Machine Learning (2025) 114:121
https://doi.org/10.1007/s10994-024-06664-2

Gender disentangled representation learning in neural rankers

Shirin Seyedsalehi[1,2] · Sara Salamat[2] · Negar Arabzadeh[3] · Sajad Ebrahimi[4] · Morteza Zihayat[2] · Ebrahim Bagheri[1]

Received: 30 May 2024 / Revised: 14 August 2024 / Accepted: 13 December 2024

Abstract
Recent studies have demonstrated that while neural ranking methods excel in retrieval effectiveness, they also tend to amplify stereotypical biases, especially those related to gender. Current mitigation strategies often focus on adjusting training methods, like adversarial techniques or data balancing, but typically overlook explicit consideration of gender as an attribute. In this paper, we introduce a systematic approach that treats gender as a distinct component within neural ranker representations. Our neural disentanglement method separates content semantics from gender information, enabling the neural ranker to evaluate document relevance based on content alone, without the interference of gender-related information during retrieval. Our extensive experiments demonstrate that: (1) our disentanglement approach matches the effectiveness of baseline models and offers more consistent performance across queries of different gender affiliations; (2) isolating gender within the representations allows the neural ranker to produce an unbiased list of documents, not favoring any specific gender; and (3) the disentangled gender component effectively and concisely captures gender information independently from the semantic content.

Keywords Neural rankers · Information retrieval · Responsible AI · Gender bias

1 Introduction

Information Retrieval (IR) methods have traditionally relied on statistical models of language, also known as *language models*, to bridge the gap between query and search spaces (Banerjee & Han, 2009). As language models grow in complexity, many IR tasks, which were considered stubborn problems, have now become increasingly softer to address (Zhao et al., 2023). For instance, ad hoc retrieval, a key IR task that involves identifying and ranking relevant documents based on a user's free-form query, previously struggled with challenges like vocabulary mismatch (Zhao & Callan, 2010). The adoption of Large Language Models (LLMs) and advanced fine-tuning techniques has resulted in '*neural rankers*', which significantly enhance retrieval effectiveness (Zhou et al., 2023). The

Editors: Kee-Eung Kim, Shou-De Lin.

Extended author information available on the last page of the article

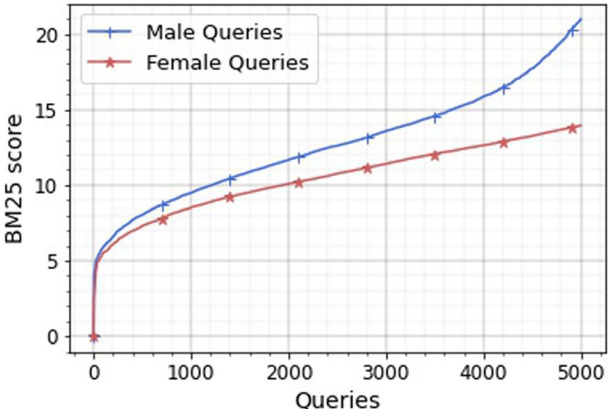

Fig. 1 The distribution of queries and their BM25 scores calculated with their relevance judgement

tangible evidence of the impact of such innovative methods can be observed on the standard MS MARCO passage retrieval benchmark (Nguyen et al., 2016) based on which the mean reciprocal rank measure has, as a result, increased from ∼ 0.19 to 0.45+, a 2.3 times increase in performance.

Despite the significant improved retrieval effectiveness, researchers have also observed that neural rankers have the potential to intensify various forms of stereotypical biases, most notably those related to gender. The observed impact in the context of gender bias has been both quantitative as well as qualitative. For instance, authors such as Rekabsaz and Schedl (2020); Bigdeli et al. (2021), and Zerveas et al. (2021), among others (Krieg et al., 2023; Seyedsalehi et al., 2022; Bigdeli et al., 2021), have shown that neural rankers (1) exhibit disproportionate affinity towards a certain gender identity during the retrieval process, and (2) the retrieval effectiveness of the results retrieved for queries affiliated with a certain gender (the male gender in particular) is far superior to others. As with many other supervised machine learning models (Feldman & Peake, 2021; Cabrera et al., 2019; Latif et al., 2023), the reason for such behavior can, at least in part, be attributed to the gender biases that are encoded in the training data used to train neural rankers.

To demonstrate biases in training data, we analyzed 10,000 evenly distributed and randomly selected male and female affiliated queries from the MS MARCO passage ranking task. Results showed that only about 16% of male queries returned documents predominantly affiliated with females, while over 22% of female queries returned predominantly male-affiliated documents, indicating a 6% higher likelihood for female queries to retrieve male-affiliated content. This suggests that neural rankers, once trained, exhibit a preference for male-affiliated documents, irrespective of the query's gender affiliation. Furthermore, male-affiliated queries consistently show a higher similarity score (BM25) with their relevant documents than female-affiliated queries, as depicted in Fig. 1, making them practically easier to retrieve and satisfy.

These observations, alongside earlier research (Rekabsaz et al., 2021; Bigdeli et al., 2022; Seyedsalehi et al., 2022) is the foundation for the work presented in this paper. The *main objective* is to mitigate gender biases in neural rankers while preserving their retrieval effectiveness. We hypothesize that isolating gender as an attribute from query and document representations could reduce bias. If gender is not encoded in the representations, it cannot intensify gender biases, given the observed disparities in relevance judgements and varying similarity scores across different gender affiliations. To this end, we

propose a neural architecture that disentangles gender from content semantics when encoding a query or a document. In the disentanglement process, the neural representation is broken down systematically into two distinct and independent components, one of which captures the semantics of the content, while the other encapsulates the gender affiliation of the query or the document. The *concrete contributions of this paper* can be enumerated as follows:

1. We propose a neural ranking architecture that disentangles content semantics from gender affiliation information and offers two independent representation components that encode each of these aspects separately;
2. Given the disentangled representations for queries and documents, we propose to use the content semantics component of the disentangled representation to rank-order documents in relation to the query and minimize the influence of gender and any associated biases in the ranking process;
3. Our extensive experiments demonstrate that: (1) the disentanglement process significantly reduces stereotypical gender biases in retrieved documents; (2) this reduction does not compromise retrieval effectiveness but rather enhances it; and (3) disentangling neural representations improves performance parity across various gender affiliations and query subsets.

2 Related work

Recent research has focused on identifying and reducing gender biases in neural rankers. Rekabsaz and Schedl (2020) were among the first to identify this issue in such systems. Their paper investigates the presence and extent of gender bias in neural ranking models. The authors develop a framework to measure gender bias, introducing two metrics to quantify gender bias in the ranked list of retrieved documents. Their work offers a dataset of gender-neutral queries and employs it to evaluate various models, including the baseline BM25 method and several neural ranking models. Their findings show that all models exhibit a male bias, but neural models, especially those using contextualized embeddings like BERT, significantly amplify this bias. The study also reveals that transfer learning with pre-trained embeddings tends to increase gender bias in neural rankers. The work by Rekabsaz can be considered a pioneering work that highlights the need to reduce gender biases in neural rankers while maintaining their retrieval effectiveness.

Subsequently, Bigdeli et al. (2021) investigate the presence of gender biases in gold standard relevance judgment datasets used for training and evaluating neural rankers. Since these relevance judgment datasets greatly influence how neural rankers learn the concept of relevance, the authors focused on quantifying and analyzing gender biases in relevance judgments. The authors use a fine-tuned BERT model to label a large collection of queries within the MS MARCO dataset (Nguyen et al., 2016), which were then used to assess the associated documents for their psychological characteristics using the Linguistic Inquiry and Word Count (LIWC) toolkit (Pennebaker et al., 2001). Their findings showed that stereotypical biases are common in relevance judgment collections, particularly with regards to affective and cognitive processes, as well as personal concerns and drives. Bigdeli et al advocate for the need for unbiased gold standard relevance judgement datasets that can avoid training biased neural rankers. Based on the findings from Rekabsaz et al. (2021) and Bigdeli et al. (2022) that showed neural rankers are susceptible to intensifying gender

biases, follow up work has been focused on developing methods that can reduce or eliminate such biases. The authors in Rekabsaz et al. (2021) attempt to eliminate gender information from the intermediate vector representation produced by BERT. Their proposed architecture comprises a BERT encoder and two classifier heads. One of the classifiers acts as an adversarial network, designed to discourage BERT from encoding gender information in its internal representations. This adversarial network is trained to predict gender, while BERT's encoder is trained to minimize this prediction accuracy by making its internal representations less informative. Using the adversarial framework, the network aims to maximize relevance prediction while minimizing the prediction of gender labels. This approach allows the encoder to gradually exclude gender information from the intermediate vector representation, preventing the gender classifier head from being able to predict gender from the vector representation.

Another relevant work (Bigdeli et al., 2021) explores the commonly held belief that reducing bias in ranker systems comes at the cost of utility (retrieval effectiveness). The authors propose a bias-aware pseudo-relevance feedback framework that aims to revise input queries to maintain or improve retrieval utility while significantly reducing bias. The paper demonstrates that it is possible to reduce bias without compromising retrieval effectiveness. This work challenges the traditional view of bias and utility as competing aspects and suggests that they can be addressed concurrently. Although the method is effective in reducing gender biases while maintaining the performance, the method is limited to non-neural models such as BM25.

Furthermore, SeyedSalehi et al. have proposed an approach to mitigate gender biases in neural ranking systems (Seyedsalehi et al., 2022). Their approach aims to mitigate gender biases in search results by introducing a bias-aware neural ranking approach. The proposed method explicitly incorporates a penalty for gender bias while maintaining retrieval effectiveness. The core idea involves ranking documents by learning their relevance to a given query while penalizing those documents that display gender biases, particularly those that are irrelevant to the query. This is achieved by incorporating a bias term into the ranking model loss function. By penalizing the relevance of irrelevant biased documents, the model learns to rank them lower while still prioritizing relevant documents, thus reducing bias in search results.

The work by Zerveas et al. (2022) introduced a novel approach to mitigate bias in neural rankers through an end-to-end differentiable, transformer-based framework called COntextual Document Embedding Reranking (CODER), which optimizes document relevance scores while simultaneously imposing neutrality regularization. CODER uses a transformer query encoder that scores a set of candidate documents collectively rather than in isolation, achieving contextual ranking. For bias mitigation, a regularization loss penalizes high-scoring documents that deviate from neutrality with respect to gender. The neutral ranking objective is achieved by comparing the distribution of scores against ideal, unbiased ranking scores. The authors show that CODER provides a smoother and more predictable bias mitigation process.

