Evaluation of bias and robustness in search and conversational systems

Abolghasemi, A.

4

MEASURING BIAS IN A RANKED LIST USING TERM-BASED REPRESENTATIONS

In most recent studies, gender bias in document ranking is evaluated with the NFaiRR metric, which measures bias in a ranked list based on an aggregation over the unbiasedness scores of each ranked document. This perspective in measuring the bias of a ranked list has a key limitation: individual documents of a ranked list might be biased while the ranked list as a whole balances the groups' representations. To address this issue, we propose a novel metric called TExFAIR (term exposure-based fairness), which is based on two new extensions to a generic fairness evaluation framework, attention-weighted ranking fairness (AWRF). TExFAIR assesses fairness based on the term-based representation of groups in a ranked list: (i) an explicit definition of associating documents to groups based on probabilistic term-level associations, and (ii) a rank-biased discounting factor (RBDF) for counting non-representative documents towards the measurement of the fairness of a ranked list. We assess TExFAIR on the task of measuring gender bias in passage ranking, and study the relationship between TExFAIR and NFaiRR. Our experiments show that there is no strong correlation between TExFAIR and NFaiRR, which indicates that TExFAIR measures a different dimension of fairness than NFaiRR. With TExFAIR, we extend the AWRF framework to allow for the evaluation of fairness in settings with term-based representations of groups in documents in a ranked list.

4.1. INTRODUCTION

Ranked result lists generated by ranking models may incorporate biased representations across different societal groups [24, 50, 146]. Societal bias (unfairness) may reinforce negative stereotypes and perpetuate inequities in the representation of groups [89, 186]. A specific type of societal bias is the biased representation of genders in ranked lists of documents. Prior work on binary gender bias in document ranking associates each group (*female*, *male*) with a predefined

set of gender-representative terms [23, 146, 147], and measures the inequality of representation between the genders in the result list using these groups of terms. While there have been efforts in optimizing rankers for mitigating gender bias [146, 156, 205], there is limited research addressing the metrics that are used for the evaluation of this bias. The commonly used metrics for gender bias evaluation are *average rank bias* (which we refer to as ARB) [147] and *normalized fairness in the ranked results* (NFaiRR) [146]. These metrics have been found to result in inconsistent fairness evaluation results [94].

There are certain characteristics of ARB and NFaiRR that limit their utility for bias evaluation of ranked result lists: ARB provides a signed and unbounded value for each query [147], and therefore the bias (unfairness) values are not properly comparable across queries. NFaiRR evaluates a ranked list by aggregating over the unbiasedness score of each ranked document. This approach may result in problematic evaluation results. Consider Figure 4.1, which shows two rankings for a single query where the unbiasedness score of all documents is zero (as each document is completely biased to one group). The fairness of these two rankings in terms of NFaiRR is zero (i.e., both have minimum fairness), while it is intuitively clear that the ranking on the left is fairer as it provides a more balanced representation of the two groups. There are metrics, however, that are not prone to the kind of problematic cases shown in Figure 4.1, but are not directly applicable to fairness evaluation based on term-based group representation off-the-shelf. In particular, *attention-weighted rank fairness* (AWRF) [51, 141, 153] works based on soft attribution of items (here, documents) to multiple groups. AWRF is a generic metric; for a specific instantiation it requires definitions of:

(i) the association of items of a ranked list with respect to each group,

(ii) a weighting schema, which determines the weights for different rank positions,

(iii) the target distribution of groups, and

(iv) a distance function to measure the difference between the target distribution of groups with their distribution in the ranked list.

We propose a new metric *TExFAIR* (term exposure-based fairness) based on the AWRF framework for measuring fairness of the representation of different groups in a ranked list. TExFAIR extends AWRF with two adaptations:

(i) an explicit definition of the association of documents to groups based on probabilistic term-level associations, and

(ii) a ranked-biased discounting factor (RBDF) for counting non-representative documents towards the measurement of the fairness of a ranked list.

Specifically, we define the concept of *term exposure* as the amount of attention each *term* receives in a ranked list, given a query and a retrieval system. Using term exposure of group-representative terms, we estimate the extent to which each group is represented in a ranked result list. We then leverage the discrepancy in the

Figure 4.1: Two ranked lists of retrieved results for "who is the best football player". Documents in blue contain only female-representative terms and documents in red contain only male-representative terms. In terms of NFaiRR, fairness of both ranked result lists is zero (minimum fairness).

representation of different groups to measure the degree of fairness in the ranked result list. Moreover, we show that the estimation of fairness may be highly impacted by whether the non-representative documents (documents that do not belong to any of the groups) are taken into account or not. To count these documents towards the estimation of fairness, we propose a rank-biased discounting factor (RBDF) in our evaluation metric. Finally, we employ counterfactual data substitution (CDS) [114] to measure the gender sensitivity of a ranking model in terms of the discrepancy between its original rankings and the ones it provides if it performs retrieval in a counterfactual setting, where the gender of each gendered term in the documents of the collection is reversed, e.g., "he" → "she," "son" → "daughter."

