Mitigating Bias in Search Results Through Contextual Document Reranking and Neutrality Regularization

George Zerveas
george_zerveas@brown.edu
Brown University, AI Lab
Providence, RI, USA

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

Daniel Cohen
daniel_cohen@brown.edu
Brown University, AI Lab
Providence, RI, USA

Carsten Eickhoff
carsten@brown.edu
Brown University, AI Lab
Providence, RI, USA

ABSTRACT

Societal biases can influence Information Retrieval system results, and conversely, search results can potentially reinforce existing societal biases. Recent research has therefore focused on developing methods for quantifying and mitigating bias in search results and applied them to contemporary retrieval systems that leverage transformer-based language models. In the present work, we expand this direction of research by considering bias mitigation within a framework for contextual document embedding reranking. In this framework, the transformer-based query encoder is optimized for relevance ranking through a list-wise objective, by jointly scoring for the same query a large set of candidate document embeddings in the context of one another, instead of in isolation. At the same time, we impose a regularization loss which penalizes highly scoring documents that deviate from neutrality with respect to a protected attribute (e.g., gender). Our approach for bias mitigation is end-to-end differentiable and efficient. Compared to the existing alternatives for deep neural retrieval architectures, which are based on adversarial training, we demonstrate that it can attain much stronger bias mitigation/fairness. At the same time, for the same amount of bias mitigation, it offers significantly better relevance performance (utility). Crucially, our method allows for a more finely controllable and predictable intensity of bias mitigation, which is essential for practical deployment in production systems.[1]

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**; • **Computing methodologies →** *Regularization*; *Natural language processing*; • **Social and professional topics → Computing / technology policy**.

KEYWORDS

information retrieval, bias mitigation, fairness, transformer-based language models, neutrality regularization, list-wise ranking, set-based ranking, contextual document reranking

ACM Reference Format:
George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022. Mitigating Bias in Search Results Through Contextual Document Reranking and Neutrality Regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3477495.3531891

1 INTRODUCTION

Information Retrieval (IR) systems reflect and may even exaggerate societal biases and stereotypes in their results [6, 15, 18, 28, 33]. If optimization of the underlying IR models is left to exclusively utility-oriented objectives, search engines, through the continual feedback loop of user interactions, can reinforce these biases in society (and hence back in IR systems) [1, 11, 12, 29, 41]. This accentuates the need for bias-aware IR models, in which fairness constraints are imposed with an adjustable degree of effect to control the fairness-utility trade-off in retrieval results [3, 4, 10, 25, 31, 38, 44]. To this end, Rekabsaz et al. [31] recently introduced $\text{MSMARCO}_{\text{FAIR}}$, a reproducible evaluation framework to measure gender bias in the text contents of search results, particularly suited to assessing the interplay of bias mitigation and utility in deep IR models. The framework identifies a set of *bias-sensitive queries*, singled out from the queries of the MS MARCO dataset [2]. Given such a query, an effective and bias-aware IR model should highly rank relevant documents with a balanced or neutral representation of genders in their text contents. Beyond the notion of bias with respect to so-called *protected attributes* such as gender or ethnicity, we note that neutrality of documents can also refer to opinion or political biases: if a query can be answered by a relevant document with purely factual or unbiased content, this document is preferable to a similarly relevant but overtly biased document, which, notwithstanding reliability, may contribute to polarization.

As an example, a query such as "what is the role of a governor?" can and should be answered in a gender-neutral way. A document reading *"The governor is the chief executive of the state.* His *duties include ... /* he *is responsible for ..."* contains words charged/biased

