T-Retrievability: A Topic-Focused Approach to Measure Fair Document Exposure in Information Retrieval

Xuejun Chang
University of Glasgow
Glasgow, UK
x.chang.2@research.gla.ac.uk

Zaiqiao Meng
University of Glasgow
Glasgow, UK
Zaiqiao.Meng@glasgow.ac.uk

Debasis Ganguly
University of Glasgow
Glasgow, UK
Debasis.Ganguly@glasgow.ac.uk

Abstract

Retrievability of a document is a collection-based statistic that measures its expected (reciprocal) rank of being retrieved within a specific rank cut-off. A collection with uniformly distributed retrievability scores across documents is an indicator of fair document exposure. While retrievability scores have been used to quantify the fairness of exposure for a collection, in our work, we use the distribution of retrievability scores to measure the exposure bias of retrieval models. We hypothesise that an uneven distribution of retrievability scores across the entire collection may not accurately reflect exposure bias but rather indicate variations in topical relevance. As a solution, we propose a topic-focused localised retrievability measure, which we call *T-Retrievability* (topic-retrievability), which first computes retrievability scores over multiple groups of topically-related documents, and then aggregates these localised values to obtain the collection-level statistics. Our analysis using this proposed T-Retrievability measure uncovers new insights into the exposure characteristics of various neural ranking models. The findings suggest that this localised measure provides a more nuanced understanding of exposure fairness, offering a more reliable approach for assessing document accessibility in IR systems.

CCS Concepts

• **Information systems → Retrieval models and ranking**.

Keywords

Retrievability Measure; Group-Exposure Fairness; Neural Rankers.

1 Introduction

Supervised information retrieval (IR) models based on pre-trained transformer architectures, such as ColBERT [18] and Mono-T5 [25], have demonstrated superior relevance ranking performance compared to their unsupervised counterparts, such as BM25 [31].

However, this paradigm shift necessitates the consideration of additional dimensions of effectiveness, including those of efficiency [7, 9], explainability [26, 32, 36], and exposure fairness [15, 27, 33]. In this study, we specifically examine the exposure fairness criterion of search systems, which has significant societal implications. For instance, ensuring a balanced distribution of topically relevant documents from diverse sources and ideological viewpoints is essential for preserving the integrity of information ecosystems [27], fostering an informed and unbiased public opinion, and supporting democratic decision-making [1, 19, 20].

The data-driven learning of relevance in neural models makes it challenging to identify and mitigate latent biases present in the training data, which can manifest as algorithmic preferences leading to exposure biases [22, 24, 27]. These exposure biases are expected ranks of retrieved documents computed over a large set of queries that several users may execute on a collection. In the case of queries with similar information needs and overlapping sets of relevant documents, an IR model that aims to ensure exposure fairness should distribute the rankings of these documents more evenly by shuffling their positions across queries [10, 11]. Assessing whether IR models achieve this fairness objective is crucial for understanding and addressing potential biases in retrieval systems. We restrict the scope of our investigation of exposure bias to only document collections without categorical meta-data attributes, e.g., gender, polarity etc., as studied in [30].

Previous research has introduced the concept of *retrievability* as a measure of document accessibility [4, 37–39]. This collection-based statistic quantifies a document's expected (reciprocal) rank of retrieval within a predefined rank cut-off, averaged over a sufficiently large and topically diverse set of queries. While retrievability analysis has traditionally been used to assess document accessibility across different collections [37], more recent studies have applied it to examine exposure bias across various IR models [2]. However, two key considerations must be taken into account when using retrievability for such analyses. First, most existing studies, including those evaluating exposure fairness in IR models [2], rely on synthetic queries generated via word *n*-gram sampling. Synthetic queries used in prior research predominantly consist of unigrams or bigrams, whereas real-world user queries tend to be longer on average. Second, non-uniform retrievability values often arise due to variations in relevance priors, where certain high-quality documents are inherently more likely to be relevant than others. In such cases, disparities in retrievability should be accounted for by aggregating locally computed retrievability measures over groups of similar information needs.

The contributions of this paper are two-fold: a) addressing the limitation of synthetic queries by conducting retrievability analysis

on a large set of real queries (MS MARCO dev set), aligning with real-world information needs; and b) proposing a topic-focused localized retrievability measure, named *T-Retrievability* (topical retrievability), to conduct more fine-grained analysis of document accessibility of various ranking models.