In summary, our main contributions are as follows:

- We define an extension of the AWRF evaluation framework with the metric *TExFAIR*, which explicitly defines the association of each document to the groups based on a probabilistic term-level association.
- We show that non-representative documents, i.e., documents without any representative terms, may have a high impact in the evaluation of fairness with group-representative terms and to address this issue we define a rank-biased discounting factor (RBDF) in our proposed metric.
- We evaluate a set of ranking models in terms of gender bias and show that the correlation between TExFAIR and NFaiRR is not strong, indicating that TExFAIR measures a different dimension of fairness than NFaiRR.

4.2. BACKGROUND

Fairness in rankings. Fairness is a subjective and context-specific constraint and there is no unique definition when it comes to defining fairness for rankings [6, 68, 115, 161, 203]. The focus of this chapter is on measuring fairness in the representation of groups in rankings [60, 120, 140, 146, 202], and, specifically, the setting in which each group can be represented by a predefined set of group-representative terms. We particularly investigate gender bias in document ranking and follow prior work [22, 67, 146, 147, 205] on gender bias in the binary setting of two groups: female and male. In this setup, each gender is defined by a set of gender-representative terms (words), which we adopt from prior work [146].

Previous studies on evaluating gender bias [24, 146, 147, 205] mostly use the ARB [147] and NFaiRR [146] metrics. Since the ARB metric has undesirable properties (e.g., being unbounded), for the purposes of this chapter we will focus on comparing our newly proposed metric to NFaiRR as the most used and most recent of the two metrics [146, 205]. Additionally, there is a body of prior work addressing the evaluation of fairness based on different aspects [21, 46, 64, 161, 193, 201]. The metrics used in these works vary in different dimensions including

(i) the goal of fairness, i.e., what does it mean to be fair,

(ii) whether the metric considers relevance score as part of the fairness evaluation,

(iii) binary or non-binary group association of each document,

(iv) the weighting decay factor for different positions, and

(v) evaluation of fairness in an individual ranked list or multiple rankings [140, 141].

In light of the sensitivity of gender fairness, which poses a constraint where each ranked list is supposed to represent different gender groups in a ranked list equally [24, 146, 205], we adopt attention-weighted rank fairness (AWRF) [153] as a framework for the evaluation of group fairness in an *individual ranked list* with soft attribution of documents to multiple groups.

Normalized fairness of retrieval results (NFaiRR). In the following, q is a query, $\text{tf}(t,d)$ stands for the frequency of term t in document d, G is the set of N groups where G_i is the i-th group with $i \in \{1,\dots,N\}$, V_{G_i} is the set of group-representative terms for group G_i, d_q^r is the retrieved document at rank r for query q, and k is the ranking cut-off. $M^{G_i}(d)$ represents the magnitude of group G_i, which is equal to the frequency of G_i's representative terms in document d, i.e., $M^{G_i}(d) = \sum_{t \in V_{G_i}} \text{tf}(t,d)$. τ sets a threshold for considering a document as neutral based on $M^{G_i}(d)$ of all groups in G. Finally, J_{G_i} is the expected proportion of group G_i in a balanced representation of groups in a document, e.g., $J_{G_i} = \frac{1}{2}$ in equal representation for $G_i \in \{female, male\}$ [24, 146, 205].

Depending on $M^{G_i}(d)$ for all $G_i \in G$, document d is assigned with a neutrality (unbiasedness) score $\omega(d)$:

$$\omega(d) = \begin{cases} 1, & \text{if } \sum\limits_{G_i \in G} M^{G_i}(d) \leq \tau \\ 1 - \sum\limits_{G_i \in G} \left| \frac{M^{G_i}(d)}{\sum\limits_{G_x \in G} M^{G_x}(d)} - J_{G_i} \right|, & \text{otherwise.} \end{cases} \tag{4.1}$$

