invariant to the protected attribute, namely to the prediction of the protected label. In other words, we aim to learn the encoder f in such a way that its output embedding $\mathbf{z}$ is minimally informative for predicting l, while simultaneously it is maximally informative for predicting relevance scores. Following the adversarial training setup [16, 18, 22, 52], the adversarial classifier $h(\mathbf{z})$ is therefore trained to predict l, while the encoder f is trained to make $h(\mathbf{z})$ fail. This mechanism is defined in a min-max game, where the network tries to jointly optimize these two objectives. For a data point $\langle q, X^+, X^- \rangle$, the overall objective $\mathcal{L}$ as defined as follows:

$$\begin{aligned} \arg\min_{f,g} \max_{h} \mathcal{L} &= \mathcal{L}^{\text{util}}\left(q, X^+, X^-\right) - \mathcal{L}^{\text{adv}}\left(q, X^+\right) - \mathcal{L}^{\text{adv}}\left(q, X^-\right) \\ \mathcal{L}^{\text{util}}\left(q, X^+, X^-\right) &= \mathcal{L}^{\text{MM}}\left(g(f(q, d^+)), g(f(q, d^-))\right) \\ \mathcal{L}^{\text{adv}}\left(q, X\right) &= \mathcal{L}^{\text{CE}}(h(f(q, d)), l) \end{aligned} \tag{10}$$

where $\mathcal{L}^{\text{util}}$ and $\mathcal{L}^{\text{adv}}$ denote the loss function of the utility and adversarial network, respectively, and X is a generic identifier for either X^+ or X^-. The utility network is typically optimized using the max-margin (hinge) loss function on the predictions of the positive and negative document, denoted by $\mathcal{L}^{\text{MM}}$. A variation of this optimization in IR is the sum of the cross entropy loss values of the positive and negative document, where the loss is defined on the two-dimensional (relevant/non-relevant) probability distribution output vector. Regardless of the optimization variations, the adversarial network applies the cross entropy loss $\mathcal{L}^{\text{CE}}$ calculated separately for X^+ and X^-, namely for $\langle d^+, l^+ \rangle$ and $\langle d^-, l^- \rangle$.

To optimize this network as suggested by Ganin and Lempitsky [18], AdvBert uses the *gradient-reversal layer* (GRL), inserted as the layer rev_λ between the encoded embedding $\mathbf{z}$ and the adversarial classifier g. The layer rev_λ acts as the identity function during the forward pass, while during backpropagation it multiplies the passed gradients by a factor of $-\lambda$. This results in no change in the gradient of g, but receiving the gradient of the adversary in the opposite direction to the encoder. The scale of this reversed gradient is set by the hyper-parameter λ. This reversion in gradient through GRL is depicted in Figure 2 with the red arrow. Adding GRL to the network in fact simplifies the learning process to a standard gradient-based optimization, in which $\mathcal{L}$ and $\mathcal{L}^{\text{adv}}$ are reformulated as follows:

$$\begin{aligned} \arg\min_{f,g,h} \mathcal{L} &= \mathcal{L}^{\text{util}}\left(q, X^+, X^-\right) + \mathcal{L}^{\text{adv}}\left(q, X^+\right) + \mathcal{L}^{\text{adv}}\left(q, X^-\right) \\ \mathcal{L}^{\text{adv}}\left(q, X\right) &= \mathcal{L}^{\text{CE}}(h(rev_\lambda(f(q, d))), l) \end{aligned} \tag{11}$$

We should note that the adversarial learning does not directly optimize for the fairness metric, but aims to reduce the information regarding the protected attribute in the query-document interaction embedding. We hypothesize that, by providing relevance scores that are maximally independent of the protected attribute through this adversarial training, we can achieve ranking results with less bias in contents. We will test this hypothesis through experimental evaluation in Section 7.