[1] Code repository: https://github.com/gzerveas/CODER

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531891

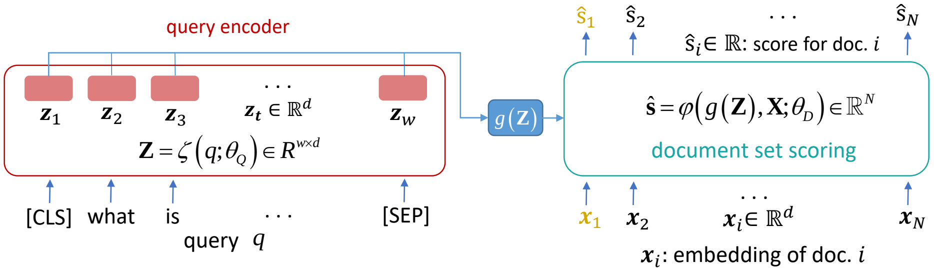
query encoder
z_1
z_2
z_3
. . .
$z_t \in \mathbb{R}^d$
z_w
$\mathbf{Z} = \zeta(q;\theta_Q) \in R^{w \times d}$
[CLS]
what
is
. . .
[SEP]
query q
$g(\mathbf{Z})$
$\hat{s}_1$
$\hat{s}_2$
. . .
$\hat{s}_N$
$\hat{s}_i \in \mathbb{R}$: score for doc. i
$\hat{\mathbf{s}} = \varphi(g(\mathbf{Z}), \mathbf{X}; \theta_D) \in \mathbb{R}^N$
document set scoring
x_1
x_2
. . .
$x_i \in \mathbb{R}^d$
x_N
x_i: embedding of doc. i

Figure 1: Schematic diagram of the contextual document embedding reranking (CODER) framework.

with respect to gender and induces an unnecessarily gender-biased exposition. Thus, it would be desirable to rank higher a document offering equally relevant information but using either gender-neutral words or *gender-representative* words in a balanced way (e.g. "he or she"). In this example, gender-representative words are pronouns, but additional instances include *"man"/"woman"*, *"father"/"mother"*, *"actor"/"actress"*, male/female names, etc. Similarly, a document answering a purely factual query about the current state of the economy (e.g. "is inflation now higher than in 70's") could be penalized if it contained words indicating political bias, such as *"trumpism"*, *"wokeness"*, *"MAGA"*, *"neoliberal"*, *"leftist"*, etc.

We note that this notion of *fairness through neutrality* is not applicable to all queries; rather, we measure it on a set of expert-curated queries for which bias is considered "socially problematic" [20, 21]. Furthermore, it is not applicable to all retrieval use cases; for example, when ranking job applicants, who are inevitably gendered, a more suitable framework would optimize for *fairness of exposure* [10] (e.g. making sure that, given the same level of fitness, female applicants are ranked higher than male applicants as often as male applicants are ranked higher than female applicants).

A variety of past works have proposed effective methods to address bias mitigation, often using some sort of list-wise optimization such as Gumbel-Softmax, stochastic optimization, or reinforcement learning [10, 25, 27, 37, 39]. However, to the best of our knowledge, almost all prior works operate solely on extremely shallow ranking models. As Cohen [7] shows, naively applying shallow methods to deep transformer based architectures often presents significant challenges. Given the scope of this contribution, we therefore consider the relevant work of Rekabsaz et al. [31], where the authors propose integrating adversarial training in deep ranking models in order to improve bias mitigation. Their adversarial method aims to remove gender-related information from the model's internal embeddings, making the predictions (potentially) agnostic to the explicit/implicit presence of gender concepts in a given query-document pair.

The inherent point-wise nature of this adversarial method fits well to the current dominating paradigm of point-/pair-wise optimization [19, 22, 32, 47], embraced mainly due to the practical and conceptual complexities of training deep IR models [7]. Despite the benefits of this approach in mitigating gender bias, previous work [8, 13, 23, 31] and our own experiments show that adversarial training can be highly unstable, with unabated fluctuations over fairness and utility metrics. This issue significantly impedes identifying a model checkpoint that reliably yields generalizable performance and operates within a desired range in the fairness-utility trade-off – a necessary aspect for the wide adoption of bias-aware IR models in practice.

We approach this topic by introducing a novel list-wise bias mitigation method that leverages the recently introduced *COntextual Document Embedding Reranking (CODER)* [45], a neural retrieval framework which enables set-based training of any pretrained dual-encoder deep IR model. The latter is a class of models that represents the state of the art in dense retrieval (e.g. [16, 30, 35]). The guiding principle of bias mitigation is that neutral/unbiased documents should be ranked higher than biased documents of the same or similar relevance. To this end, we extend CODER with a novel list-wise regularization term defined to support neutral documents in final relevance predictions.