Different from other work that focus on adjusting the training model or the model architecture, the recent work by Bigdeli et al. (2022) addresses the problem of gender bias in neural retrieval models by proposing a simple and effective training data sampling strategy. The authors suggest incorporating the degree of gender bias when sampling documents for training neural rankers, allowing these models to maintain retrieval effectiveness while reducing gender biases. This strategy involves a systematic negative sampling approach that exposes neural rankers to biased documents, teaching them to avoid gender biases without architectural changes to neural rankers. This approach is notable for its simplicity

and efficacy, offering a practical solution for reducing gender biases while being applicable to a range of neural rankers.

These existing methods can be broadly categorized into three main classes: 1) *Data-Driven Debiasing*, 2) *Loss Function Regularization*, and 3) *Adversarial Training*. The approach presented by Bigdeli et al. (2022) exemplifies a data-driven debiasing strategy, where biases are directly addressed within the training data itself. Loss function regularization techniques can be further divided into two subcategories: the first subcategory includes methods such as the one proposed by Zerveas et al. (2021), which introduces an interpolated loss in order to consider bias during training. The second subcategory involves methods such as the one by Seyedsalehi et al. (2022), where a regularizer is applied to the loss function to mitigate biases. In the realm of adversarial training, the method proposed by Rekabsaz et al. (2021) is among the first to consider adversarial models for mitigating bias. It aims to completely remove the gender attribute from the intermediate representation of query-document pairs in neural rankers. Our proposed approach has close affinity with this paper. However, our proposed goes beyond the work by Rekabsaz et al. by introducing a novel multi-tasking loss. This loss function adaptively disentangles gender from the ranking representation during the multi-task training process, enhancing the model's ability to ensure gender neutrality in the ranked results.

Unlike existing solutions that often modify training data or algorithmic behavior, our approach ensures that gender information is isolated and not used in the relevance determination process. Consequently, our neural rankers operate without leveraging gendered assumptions and mitigate biases without sacrificing retrieval effectiveness

3 Proposed approach

3.1 Preliminaries

Neural rankers: Given a set of queries, denoted by $Q = \{q_1, q_2, \dots, q_n\}$, and a corresponding pool of documents represented by $D = \{d_1, d_2, \dots, d_m\}$, a neural ranker, Φ, employs a neural network architecture with a set of parameters θ to rank documents in D in relation to queries in Q. The neural ranker generates a ranked list R of documents by evaluating the relevance of each document to a given query. This is achieved by calculating a relevance score $s = \Phi(q, d)$ for each query-document pair (q_i, d), where $q_i \in Q$ and $d \in D$. Since neural rankers are supervised methods, during the training process, their parameters θ are optimized to improve the ranker's ability to accurately reflect the relevance of documents in relation to input queries.

Neural ranking architectures: A neural ranker Φ often consists of two components: (i) an encoder, and (ii) a scoring mechanism. The encoder, which is typically a large language model (LLM), processes the inputs to generate vector representations for queries and documents. Within a cross-encoder architecture (Reimers & Gurevych, 2019), the vector representation of the query q and document d are often concatenated, which can be expressed as:

$$E = \text{encoder}(q \oplus d) \tag{1}$$

where $\oplus$ denotes the concatenation operator. Subsequently, a multi-layer feedforward network is employed as the scoring mechanism. It takes the vector E and computes the relevance scores used to rank documents in relation to the input query q.

Training neural rankersL: A neural ranker Φ is often trained using a pairwise training process (Burges et al., 2005), which adopts a contrastive learning strategy (Zou et al., 2013). This strategy ensures that vectors representing queries are placed closer to those of their relevant documents and placed furthest away from those of their irrelevant documents within the vector space. This objective is achieved through a *marginal ranking loss* function, as follows:

$$L = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \Phi(q, d_i^+) + \Phi(q, d_j^-)) \tag{2}$$

where d_i^+ and d_j^- denote relevant and an irrelevant document, respectively, relative to the query q. Furthermore, N^+ and N^- represent the total number of relevant and irrelevant documents, and n is the total number of training samples across all queries. This loss function helps guide the training process, enabling the neural ranker Φ to better distinguish between relevant and irrelevant documents for each query.

3.2 Problem definition

When considering the issue of gender, user queries can be broadly categorized in two classes, namely (i) *gender-neutral queries*, and (ii) *gender-specific queries*. Gender-neutral queries are those queries, which seek information that can be answered independently of gender considerations and include examples such as 'what happened in cabo shooting' and 'what is early childhood studies'. In contrast, gender-specific queries will need to take gender as a consideration when effectively addressing the query. Examples of such queries include 'when can you feel signs of pregnancy' (female-affiliated query) and 'what is age for prostate cancer' (male-affiliated query). Table 1 provides further examples of gender-specific and gender-neutral queries as provided by Rekabsaz and Schedl (2020). The objective of our work is to ensure that a neural ranker Φ is fair when dealing with these two different types of queries. We adopt the definition of gender fairness as laid out by earlier work (Seyedsalehi et al., 2022; Bigdeli et al., 2023; Krieg et al., 2022), and formulate them as follows:

Fairness for gender-neutral queries: A neural ranker would be deemed fair when processing gender-neutral queries if the retrieved ranked list of documents would not exhibit any predispositions towards any specific gender. This principle is predicated on the understanding that a query devoid of gender-related cues should yield a set of documents whose relevance is determined independently of gender implications (Kopeinik et al., 2023; Rekabsaz and Schedl, 2020; Bigdeli et al., 2021). For instance, the query 'how can one become

Table 1 Sample queries and their gender affiliations from Rekabsaz and Schedl (2020)

Query gender affiliation	Query
Female	actress who born at litrhuania
	what does it mean when you bleed before period
Male	who is king philip of spain
	is king arthur real or legend?
Neutral	where is kobenhavn
	what is hemianopsia

an engineer?' is gender-neutral, as the path to becoming an engineer is independent of the individual's gender. In such a case, one would expect to receive a ranked list of documents that does not carry any preconceived notions of gender preference in relation to the engineering profession. In order to quantitatively assess the fairness of a ranked list of documents, denoted as R_q, in relation to a gender-neutral query q, researchers have assumed that a function $\Psi(R_q)$ can be formulated for measuring the extent of gender bias manifested by R_q (Rekabsaz & Schedl, 2020; Rekabsaz et al., 2021; Abolghasemi et al., 2024) where lower values of $\Psi(R)$ depict increased degrees of fairness. For a ranking R_q to be considered fair in response to a gender-neutral query, the ideal outcome would be:

$$\Psi(R_q) \rightarrow 0 \tag{3}$$

This symbolizes the expectation that the ranked list of documents for a gender-neutral query q should approach a state of gender parity, where *ideally* no discernible bias in favor of any gender is observable.

Fairness for gender-specific queries: When processing gender-specific queries, a neural ranker would be deemed fair if its ability to effectively rank documents does not vary based on the gender of the query. This concept posits that the performance of a neural ranker, can be quantitatively assessed using a performance metric $\lambda(Q)$, where a higher $\lambda(Q)$ indicates superior model performance on the query set Q. For a neural ranker to be considered fair under this definition, it must exhibit comparable performance levels for queries belonging to different gender affiliations, such as male-affiliated queries (Q_m), or female-affiliated queries (Q_f), essentially satisfying the following condition:

$$\lambda(Q_m) \approx \lambda(Q_f) \tag{4}$$

This definition emphasizes the need for promoting a system that treats all queries equally without bias towards any gender association.

In summary, a fair ranker should not exhibit stereotypical biases toward both *gender-neutral* and *gender-specific* queries. For gender-neutral queries, the goal is to reduce, and ideally remove, biases in the document retrieval process, as shown in Eq. 3: $\Psi(R_q) \rightarrow 0$. For gender-specific queries, the ranker should demonstrate comparable retrieval effectiveness across different gendered queries, as detailed in Eq. 4: $\lambda(Q_m) \approx \lambda(Q_f)$.

3.3 Overview of the disentanglement approach

Neural rankers order documents by the similarity between their vector representations and those of user queries. It has been empirically demonstrated (Rekabsaz and Schedl, 2020) that these vectors often encode gender preferences, which can intensify biases. Consequently, these biases are implicitly considered during the ranking process. Therefore, our hypothesis is that by excluding gender information from the vector representations of queries and documents, the neural ranker will be unable to access, even implicitly, any gender data during the ranking process, thereby preventing the intensification of biases.

To this end, we propose to *'disentangle'* query and document vector representations into discernible components dedicate to gender and semantic sub-vectors. The gender sub-vector would be responsible for capturing possible gender information in the query or document, whereas the semantic component would, independently of the gender information, only represent the content value of the query or document. Disentangled representation learning (Bengio et al., 2013) aims to create factorized representations that isolate

underlying factors of input data. In our work, we focus on separating content gender from content semantics. We propose that separating gender-related information from E can debias neural ranking outputs. Thus, we disentangle E into two components: E_r, containing all content semantics and non-gender factors, and E_g, representing gender-related information. Let Γ be a function that disentangles a vector E of size d into two components of sizes m and n such that:

$$E_r, E_g = \Gamma(E, m, n), \quad d = m + n \tag{5}$$

Based on Eq. 5, we propose using E_r for ranking while excluding E_g, potentially reducing the gender biases inherent in neural rankers. By omitting E_g from the ranking process, we suggest that gender will no longer influence query performance or the makeup of the ranked results.

3.4 Neural architecture for gender disentanglement

Figure 2 shows our architecture for disentangling gender from content semantics in neural rankers. It includes two distinct networks: the *Ranking Network* and the *Gender Network*. The Ranking Network processes only the semantic subvector, E_r, learning the relevance between queries and documents. The Gender Network refines the gender-specific subvector, E_g, to predict gender attributes accurately. When trained together, these networks separate content into E_r and gender information into E_g, effectively disentangling the two.