2 Proposed Measure for Exposure Fairness

Background on retrievability. As per the originally proposed definition of retrievability [4], its value for a document $D \in C$ (a collection of documents) indicates the likelihood of the document to be retrieved within a specified cut-off rank. As such, the retrievability value of a document depends on the retrieval model used to induce the ranking itself, and a cut-off rank is usually set to a small number, such as 10, indicating the size of a typical search results page [17]. More formally, $r(D, C, Q, \theta, k) = \frac{1}{|Q|} \sum_{Q \in Q} \mathbb{I}[\rho(D; Q, \theta) \leq k]$, where D is a document of the collection C, Q denotes a set of queries, $\theta : D \times Q \mapsto \mathbb{R}$ denotes an IR model that assigns scores to documents thereby inducing a rank for D denoted as $\rho(D; Q, \theta)$, which is a function of both a particular query Q and the IR model θ. The parameter k denotes a cut-off rank and $\mathbb{I}[X]$ denotes an indicator function which is 1 if the condition X is true, otherwise 0.

Removing dependence on rank cut-off. In this paper, we work with a slightly modified version of the retrievability measure, where we remove the dependence of the cut-off rank k. More specifically, instead of computing the likelihood of a document to be retrieved within a top-k, we rather compute the expected rank of a document [35]. In particular, to interpret retrievability as a score, i.e., the higher the better, we work with reciprocals of ranks instead of the ranks themselves. Formally,

$$r(D, C, Q, \theta) = \frac{1}{|Q|} \sum_{Q \in Q} \frac{1}{\log(1 + \rho(D; Q, \theta))}, \quad (1)$$

As a practical consideration, we restrict the expected rank computation to the top-100 set for each query. The advantage of Equation 1 is that it is a rank-based measure, in contrast to being a set-based measure within top-k, and hence is also less likely to be sensitive to particular choices of the rank cut-off (100 in our case).

The bias in information access is computed by the Gini coefficient [14] over the retrievability distribution of each document in the collection - a high value of the coefficient indicating a high disparity between document accessibility.

From simulated to human-formulated queries. An adequately large set of representative queries Q is used to measure the accessibility of a document in a collection C (Equation 1). All previous work on retrievability analysis [3, 4, 13, 37, 39] uses simulation to create this set Q - specifically via sampling words and their bigrams from the collection. The $r(D)$ values can thus be affected by sampling biases resulting from the simulation. Instead, in our work, we conduct our experiments on real queries formulated by humans. Specifically, we employ the MS MARCO dev set as Q comprising a total of over 100K queries (collected from the Bing search log [5]).

The retrievability distribution thus computed over a large set of real queries for a number of different IR models reflects a more realistic information need. Furthermore, another reason to employ real queries is that it makes it possible to analyse the effects of

document accessibility of neural rankers on the same dataset on which they are actually trained.

A likely reason why real-world query logs are often avoided in retrievability studies [3, 4, 13, 37, 39] is that they tend to be biased toward specific topics or user interests. However, employing a sufficiently large number of such human-formulated queries from search logs, as in our work, can mitigate this bias, as the topical skew diminishes when aggregated across a broader query set.

Problem with non-uniform relevance-priors. Previous research has shown that the prior distribution of relevance in a collection is usually non-uniform [9], meaning that some documents due to their inherent characteristics, such as their lengths [34], citation counts [28] etc., are more likely to be relevant for a higher number of queries than others. However, this means that this non-uniformity in relevance-priors is also likely to be manifested as non-uniformities in the collection-level retrievability scores (see Equation 1) computed for a retrieval model that effectively models relevance, thereby falsely penalising the model.

Localized Retrievability. To address the influence of relevance priors on a collection-level document accessibility measure, such as retrievability, we propose to measure document accessibility for groups of topically related queries. This idea aligns with the per-query analysis of retrievability over a pool of retrieved documents as studied in [39]. This idea also aligns with the existing research on measuring exposure fairness for groups defined by document metadata (attribute value) combinations, such as demographics [30], location [11], and polarities [20]. In our setup, instead of measuring disparities across groups of attribute values, we investigate disparities in document accessibility.