We apply our bias-aware optimization method using CODER with TAS-B [16] as the base transformer encoder model. We compare our method with adversarial training for bias mitigation of TAS-B and a cross-encoder (query-document term interaction) BERT Reranker [26], the current SOTA for bias mitigation proposed by Rekabsaz et al. [31]. Ranking results as well as training dynamics are evaluated in terms of utility (MRR, and Recall) and neutrality/fairness (NFaiRR) metrics on the MSMARCO$_{\text{FAIR}}$ collection. Our results show that besides achieving state-of-the-art performance in terms of fairness for the same utility, our set-based neutrality regularization method, in contrast to adversarial alternatives, provides a stable optimization of the network in a short training time, and allows predictably adjusting the intensity of the trade-off between fairness and utility, in a far wider range.

Our contribution thus represents a significant step towards advancing bias mitigation in search, from a theoretical discussion point or a largely experimental research topic, to a practical approach that can be readily integrated in state-of-the-art retrieval systems deployed in industry.

2 METHOD

Our approach optimizes the parameters of a retrieval model such that it assigns scores to documents in proportion to their relevance to a query, while at the same time directly imposing a neutrality constraint on the top-ranked documents. To achieve this, we need a training setup where a large number of documents is simultaneously scored for the same query, and therefore the standard training setup using *(query, positive document, negative document)* triplets is unsuitable. Making use of in-batch documents (e.g., [17, 24, 30]) can indeed provide a large number of random documents as negatives; however, random negatives are only very rarely related to the query or each other, and have been convincingly shown to be less effective than retrieved negatives [30, 42, 45, 46]. Importantly, we wish to impose neutrality on the *top-ranked* documents retrieved by a system, whereas randomly sampled documents are extremely unlikely to end up in high-ranking positions and are thus poor targets for regularization. For these reasons, we instead utilize the recently introduced contextual document embedding reranking (CODER) framework [45], which, given a query, jointly scores a large set of candidate documents that together constitute a ranking *context*.

A schematic diagram of the framework is shown in Figure 1. A pre-trained transformer encoder ζ first transforms a tokenized query q of length w into a sequence of d-dimensional embedding vectors: $\mathbf{Z} = [\mathbf{z}_1; \ldots; \mathbf{z}_w] = \zeta\left(q; \theta_Q\right) \in \mathbb{R}^{w \times d}$, out of which an aggregator function $g(\mathbf{Z})$ extracts a single vector. In this work, as ζ we choose the query encoder from TAS-B [16], which is based on DistilBERT [36] and was the most effective base model evaluated by Zerveas et al. [45]. The aggregation function here simply selects the output embedding corresponding to the first query token, i.e., `[CLS]`: $g(\mathbf{Z}) = \mathbf{z}_1 \in \mathbb{R}^d$.

A scoring function φ computes a scalar relevance score $\hat{s}_i$ for each document embedding $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, N$, based on their similarity to the query embedding $g(\mathbf{Z})$. The set of N documents consists of the ground-truth relevant document(s) and the top candidates retrieved for the same query by an arbitrary retrieval method (here, BM25 [9] by Anserini [43]) in advance. Their embeddings have been precomputed by the document encoder of a dual-encoder model (here TAS-B). A large number of negative documents is essential for providing adequate signal to effectively capture relevance [30, 45], and we use $N = 1000$, following Zerveas et al. [45]. Although the scoring function can in general be parametric, here we simply use the dot-product, which is commonly used for evaluating similarity and was shown by Zerveas et al. [45] to be effective:

$$\hat{\mathbf{s}} = \varphi\left(g(\mathbf{Z}), \mathbf{X}\right) = \mathbf{X} \cdot g(\mathbf{Z}) \in \mathbb{R}^N \tag{1}$$

Throughout training, the parameters of the query encoder are fine-tuned through the ListNet loss [5], which for a given query is equivalent to the KL-divergence between a distribution over the target (ground-truth) relevance labels $\mathbf{y} \in \mathbb{R}^N$, defined for the set of N candidate documents (where the relevance of all documents not explicitly defined is assumed to be 0), and a distribution over the corresponding predicted scores $\hat{\mathbf{s}}$:

$$\mathcal{L}_u\left(\mathbf{y}, \hat{\mathbf{s}}\right) = \mathrm{D}_{\mathrm{KL}}\left(\sigma(\mathbf{y}) \mid\mid \sigma(\hat{\mathbf{s}})\right) = -\sum_{i=1}^{N} \sigma(\mathbf{y})_i \log \frac{\sigma(\hat{\mathbf{s}})_i}{\sigma(\mathbf{y})_i} \tag{2}$$

where σ denotes the softmax function.