The ranking network: This network is designed to capture and learn the concept of relevance between queries and documents. It operates by only processing the semantic component of the original vector, E_r, using a Multi-Layer Perceptron (MLP), denoted as MLP_r, to predict relevance scores for query-document pairs. During the training process, this network generates relevance scores for both relevant and irrelevant documents associated with each query in the dataset, as defined in the following:

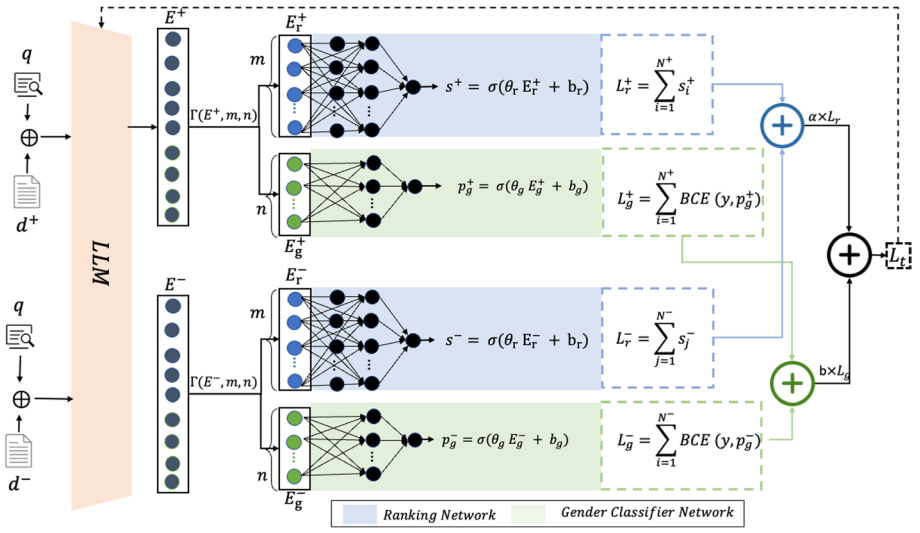

Fig. 2 Overview of the proposed neural disentanglement architecture

$$s = \sigma(\theta_r E_r + b_r), \tag{6}$$

where σ is the activation function, and $\Theta = \theta_r \cup b_r$ are the parameters of the MLP responsible for predicting the relevance score s. To enhance the model's discrimination capabilities, we embed a contrastive loss function that aims to increase the relevance scores for matches between queries and their corresponding relevant documents while reducing the scores for mismatches with irrelevant documents. The loss function can be formulated as follows:

$$L^+ = \sum_{i=1}^{N^+} \sigma(\theta_r E^+_{r_i} + b_r), \quad L^- = \sum_{j=1}^{N^-} \sigma(\theta_r E^-_{r_j} + b_r) \tag{7}$$

Therefore,

$$L_r = \frac{1}{n} \sum \max(0, m - L^+ + L^-) \tag{8}$$

Given L_r, θ_r and b_r are updated as follows:

$$\theta_r^{(t+1)} = \theta_r^{(t)} - \eta \frac{\partial L_r}{\partial \theta_r}, \quad b_r^{(t+1)} = b_r^{(t)} - \eta \frac{\partial L_r}{\partial b_r} \tag{9}$$

where η is the learning rate, and t denotes the iteration number. This approach allows the network to accurately identify and enhance the relevance of query-document pairs, thus optimizing the performance of the neural ranker.

The gender network: This network specifically targets the gender-specific subvector of the neural ranker's intermediate vector, E_g, to predict gender attributes. The network utilizes a Multilayer Perceptron (MLP), denoted as MLP_g, which processes E_g to estimate the probability of document or query gender affiliation as follows:

$$p_g = \sigma(\theta_g E_g + b_g), \tag{10}$$

where σ is the activation function, θ_g are the weights, and b_g the bias of the MLP_g.

To obtain accurate gender affiliations, a function Λ is assumed, which can identify the gender affiliation of a text t:

$$g = \Lambda(t) \tag{11}$$

Here, g represents the gender affiliation of the text, t. The training of the Gender Network is governed by a Binary Cross Entropy loss function, formulated as:

$$L_g = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(p_{g_i}) + (1 - y_i) \log(1 - p_{g_i}) \right] \tag{12}$$

where N is the total number of instances, p_{g_i} is the predicted probability that the i-th instance belongs to a particular gender, and $y_i = \Lambda(q_i \oplus d_i)$ is the true label derived from the function Λ.

Simultaneous training of the ranking and gender networks disentangles gender and semantic information. This dual-training strategy divides the encoder's output, E, into two components: E_r for semantics and E_g for gender information. Consequently, the architecture assesses relevance using E_r and remains unbiased by gender influences

from E_g. The total loss function L_t is a linear combination of the ranking loss L_r and the gender classification loss L_g:

$$\begin{aligned} L_t =& \alpha \times \left(\frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \sigma(\theta_r E_{r_i}^+ + b_r) + \sigma(\theta_r E_{r_j}^- + b_r)) \right) \\ &+ \beta \times \left(-\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(p_{g_i}) + (1 - y_i) \log(1 - p_{g_i}) \right] \right) \end{aligned} \tag{13}$$

Algorithm 1 Training of the Disentangled Ranking Network

1: **Data:** $\{(q, d^+, d^-)\}$, number of training iterations T.
2: **Initialize:** θ_r, θ_g, b_r, b_g randomly.
3: $g^+ \leftarrow \Lambda(q \oplus d^+)$
4: $g^- \leftarrow \Lambda(q \oplus d^-)$
5: **for** $t = 1$ **to** T **do**
6: **for** each sample (q, d^+, d^-, g^+, g^-) in the batch **do**
7: $E^+ \leftarrow \text{encoder}(q \oplus d^+)$
8: $E^- \leftarrow \text{encoder}(q \oplus d^-)$
9: $E_r^+, E_g^+ \leftarrow \Gamma(E^+, m, n)$
10: $E_r^-, E_g^- \leftarrow \Gamma(E^-, m, n)$
11: $s^+ \leftarrow \sigma(\theta_r E_r^+ + b_r)$
12: $s^- \leftarrow \sigma(\theta_r E_r^- + b_r)$
13: $p_g^+ \leftarrow \sigma(\theta_g E_g^+ + b_g)$
14: $p_g^- \leftarrow \sigma(\theta_g E_g^- + b_g)$
15: $L_r \leftarrow \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - s^+ + s^-)$
16: $L_g \leftarrow -\frac{1}{N} \sum_{i=1}^{N} \left[g_i \log(p_{g_i}) + (1 - g_i) \log(1 - p_{g_i}) \right]$
17: $L_t \leftarrow \alpha \times L_r + \beta \times L_g$
18: $\theta^{(t+1)} = \theta^{(t)} - \eta \frac{\partial L_t}{\partial \theta}, \quad b^{(t+1)} = b^{(t)} - \eta \frac{\partial L_t}{\partial b}$
19: **end for**
20: **end for**

The final formulation of L_t clearly illustrates the model's balance between optimizing ranking performance and promoting gender fairness. Adjustable weights α and β enable fine-tuning to meet specific performance and fairness objectives.

3.5 Model training

To effectively train the model, we form a dataset comprising of samples $(q, d^+, d^-, gender^+, gender^-)$, where $gender^+$, and $gender^-$ denote the gender affiliation of the relevant, and irrelevant documents, respectively. The training procedure is illustrated in Algorithm 1, which begins by initializing the dataset with query-document pairs (q, d^+, d^-) and settings the number of training iterations T. Initially, the network parameters θ_r, θ_g, b_r, and b_g, are initialized randomly. Subsequently, true gender affiliations are computed for both relevant and irrelevant documents by applying function Λ to the concatenated pairs, resulting in g^+ and g^- as described in Lines 3 and 4 of the algorithm. The primary training loop executes over T iterations. Each iteration processes a batch of samples, which includes both query-document pairs and their corresponding gender affiliations. In each iteration the following steps are taken: **(Step 1)** The query-document pairs are transformed into vector representations E^+ and E^- (Lines 7 and 8). **(Step 2)** The representations are then split

into components dedicated to ranking (E_r^+ and E_r^-) and gender (E_g^+ and E_g^-) (Lines 9 and 10). **(Step 3)** Relevance scores are computed from the ranking components (Lines 11 and 12), while gender affiliations are estimated from the gender components (Lines 13 and 14). **(Step 4)** The algorithm computes the ranking loss L_r using a hinge loss in Line 15. The gender loss L_g is calculated using binary cross-entropy (Line 16), and **(Step 5)** The total loss L_t is computed as a linear interpolation of both losses, as shown on Line 17. Finally, the parameters θ_r, θ_g, b_r, and b_g are updated on Line 18 by minimizing the total loss.

4 Experiments

4.1 Datasets and setup

To train our neural rankers, we utilize the MS MARCO passage ranking dataset (Nguyen et al., 2016), which contains approximately 200,000 queries and 8.8 million passages. For training purposes, we select 200,000 random samples of query triples $(query, doc^+, doc^-)$ to teach the model effectively. The models are trained for five epochs using the Adam optimizer and a sigmoid activation function. Additional implementation details, as well as the code, are available in our publicly available GitHub repository.[1]

To evaluate model performance and measure the proposed model's effectiveness in reducing gender biases we require two sets of queries:

Gender-neutral queries: (Q_n): This set allows us to evaluate stereotypical gender biases for gender-neutral queries. In particular to test the condition set out in Eq. 3. When a gender-neutral query is fed to the model, it is expected that the ranked list does not show any inclination towards male, or female. We use two query sets proposed by Rekabsaz et al. The primary set (Rekabsaz et al., 2021) comprises 1,765 gender-neutral queries, annotated from a pool of 55,578 MS MARCO queries by three Amazon Mechanical Turk workers. The annotators flagged queries with words or phrases related to gendered concepts. The second set, contains 215 socially problematic queries that could potentially reinforce existing gender norms and propagate gender inequality if the search results are biased.

Gender-specific queries: (Q_g): This set is employed to evaluate the fairness for the gender-specific queries. In particular this query set is used to evaluate the condition set out in Eq. 4. We use the dataset labeled by Bigdeli et al. (2021). Their work involved training a BERT classifier with human-annotated queries, which they applied to the MS MARCO passage ranking development set. This labeling effort resulted in two sets: 1,405 male affiliated queries and 1,405 female affiliated queries. We evaluate the model performance on both male affiliated, and female affiliated queries in order to asses whether the model shows comparable performance over different genders.