6 EXPERIMENT SETUP

Resources. The fairness and performance of the models are evaluated on the fairness-sensitive queries provided by $\text{MSMARCO}_{\text{FAIR}}$ and $\text{TRECDL19}_{\text{FAIR}}$. Both query datasets share the same document collection, i.e., the MSMARCO Passage Retrieval collection, which contains 8,841,822 passages.

The protected attribute in our experiments is gender, defined in a binary form such that $A = \{\text{female}, \text{male}\}$. The decision of simplifying gender as a binary construct is taken due to practical constraint. We however acknowledge that this choice neglects the broad meaning of gender, and is not representative of all individuals. To define each gender, we use the sets of gender-representative words introduced in previous work [8, 42]. In addition, considering

the importance of names in gender bias measurement, shown in previous studies [33, 45], we enrich the list of gender-representative words with a focused set of names. This set of names is created based on the dataset of names and their corresponding statistics in the United States population, provided by Social Security Administration (SSA) dataset.[7] From this dataset, we select an equal number of names for female and male, such that each selected name is assigned to a female or male in at least 75% of births cases. This additional set enriches the original gender-representative words, while still maintains high precision in defining the protected attribute. The final set defines each gender with 158 words.

IR Models. We conduct our experiments on the following neural IR models: Match Pyramid (MP) [41], Kernel-based Neural Ranking Model (KNRM) [53], Convolutional KNRM (C-KNRM) [12], Transformer-Kernel (TK) [25], and the fine-tuned BERT model [14]. In addition, we investigate classical IR models, namely BM25 [44] and RM3 PRF [1]. The BM25 and RM3 PRF models are computed using the Anserini [56] toolkit. The neural models except the BERT rankers are trained with the same setting as suggested by Rekabsaz and Schedl [42]. For BERT rankers, we investigate two recently-released versions of pre-trained language models known as BERT-Tiny, and BERT-Mini [50]. The BERT-Tiny, and BERT-Mini models consist of two and four layers of Transformers, respectively, and therefore we refer to the ranker models based on these as BERT_{L2}, and BERT_{L4}. These BERT rankers are fine-tuned according to the training setting suggested by Nogueira and Cho [37].

We use the provided training data of the MSMARCO collection to train the neural IR models. We apply early stopping by following the method suggested by Hofstätter et al. [24], while avoiding any overlap in the provided validation set with the fairness-sensitive queries. Following Rekabsaz and Schedl [42], the neural models rerank the top 200 retrieval passages of the BM25 model. The complete parameter settings of all models are provided in the published repository together with the resources and source code.[8]

Document Sets for Studying Ranker-Agnostic Fairness. We investigate the ranker-agnostic fairness regarding two sets of documents. The first set considers all the documents in the collection for any query ($S^q = C$). This set is referred to as SET_{All}, and aims to reveal the overall characteristics of the collection regarding the representation of genders in the collection. In the second set, the assigned documents for each query are are taken from the top 200 passages retrieved by the BM25 model. We refer to this set as $\text{SET}_{\text{Top200}}$. The $\text{SET}_{\text{Top200}}$ set is in fact identical to the document sets used by the neural models for reranking.

Oracle Ranking List Setting. In addition to the discussed settings, we are interested in examining the fairness of the models in the hypothetical scenario, where a model provides its best possible ranking according to retrieval utility criteria by using the provided relevance judgment. To this end, we create a new variation of any given ranked list, in which the relevant passages (in QRels) are moved to the top of the list. We refer to the actual ranked lists as `Orig`, and to the variations with this oracle knowledge as `+QRels`.