The loss function above guides parameter optimization towards maximizing the relevance of top-ranked documents, i.e., the *utility* for the user. To impose neutrality on the top-ranked documents, we add the following *neutrality* loss term to obtain the total loss:

$$\mathcal{L}_n\left(\mathbf{y}_n, \hat{\mathbf{s}}\right) = \mathrm{D}_{\mathrm{KL}}\left(\sigma(\hat{\mathbf{s}}) \mid\mid \sigma(\mathbf{y}_\mathbf{n})\right) = -\sum_{i=1}^{C} \sigma(\hat{\mathbf{s}})_i \log \frac{\sigma(\mathbf{y}_\mathbf{n})_i}{\sigma(\hat{\mathbf{s}})_i} \tag{3}$$

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_u + \lambda_r \mathcal{L}_n \tag{4}$$

where λ_r is the regularization coefficient, C is the cut-off rank for considering neutrality (we use $C = 10$), and $\mathbf{y}_\mathbf{n}$ are the neutrality scores for each document. These are computed following Rekabsaz et al. [31], and are based on the frequency of occurrence of terms indicative of bias with respect to the protected attribute.

We note that the order of distributions in the asymmetric KL-divergence is reversed in the two loss terms: in the utility loss, we primarily penalize assigning a low score to ground-truth relevant documents (rather than the case of assigning a high score to documents which have not been annotated as relevant). This is desirable, among other reasons, because relevance annotations are sparse and many candidate documents can be relevant without having been marked as such [30]. By contrast, the neutrality loss primarily penalizes assigning high scores to documents with low neutrality scores, rather than the case of assigning a low score to neutral documents (since neutrality alone is not indicative of relevance).

3 EXPERIMENT SETUP

Resources. Our experimental setting closely follows Rekabsaz et al. [31]. Fairness and utility of the models are evaluated on 215 curated bias-sensitive queries, i. e., "gender-neutral queries for which biases in their retrieval results are considered as socially problematic" [20, 21], provided by $\text{MSMARCO}_{\text{FAIR}}$. The models are trained on the data provided by the MS MARCO Passage Retrieval collection [2], and retrieval is conducted on the collection's passages. The protected attribute is gender, defined in binary fashion using the gender-representative words (158 words for each gender) provided in previous work [31, 34]. These words are used to calculate the neutrality score for each document with the term occurrence threshold set to 1 (c. f. Rekabsaz et al. [31]).

Retrieval Models. All models are trained and evaluated by reranking candidates first retrieved by BM25 [9]. **CODER(TAS-B)** is our proposed bias mitigation approach explained in Section 2. We select TAS-B [16] as a base transformer encoder for CODER, due to its superior retrieval performance in reranking and dense retrieval scenarios. **AdvBERT** is the model introduced by Rekabsaz et al. [31] which applies adversarial training to a BERT Reranker [26]. The adversarial network in AdvBERT is defined on the output vector of the `[CLS]` token when both query and documents are passed to the BERT model. Following Rekabsaz et al. [31], the prediction label of the adversarial network is defined as a binary variable, which is set to 1 (gendered) if either the given query or document are not fully gender neutral texts, and 0 otherwise. AdvBERT approaches removing gender-related information in models using a gradient reversal mechanism [14]. The gradient of the loss corresponding to the adversarial "gender detector" is scaled by the *adversarial factor* λ_a, which allows tuning the intensity of bias mitigation. AdvBERT is the best performing model reported in Rekabsaz et al. [31], achieved by fine-tuning the BERT-Mini [40] model. **AdvTAS-B** applies a similar adversarial training procedure as AdvBERT to the TAS-B model, providing an adversarial baseline directly comparable with CODER(TAS-B). Because AdvBERT is a cross-encoder model, while TAS-B is a dual encoder, in AdvTAS-B the adversarial network is defined over the concatenation of the query and document embeddings (the `[CLS]` output of the corresponding encoder), and training aims to remove gender-related information in both query and document encoders.[2]

[2] We additionally conducted experiments on other variations, such as exclusively training the query or document encoder. We observed that the chosen AdvTAS-B model shows the best and most stable performance in terms of both fairness and utility.