To estimate the fairness of the top-k documents retrieved for query q, first, the neutrality score of each ranked document d_q^r is discounted with its corresponding position bias, i.e., $(\log(r+1))^{-1}$, and then, an aggregation over top-k documents is applied (Eq. 4.2). The resulting score is referred to as the *fairness of retrieval results*

(FaiRR) for query q:

$$\text{FaiRR}(q,k) = \sum_{r=1}^{k} \frac{\omega(d_q^r)}{\log(r+1)}. \tag{4.2}$$

As FaiRR scores of different queries may end up in different value ranges (and consequently are not comparable across queries), a background set of documents S is employed to normalize the fairness scores with the *ideal FaiRR* (IFaiRR) of S for query q [146]. IFaiRR(q,S) is the best possible fairness result that can be achieved from reordering the documents in the background set S [146]. The NFaiRR score for a query is formulated as follows:

$$\text{NFaiRR}(q,k,S) = \frac{\text{FaiRR}(q,k)}{\text{IFaiRR}(q,S)}. \tag{4.3}$$

Attention-weighted rank fairness (AWRF). Initially proposed by Sapiezynski et al. [153], AWRF measures the unfairness of a ranked list based on the difference between the exposure of groups and their target exposure. To this end, it first computes a vector E_{L_q} of the accumulated exposure that a list of k documents L retrieved for query q gives to each group:

$$E_{L_q} = \sum_{r=1}^{k} v_r a_{d_q^r}. \tag{4.4}$$

Here, v_r represents the attention weight, i.e., position bias corresponding to the rank r, e.g., $(\log(r+1))^{-1}$ [51, 153], and $a_{d_q^r} \in [0,1]^{|G|}$ stands for the alignment vector of document d_q^r with respect to different groups in the set of all groups G. Each entity in the alignment vector $a_{d_q^r}$ determines the association of d_q^r to one group, i.e., $a_{d_q^r}^{G_i}$. To convert E_{L_q} to a distribution, a normalization is applied:

$$nE_{L_q} = \frac{E_{L_q}}{\|E_{L_q}\|_1}. \tag{4.5}$$

Finally, a distance metric is employed to measure the difference between the desired target distribution $\hat{E}$ and the nE_{L_q}, the distribution of groups in the ranked list retrieved for query q:

$$\text{AWRF}(L_q) = \Delta(nE_{L_q}, \hat{E}). \tag{4.6}$$

4.3. METHODOLOGY

As explained in Section 4.1 and 4.2, NFaiRR measures fairness based on document-level unbiasedness scores. However, in measuring the fairness of a ranked list, individual documents might be biased while the ranked list as a whole balances the groups' representations. Hence, fairness in the representation of groups in a ranked list should not be defined as an aggregation of document-level scores.

We, therefore, propose to measure group representation for a top-k ranking using term exposure in the ranked list as a whole. We adopt the weighting

approach of AWRF, and explicitly define the association of documents on a term-level. Additionally, as we show in Section 4.5, the effect of documents without any group-representative terms, i.e., non-representative documents, could result in under-estimating the fairness of ranked lists. To address this issue, we introduce a rank-biased discounting factor in our metric. Other measures for group fairness exist, and some of these measures also make use of exposure [46, 161].[1] However, these measures are not at the term-level, but at the document-level. In contrast, we perform a finer measurement and quantify the amount of attention a *term* (instead of document) receives.

Term exposure. In order to quantify the amount of attention a specific term t receives given a ranked list of k documents retrieved for a query q, we formally define *term exposure* of term t in the list of k documents L_q as follows:

$$\text{TE@}k(t, q, L_q) = \sum_{r=1}^{k} p_o(t \mid d_q^r) \cdot p_o(d_q^r). \tag{4.7}$$

Here, d_q^r is a document ranked at rank r in the ranked result retrieved for query q. $p_o(t \mid d_q^r)$ is the probability of observing term t in document d_q^r, and $p_o(d_q^r)$ is the probability of document d at rank r being observed by user. We can perceive $p_o(t \mid d_q^r)$ as the probability of term t occurring in document d_q^r. Therefore, using maximum likelihood estimation, we estimate $p_o(t \mid d_q^r)$ with the frequency of term t in document d_q^r divided by the total number of terms in d_q^r, i.e., $\text{tf}(t, d_q^r) \cdot |d_q^r|^{-1}$. Additionally, following [115, 161], we assume that the observation probability $p_o(d_q^r)$ only depends on the rank position of the document, and therefore can be estimated using the position bias at rank r. Following [146, 161], we define the position bias as $(\log(r+1))^{-1}$. Accordingly, Eq. 4.7 can be reformulated as follows:

$$\text{TE@}k(t, q) = \sum_{r=1}^{k} \frac{\frac{\text{tf}(t, d_q^r)}{|d_q^r|}}{\log(r+1)}. \tag{4.8}$$