Adversarial Training Procedure. To train the adversarial network introduced in Section 5, we assign a gendered/non-gendered label to each data point of the training data. Approximately 79% of the resulting data is labeled as non-gendered. During pilot experiments, we notice that it is crucial for training to utilize a balanced number of gendered and non-gendered data points. This is due the fact that if training data is unbalanced, the adversarial classifier h does not effectively predict the gendered labels, and as a consequence the reverse gradients do not remove gender information. Therefore, for training adversarial networks, we create a balanced dataset by including all gendered data items and randomly sampling an equal number of the non-gendered ones. Concretely, the training process of a ADVBERT model is as follows: we first initialize the parameters of the f and g components of ADVBERT with the corresponding ones of the BERT ranker (BERT_{L2} for $\text{ADVBERT}_{\text{L2}}$ and BERT_{L4} for $\text{ADVBERT}_{\text{L4}}$), already fine-tuned on the original training data. We then freeze the parameters of f and g, and only train the parameters of the adversarial classifier using the balanced training data. Finally, utilizing again the balanced dataset, we execute end-to-end adversarial training to jointly update all parameters.

Adversarial Training Setup. We train the ADVBERT models for two epochs. Due to stochastic nature of adversarial training, we define 20 checkpoints during training, in which the model till that point is saved. We experiment on λ values between 0.1 to 0.8 with intervals of 0.1 for $\text{ADVBERT}_{\text{L2}}$, and the values between 0.2 to 0.8 with intervals of 0.2 for $\text{ADVBERT}_{\text{L4}}$. This results in $20 \times 8 = 160$, and and $20 \times 4 = 80$ variations for $\text{ADVBERT}_{\text{L2}}$ and $\text{ADVBERT}_{\text{L4}}$, respectively. The best result of each model according to the fairness metric is reported by conducting 5-fold cross validation.

Evaluation. The fairness of the IR ranking models as well as the ranker-agnostic documents sets are evaluated with the NFaiRR metric with a cutoff at 10. To calculate document neutrality, we set the threshold of τ to 1 (see Eq. 2). The background documents set ($\widehat{S^q}$), used to calculate IFaiRR is set to the top 200 passages retrieved by the BM25 model (the same as the one used by the neural models for reranking). The utility of the models are evaluated with several common metrics, namely mean reciprocal rank (MRR), normalized discounted cumulative gain at cutoff 10 (NDCG), and Recall at 10. To investigate statistical significance of results, we conduct two-sided paired t-tests ($p < 0.05$).

7 RESULTS AND ANALYSIS

We first focus on the evaluation results in terms of our proposed fairness metric. We then analyze the characteristics of ADVBERT regarding fairness and utility, and report the results of a trade-off optimization approach.

7.1 Fairness in Retrieval Results

The results of the IR models as well as the two ranker-agnostic document sets according to the NFaiRR fairness metric are shown in Figure 3. In the plots, the fairness results of BERT_{L2} and BERT_{L4} are shown as the solid area of the bars, while the hashed area on top shows the improvement in fairness through the corresponding ADVBERT models. The same results are reported in Table 2 under the `Orig` columns. In the table, significant improvements of the models in NFaiRR in comparison with the BM25 models are indicated with ‡, and significant improvements of $\text{ADVBERT}_{\text{L2}}$ over BERT_{L2}, and $\text{ADVBERT}_{\text{L4}}$ over BERT_{L4} are shown with the † sign.

[7] https://www.ssa.gov/oact/babynames/background.html
[8] https://github.com/CPJKU/FairnessRetrievalResults

Investigating first the ranker-agnostic document sets, we observe that in the MSMARCO$_{\text{FAIR}}$ collection the NFaiRR of SET$_{\text{Top200}}$ is slightly lower than SET$_{\text{All}}$; in other words the ranker-agnostic fairness results of the top retrieved documents are more biased in comparison with the whole documents in the collection. This indicates that the MSMARCO$_{\text{FAIR}}$ queries tend to retrieve documents from some subspaces of the collection, which contain higher gender biases in comparison with the average bias of the collection. In contrast, this is the other way around in TRECDL19$_{\text{FAIR}}$, such that SET$_{\text{Top200}}$ is more fair than SET$_{\text{All}}$ and almost reaching the maximum (or ideal) fairness value. This indicates that the gender fairness-sensitive queries of the TRECDL19 task lead to either (almost) balanced or no representation of genders in retrieval results.