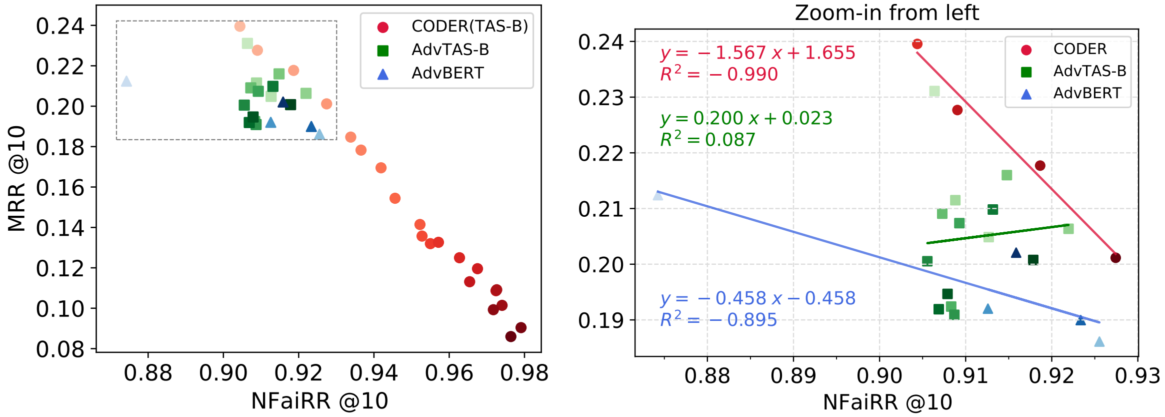

Figure 2: Utility (MRR@10) versus fairness (NFaiRR@10). The intensity of color corresponds to an increasing adversarial or regularization factor. Compared to adversarial baselines, regularization with CODER allows modulating fairness to much higher values, while for the same values of fairness, utility is higher.

Evaluation of Bias Mitigation and Utility. The fairness of the ranking models is evaluated in terms of the *Normalized Fairness of Retrieval Results (NFaiRR)* metric [31]. The NFaiRR metric measures to what extent the contents of the retrieved documents show a balanced representation of a protected attribute (gender in our experiments). This is done by first calculating FaiRR scores as the sum over the neutrality scores of the top retrieved documents, weighted by their ranking positions. The NFaiRR metric provides comparable results across queries by normalizing the per-query FaiRR scores over the ideal FaiRR inferred from a background set of documents (e.g. the top documents retrieved by a baseline BM25 model [31]). We calculate the NFaiRR metric with a cutoff at 10 for each bias-sensitive query, and report the average results over queries. The utility of the models is evaluated with common metrics for MS MARCO (which defines relevance in a binary fashion and is a sparsely annotated collection, most often with a single relevant passage per query), namely mean reciprocal rank (MRR) and Recall, both at cutoff 10.

Training, Model Selection and Hyperparameters. When tuning the intensity of bias mitigation, we need to train a model for each value of the regularization coefficient or adversarial factor. Which time-dependent model instance (checkpoint) should we choose as "best", in order to evaluate the method's performance? Since there is a trade-off between utility and fairness, we follow the principled approach proposed by Rekabsaz et al. [31]: we max-min normalize MRR and NFaiRR to a range between 0 and 1, and choose the instance where their harmonic mean (F1 score) is maximum over the entire training session. In the case of adversarial methods, because validation performance fluctuates persistently during training, it is not clear when to stop training. By contrast, CODER shows a smooth convergence behavior (see Fig.4) and in practice would benefit from stopping criteria based on the relative improvement of metrics. However, to avoid giving this advantage to CODER, while still reflecting practical concerns for model training and selection, we fix the maximum training time of all models to a value that we estimated to be sufficient for each model to achieve its "best" performance (as defined above), after running a few tentative training sessions. Consequently, regardless of the regularization/adversarial factor, this value was set to 10 hours for CODER, 12 hours for AdvTAS-B, and 8 days for AdvBERT[3]. We nevertheless note that, unlike in the case of the adversarial models, in the case of CODER the "best" performance was most often reported at or close to the very end of training, which indicates that its maximum performance is likely underestimated and that it would benefit from a training time dependent on the regularization coefficient. To train CODER, we use the same hyperparameters as in Zerveas et al. [45], but increase batch size from 32 to 64 to accelerate convergence.