Group representation. We leverage the term exposure (Eq. 4.8) to estimate the representation of each group using the exposure of its representative terms as follows:

$$p(G_i \mid q, k) = \frac{\sum_{t \in V_{G_i}} \text{TE@}k(t, q)}{\sum_{G_x \in G} \sum_{t \in V_{G_x}} \text{TE@}k(t, q)}. \tag{4.9}$$

Here, G_i represents the group i in the set of N groups indicated with G (e.g., $G = \{female, male\}$), and V_{G_i} stands for the set of terms representing group G_i. The component $\sum_{G_x \in G} \sum_{t \in V_{G_x}} \text{TE@}k(t, q)$ can be interpreted as the total amount of attention that users spend on the representative terms in the ranking for query q. This formulation of the group representation corresponds to the normalization step in AWRF (Eq. 4.5).

Term exposure-based divergence. To evaluate the fairness based on the representation of different groups, we define a fairness criterion built upon our

[1] Referring to the amount of attention an item (document) receives from users in the ranking.

term-level perspective in the representation of groups: in a fairer ranking – one that is less biased – each group of terms receives an amount of attention proportional to their corresponding desired target representation. Put differently, a divergence from the target representations of groups can be used as a means to measure the bias in the ranking. This divergence corresponds to the distance function in Eq. 4.6. Let $\hat{p}_{Gi}$ be the target group representation for each group G_i (e.g., $\hat{p}_{G_i} = \frac{1}{2}$ for $G_i \in \{female, male\}$ for equal representation of male and female), then we can compute the bias in the ranked results retrieved for the query q as the absolute divergence between the groups' representation and their corresponding target representation. We refer to this bias as the *term exposure-based divergence* (TED) for query q:

$$\text{TED}(q,k) = \sum_{G_i \in G} |p(G_i \mid q,k) - \hat{p}_{G_i}|. \tag{4.10}$$

Rank-biased discounting factor (RBDF). With the current formulation of group representation in Eq. 4.9, non-representative documents, i.e., the documents that do not include any group-representative terms, will not contribute to the estimation of bias in TED (Eq. 4.10). To address this issue, we discount the bias in Eq. 4.10 with the proportionality of those documents that count towards the bias estimation, i.e., documents which include at least one group-representative term. To take into account each of these documents with respect to their position in the ranked list, we leverage their corresponding position bias, i.e., $(\log(1+r))^{-1}$ for a document at rank r, to compute the proportionality. The resulting proportionality factor which we refer to as *rank-biased discounting factor* (RBDF) is estimated as follows:

$$\text{RBDF}(q,k) = \frac{\sum_{r=1}^{k} \frac{\mathbb{1}[d_q^r \in S_R]}{\log(1+r)}}{\sum_{r=1}^{k} \frac{1}{\log(1+r)}}. \tag{4.11}$$

Here, S_R stands for the set of representative documents in top-k ranked list of query q, i.e., documents that include at least one group-representative term. Besides, $\mathbb{1}[d_q^r \in S_R]$ is equal to 1 if $d_q^r \in S_R$, otherwise, 0. Accordingly, we incorporate RDBF(q,k) into Eq. 4.10 and reformulate it as:

$$\text{TED}(q,k) = \sum_{G_i \in G} |p(G_i \mid q,k) - \hat{p}_{G_i}| \cdot \frac{\sum_{r=1}^{k} \frac{\mathbb{1}[d_q^r \in S_R]}{\log(1+r)}}{\sum_{r=1}^{k} \frac{1}{\log(1+r)}}. \tag{4.12}$$

Alternatively, as TED(q,k) is bounded, we can leverage the maximum value of TED to quantify the fairness of the rank list of query q. We refer to this quantity as *term exposure-based fairness* (TExFAIR) of query q:

$$\text{TExFAIR}(q,k) = \max(\text{TED}) - \text{TED}(q,k). \tag{4.13}$$

In the following, we use TExFAIR to refer to TExFAIR with proportionality (RBDF), unless otherwise stated. With $\hat{p}_{G_i} = \frac{1}{2}$ for $G_i \in \{female, male\}$, TED (Eq. 4.10 and 4.12) falls into the range of [0,1], therefore TExFAIR$(q,k) = 1 - \text{TED}(q,k)$.