Looking at the results of the ranking models in MSMARCO$_{\text{FAIR}}$, we observe considerable differences across the IR models, while all show lower fairness scores when compared to SET$_{\text{Top200}}$. In particular, the classical IR models (BM25 and RM3 PRF) show the lowest fairness, whereas NFaiRR is significantly higher for all neural models. We root the cause of these differences in the more noisy results of the two classical models, which on these specific queries, result in higher degrees of gender bias. Among the neural models, the BERT rankers show the overall best fairness results. As reported in Table 2, the NFaiRR results of both BERT models significantly improve in the ADVBERT models, by a considerable margin (7% and 9% in ADVBERT$_{\text{L2}}$ and ADVBERT$_{\text{L4}}$ respectively), showing the effectiveness of the adversarial training method in providing more balanced rankings. Considering the results of TRECDL19$_{\text{FAIR}}$, we observe much less nuances as the starting document set SET$_{\text{Top200}}$ already provides almost fair results. Despite this, we still observe a marginal improvement in NFaiRR when applying our proposed adversarial training method, such that the ADVBERT$_{\text{L4}}$ model offers the best fairness results in both collections.

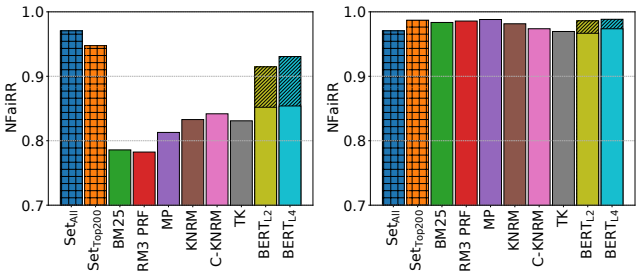
Let us now investigate the changes in NFaiRR by reordering the ranked lists with relevance judgements, as shown in the `+QRels` columns of Table 2. The results show both increase and decrease in fairness scores, where ADVBERT models are the only consistent ones (decreasing). Considering these results, we can not conclude any particular pattern regarding the relation between fairness and oracle rankings in respect to utility. This is, however, a fairly expected behavior, as the topic of fairness is not considered during the process of creating relevance judgements in these collections.

For the sake of completeness, Table 3 reports the evaluation results of the models according to the utility metrics, in which the significant improvements in comparison with the BM25 models are indicated with ‡. In the following section, we discuss in detail the effect of fairness and utility in ADVBERT models.

7.2 Fairness – Utility Trade-off

The reported fairness results of the ADVBERT models so far only consider the models with the best NFaiRR scores, selected through cross validation over all variations of ADVBERT models (see Section 6 – *Adversarial Training Setup*). Despite the fact that the adversarial training method jointly optimizes both fairness and utility objectives, as reported in Table 3, the improvement in fairness is achieved with the cost of a relative decrease in the utility metrics.

Figure 4 shows the NFaiRR and NDCG scores of all model variations of ADVBERT$_{\text{L2}}$ and ADVBERT$_{\text{L4}}$ on MSMARCO$_{\text{FAIR}}$, where each point on the plot corresponds to one variation of each model.[9] The dashed lines show a linear regression fitted to the scores of all



(a) MSMARCO$_{\text{FAIR}}$ **(b) TRECDL19$_{\text{FAIR}}$**

Figure 3: Evaluation results of fairness in retrieval results. The hashed areas on top of the BERT rankers indicate the improvements achieved through adversarial training with the corresponding ADVBERT models.

Table 2: NFaiRR results on `Orig` and the changes after applying the `+QRels` setting. ‡ indicates significant improvement over BM25; † indicates significant improvement of the ADVBERT models over their corresponding BERT models.