4 RESULTS

Figure 2 depicts how retrieval performance changes in terms of utility and fairness/neutrality (measured as described in Section 3 in terms of MRR@10 and NFaiRR@10 respectively) as we progressively increase the intensity of bias mitigation (shown by the intensity of marker color), starting from 0. Compared to the adversarial baselines, it is evident that regularization with CODER allows modulating fairness to much higher values. Importantly, our method yields an approximately linear trade-off between utility and fairness and allows finely controlling it through the regularization coefficient λ_r. This is more directly shown in Figure 3. By contrast, in the case of adversarial training, although fairness can be increased to some extent, the dependence of utility and fairness on the adv. gradient factor λ_a is complicated and unpredictable. It is thus difficult to select a desired point in the trade-off, and the corresponding evaluations appear as a disorderly point cloud on Figure 2.

Furthermore, for the same values of fairness, Figure 2 shows that CODER can achieve substantially higher utility (evaluation points lie higher and to the right of all baseline points). Given that AdvBERT is the hitherto state-of-the-art method for this task and dataset, our neutrality regularization method based on CODER therefore achieves the new state-of-the-art performance.

[3] AdvBERT requires a much longer time because it is a model based on much slower self-attention over the concatenated query and document, and because it is fine-tuned from the standard NLP version of BERT-Mini, as opposed to TAS-B, an encoder pretrained for retrieval on MS MARCO.

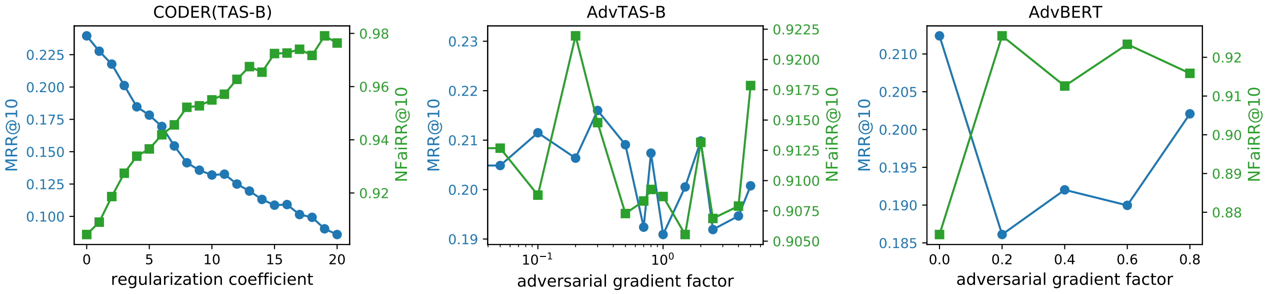

Figure 3: Modulation of utility and fairness by controlling the regularization factor (CODER) or the factor of adversarial gradient (AdvBERT, AdvTAS-B). Regularization with CODER allows to select the utility vs. fairness trade-off in a finely controllable and predictable manner.

In the limited range of NFAiRR@10 ∈ [0.90, 0.93] we find the top 4 points in terms of F1 scores (harmonic mean between MRR@10 and NFaiRR@10) for each method and display statistics of performance metrics in Table 1. NFaiRR itself is provided only as a reference in this table, since it acts as the control parameter and the range is merely chosen to have overlapping values.

Besides unpredictability with respect to the bias mitigation control factor, the performance of models undergoing adversarial training fluctuates haphazardly during training and it is thus very difficult to know when the model has a potential to further improve its performance or when to stop training. In practice, one resorts to using a fixed training time. By contrast, in Figure 4 we observe that regularization with CODER shows smooth convergence patterns and allows setting a stopping criterion based on monitoring performance, which in turn allows an adaptable training time for each regularization coefficient and can yield better performance (although for the sake of comparison we didn't use this technique in this work).

Finally, we note that Rekabsaz et al. [31] also introduced $\text{TRECDL}_{\text{FAIR}}$, a subset of the TREC Deep Learning Track 2019 queries, but discovered that this set is not very challenging, and all methods they examined performed well. Indeed, we find that CODER(TAS-B) with $\lambda_r = 1$ can attain an MRR@10 score of 1.0 at a NFaiRR@10 score of 0.967, and when boosting NFaiRR@10 to 0.985 with $\lambda_r = 8$, MRR@10 only drops to 0.944, which is a stronger performance than all reported methods in Rekabsaz et al. [31].