Grouping the Queries. For a fine-grained computation of retrievability scores, we partition the query set Q (in our experiments, the MS MARCO development queries) into K clusters [23]. Specifically, we apply K-means clustering using two types of document vector representations: sparse (lexical) embeddings and dense (semantic) embeddings. These representations yield query groupings with distinct characteristics—one emphasizing precision and the other recall. The lexical approach is more conservative, grouping queries into the same cluster only when they share exact term matches, thereby achieving higher precision. In contrast, the semantic approach expands clusters based on term-level semantics, which increases recall but introduces a greater risk of reduced precision within each cluster.

For the sparse representation, we employ the standard TF–IDF vectorization of each query. For the dense representation, we use the [CLS] token's vector computed with the Sentence-BERT (SBERT) model [29]. To ensure fair comparison across neural models, we adopt the all-MiniLM-L6-v2 variant of SBERT – a model pretrained to capture semantic similarities [6].

From localised measures to a collection-level measure. On the partitioned query set $Q = \cup_{i=1}^{K} Q_i$, we then obtain K such localised T-Retrievability scores - one for each group, i.e.,

$$r(D, C, Q_i, \theta) = \frac{1}{|Q_i|} \sum_{Q \in Q_i} \frac{1}{\log(1 + \rho(D; Q, \theta))}. \quad (2)$$

Table 1: A comparative analysis of the exposure fairness of different retrieval models on MS MARCO dev set. 'G' denotes the Gini coefficient computed over the collection-level retrievability scores, whereas $G_\oplus$ denotes an aggregation ($\oplus \in \{\min, \text{avg}, \max\}$) of the topical retrievability values (see Equations 2 and 3), calculated by K-means clustering with dense vector representations of the dev set. The number of groups (K) for the queries of the dev set is set to 5000.

IR Model	Exposure fairness↓				Relevance↑	
	G	$G_{\min}$	G_{avg}	$G_{\max}$	nDCG@10	MAP@100
BM25	.4731	.1843	.2878	.7412	.2131	.1781
SPLADE	**.3948**	.1843	.3057	.5799	**.4460**	**.3863**
TCT	.3994	.1843	.2970	**.5417**	.4210	.3648
BM25»TCT	.4473	.1843	.2832	.7202	.3713	.3226
Mono-T5	.4428	.1843	**.2818**	.7122	.3962	.3469

After computing the localised retrievability measures for each topic group (as shown in Equation 2), the next step is to compute the document exposure fairness for each of these topic groups by computing K different Gini coefficients. These can then be aggregated (minimum, maximum, or average) to obtain an overall document accessibility measure for a specific retrieval model θ on a collection $\mathcal{C}$. Each way of aggregating corresponds to an analysis of the best case (minimum), average case, or worst case (maximum) exposure fairness of documents. More formally speaking,

$$G[(D, \mathcal{C}, \mathcal{Q}, \theta), \oplus] = \oplus_{i=1}^{K} \bar{G}[r(D, \mathcal{C}, \mathcal{Q}_i, \theta)], \tag{3}$$

where $\bar{G}[\cdot]$ represents the Gini coefficient of a set of localised retrievability values, $\oplus \in \{\min, \text{avg}, \max\}$ denotes an aggregation operator. We call this new collection-level measure of document accessibility Equation 3 by the name **Topical-Retrievability**, abbreviated as **T-Retrievability**[1].

3 Experiment Setup

The objective of our experiments is to compare the exposure biases of various types of ranking models, and also explore how these biases relate to the relevance-based effectiveness measures of these models. Specifically, our first research question is:

• **RQ-1**: How sensitive is exposure bias to a ranking model?

Next, regarding measuring exposure bias, we investigate whether our proposed fine-grained approach with different aggregating mechanisms and the collection-level measure are correlated with each other, or they yield different system preferences in terms of exposure fairness. Explicitly,

• **RQ-2**: How correlated is the topic-based exposure bias measure with the collection-level one?

It is worth noting that our proposed localized retrievability scores depend on the granularity of the topical relatedness of queries, i.e., too coarse topics may lead to relevance prior biases, as is expected to be the case with the standard collection-level retrievability [4], whereas too fine-grained topics may fail to capture a collection-level accessibility [39]. Our third research question is thus:

• **RQ-3**: How sensitive is the proposed T-Retrievability measure on the granularity of the query groups?