Model	MSMARCO$_{\text{FAIR}}$		TRECDL19$_{\text{FAIR}}$	
	`Orig`	`+QRels`	`Orig`	`+QRels`
SET$_{\text{All}}$	0.971	-	0.971	-
SET$_{\text{Top200}}$	0.948	-	0.987	-
BM25	0.786	+0.023	0.984	−0.019
RM3 PRF	0.782	+0.024	0.986	−0.021
MP	0.813‡	+0.013	**0.988**	−0.024
KNRM	0.833‡	+0.011	0.981	−0.017
C-KNRM	0.842‡	+0.010	0.974	−0.009
TK	0.831‡	+0.011	0.970	−0.005
BERT$_{\text{L2}}$	0.852‡	+0.006	0.967	−0.002
ADVBERT$_{\text{L2}}$	0.915‡†	−0.008	0.986	−0.021
BERT$_{\text{L4}}$	0.854‡	+0.006	0.974	−0.009
ADVBERT$_{\text{L4}}$	**0.931**‡†	−0.014	**0.988**	−0.024

Table 3: Utility evaluation of the ranking models. ‡ indicates significant improvement over BM25.

Model	MSMARCO$_{\text{FAIR}}$			TRECDL19$_{\text{FAIR}}$		
	MRR	NDCG	Recall	MRR	NDCG	Recall
BM25	0.107	0.125	0.230	0.850	0.534	0.133
RM3 PRF	0.085	0.104	0.209	0.841	0.556	0.141
MP	0.169‡	0.191‡	0.297‡	0.961	0.578	0.136
KNRM	0.141‡	0.167‡	0.295‡	0.849	0.552	0.137
C-KNRM	0.170‡	0.197‡	0.325‡	0.877	0.595	0.144
TK	0.212‡	0.231‡	0.360‡	0.903	0.679‡	0.149
BERT$_{\text{L2}}$	0.188‡	0.211‡	0.338‡	**0.939**	**0.684**‡	0.150
ADVBERT$_{\text{L2}}$	0.149‡	0.173‡	0.301‡	0.917	0.645‡	0.144
BERT$_{\text{L4}}$	**0.216**‡	**0.252**‡	**0.406**‡	0.933	0.672‡	**0.154**
ADVBERT$_{\text{L4}}$	0.160‡	0.197‡	0.360‡	0.903	0.636‡	0.140

[9] In this section, we only report the results of MSMARCO$_{\text{FAIR}}$ due to the lack of space and since this collection provides a more challenging task for studying fairness.

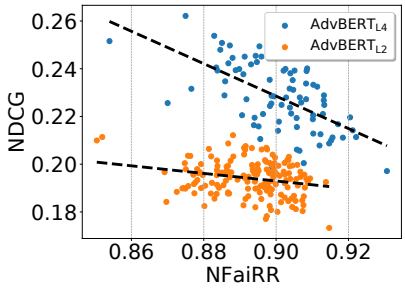


Figure 4: Fairness versus utility results on MSMARCO$_{\text{FAIR}}$ for the model variations of ADVBERT.

variations of a model, which suggests the overall trend: As NFaiRR increases, NDCG generally decreases; and this decrease is more steep for ADVBERT$_{L4}$ than for ADVBERT$_{L2}$.

7.2.1 A Combinatorial Metric for Model Selection. In practice, it is crucial to follow a principled approach for model selection when it comes to a trade-off between fairness and utility. To this end, we first define a combinatorial metric based on the formulation of the well-known F_β score, to combine the fairness metric (NFaiRR) with a utility one (NDCG), as formulated in the following:

$$F_\beta = (1+\beta^2)\frac{\Delta\text{NDCG} \cdot \Delta\text{NFaiRR}}{(\beta^2 \cdot \Delta\text{NDCG}) + \Delta\text{NFaiRR}} \tag{12}$$

where ΔNDCG for a variation of a model is the difference of the NDCG of the variation from the minimum NDCG score among all variations of the model, scaled by min-max normalization (the same is applied to ΔNFaiRR). The β hyper-parameter acts as a leverage (in the hand of practitioners) to impose the preference in the model selection process, inclined towards fairness or utility. Specifically, $\beta = 0$ indicates full preference towards utility, where the selected ADVBERT model most probably becomes the same as BERT (no bias mitigation). $\beta = 1$ gives equal importance to fairness and utility. Higher β values give proportionally more importance to fairness than to utility. Setting β to ∞ (or in fact in practice to a very large number) results in the selection of the variation of the ADVBERT$_{L2}$ or ADVBERT$_{L4}$ model with the highest NFaiRR score, which are equivalent to the ones reported in Section 7.1.