4.2 Baselines and metrics

To evaluate our work against robust state-of-the-art baselines, we use five distinct methods: (1) Original Model: A cross-encoder model trained for the passage re-ranking task, using the OpenMatch implementation (Liu et al., 2021). (2) AdvBert(Rekabsaz et al.,

[1] https://github.com/genderdisen/genderdisen

2021)[2] Utilizes an adversarial strategy to eliminate gender data from the neural rankers' intermediate representations, replicated from their GitHub repository. (3) Bias-aware Penalty(Seyedsalehi et al., 2022): Incorporates a direct bias penalty in the neural ranker's loss function to explicitly address gender biases during training. (4) **CODER**(Zerveas et al., 2021)[2]: A transformer-based framework that evaluates document relevance collectively rather than individually and includes neutrality regularization to penalize deviations from gender neutrality. (5) Light-weight Sampling (Bigdeli et al., 2022): Employs a negative sampling strategy that selects the most biased documents as negative samples to train the model, teaching it to recognize and reduce bias.

To evaluate model performance, we assess ranking effectiveness and the degree of gender biases: i) **Ranking Effectiveness.** We use Mean Reciprocal Rank (MRR) to gauge the baseline models' performance. MRR calculates the average of reciprocal ranks for all queries, focusing on the rank of the first relevant result, with MRR@10 being the standard metric for the MS MARCO passage ranking task (Nguyen et al., 2016). ii) **Measuring Gender Biases.** We use three metrics to quantitatively assess each model's gender biases: a) **Average Rank Bias (ARaB)** (Rekabsaz & Schedl, 2020) measures the presence of gender-specific words in documents, using Term Frequency (TF) and Boolean methods to calculate gendered terms. b) **NFaiRR Metric** (Rekabsaz et al., 2021) evaluates fairness at the document level within ranked lists and across all queries, based on the concept of 'document neutrality', where a higher NFaiRR indicates a fairer ranking. c) **Linguistic Inquiry and Word Count (LIWC)** (Pennebaker et al., 2001) is employed to determine the gender affiliation of text using the social referents category, specifically the male and female reference subcategories, as outlined in Bigdeli et al. (2021).

4.3 Ranking effectiveness and bias mitigation evaluation

In our experiments, we evaluate the effectiveness of our proposed disentanglement approach in reducing stereotypical gender biases in neural rankers. We conduct experiments using two sets of gender-neutral queries, comprised of 215 and 1, 765 queries introduced in Sect. 4.1, respectively. To demonstrate the generalizability of our approach, we report the results based on the MiniLM Wang et al. (2020) language model in Tables 2 and 3, and BERT-Mini language model Bhargava et al. (2021) in Tables 4 and 5. Fig. 3 shows the training loss, and MRR on the development set queries for both of the base models. We can infer from the figure that as training goes on, the training losses are decreased, while the MRR of the development sets is increases, which shows that the model is trained properly, and it is not overfitted on the training data. As shown in Table 2, our model significantly outperforms the original model in the 215 query set, achieving a higher MRR of 0.1877 at Cut-off 10 compared to the original's 0.1602. Our model also shows considerable reductions in ARaB metrics: ARaB-tc decreases from 0.3183 to 0.0737, ARaB-tf from 0.1374 to 0.046, and ARaB-bool from 0.1101 to 0.0567, with the NFaIR score improving from 0.8107 to 0.8664. In contrast, the CODER model, while achieving lower ARaB values, only reaches an MRR of 0.0152, and the AdvBert model, despite lower ARaB values, significantly compromises retrieval effectiveness with an MRR of only 0.0093 at Cut-off

[2] The numbers reported in the table represent the best results obtained from their implementation. We verified these results with the authors through multiple meetings, during which they confirmed the accuracy of the very low MRR values.

Table 2 Gender bias measures for 215 neutral queries with MiniLM base model

	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFalR↓	LIWC↓
Original Model	0.1602	0.3183	0.1374	0.1101	0.8107	1.023
AdvBert	0.0093	0.0304	0.0022	0.0008	0.9854	0.1134
Bias-aware Penalty	0.167	0.1994	0.0867	0.0689	0.8453	0.8545
CODER	0.0152	0.0254	0.0240	0.0313	0.9336	0.4114
Ours	0.1877	0.0737	0.046	0.0567	0.8664	0.8404
Cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFalR↓	LIWC↓
Original Model	0.1658	0.2635	0.1142	0.092	0.8274	0.7966
AdvBert	0.0103	0.0289	0.0028	0.0016	0.9824	0.1101
Bias-aware Penalty	0.1722	0.1674	0.0725	0.0576	0.8564	0.6426
CODER	0.0155	0.0351	0.0263	0.0311	0.9348	0.3491
Ours	0.1941	0.0574	0.035	0.0422	0.8722	0.717

Table 3 Gender bias measures for 1765 neutral queries with MiniLM base model

	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFalR↓	LIWC↓
Original Model	0.2673	0.1535	0.0721	0.0611	0.7066	1.5599
AdvBert	0.0081	0.0051	0.0018	0.000267	0.9657	0.2374
Bias-aware Penalty	0.2814	0.0506	0.0256	0.0218	0.7396	1.4368
CODER	0.0021	0.1507	0.0721	0.0663	0.8404	0.7199
Our Approach	0.2969	0.0805	0.0131	0.0178	0.7623	1.4521
Cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFalR↓	LIWC↓
Original Model	0.2726	0.0721	0.0641	0.0538	0.722	1.3001
AdvBert	0.0099	0.0025	0.0003	0.0014	0.9642	0.2283
Bias-aware Penalty	0.2868	0.0256	0.0192	0.0152	0.7527	1.1809
CODER	0.0025	0.1490	0.0716	0.0663	0.8407	0.6467
Our Approach	0.3023	0.0131	0.0313	0.0052	0.7658	1.2767

10. This table demonstrates that our approach not only reduces bias but also enhances ranking effectiveness. Furthermore, at Cut-off 20, our model continues to show improvement, with an MRR of 0.1941 compared to the original model at 0.1658. The ARaB-tc value further decreased to 0.0574, and ARaB-tf to 0.035. The NFaIR score increased to 0.8722, again indicating reduced bias. Despite AdvBert's superior bias reduction, it's MRR remained noticeably low at 0.0103, reinforcing the trade-off between bias reduction and retrieval effectiveness in their work.

In the 1765 query set, as shown in Table 3, our model achieves a superior MRR of 0.2969 at Cut-off 10, outperforming the original model's 0.2673, illustrating our approach's effectiveness in enhancing retrieval while reducing biases. Notable improvements in bias metrics include ARaB-tc decreasing from 0.1535 to 0.0805, ARaB-tf from 0.0721 to 0.0131, and ARaB-bool from 0.0611 to 0.0178, with the NFaIR score rising from 0.7066

Table 4 Gender bias measures for 215 neutral queries with BERT-Mini base model

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.1662	0.2544	0.1058	0.0751	0.8273	0.8467
Advbert	0.0431	0.0308	0.0159	0.0155	0.9644	0.2943
Bias-aware Penalty	0.1714	0.2472	0.1025	0.0756	0.8389	0.8217
CODER	0.0014	0.0260	0.0171	0.0205	0.9649	0.2998
Our Approach	0.1399	0.0376	0.0132	0.0075	0.8583	0.6969
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR	LIWC↓
Original Model	0.1742	0.2318	0.0929	0.0646	0.8457	0.6964
Advbert	0.0487	0.0289	0.0151	0.015	0.9657	0.2474
EDBT	0.181	0.2331	0.0928	0.0662	0.8563	0.6448
CODER	0.0014	0.0228	0.0148	0.0178	0.9650	0.2828
Our Approach	0.1455	0.047	0.0158	0.0083	0.8691	0.5674

Table 5 Gender bias measures for 1765 neutral queries with BERT-Mini base model

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2475	0.1387	0.056	0.0369	0.7304	1.4942
AdvBert	0.0081	0.0051	0.0018	0.0003	0.9657	0.4403
Bias-aware Penalty	0.244	0.1374	0.0536	0.0334	0.7384	1.4474
CODER	0.7082e−4	0.0646	0.0371	0.0421	0.9093	0.5713
Our Approach	0.1922	0.0928	0.0354	0.026	0.7565	1.3468
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2548	0.1262	0.505	0.0329	0.7451	1.2592
Advbert	0.0099	0.0025	0.0003	0.0014	0.9642	0.4043
Bias-aware Penalty	0.2505	0.1138	0.0441	0.027	0.7583	1.1984
CODER	0.0001	0.0674	0.0388	0.0440	0.9096	0.4858
Our Approach	0.1996	0.0928	0.037	0.0285	0.7672	1.1682

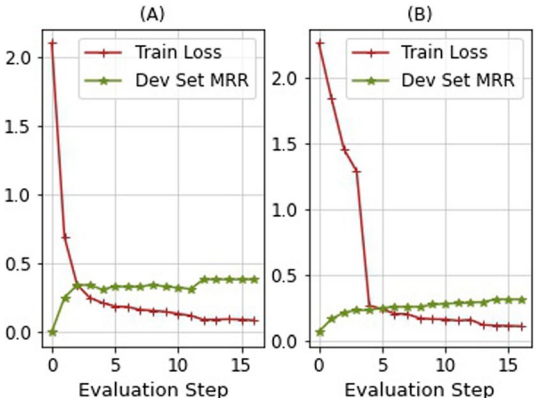

Fig. 3 The train loss, and MRR of the development set queries for (A) MiniLM base model, and (B) BERT-Mini base model

to 0.7623. The CODER model records lower ARaB but a significantly reduced MRR of 0.0021, while the AdvBert model, despite achieving low bias scores, suffers in performance with an MRR of only 0.0081 at Cut-off 10. At Cut-off 20, our model maintains its performance with an MRR of 0.3023, further reducing ARaB-tc to 0.0131 and ARaB-tf to 0.0313, with an increased NFaIR score of 0.7658, highlighting continued bias reduction. AdvBert's low bias metrics come with a trade-off in retrieval effectiveness, indicated by an MRR of just 0.0099.