Since it is possible that the query grouping itself (conservative grouping by lexical representation vs. more aggressive grouping by a semantic approach) may lead to different trends in exposure fairness as computed by the T-Retrievability measure, our next research question thus explores the sensitivity of the T-Retrievability measure to the query representation used to construct the topics.

• **RQ-4**: How sensitive is the T-Retrievability measure on different embeddings (sparse vs. dense) used to cluster the queries?

Dataset. All our experiments corresponding to retrievability analysis and evaluating the relevance of different rankers are conducted based on the MS MARCO dev dataset, which is comprised of 101,093 queries and their associated relevant documents. The underlying collection of documents used is the MS MARCO passage collection, comprised of over 8.8 million passages [5].

Although it is common practice to evaluate the effectiveness of IR models using standard test collections with depth-pooled relevance assessments (e.g., DL, Robust), such collections are not suitable for assessing exposure fairness. By definition, retrievability is the expected (reciprocal) rank of documents across queries (Equation 1). According to the law of large numbers, this expectation is meaningful only when computed over a sufficiently large set of queries, which small benchmark collections cannot provide.

IR Models. To investigate what characteristic differences may cause variations in document accessibility, we experiment with different classes of retrieval models - namely, sparse, learned-sparse, sparse with reranking, and dense end-to-end, as listed below.

• **BM25** [31]: BM25 is a member of the *sparse* family of IR models; the term weighting scheme is a combination of the relative importance of terms within documents, term informativeness and document lengths.

• **SPLADE** [12]: SPLADE, which belongs to the *learned-sparse* family of rankers, combines the posterior likelihoods of the masked language model (MLM) objective of BERT, and the noise contrastive estimation likelihoods to derive the token weights.

• **TCT-ColBERT** (abbreviated in Table 1 as 'TCT') [21]: This is a *dense end-to-end* model distilled from ColBERT that learns effective document encoded vectors. An approximate nearest neighbour search on these embedded representations has been reported to yield better performance than the [CLS] pooling or mean pooling of the ColBERT model. In our experiments, we use the pyterrier_dr library to construct the dense index.

• **BM25»TCT-ColBERT** (abbreviated in Table 1 as 'BM25»TCT') [21]: the *retrieve-and-rerank* version of TCT-ColBERT, where we rerank the BM25 top-100 documents with the TCT-ColBERT scores for each query-document pair.

• **BM25»Mono-T5** (abbreviated in Table 1 as '**Mono-T5**') [25]: A *retrieve-and-rerank* class of model, where the top 100 documents retrieved by BM25 are subsequently reranked using a cross-encoder model MonoT5, which assigns relevance scores to each query-document pair.

Evaluation Metrics. As evaluation metrics for document exposure bias, we employ the Gini coefficient of the retrievability scores (lower scores indicating fair document exposure). As the measure

(a) Min across Topics

(b) Mean across Topics
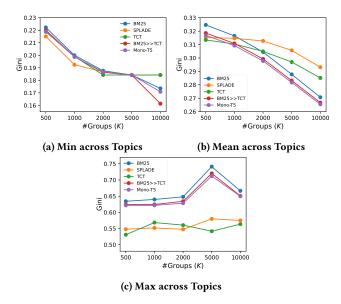
(c) Max across Topics

Figure 1: Variations in document exposure fairness (as measured by Gini coefficients of T-retrievability) for different granularity of topics (query groups) obtained with K-means on dense representations of the queries.

(a) Min across Topics

(b) Mean across Topics

(c) Max across Topics

Figure 2: Similar to Figure 1 – the only difference being that sparse (tf-idf) representation is used to cluster the queries.

for relevance, we employ the common measure of nDCG@10 [16] and MAP [8].

It is important to note that, unlike per-query retrieval effectiveness measures, statistical significance testing (e.g., paired t-test) is not applicable to the Gini coefficient of retrievability or T-retrievability, since these are collection-level measures.

4 Results

Table 1 shows a comparison between the collection-level and the topic-focused Gini coefficients of the retrievability scores for the different IR models. In relation to RQ-1, we observe that there is considerable variance in the exposure biases across different IR models. As can be seen from the collection-level Ginis in Table 1, the variance ranges from the highest Gini observed for BM25 (0.4731) to the lowest for Splade (0.3948).