4.4. EXPERIMENTAL SETUP

Query sets and collection. We use the MS MARCO Passage Ranking collection [17], and evaluate the fairness on two sets of queries from prior work [24, 147, 205]:

(i) QS1 which consists of 1756 non-gendered queries [147], and

(ii) QS2 which includes 215 bias-sensitive queries [146] (see [147] and [146] respectively for examples).

Ranking models. Following the most relevant related work [146, 205], we evaluate a set of ranking models which work based on pre-trained language models (PLMs). Ranking with PLMs can be classified into three main categories: sparse retrieval, dense retrieval, and re-rankers. In our experiments we compare the following models:

(i) two *sparse retrieval* models: uniCOIL[106] and DeepImpact [113];

(ii) five *dense retrieval* models: ANCE [192], TCT-ColBERTv1 [109], SBERT [145], distilBERT-KD [70], and distilBERT-TASB [71];

(iii) three commonly used cross-encoder *re-rankers*: BERT [126], MiniLM$_{KD}$ [179] and TinyBERT$_{KD}$ [86]. Additionally, we evaluate BM25 [149] as a widely-used traditional lexical ranker [1, 108].

For sparse and dense retrieval models we employ the pre-built indexes, and their corresponding query encoders provided by the Pyserini toolkit [107]. For re-rankers, we use the pre-trained cross-encoders provided by the sentence-transformers library [145].[2] For ease of fairness evaluation in future work, we make our code publicly available at https://github.com/aminvenv/texfair.

Evaluation details. We use the official code available for NFaiRR.[3] Following suggestions in prior work [146], we utilize the whole collection as the background set S (Eq. 4.3) to be able to do the comparison across rankers and re-rankers (which re-rank top-1000 passages from BM25). Since previous instantiations of AWRF cannot be used for the evaluation of term-based fairness of group representations out-of-the-box, we compare TExFAIR to NFaiRR.

4.5. RESULTS

Table 4.1 shows the evaluation of the rankers in terms of effectiveness (MRR and nDCG) and fairness (NFaiRR and TExFAIR). The table shows that almost all PLM-based rankers are significantly fairer than BM25 on both query sets at ranking cut-off 10. In the remainder of this section we address three questions:

(i) What is the correlation between the proposed TExFAIR metric and the commonly used NFaiRR metric?

[2] https://www.sbert.net/docs/pretrained-models/ce-msmarco.html
[3] https://github.com/CPJKU/FairnessRetrievalResults

Method	QS1					QS2				
	MRR	nDCG	NFAIRR	TExFAIR	r	MRR	nDCG	NFAIRR	TExFAIR	r
Sparse retrieval										
BM25	0.1544	0.1958	0.7227	0.7475	0.4823†	0.0937	0.1252	0.8069	0.8454	0.5237†
UniCOIL	0.3276‡	0.3892‡	0.7819‡	0.7629‡	0.5166†	0.2288‡	0.2726‡	0.8930‡	0.8851‡	0.4049†
DeepImpact	0.2690‡	0.3266‡	0.7721‡	0.7633‡	0.5487†	0.1788‡	0.2200‡	0.8825‡	0.8851‡	0.4971†
Dense retrieval										
ANCE	0.3056‡	0.3640‡	**0.7989**‡	**0.7725**‡	0.5181†	0.2284‡	0.2763‡	0.9093‡	**0.9060**‡	0.4161†
DistillBERT$_{KD}$	0.2906‡	0.3488‡	0.7913‡	0.7683‡	0.5525†	0.2306‡	0.2653‡	**0.9149**‡	0.9044‡	0.4257†
DistillBERT$_{TASB}$	0.3209‡	0.3851‡	0.7898‡	0.7613‡	0.5091†	0.2250‡	0.2725‡	0.9088‡	0.8960‡	0.4073†
TCT-ColBERTv1	0.3138‡	0.3712‡	0.7962‡	0.7688‡	0.5253†	0.2300‡	0.2732‡	0.9116‡	0.9056‡	0.4249†
SBERT	0.3104‡	0.3693‡	0.7880‡	0.7637‡	0.5217†	0.2197‡	0.2638‡	0.8943‡	0.8999‡	0.3438†
Re-rankers										
BERT	0.3415‡	0.4022‡	0.7790‡	0.7584‡	0.5135†	0.2548‡	0.2950‡	0.8896‡	0.8807‡	0.4323†
MiniLM$_{KD}$	**0.3832**‡	**0.4402**‡	0.7702‡	0.7516	0.5257†	**0.2872**‡	**0.3323**‡	0.8863‡	0.8865‡	0.3880†
TinyBERT$_{KD}$	0.3482‡	0.4093‡	0.7799‡	0.7645‡	0.5437†	0.2485‡	0.3011‡	0.8848‡	0.8952‡	0.4039†