Upon deciding on a β value, the standard model selection process is conducted: the F_β score is calculated for all the variations of a given model, and the model variation with the highest F_β score is selected. In the following experiments, we report the corresponding fairness and utility scores achieved through this model selection by applying 5-fold cross validation on the MSMARCO$_{\text{FAIR}}$ queries.

7.2.2 Final Results. Table 4 shows the evaluation results of the fairness and utility metrics of ADVBERT$_{L2}$ and ADVBERT$_{L4}$ for a range of β values from 0.0 (no bias mitigation – highest utility) to ∞ (maximum fairness). In the table, the results of $\beta = 0.0$ are in fact the same as the ones of the BERT models, which we consider as baselines. For each adversarial model, The † sign on NFaiRR indicates a significant increase in fairness, while on utility metrics (MRR, NDCG, Recall) it shows a significant decrease in scores, when compared with the results of the corresponding BERT model.

As shown, in both models the highest fairness scores result in a significant loss in utility metrics. However, for several lower β values, we observe significant improvements in NFaiRR with no significant loss in utility metrics. In particular, the proper ranges of β that satisfy both fairness and utility are indicated in Table 4 between the dashed lines for each model. These results highlight the effectiveness of our adversarial training approach, which – when combined with the discussed model selection method – provide significantly more fair models without significant loss in utility.

Table 4: Changes in fairness and utility metrics with different values of β. The significant improvements of NFaiRR and the significant loss in utility metrics in comparison with the BERT models are indicated with the † sign.

Model	β	NFaiRR	MRR	NDCG	Recall
BERT$_{L2}$ →	0.0	0.850	0.188	0.211	0.338
ADVBERT$_{L2}$	0.2	0.888†	0.180	0.212	0.364
	0.5	0.888†	0.180	0.212	0.364
	1.0	0.904†	0.167	0.203	0.357
	2.0	0.914†	0.171	0.193	0.319
	5.0	0.914†	0.171	0.193	0.319
	∞	0.915†	0.149†	0.173†	0.301
BERT$_{L4}$ →	0.0	0.854	0.216	0.252	0.406
ADVBERT$_{L4}$	0.2	0.900†	0.215	0.258	0.434
	0.5	0.900†	0.215	0.258	0.434
	1.0	0.900†	0.215	0.258	0.434
	2.0	0.909†	0.214	0.240	0.369
	5.0	0.920†	0.184†	0.220†	0.376
	∞	0.931†	0.160†	0.197†	0.360

8 CONCLUSION AND OUTLOOK

This work provides a standard benchmark for measuring fairness in retrieval results, and proposes an adversarial training method to mitigate bias in BERT ranking models. The benchmark puts forward a generic framework, consisting of metrics for measuring fairness in a given ranked list as well as in a subset of collection's documents. Through human annotation, we provide fairness-sensitive subsets of the queries of two recent passage retrieval collections, MSMARCO$_{\text{FAIR}}$ and TRECDL19$_{\text{FAIR}}$, enabling reproducible research on fairness of IR models together with their utilities. Our experimental results show that, in the more fairness-challenging MSMARCO$_{\text{FAIR}}$ collection, the results of IR rankers are more gender-biased in comparison with the ranker-agnostic baselines, while the fairness of BERT rankers significantly improves by applying our proposed adversarial training. Finally, through a principled model selection method, we show that the resulting ADVBERT models can effectively maintain the significant improvements in fairness with no significant loss in the utility metrics. Future research will investigate the relations between the introduced fairness metrics and the human perception of bias and fairness in retrieved contents, as well as other algorithmic bias mitigation approaches.