It's worth noting that while the AdvBert model greatly reduces biases, it suffers a significant performance drop, limiting its practical use. The Bias-aware Penalty baseline offers moderate bias reduction with good performance, yet our model exceeds it in both bias reduction and ranking effectiveness. Similarly, the CODER baseline significantly reduces bias but has markedly lower ranking performance compared to our approach.

Using the BERT-Mini model, shown in Tables 4 and 5, similar trends are observed. In the 215 query set, our model achieves an MRR of 0.1399 at Cut-off 10, compared to the original's 0.1662. Despite a slight drop in ranking effectiveness, our model significantly mitigates bias, with ARaB-tc dropping from 0.2544 to 0.0376, ARaB-tf from 0.1058 to 0.0132, and ARaB-bool from 0.0751 to 0.0075. The NFaIR score improved from 0.8273 to 0.8583. The AdvBert model, though achieving lower ARaB values, suffers from drastically reduced performance, with an MRR as low as 0.0431 at Cut-off 10. At Cut-off 20, our model continues to improve, reducing ARaB-tc to 0.047 and ARaB-tf to 0.0158, and increasing the NFaIR score to 0.8691. However, AdvBert's performance remains low with an MRR of 0.0487, underscoring a substantial trade-off between bias reduction and retrieval effectiveness. In the 1765 query set, our model significantly improves bias metrics with ARaB-tc decreasing from 0.1387 to 0.0928, ARaB-tf to 0.0354, and ARaB-bool to 0.026. The NFaIR score rose from 0.7304 to 0.7565. Despite AdvBert achieving the lowest bias scores, its retrieval effectiveness is compromised, showing an MRR of only 0.0081 at Cut-off 10. At Cut-off 20, further reductions in bias metrics and an increase in NFaIR to 0.7672 continue, yet AdvBert's MRR remains low at 0.0099, highlighting its limited practical utility.

We point out that while the AdvBert model significantly reduces biases across all metrics, it does so with a marked decline in retrieval effectiveness. The Bias-aware Penalty baseline shows a moderate reduction in bias with relatively good performance. However, our model outperforms it in both bias reduction and retrieval effectiveness. In Appendix A, we have incorporated an adversary network into our architecture to intensify the penalization of gender information within the ranking representation. Detailed explanations, results, and discussions are provided there. Additionally, a case study example is presented in Appendix B to further illustrate the effectiveness of our proposed model in mitigating stereotypical gender biases.

4.4 Light-weight sampling strategy

As an additional baseline, Bigdeli et al. (2022) have suggested a negative sampling strategy, in which the negative documents are selected such that they exhibit large amount of bias. By doing so, the model will implicitly recognize bias as a negative factor during the training; therefore, biased documents will be sorted lower when re-ranked with the trained model. We adopt this negative sampling strategy, and select two negative samples from the proposed dataset, and train our gender disentanglement

Table 6 Bias measures for the light weight(LW) random samples proposed in Bigdeli et al. (2022) on 215 queries

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.1602	0.3183	0.1374	0.1101	0.8107	1.023
Original LW	0.1604	0.0269	0.0147	0.0141	0.9596	0.2913
Disentangled LW	0.1466	0.0164	0.0103	0.0133	0.983	0.1292
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.1658	0.2635	0.1142	0.092	0.8274	0.7966
Original LW	0.1659	0.0218	0.0123	0.0125	0.9595	0.3004
Disentangled LW	0.1554	0.0137	0.0076	0.0086	0.9788	0.1749

Table 7 Bias measures for the light weight (LW) random samples proposed in Bigdeli et al. (2022) on 1765 queries

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2673	0.1535	0.0721	0.0611	0.7066	1.5599
Original LW	0.2737	0.027	0.0136	0.0124	0.8810	0.8192
Disentangled LW	0.2393	0.0157	0.0032	0.0028	0.915	0.5915
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2726	0.0721	0.0641	0.0538	0.722	1.3001
Original LW	0.2795	0.0214	0.0109	0.0101	0.8802	0.7268
Disentangled LW	0.2464	0.0177	0.0049	0.0003	0.9091	0.5621

model with this bias-aware negative strategy. Given limited space, we report the results on the MiniLM language model in Tables 6, and 7 for the 215, and 1,765 queries.

When comparing the results of the original and disentangled models in both tables before and after the light-weight negative sampling strategy is used, we make three consistent improvements: (1) the negative sampling strategy does not lead to a drop in retrieval effectiveness on the base retrieval method but decrease in our disentanglement method is more pronounced on the MRR metric. This shows that When shown severely biased negative samples, the proposed disentanglement model cannot learn the concept of relevance as well as when random negative samples were selected; (2) on the other hand, the negative sampling strategy leads to notable reduction in bias in our proposed approach, which is superior to both the original base model as well as when negative sampling strategy was applied to the base retrieval method. This suggests, as also reported by Bigdeli et al. (2022) that the selection of the negative samples can lead to reduced bias. In summary, while the disentangled model consistently reduces bias metrics (ARaB-tc, ARaB-tf, ARaB-bool) and improves NFaIR and LIWC scores compared to the original model, this often comes at the cost of a slight decrease in MRR, which may be tolerable depending on the application area and the significance of the observed bias reduction.

4.5 Performance disparities

Besides stereotypical gender biases, disparities in retrieval effectiveness between male and female-affiliated queries are notable. As shown in the top row of Table 8, the original neural ranker performs significantly better on male queries than on female queries, with a 19% higher effectiveness at cut-off 10 and a similar disparity at cut-off 20. However, the results from our disentanglement approach, detailed in the second row of Table 8, offer the following observations: (1) Our approach reduces the performance disparity between male and female affiliated queries from 19% to 14%, a 5% improvement. This significant progress suggests the importance of addressing not only the stereotypical gender biases in document retrieval but also the disparities in retrieval effectiveness across different gendered queries. Otherwise, merely reducing gender biases without improving retrieval effectiveness could result in less biased but potentially irrelevant documents being retrieved. (2) Our approach reduces the performance disparity between male and female queries without sacrificing the performance of either group. Contrary to concerns raised in earlier studies (Seyedsalehi et al., 2022) that reducing gender biases might decrease retrieval effectiveness, our method actually enhances performance for both groups. Specifically, male-affiliated queries experience more than a 3.5% improvement, and female-affiliated queries improve by over 10%. We have also reported the standard deviation of the reciprocal ranks of the male, and female queries. The consistent standard deviation for both the original, and the disentangled models in cut-offs 10, and 20 implies that the system has become better at ranking relevant results higher on average (as indicated by the increased MRR), but the degree to which these rankings vary across different queries is the same as before. This could mean that the improvement in MRR is consistent across many queries.

4.6 Gender Disentanglement Quality

Evaluating gender bias in neural embeddings is challenging due to the absence of standardized measures, particularly for complex contextualized embeddings (Zhao et al., 2019; Basta et al., 2019). Previous research has utilized clustering and classification to identify gender information in embeddings, comparing debiased and non-debiased versions (Gonen & Goldberg, 2019; Bolukbasi et al., 2016). In our evaluation, we test the efficacy of our approach to separate gender from semantics using three established strategies: (1) analyzing occupational stereotypes (Basta et al., 2019; Zhao et al., 2019) and (2) detecting

Table 8 Performance on gender-specific queries

MRR@10	Male	Female	Δ
Original Model	0.3939 (std=0.3528)	0.3178 (std=0.3445)	19.3196
Disentangled Model	0.4093 (std=0.3485)	0.3511(std=0.3474)	14.2194
Improvement	3.90%	10.47%	−26.39%
MRR@20			
Original Model	0.4002 (std=0.3682)	0.3242 (std=0.3582)	0.1899
Disentangled Model	0.4146 (std=0.3597)	0.3572 (std=0.3610)	0.1384
Improvement	3.59%	10.17%	−27.11%

gender spaces in embeddings (Bolukbasi et al., 2016) and (3) Measuring Bias in Sentence Encoders.

4.6.1 Occupational stereotypes

Previous research has shown that neural embeddings often capture biases related to gender roles and professions, such as stereotypically associating engineering with men and nursing with women. In this section, we explore how gender and semantic components within these embeddings affect the representation of occupations deeply linked to gender stereotypes. We are interested in determining if our disentanglement approach effectively separates gender from the semantic aspects of occupations. To achieve this, we use the method suggested in Basta et al. (2019); Zhao et al. (2019), selecting sentence pairs from the WinoBias dataset (Zhao et al., 2018). Each pair features the same sentence with different gender-specific pronouns, for example, "[The manager] fired the cleaner because [he] was angry" and its counterpart "[The manager] fired the cleaner because [she] was angry". In Fig. 4, we illustrate the 20 occupations from Zhao et al. (2019). We applied PCA to the disentangled semantic component (in blue) and the gender component (in red) of our model. We then analyzed the difference between the first principal component of the female representation and the male representation within the same sentence. Given that we are utilizing contextualized embeddings, the embedding of the pronoun token will vary depending on the context. As depicted in Figure 4, the semantic component shows a smaller difference between PCA components across the same occupations when represented in the same sentence with different gender pronouns. This indicates a reduced dependency on gendered pronouns within the semantic component. In other words, the difference between the first PCA components in sentences with gendered pronouns is smaller when analyzing the semantic component than when analyzing the gender component. In contrast, the gender component displays greater variation, suggesting that it more effectively captures the gender associations tied to different occupations. This provides a clearer distinction between

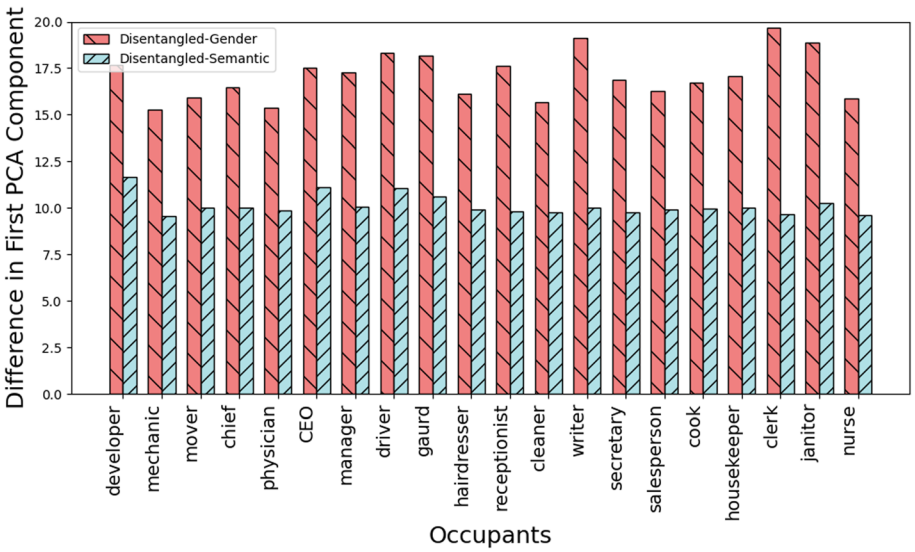

Fig. 4 PCA of stereotyped occupations by pronouns, using gender and semantic disentanglement

the representations of gendered pronouns across different occupations. This contrast demonstrates that the semantic component exhibits a more uniform representation of pronouns across occupations, meaning that pronoun differences are less influenced by occupation titles. Conversely, the gender component highlights these differences, indicating that it successfully isolates gender characteristics within the gender component, as intended by our disentanglement approach.