5 CONCLUSION

We introduce a novel method for reducing bias in search results, which is based on directly imposing a neutrality regularization loss to the documents most highly scored for relevance. To achieve this, we leverage a contextual document embedding reranking framework, which, for the same query, jointly scores a large set of retrieved candidate documents that together constitute a retrieval context. We demonstrate that our method can lead to much stronger bias mitigation/fairness compared to the existing alternatives for deep neural retrieval architectures, which are based on adversarial training. At the same time, it achieves the state-of-the-art performance with respect to utility (relevance) for the same amount of bias mitigation. Finally, our method allows for a more finely controllable and predictable intensity of bias mitigation, which is of paramount importance with respect to widespread adoption.

		F1	NFaiRR@10	MRR@10	Recall@10
Mean	CODER(TAS-B)	**0.356**	0.915	0.222	0.429
	AdvTAS-B	0.351	0.911	0.217	0.381
	AdvBERT	0.318	0.919	0.193	0.402
Median	CODER(TAS-B)	**0.358**	0.914	0.223	0.428
	AdvTAS-B	0.346	0.911	0.214	0.386
	AdvBERT	0.316	0.920	0.191	0.401
Max	CODER(TAS-B)	**0.379**	0.927	0.240	0.450
	AdvTAS-B	0.348	0.915	0.231	0.394
	AdvBERT	0.331	0.926	0.202	0.435

Table 1: Comparison of ranking performance based on top 4 F1 scores within the NFaiRR range [0.9 - 0.93].

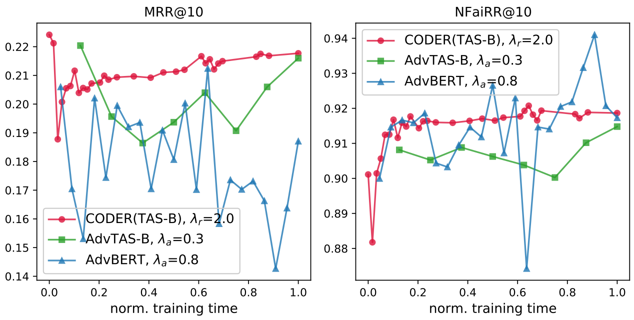

Figure 4: Utility (left) and fairness (right) on `MSMARCO dev` as they evolve during training, for a particular setting of regularization or adversarial factor, chosen such that the models perform comparably. Training times are significantly shorter for CODER and thus the horizontal axis is normalized in the same range.

ACKNOWLEDGMENTS

G. Zerveas would like to acknowledge the Onassis Foundation for financial support. This research is supported in part by the NSF (IIS-1956221), the State of Upper Austria and the Austrian Federal Ministry of Education, Science, and Research (LIT-2021-YOU-215). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, the U.S. Government, or the Austrian Government.

REFERENCES

[1] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 27–37. https://doi.org/10.1145/3406522.3446023

[2] 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. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268 [cs]* (Oct. 2018). http://arxiv.org/abs/1611.09268 arXiv: 1611.09268.

[3] 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.

[4] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. 2021. On the Orthogonality of Bias and Utility in Ad hoc Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1748–1752.

[5] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning (ICML '07)*. Association for Computing Machinery, New York, NY, USA, 129–136. https://doi.org/10.1145/1273496.1273513

[6] 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.

[7] Daniel Cohen. 2021. Allowing for The Grounded Use of Temporal Difference Learning in Large Ranking Models via Substate Updates. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 438–448. https://doi.org/10.1145/3404835.3462952

[8] Daniel Cohen, Bhaskar Mitra, Katja Hofmann, and W. Bruce Croft. 2018. Cross Domain Regularization for Neural Ranking Models Using Adversarial Learning. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 1025–1028. https://doi.org/10.1145/3209978.3210141

[9] Fabio Crestani, Mounia Lalmas, Cornelis J. Van Rijsbergen, and Iain Campbell. 1998. "Is This Document Relevant?…Probably": A Survey of Probabilistic Models in Information Retrieval. *ACM Comput. Surv.* 30, 4 (dec 1998), 528–552. https://doi.org/10.1145/299917.299920

[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* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3340531.3411962

[11] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3404835.3462851

[12] Robert Epstein and Ronald E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (Aug. 2015), E4512–E4521. https://doi.org/10.1073/pnas.1419828112 Publisher: Proceedings of the National Academy of Sciences.