Table 4.1: Effectiveness and fairness results at ranking cut-off = 10. r denotes the correlation between TExFAIR and NFaiRR. Higher values of TExFAIR and NFaiRR correspond to higher fairness. † denotes statistical significance for correlations with ($p < 0.05$). ‡ indicates statistically significant improvement over BM25 according to a paired t-test ($p < 0.05$). Bonferroni correction is used for multiple testing.

(ii) What is the sensitivity of the metrics to the ranking cut-off?

(iii) What is the relationship between the bias in ranked result lists of rankers, and how sensitive they are towards the concept of gender?

(i) Correlation between metrics. To investigate the correlation between the TExFAIR and NFaiRR metrics, we employ Pearson's correlation coefficient on the query level. As the values in Table 4.1 indicate, the two metrics are significantly correlated, but the relationship is not strong ($0.34 < r < 0.55$). This is likely due to the fact that NFaiRR and TExFAIR are structurally different: NFaiRR is document-centric: it estimates the fairness in the representation of groups on a document-level and then aggregates the fairness values over top-k documents. TExFAIR, on the other hand, is ranking-centric: each group's representation is measured based on the whole ranking, instead of individual documents. As a result, in a ranked list of k documents, the occurrences of the terms from one group at rank i, with $i \in \{1,\dots,k\}$, can balance and make up for the occurrences of the other group's terms at rank j, with $j \in \{1,\dots,k\}$. This is in contrast to NFaiRR in which the occurrences of the terms from one group at rank i, with $i \in \{1,\dots,k\}$, can only balance and make up for the occurrences of other group's terms at rank i. Thus, TExFAIR measures a different dimension of fairness than NFaiRR.

(ii) Sensitivity to ranking cut-off k. Figure 4.2 depicts the fairness results at various

cut-offs using TExFAIR with and without proportionality (RBDF) as well as the results using NFaiRR. The results using TExFAIR without proportionality show a high sensitivity to the ranking cut-off k in comparison to the other two metrics. The reason is that without proportionality factor RDBF, the unbiased documents with zero group-representative term, i.e., non-representative documents, do not count towards the fairness evaluation. As a result, regardless of the number of this kind of unbiased documents, documents that include group-representative terms potentially can highly affect the fairness of the ranked list. On the contrary, NFaiRR and TExFAIR with proportionality factor are less sensitive to the ranking cut-off: the effect of unbiased documents with zero group-representative term is addressed in NFaiRR with a maximum neutrality for these documents (Eq. 4.1), and in TExFAIR with proportionality factor RBDF by discounting the bias using the proportion of documents that include group-representative terms (Eq. 4.12).

(iii) Counterfactual evaluation of gender fairness. TExFAIR and NFaiRR both measure the fairness of ranked lists produced by ranking models. Next, we perform an analytical evaluation to measure the extent to which a ranking model acts indifferently (unbiasedly) towards the genders, regardless of the fairness of the ranked list it provides. Our evaluation is related to counterfactual fairness measurements which require that the same outcome should be achieved in the real world as in the term-based counterfactual world [134, 188]. Here, the results of the real world correspond to the ranked lists that are returned using the original documents, and results of the counterfactual world correspond to the ranked lists that are returned using counterfactual documents.

In order to construct counterfactual documents, we employ counterfactual data substitution (CDS) [110, 114], in which we replace terms in the collection with their counterpart in the opposite gender-representative terms, e.g., "he" → "she," "son" → "daughter," etc. For names, e.g., Elizabeth or John, we substitute them with a name from the opposite gender name in the gender-representative terms [114]. Additionally, we utilize POS information to avoid ungrammatically assigning "her" as a personal pronoun or possessive determiner [114]. We then measure how the ranked result lists of a ranking model on a query set Q would diverge if a ranker performs the retrieval on the counterfactual collection rather than the original collection.

In order to measure the divergence, we employ rank-biased overlap (RBO) [182] as a measure to quantify the similarities between two ranked lists. We refer to this quantity as *counterfactually-estimated rank-biased overlap* (CRBO). RBO ranges from 0 to 1, where 0 represents disjoint and 1 represents identical ranked lists. RBO has a parameter $0 < p \leq 1$ which regulates the degree of top-weightedness in estimating the similarity. From another perspective, p represents searcher patience or persistence and larger values of p stand for more persistent searching [38]. Since we focus on top-10 ranked results, we follow the original work [182] for a reasonable choice of p, and set it to 0.9 (see [182] for more discussion).