4.6.2 Detecting gender spaces

Bolukbasi et al. (2016) introduced a method to identify gender biases in embeddings by defining a gender space through directional differences between gender-related word pair vectors. Adopting this method, we analyzed the Principal Component Analysis (PCA) of male and female word pairs' vector differences, as illustrated in Fig. 5. This analysis extends to both the original and two disentangled models-one emphasizing semantic aspects, the other on gender components-providing insights into the inherent gender biases. We initiated our experiments by calculating directional vectors for pairs like 'he-she' and 'man-woman', as recommended by Gonen and Goldberg (2019), and applied PCA to these vectors to detect and quantify gender biases. Notably, our evaluation includes contextualized embeddings, allowing word representations to adapt based on sentence context. Comparing the principal components from the original and disentangled models, we assessed whether our disentanglement approach effectively reduces biased gender representation in embeddings. These comparisons are vital for validating the effectiveness of disentangling semantic content from gender information in mitigating gender bias within neural embeddings. Our observations can be enumerated as follows: (1) When comparing the principal components for gendered terms to those for random terms, we observed a higher variance among the gendered terms. This finding validates our experimental approach by highlighting the fact that gendered terms have a more pronounced first principal component compared to random terms, which can be an indication that this principal component is capturing aspects related to gender. (2) The disentangled semantics model shows a higher first component percentage compared to the original model, suggesting a more significant separation of semantic information from gender influences. The disentangled gender model

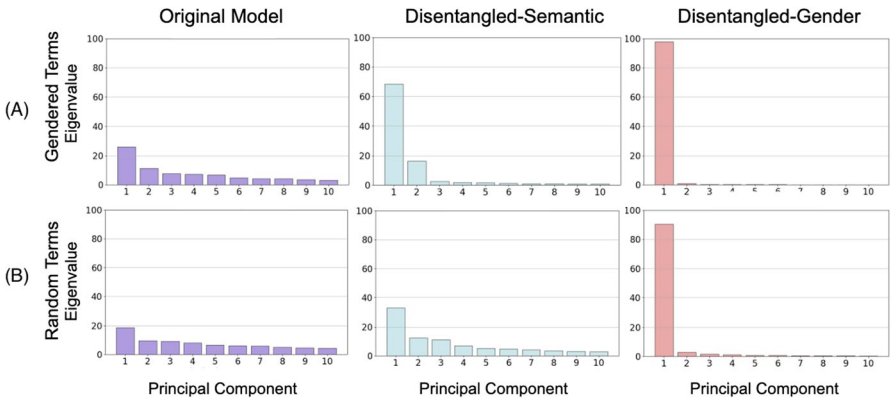

Fig. 5 (A) Variance percentages in the principal components for the original, disentangled semantics, and disentangled gender models. (B) Corresponding percentages for random vectors

exhibits an even higher first eigenvalue than the disentangled semantics model. Given that the disentangled semantics model focuses on detecting gender spaces, the predominance of a single principal component explaining a large variance aligns with our expectations. This principal component is likely capturing the primary axis of gender differentiation, further confirming that our model effectively disentangles gender information. (3) Although similar trends were observed with random terms, the intensity of the variance was much less pronounced. This is likely due to the shorter context and a higher occurrence of pronouns or stereotypically gendered occupations within these samples. Both the content and gender representations in the disentangled models showed higher first principal components compared to the original model, yet these were substantially lower than those observed for gendered terms. These results underscore the efficacy of our disentangled models in isolating and analyzing gender-related components.

It is worth mentioning that in Fig. 5, the principal component, which is interpreted here as the gender component, shows a significantly higher variance for both gendered and random terms. This observation aligns with trends observed in other studies that apply PCA to analyze bias in representations (Zhao et al., 2018, 2019; Gonen & Goldberg, 2019). Specifically, when applying PCA to data where certain dimensions are prominent or where the variance is heavily influenced by specific attributes, the first principal component tends to capture the majority of this variance. In this context, since the sentences are all focused on stereotypical occupations, it makes sense to have one bold principal component. However, by comparing it to the gendered terms, we see that it is less pronounced; the components in the first row and for gendered terms are significantly greater than those in the second row, which correspond to random terms.

4.6.3 Measuring bias in sentence encoders

May et al. (2019) have proposed the Sentence Encoder Association Test (SEAT) to measure social biases in sentence encoders. SEAT is a generalization of the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), which was originally designed to measure biases in word embeddings by comparing the associations between sets of words (target concepts) and sets of attributes.

SEAT relies on cosine similarity to measure the association between sentence embeddings. Bias is quantified using a test statistic, $s(X, Y, A, B)$, which measures the difference in cosine similarity between the embeddings of target concepts X and Y the embeddings of attributes A and B. The magnitude of the association between the target concepts and the attributes is measured by the effect size. It is calculated as the difference in mean cosine similarity scores for the target concepts with respect to the attribute sets, normalized by the standard deviation as follows:

$$EffectSize = \frac{\mu_{x\in X}\, s(x, A, B) - \mu_{y\in Y}\, s(y, A, B)}{\sigma_{w\in X\cup Y}\, s(w, A, B)} \tag{14}$$

where μ and σ indicate mean and standard deviation and $s(w, A, B)$ is the difference in mean of the cosine similarities:

$$s(w, A, B) = \text{mean}_{a\in A}\ \cos(w, a) - \text{mean}_{b\in B}\ \cos(w, b) \tag{15}$$

A larger effect size indicates stronger bias. To apply this bias measurement, specific sentences are crafted using templates that incorporate target concepts (e.g., names associated with a particular race or gender) and attributes (e.g., pleasant or unpleasant adjectives).

Table 9 Comparison of Gender bias in terms of Effect Size for MiniLM and BERT-Mini with Original and Disentangled-Semantic Representations

	MiniLM	Δ	BERT-Mini	Δ
Original	0.482	–	0.256	–
Disentangled-Semantic	0.368	−23.52%	0.214	−16.55%

Table 10 Training and inference time of the original and disentangled model

	MiniLM		BERT-Mini	
	Original	Disentangled	Original	Disentangled
Training time	01:36':16"	01:39':07"	01:00':42"	01:01':37"
Inference time	21.96 μs	23.07 μs	15.97 μs	17.97 μs

The sentence embeddings generated by the encoder are then tested for bias by comparing the cosine similarity of the embeddings for different combinations of target concepts and attributes. For example, a biased sentence encoder might show higher similarity between Male names and career-related words compared to Female names reflecting stereotypical gender bias.

Based on SEAT and in our experiments, we adopt and measure effect size as an indicator of bias (May et al., 2019), as explained in Eq. 14. In this test, the target groups consist of female-related terms and male-related terms, embedded in sentences like "This is a mother" and "This is a father," respectively. The attributes are represented by sentences associated with science and arts, such as "This is a dance" and "This is chemistry," respectively. Each target group contains 80 sentences, and there are over 55 attribute sentences in both the science and arts groups.

Our objective is to explore the extent to which contextualized sentence representations carry stereotypical gender biases, as shown in Table 9. We report the effect sizes for two different pre-trained language models, MiniLM and BERT-Mini, using both the original embedding representations and the representations where gender has been disentangled using our proposed approach. A higher effect size in any of the embeddings indicates a greater degree of bias, suggesting a stronger association of women with the arts and men with science. As illustrated in the table, when the semantic component of the original embeddings is disentangled from gender, both language models demonstrate a lower degree of stereotypical gender biases. The results show that our approach has been able to reduce biases as a result of the disentanglement process, namely MiniLM shows a reduction of over 23% in gender bias in terms of effect size, while BERT-Mini exhibits a reduction of over 16%.

5 Discussion

5.1 Scalability

In real-world and large-scale applications, the effectiveness and efficiency of a model are both crucial factors that determine its practical usability. The sheer volume of data, the computational resources required, and the time needed for training can all pose significant

challenges when deploying a model in real environments. Therefore, it is imperative that the proposed model for eliminating stereotypical gender biases does not introduce prohibitive levels of complexity or excessive time demands, as these could impede its application in real-world scenarios.

To assess the scalability of our model in large-scale and real-world applications, we evaluated the training and inference times of the proposed disentanglement model. As shown in Table 10, we compare these running times for both the original and disentangled models across two base models. All experiments were conducted using an NVIDIA RTX A6000 GPU, which is well-suited for handling high-performance deep learning tasks. The results indicate that the disentanglement approach does not significantly increase the model's running time, as evidenced by the minimal differences in training times (less than three minutes) and inference times (around 1 microsecond) between the original and disentangled models.

These findings confirm that our proposed model remains scalable, making it well-suited for deployment in large-scale applications without compromising performance.

5.2 Interpretability

In real-world applications, ensuring that a model is interpretable is crucial, particularly in scenarios where the decision-making processes must be transparent and trackable. Interpretability becomes even more significant in contexts like information retrieval systems, where users and stakeholders need to understand how and why certain results are ranked or presented. One effective approach to enhancing interpretability in these models is through representation disentanglement. By decomposing vector representations, which often encode a mixture of various information, into more interpretable and meaningful components, disentanglement allows us to better understand the underlying factors that influence the model's decisions. This process of separating distinct attributes within the representations makes it easier to track and explain the model's behavior. Representation disentanglement has been successfully employed in various areas to enhance model interpretability (Zhu et al., 2021; Hsu et al., 2017; Sarhan et al., 2019; Tsang et al., 2018; Wang et al., 2023; Du et al., 2020).