ACKNOWLEDGEMENTS

Thanks to Klara Krieg for her help on creating the dataset. This work was funded by the Know-Center GmbH (FFG COMET program) and the H2020 projects TRIPLE (GA: 863420) and AI4EU (GA: 825619).

REFERENCES

[1] Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. *Computer Science Department Faculty Publication Series* (2004), 189.

[2] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* (2018).

[3] Ricardo Baeza-Yates. 2020. Bias in search and recommender systems. In *Proceedings of the ACM Conference on Recommender Systems*. 2–2.

[4] Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6331–6336.

[5] 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.

[6] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[7] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems* (2016).

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

[9] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2018. Ranking with Fairness Constraints. In *Proceedings of the 45th International Colloquium on Automata, Languages, and Programming (ICALP)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[10] 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.

[11] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820* (2020).

[12] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. 126–134.

[13] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[15] Michael D Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and discrimination in retrieval and recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1403–1404.

[16] Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 11–21.

[17] 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* (2020).

[18] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1180–1189.

[19] Ruoyuan Gao and Chirag Shah. 2020. Toward creating a fairer ranking in search engine results. *Information Processing & Management* (2020), 102138.

[20] 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.

[21] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2221–2231.

[22] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*. 2672–2680.

[23] UN Women Headquarters. 2020. UN Women Gender equality: Women's rights in review 25 years after Beijing. https://www.unwomen.org/en/digital-library/publications/2020/03/womens-rights-in-review. Accessed: 2021-02-06.

[24] Sebastian Hofstätter, Navid Rekabsaz, Carsten Eickhoff, and Allan Hanbury. 2019. On the effect of low-frequency terms on neural-IR models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1137–1140.

[25] Sebastian Hofstätter, Hamed Zamani, Bhaskar Mitra, Nick Craswell, and Allan Hanbury. 2020. Local Self-Attention over Long Text for Efficient Document Retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

[26] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* (2002), 422–446.

[27] 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.

[28] Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 39–48.

[29] 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.

[30] Judith Lorber. 2005. *Breaking the bowls: Degendering and feminist change*.

[31] Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Efficient document re-ranking for transformers by precomputing term representations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 49–58.

[32] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3384–3393.

[33] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5270–5278.

[34] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[35] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 429–438.

[36] 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. *arXiv preprint arXiv:1611.09268* (2016).

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

[38] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 6620–6631.

[39] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 933–936.

[40] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication* (2007), 801–823.

[41] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[42] 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.

[43] Navid Rekabsaz, Robert West, James Henderson, and Allan Hanbury. 2021. Measuring Societal Biases in Text Corpora via First-Order Co-occurrence. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)* (2021).

[44] Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in IR* (2009).

[45] Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4187–4195.

[46] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of SIGKDD*.

[47] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *Conference on Neural Information Processing Systems (NeurIPS)*.
[48] K Sparck Jones and C Van Rijsbergen. 1975. Report on the Need for and Provision of an 'ideal' information retrieval test collection. *British Library Research and Development Report 5266* (1975).
[49] Tom Sühr, Asia J Biega, Meike Zehlike, Krishna P Gummadi, and Abhijnan Chakraborty. 2019. Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3082–3092.
[50] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. arXiv:1908.08962 [cs.CL]
[51] Ruth A Wienclaw. 2011. Gender roles. *Sociology Reference Guide: Gender Roles and Equality* (2011), 33–40.
[52] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 585–596.
[53] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 55–64.
[54] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwikj. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *Proceedings of the International Conference on Learning Representations*.
[55] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of Conference on Scientific and Statistical Database Management*.
[56] 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*. 1253–1256.
[57] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.
[58] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference*. 2849–2855.
[59] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the International conference on Machine Learning*. PMLR, 325–333.
[60] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.
[61] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*. 307–314.