For RQ-2, we observe that the collection-level exposure fairness measure (Equation 1) is not always correlated with the more fine-grained measures (Equation 2). For example, Table 1 shows that Splade yields the lowest Gini coefficient at the collection level (0.3948), whereas Mono-T5 achieves the lowest average Gini (0.2818) when using fine-grained measures. The key observation is that our proposed fine-grained exposure fairness measures reveal further insights into a model's best, average, and worst-case behaviour.

In relation to RQ-3, from Figure 1, we observe that both the best-case and average-case exposure fairness metrics generally decrease as the number of groups of queries increases. An exception to this trend, illustrated in Figure 2, is that the best-case exposure fairness for Splade and TCT remains stable, and the average-case fairness for all models exhibits a slight increase in the intermediate range (i.e., with group sizes between 1000 and 2000).
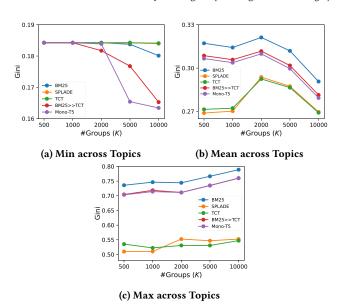
As can be seen from Figure 1, the topic-based retrievability analysis exhibits variations in terms of model performance for exposure fairness across a range of different topic granularities. For instance, over fine-grained topics, Splade is no longer the most fair model (as observed in Table 1). Additional insights are revealed in the worst-case scenarios across both graphs, where Splade and TCT demonstrate notably higher exposure fairness (i.e., lower Gini values) as compared to other retrieval methods. The worst-case exposure fairness, mostly, tends to decrease (i.e., Gini values mostly tend to increase) when the topics become more fine-grained (i.e., as K increases) thus suggesting that ranking models that appear to ensure relatively balanced document accessibility may, in fact, exhibit disparities when analysed at a finer level of query topics.

In relation to RQ-4, it can be observed that the query grouping strategy (which in turn depends on the query representations themselves) exhibits different trends in the T-Retrievability values, e.g., the exposure bias of Splade in Figures 2b and 2c relative to other models is much lower than those observed in Figures 1b and 1c.

Concluding Remarks. In this paper, we analysed the exposure fairness of various retrieval models using a modified retrievability measure. We argued that measuring exposure biases by the standard retrievability-based Gini coefficient may disproportionately penalise effective ranking models, as some degree of non-uniformity can be attributed to inherent disparities in the relevance priors. Our proposed measure mitigates this effect by aggregating disparities in retrievability scores across topically focused subsets of queries. Experimental results on a standard benchmark set of a large number of queries (MS MARCO dev) indicate substantial variations in the exposure biases of different neural models.

In future, we plan to leverage a historical set of queries (e.g., the MS MARCO training set) to adjust document rankings based on prior exposure levels.

GenAI Usage Disclosure

Generative AI tools were not used for core idea generation or experimental design. Its use was limited to minor writing and formatting.

References

[1] Amin Abolghasemi, Leif Azzopardi, Arian Askari, Maarten de Rijke, and Suzan Verberne. 2024. Measuring Bias in a Ranked List Using Term-Based Representations. In *European Conference on Information Retrieval*. Springer, 3–19.

[2] Amin Abolghasemi, Suzan Verberne, Arian Askari, and Leif Azzopardi. 2023. Retrievability Bias Estimation Using Synthetically Generated Queries. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. 3712–3716.

[3] Leif Azzopardi. 2015. Theory of Retrieval: The Retrievability of Information. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. 3–6.

[4] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: an evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. 561–570.

[5] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Neural Information Processing Systems (NIPS) Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*.

[6] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lluís Màrquez, Chris Callison-Burch, and Jian Su (Eds.). 632–642.

[7] Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. 2024. Efficient Inverted Indexes for Approximate Retrieval over Learned Sparse Representations. In *SIGIR*. 152–162.

[8] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00)*. Association for Computing Machinery, New York, NY, USA, 33–40. doi:10.1145/345508.345543

[9] Xuejun Chang, Debabrata Mishra, Craig Macdonald, and Sean MacAvaney. 2024. Neural Passage Quality Estimation for Static Pruning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 174–185.