In the context of gender bias in information retrieval systems, disentangling gender-related information from other aspects of the data can significantly contribute to the interpretability of the ranking model. When gender information is disentangled, it allows researchers and practitioners to isolate and examine the impact of gender on the ranking process, providing clearer insights into whether and how gender biases are influencing the model's outputs. This level of transparency not only helps in identifying potential biases but also aids in the development of more fair and balanced models. By leveraging disentanglement techniques, it becomes possible to create systems that not only perform well but also offer interpretable, bias-aware decision-making processes, which is essential for ethical AI deployment.

5.3 Ethical implications

Gender is considered to be a sensitive attribute, and any attempt to manipulate or alter gender information in machine learning models can lead to significant ethical concerns. This is particularly true in information retrieval systems, where fairness and transparency are paramount. In our work, we emphasize that our approach does not involve changing or manipulating

gender attributes in any way. Instead, we focus on disentangling gender from the intermediate representations of query-document pairs. This process ensures that gender influences the ranking decisions for neutral queries to the extent to which it is relevant in the context of the search query, thereby avoiding the introduction of bias or unfair treatment based on gender.

Disentangling gender in this manner does not alter the gender attribute itself. Rather, it aims to create a more unbiased and fair model by ensuring that gender does not unintentionally affect the outcomes of the model's decision-making process. This approach is especially important in contexts where the objective is to achieve gender fairness.

6 Concluding remarks

We introduce a novel method for mitigating gender bias in neural ranker representations by disentangling content semantics from gender associations. Our approach isolates gender-related information, enabling the ranker to assess document relevance based solely on semantic content. Experimental results show our method outperforms state-of-the-art baselines in reducing gender bias while maintaining ranking effectiveness, decreasing the performance gap between male and female queries by around 27% at cut-offs 10 and 20. Our disentanglement strategy effectively weakens gender information in the intermediate vector representation of the cross-encoder. This balance between fairness and performance is crucial for developing unbiased neural rankers. Our methodology, while focused on the gender attribute, can be applied to other sensitive attributes like ethnicity and race for future work, promoting fairer information retrieval systems. It integrates seamlessly with existing neural rankers, allowing for immediate deployment without sacrificing performance, scalability, or accuracy, while promoting unbiased ranking of search results.

Appendix A

In this appendix, we explore the question: "What if we further penalize the presence of gender information in the ranking component of the representation to ensure it is entirely gender-neutral?" To address this, we employ an adversarial strategy. We introduce an adversary network specifically designed to detect gender in the ranking representation and aimed to alter the representation to remove any gender-related information, rendering the adversary network incapable of detecting gender from the ranking part.

To this end, we trained a gender classifier network with parameters Θ_c, which takes the ranking representation (E_r) as input and attempts to classify the gender into two categories: male or female. This process is formalized as follows:

$$p_c = \sigma(\theta_c E_r + b_c), \tag{A1}$$

where σ is the activation function, and p_c represents the predicted probability that the ranking representation is male. We use a binary cross-entropy loss for training, defined as:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_{c_i}) + (1 - y_i)\log(1 - p_{c_i})\right] \tag{A2}$$

where y_i is the true gender label obtained from the function Λ.

To ensure the ranking representation (E_r) does not contain gender information, we add an adversary loss L_{adv} to the total network loss L_t. This adversary loss maximizes the entropy of the predicted gender probability p_c, making gender information unpredictable:

$$\begin{aligned} L_adv(\theta_E) &= \mathcal{H}(p_c|E_r;\theta_c), \\ \mathcal{H}(p) &= -\sum_{i \in lables} p_i \log(p_i) \end{aligned} \tag{A3}$$

By maximizing the entropy of p_c, we modify the ranking representation during training to exclude gender information, making it challenging for the adversary network to predict gender. The total loss L_t is defined as an interpolation of three losses: 1) the ranking loss L_r, 2) the gender classification loss L_g, and 3) the adversary loss L_{adv}:

$$\begin{aligned} L_t =& \alpha \times \left(\frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \sigma(\theta_r E_{r_i}^+ + b_r) + \sigma(\theta_r E_{r_j}^- + b_r)) \right) \\ &+ \beta \times \left(-\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(p_{g_i}) + (1 - y_i) \log(1 - p_{g_i}) \right] \right) \\ &- \gamma \times \frac{1}{N} \sum_{i=1}^{N} p_{c_i} \log(p_{c_i}) \end{aligned} \tag{A4}$$

The network architecture, including the adversary network, is illustrated in Fig. 6. During training, we first optimize the adversary network parameters (Θ_c). In this stage, only the adversary network parameters are updated, and the encoder parameters remain unchanged. Then, while optimizing the total loss (L_t), the parameters of the encoder, ranking network, and gender classifier network are updated.

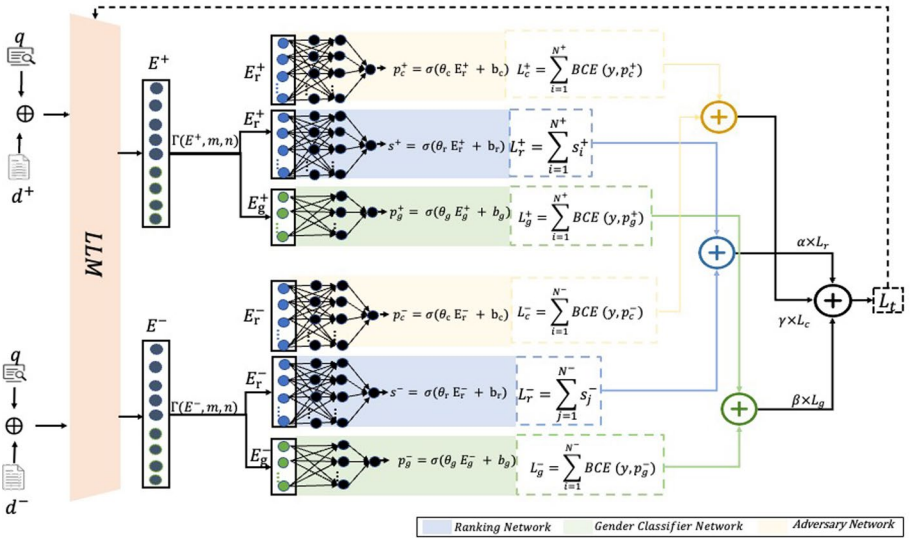

Fig. 6 Overview of the proposed neural disentanglement architecture with the adversary network

We trained this adversarial network to compare the results with the original model and our proposed disentanglement network. The results for 215 and 1,765 queries with the two base models are presented in Tables 11, 12, 13, and 14, respectively.

From the tables, we observe that the adversarial strategy is not effective in improving ranking performance or reducing bias. For both the MiniLM and BERT-Mini base models, our disentangled model consistently outperforms the original and adversarial models. For instance, in Table 11, the Disentangled Model+Adv achieves an NFaIR score of 0.8881, which is better than the Original Model's 0.8107 but still lower than the Disentangled

Table 11 Gender bias measures for 215 neutral queries with MiniLM base model for the adversarial training strategy

	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Cut-off 10						
Original Model	0.1602	0.3183	0.1374	0.1101	0.8107	1.023
Disentangled Model	0.1877	0.0737	0.046	0.0567	0.8664	0.8404
Disentangled Model+Adv	0.1657	0.2298	0.1237	0.1295	0.8881	0.7624
Cut-off 20						
Original Model	0.1658	0.2635	0.1142	0.092	0.8274	0.7966
Disentangled Model	0.1941	0.0574	0.035	0.0422	0.8722	0.717
Disentangled Model+Adv	0.1712	0.2320	0.1185	0.1185	0.8853	0.7624

Table 12 Gender bias measures for 1765 neutral queries with MiniLM base model for the adversarial training strategy

	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Cut-off 10						
Original Model	0.2673	0.1535	0.0721	0.0611	0.7066	1.5599
Disentangled Model	0.2969	0.0805	0.0131	0.0178	0.7623	1.4521
Disentangled Model+Adv	0.2724	0.1337	0.0824	0.0952	0.7915	1.307
Cut-off 20						
Original Model	0.2726	0.0721	0.0641	0.0538	0.722	1.3001
Disentangled Model	0.3023	0.0131	0.0313	0.0052	0.7658	1.2767
Disentangled Model+Adv	0.2783	0.1672	0.0911	0.0957	0.7782	1.2912

Table 13 Gender bias measures for 215 neutral queries with BERT-Mini base model for the adversarial training strategy

	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Cut-off 10						
Original Model	0.1662	0.2544	0.1058	0.0751	0.8273	0.8467
Disentangled Model	0.1399	0.0376	0.0132	0.0075	0.8583	0.6969
Disentangled Model+Adv	0.1472	0.3889	0.2082	0.2145	0.8313	1.2266
Cut-off 20						
Original Model	0.1742	0.2318	0.0929	0.0646	0.8457	0.6964
Disentangled Model	0.1455	0.047	0.0158	0.0083	0.8691	0.5674
Disentangled Model+Adv	0.1544	0.3794	0.1953	0.1960	0.8323	1.0849

Table 14 Gender bias measures for 1765 neutral queries with BERT-Mini base model

	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Cut-off 10						
Original Model	0.2475	0.1387	0.056	0.0369	0.7304	1.4942
Disentangled Model	0.1922	0.0928	0.0354	0.026	0.7565	1.3468
Disentangled Model+Adv	0.1472	0.2354	0.1419	0.1621	0.7329	1.6942
Cut-off 20						
Original Model	0.2548	0.1262	0.505	0.0329	0.7451	1.2592
Disentangled Model	0.1996	0.0928	0.037	0.0285	0.7672	1.1682
Disentangled Model+Adv	0.1544	0.2632	0.14607	0.1562	0.7299	1.5616

Model's 0.8664. Similarly, the adversarial strategy results in higher ARaB-tc values (e.g., 0.2298 in the Disentangled Model+Adv vs. 0.0737 in the Disentangled Model), indicating less bias reduction.