Table 4.2 shows the CRBO results. While there is a substantial difference in the fairness of ranked results between the BM25 and the PLM-based rankers, the CRBO results of these models are highly comparable, and even BM25, as the model which provides the most biased ranked results, is the least biased model in terms of CRBO

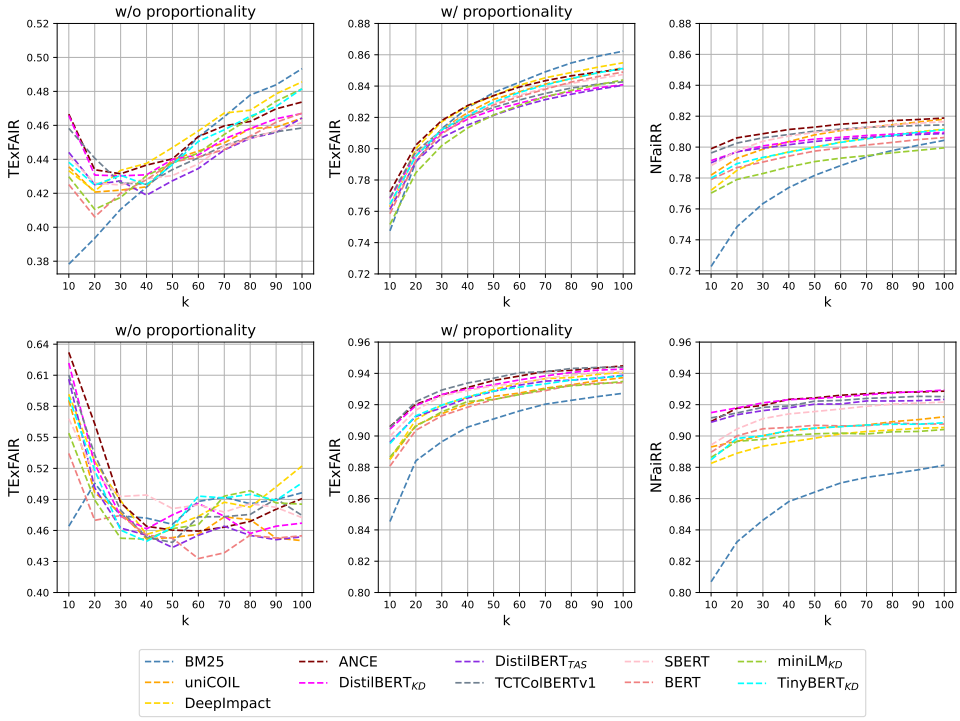

Figure 4.2: Fairness results on QS1 (first row) and QS2 (second row) at different ranking cut-off values (k).

on QS1. Additionally, among PLM-based rankers, the ones with higher TExFAIR or NFaiRR scores do not necessarily provide higher CRBO. This discrepancy between {NFaiRR, TExFAIR} and CRBO disentangles the bias of a model towards genders from the bias of the ranked results it provides. However, it should be noted that we indeed cannot come to a conclusion as to whether the bias that exists in the PLM-based rankers (the one that is reflected by CRBO) does not contribute to their superior fairness of ranked results (the one that is reflected by {NFaiRR, TExFAIR}). We leave further investigation of the quantification of inherent bias of PLM-based rankers and its relation with the bias of their ranked results for future work.

4.6. DISCUSSION

The role of non-representative documents. As explained in Section 4.3, and based on the results in Section 4.5, discounting seems to be necessary for the evaluation of gender fairness in document ranking with group-representative terms, due to the effect of non-representative documents. Here, one could argue that without our proposed proportionality discounting factor (Section 4.3), it is possible to use an association value for d_q^r to group G_i, i.e., $a_{d_q^r}^{G_i}$ in the formulation of AWRF (Section

	QS1			QS2		
Models	CRBO	TExFAIR	NFaiRR	CRBO	TExFAIR	NFaiRR
BM25	**0.9733**	0.8454	0.8069	**0.9761**	0.7475	0.7227
BERT	0.9506	0.8807	**0.8896**	0.9735	0.7629	0.7790
MiniLM	0.9597	0.8865	0.8863	0.9753	0.7516	0.7702
TinyBERT	0.9519	**0.8952**	0.8848	0.9714	**0.7645**	**0.7799**

Table 4.2: Counterfactually-estimated RBO results. For ease of comparison, TExFAIR and NFaiRR results are included from Table 4.1.