[13] Christian Ganhör, David Penz, Navid Rekabsaz, Oleg Lesota, and Markus Schedl. 2022. Mitigating Consumer Biases in Recommendations with Adversarial Training. In *Proceedings of the 45th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2022*. ACM.

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

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

[16] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy J. Lin, and A. Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. *SIGIR* (2021). https://doi.org/10.1145/3404835.3462891

[17] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[18] 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.

[19] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *arXiv:2004.12832 [cs]* (June 2020). http://arxiv.org/abs/2004.12832 arXiv: 2004.12832.

[20] Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2022. Grep-BiasIR: A Dataset for Investigating Gender Representation-Bias in Information Retrieval Results. *arXiv preprint arXiv:2201.07754* (2022).

[21] Klara Krieg, Emilia Parada-Cabaleiro, Markus Schedl, and Navid Rekabsaz. 2022. Do Perceived Gender Biases in Retrieval Results Affect Relevance Judgements?. In *Proceedings of the Workshop on Algorithmic Bias in Search and Recommendation at the European Conference on Information Retrieval (ECIR-BIAS 2022)*.

[22] Oleg Lesota, Navid Rekabsaz, Daniel Cohen, Klaus Antonius Grasserbauer, Carsten Eickhoff, and Markus Schedl. 2021. A modern perspective on query likelihood with deep generative retrieval models. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 185–195.

[23] Chen Liu, Mathieu Salzmann, Tao Lin, Ryota Tomioka, and Sabine Süsstrunk. 2020. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. *Advances in Neural Information Processing Systems* 33 (2020), 21476–21487.

[24] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, Dense, and Attentional Representations for Text Retrieval. *Transactions of the Association for Computational Linguistics* 9 (April 2021), 329–345. https://doi.org/10.1162/tacl_a_00369

[25] 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, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. ACM, 429–438. https://doi.org/10.1145/3397271.3401100

[26] Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage Re-ranking with BERT. *arXiv:1901.04085 [cs]* (April 2020). http://arxiv.org/abs/1901.04085 arXiv: 1901.04085.

[27] Harrie Oosterhuis. 2021. Computationally Efficient Optimization of Plackett-Luce Ranking Models for Relevance and Fairness. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[28] 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.

[29] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 209–216. https://doi.org/10.1145/3121050.3121074

[30] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. *arXiv:2010.08191 [cs]* (May 2021). http://arxiv.org/abs/2010.08191 arXiv: 2010.08191.

[31] 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*.

[32] Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. 2021. TripClick: The Log Files of a Large Health Web Search Engine. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 2507–2513. https://doi.org/10.1145/3404835.3463242

[33] 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.

[34] Navid Rekabsaz, Robert West, James Henderson, and Allan Hanbury. 2021. Measuring Societal Biases from Text Corpora with Smoothed First-Order Co-occurrence. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*. AAAI Press, 549–560.

[35] Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (2021), 2173–2183. https://doi.org/10.18653/v1/2021.findings-acl.191 arXiv: 2108.06027.

[36] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]* (Feb. 2020). http://arxiv.org/abs/1910.01108 arXiv:

1910.01108.

[37] 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.

[38] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[39] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 5427–5437. https://proceedings.neurips.cc/paper/2019/hash/9e82757e9a1c12cb710ad680db11f6f1-Abstract.html

[40] 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]

[41] Ryen W. White and Eric Horvitz. 2015. Belief Dynamics and Biases in Web Search. *ACM Transactions on Information Systems* 33, 4 (May 2015), 18:1–18:46. https://doi.org/10.1145/2746229

[42] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *arXiv:2007.00808 [cs]* (Oct. 2020). http://arxiv.org/abs/2007.00808 arXiv: 2007.00808.

[43] 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 (SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 1253–1256. https://doi.org/10.1145/3077136.3080721

[44] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference*. 2849–2855.

[45] George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2021. CODER: An efficient framework for improving retrieval through COntextual Document Embedding Reranking. arXiv:2112.08766 [cs.IR]

[46] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing Dense Retrieval Model Training with Hard Negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Virtual Event Canada, 1503–1512. https://doi.org/10.1145/3404835.3462880

[47] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. RepBERT: Contextualized Text Embeddings for First-Stage Retrieval. *arXiv:2006.15498 [cs]* (July 2020). http://arxiv.org/abs/2006.15498 arXiv: 2006.15498.