[10] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. 275–284.

[11] Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. 2023. Overview of the TREC 2022 Fair Ranking Track. arXiv:2302.05558 [cs.IR]

[12] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2288–2292.

[13] Debasis Ganguly, Ayan Bandyopadhyay, Mandar Mitra, and Gareth J.F. Jones. 2016. Retrievability of Code Mixed Microblogs. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. 973–976.

[14] Corrado Gini. 1921. Measurement of Inequality of Incomes. *The Economic Journal* 31, 121 (1921), 124–126.

[15] Maria Heuss, Fatemeh Sarvi, and Maarten de Rijke. 2022. Fairness of Exposure in Light of Incomplete Exposure Estimation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. 759–769.

[16] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446. doi:10.1145/582415.582418

[17] Diane Kelly and Leif Azzopardi. 2015. How many results per page? A Study of SERP Size, Search Behavior and User Experience. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. 183–192.

[18] 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 (SIGIR '20)*. 39–48.

[19] Thorsten Krause, Alina Deriyeva, Jan Heinrich Beinke, Gerrit York Bartels, and Oliver Thomas. 2024. The Relevance of Item-Co-Exposure For Exposure Bias Mitigation. *arXiv preprint arXiv:2409.12912* (2024).

[20] 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. *CoRR* abs/1704.01347. arXiv:1704.01347

[21] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. 163–173.

[22] Zhongzhou Liu, Yuan Fang, and Min Wu. 2023. Mitigating Popularity Bias for Users and Items with Fairness-centric Adaptive Recommendation. *ACM Trans. Inf. Syst.* 41, 3, Article 55 (Feb. 2023), 27 pages.

[23] Simon Lupart, Thibault Formal, and Stéphane Clinchant. 2023. MS-Shift: An Analysis of MS MARCO Distribution Shifts on Neural Retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*. 636–652.

[24] Cataldo Musto, Pasquale Lops, Giovanni Semeraro, et al. 2021. Fairness and Popularity Bias in Recommender Systems: an Empirical Evaluation.. In *DP@ AI*IA*. 77–91.

[25] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. *arXiv preprint arXiv:2003.06713* (2020).

[26] Saran Pandian, Debasis Ganguly, and Sean MacAvaney. 2024. Evaluating the Explainability of Neural Rankers. In *ECIR (4) (Lecture Notes in Computer Science, Vol. 14611)*. 369–383.

[27] Gourab K Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. 2022. Fair ranking: a critical review, challenges, and future directions. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 1929–1942.

[28] Jie Peng, Craig Macdonald, and Iadh Ounis. 2008. Automatic document prior feature selection for web retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. 761–762.

[29] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). 3982–3992.

[30] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 306–316.

[31] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994 (NIST Special Publication, Vol. 500-225)*, Donna K. Harman (Ed.). 109–126.

[32] Procheta Sen, Debasis Ganguly, Manisha Verma, and Gareth J. F. Jones. 2020. The Curious Case of IR Explainability: Explaining Document Scores within and across Ranking Models. In *SIGIR*. ACM, 2069–2072.

[33] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2219–2228.

[34] Amit Singhal, Chris Buckley, and Manclar Mitra. 2017. Pivoted Document Length Normalization. *SIGIR Forum* 51, 2 (Aug. 2017), 176–184.

[35] Aman Sinha, Priyanshu Raj Mall, and Dwaipayan Roy. 2023. Findability: A Novel Measure of Information Accessibility. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4289–4293.

[36] Manisha Verma and Debasis Ganguly. 2019. LIRME: Locally Interpretable Ranking Model Explanation. In *SIGIR*. 1281–1284.

[37] Colin Wilkie and Leif Azzopardi. 2014. A Retrievability Analysis: Exploring the Relationship Between Retrieval Bias and Retrieval Performance. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. 81–90.

[38] Colin Wilkie and Leif Azzopardi. 2015. Retrievability and retrieval bias: A comparison of inequality measures. In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29-April 2, 2015. Proceedings 37*. Springer, 209–214.

[39] Colin Wilkie and Leif Azzopardi. 2016. A Topical Approach to Retrievability Bias Estimation. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR '16)*. 119–122.