4.2) as follows:

$$a_{d_q^r}^{G_i} = \frac{M^{G_i}(d_q^r)}{\sum_{G_x \in G} M^{G_x}(d_q^r)}, \tag{4.14}$$

and simply assign equal association for each group, e.g., $a_{d_q^r}^{G_i} = \frac{1}{2}$ for $G_i \in \{female, male\}$ for documents that do not contain group-representative terms, i.e., non-representative documents. However, we argue that such formulation results in the ignorance of the frequency of group-representative terms. For instance, intuitively, a document which has only one mention of a female name as a female-representative term (therefore is completely biased towards female) and is positioned at rank i, cannot simply compensate and balance for a document with high frequency of male-representative names and pronouns (completely biased towards male) and is positioned at rank $i+1$. However, with the formulation of document associations in AWRF (Eq. 4.14) these two documents can roughly[4] balance for each other. As such, there is a need for a fairness estimation in which the frequency of terms is better counted towards the final distribution of groups. Our proposed metric TExFAIR implicitly accounts for this effect by performing the evaluation based on term-level exposure estimation and incorporating the rank biased discounting factor RBDF.

Limitations of CRBO. While measuring gender bias with counterfactual data substitution is widely used for natural language processing tasks [42, 63, 114, 152], we believe that our analysis falls short of thoroughly measuring the learned stereotypical bias. We argue that through the pre-training and fine-tuning step, specific gendered correlations could be learned in the representation space of the ranking models [183]. For instance, the representation of the word "nurse" or "babysitter" might already be associated with female group terms. In other words, the learned association of each term to different groups (either female or male), established during pre-training or fine-tuning, is a spectrum rather than binary. As a result, these kinds of words could fall at different points of this spectrum and therefore, simply replacing a limited number of gendered-terms (which are assumed to be the two end point of this spectrum) with their corresponding counterpart in the opposite gender group, might not reflect the actual inherent bias of PLM-based rankers towards different groups of gender. Moreover, while we estimate CRBO based on the divergence of the results on the original collection and a single counterfactual

[4] As they have different position bias.

collection, more stratified counterfactual setups can be studied in future work.

Reflection on evaluation with term-based representations. We acknowledge that evaluating fairness with term-based representations is limited in comparison to real-world user evaluations of fairness. However, this shortcoming exists for all natural language processing tasks where semantic evaluation from a user's perspective might not exactly match with the metrics that work based on term-based evaluation. For instance, there exists a discussion over the usage of BLEU [132] and ROUGE [105] scores in the evaluation of natural language generation [169, 208]. Nevertheless, such an imperfect evaluation method is still of great importance due to the sensitivity of the topic of societal fairness and the impact caused by the potential consequences of unfair ranking systems. We believe that this chapter addresses an important aspect of evaluation in the current research in this area and plan to work on more semantic approaches of societal fairness evaluation in the future.

4.7. CONCLUSION

In this chapter, we addressed the evaluation of societal group bias in document ranking. We pointed out an important limitation of the most commonly used group fairness metric NFaiRR, which measures fairness based on a fairness score of each ranked document. Our newly proposed metric TExFAIR integrates two extensions on top of a previously proposed generic metric AWRF: the term-based association of documents to each group, and a rank biased discounting factor that addresses the impact of non-representative documents in the ranked list. As it is structurally different, our proposed metric TExFAIR measures a different aspect of the fairness of a ranked list than NFaiRR. Hence, when fairness is taken into account in the process of model selection, e.g., with a combinatorial metric of fairness and effectiveness [146], the difference between the two metrics TExFAIR and NFaiRR could result in a different choice of model.

In addition, we conducted a counterfactual evaluation, estimating the inherent bias of ranking models towards different groups of gender. With this analysis we show a discrepancy between the measured bias in the ranked lists (with NFaiRR or TExFAIR) on the one hand and the inherent bias in the ranking models themselves on the other hand. In this regard, for our future work, we plan to study more semantic approaches of societal fairness evaluation to obtain a better understanding of the relationship between the inherent biases of ranking models and the fairness (unbiasedness) of the ranked lists they produce. Moreover, since measuring group fairness with term-based representations of groups is limited (compared with the real-world user evaluation of fairness), we intend to work on more user-oriented methods for the measurement of societal fairness in the ranked list of documents.