One reason for this is that gender information may not be entirely redundant or unnecessary for ranking. In some user queries, the presence of gender information could improve performance. For example, consider the query "rsm meaning home care" from the 215-query set by Rekabsaz et al. (2021). The relevant document is "'The mission of the Right from the Start Medical Assistance Group (RSM) is to enable children under age 19, pregnant women, low-income families, and women with breast or cervical cancer to receive comprehensive health services through Medicaid and related programs." Although this query is considered gender-neutral, the relevant document contains female-gendered information crucial for accurately answering the query. In such cases, forcing the model to remove gender information from the ranking representation is counter-productive. However, the adversarial strategy attempts to eliminate gender information, which is not always desirable.

In our proposed disentangled model, there is no external force to remove gender information from the ranking representation. The multitask training of the ranking network and the gender classification network provides the flexibility to determine how much gender information to isolate from the ranking component, optimizing both ranking performance and gender bias reduction. Consequently, our model performs better in terms of ranking performance and bias reduction. Additionally, training the adversarial network is significantly more time-consuming, taking approximately six times longer to converge.

Moreover, we observed that methods enforcing gender removal from the representation do not perform well. For example, the ADVBERT methodology, which is one of our baselines, removes gender information from the intermediate representation of query-document pairs using an adversarial strategy. From Tables 2, 3, 4, and 5, we see that the ADVBERT model significantly underperforms compared to our disentanglement approach and fails to reduce gender biases effectively.

Appendix B

To further highlight the effectiveness of our proposed model in reducing stereotypical gender biases, we provide specific examples from the set of 215 socially problematic queries. These queries are designed to be neutral, but when gender inequality appears in the ranked

Table 15 A case study example of the query "what body fat percentage is healthy", and the top 3 re-ranked documents with the original, and disentangled model

Query: what body fat percentage is healthy	Original Model	Disentangled Model
Rank 1 Document	Body Fat for **Girls**. A 5-year-old **girl** should have 14 to 21 percent body fat, while a 6-year-old **girl** is considered healthy at 14 to 22 percent. The low end of a healthy body fat range for 7- and 8-year-old **girls** is 15 percent, while the high end is 24 and 25 percent, respectively. healthy body fat percentage for an 18-year-old **girl** is between 17 and 30 percent, while a 19-year-old should fall between 19 and 31 percent. Adult **females** 20 to 39 years should strive for a body fat percentage between 21 and 32 percent	Go to Body Fat Table. The percentage of body fat in healthy humans ranges from 5 to 40 per cent. **Females** have more body fat than ***males***. Athletes vary in body fat depending on their sport. Distance runners tend to have a low fat content. While most humans have too much fat some get carried away with trying to achieve unrealistic, unhealthy low levels. For **females**, body fat should not be less than 15 percent and for ***males***, not less than 5 percent
Rank 2 Document	A healthy body fat percentage ranges from 10 to 22 percent for ***men*** and 20 to 32 percent for **women**, according to ACSM. This means a healthy percentage of lean mass is 78 to 90 percent for ***men*** and 68 to 80 percent for **women**. You'll get the most accurate assessment of your body fat levels if you consult a professional	The average healthy. adult body fat range regardless of age is 15 to 20% for ***men*** and 20 to 25% for **women**. A **woman** with. more than 32% body fat and ***males*** with more than 25% body fat are considered to be at increased risk. for disease
Rank 3 Document	Body Fat for **Girls**. A 5-year-old **girl** should have 14 to 21 percent body fat, while a 6-year-old **girl** is considered healthy at 14 to 22 percent. The low end of a healthy body fat range for 7- and 8-year-old **girls** is 15 percent, while the high end is 24 and 25 percent, respectively	As a result, different body fat percentages will be provided with the same health assessment for both genders. For **women** between age 20 and 40, 19% to 26% body fat is generally good to excellent. For **women** age 40+ to 60+, 23% to 30% is considered good to excellent.For ***men*** between age 20 and 40, 10% to 20% body fat is generally good to excellent. For ***men*** age 40+ to 60+, 19% to 23% is considered good to excellent.ctually, 5% body fat can cause serious health problems for the average person. Conversely, 25% fat can either be healthy or unhealthy depending upon your age and gender. In order to provide clarity, it's best to look at a scale of body fat percentages and what they represent

Table 16 A case study example of the query "physical health effects of stress", and the top 3 re-ranked documents with the original, and disentangled model

Query: physical health effects of stress	Original Model	Disentangled Model
Rank 1 Document	According to the American Academy of Family Physicians (AAFP), stress is an expression of the body natural instinct to protect itself. While this may warn a **woman** of immediate danger, like a fast-approaching car, prolonged stress effects can negatively affect your physical and emotional health	The Physical Effects of Long-Term Stress. Chronic stress can have a serious impact on our physical as well as psychological health due to sustained high levels of the chemicals released in the fight or flight response. Lets take a closer look at whats going on. The Role of the Nervous System
Rank 2 Document	According to the National **Womens** Health Information Center, the effects of stress on **womens** physical and emotional health can range from headaches to irritable bowel syndrome. Specific stress effects include: 1 Eating disorders	Stress, however, can affect many aspects of physical and mental health, ranging from hair, teeth, and skin to memory and concentration skills, and even how well we sleep. The good news is while these problems may seem serious, stress relief can lead to real improvements in your overall health and well-being
Rank 3 Document	According to the National **Womens** Health Information Center, the effects of stress on **womens** physical and emotional health can range from headaches to irritable bowel syndrome. Specific stress effects include: Eating disorders	It is a well-known fact that stress can affect our lives in many ways. It can even have an adverse affect on our physical health. Severe stress can actually lead to chronic health conditions. It is important to recognize symptoms of severe stress and learn how to cope with stress

Table 17 A case study example of the query "how is back pay for disability determined", and the top 3 re-ranked documents with the original, and disentangled model

Query: how is back pay for disability determined	Original Model	Disentangled Model
Rank 1 Document	VA Disability Back Pay is a payment of all the money that the veteran should have been receiving for the months in between his date of eligibility and ***his*** VA rating decision. A veteranâs date of eligibility for VA Disability Back Pay is determined in one of two ways.First, if the veteran submits ***his*** VA Disability Claim within one year of ***his*** date of separation, ***his*** date of eligibility for VA Disability Back Pay is ***his*** date of separation..	How far back Social Security will pay disability benefits to a disabled person is determined by the date you filed your disability claim when applying for Social Security and/or SSI disability. Social Security has a five-month waiting period that applies to social security disability claims for which they never pay disability benefits. Basically, the date of filing determines what month you are first entitled to begin receiving monthly Social Security disability benefits
Rank 2 Document	A veteranâs date of eligibility for VA Disability Back Pay is determined in one of two ways. First, if the veteran submits ***his*** VA Disability Claim within one year of his date of separation, ***his*** date of eligibility for VA Disability Back Pay is his date of separation. If, however, Ben submits his VA Disability Claim 13 months after he separates, and the VA takes 16 months (unfortunately not unusual) to reach their Rating Decision, ***he*** will only receive 16 months of VA Disability Back Pay	VA Disability Back Pay is a payment of all the money that the veteran should have been receiving for the months in between his date of eligibility and ***his*** VA rating decision. A veteranâs date of eligibility for VA Disability Back Pay is determined in one of two ways.First, if the veteran submits ***his*** VA Disability Claim within one year of ***his*** date of separation, ***his*** date of eligibility for VA Disability Back Pay is ***his*** date of separation..
Rank 3 Document	A veteranâs date of eligibility for VA Disability Back Pay is determined in one of two ways. First, if the veteran submits ***his*** VA Disability Claim within one year of his date of separation, ***his*** date of eligibility for VA Disability Back Pay is his date of separation. If, however, Ben submits his VA Disability Claim 13 months after he separates, and the VA takes 16 months (unfortunately not unusual) to reach their Rating Decision, ***he*** will only receive 16 months of VA Disability Back Pay	How far back Social Security will pay disability benefits to a disabled person is determined by the date you filed your disability claim when applying for Social Security and/or SSI disability

list of documents, it can inadvertently perpetuate societal biases against a particular gender. Therefore, it is crucial that the ranked documents for these queries remain impartial, showing no preference for one gender over another.

Tables 15, 16, and 17 present examples of these queries, comparing the top-3 documents ranked by the original model with those re-ranked by our proposed disentanglement approach. Our analysis reveals that the top-3 documents produced by our disentanglement model demonstrate a more balanced representation with respect to gender. Specifically, our model tends to include gender-neutral documents or documents that reference male and female terms equally, thereby mitigating the risk of reinforcing gender stereotypes. This balanced approach is essential for ensuring that search results do not unintentionally contribute to gender bias in society.

Table 15 highlights the query "what body fat percentage is healthy," where the original model's top-3 documents show a strong bias towards female representation. Specifically, the top-1 and top-3 documents include only female-related terms, indicating a clear gender bias. In contrast, the top-3 documents re-ranked by our disentangled model present a balanced representation of both male and female terms, effectively mitigating this bias. Similarly, Table 16 demonstrates a similar issue for the query "physical health effects of stress." The top-3 documents re-ranked by the original model exhibit a bias towards female representation. However, our disentangled model successfully re-ranks the documents to ensure they are gender-neutral, thus preventing any gender bias. In Table 17, the query "how is back pay for disability determined" shows a male bias in the top-3 documents re-ranked by the original model, with all documents featuring male-specific terms. On the other hand, the documents re-ranked by our disentangled model display a more neutral stance, with the top-1 and top-3 documents showing no gender inclination, significantly reducing the overall bias compared to the original model.

Author contributions Shirin Seyedsalehi: Idea, Experiments, writing Sara Salamat: Experiments Negar Arabzadeh: Brainstorming Sajad Ebrahimi: Experiments Morteza Zihayat: supervision Ebrahim Bagheri: Supervision.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interest The authors declare no competing interests.

References

Abolghasemi, A., Azzopardi, L., Askari, A., Rijke, M., & Verberne, S. (2024). Measuring bias in a ranked list using term-based representations. In: European Conference on Information Retrieval, pp. 3–19 Springer

Banerjee, P., & Han, H.-I. (2009). Language modeling approaches to information retrieval. *Journal of Computing Science and Engineering, 3*(3), 143–164.

Basta, C., Costa-jussà, M.R., & Casas, N. (2019). Evaluating the underlying gender bias in contextualized word embeddings. CoRR arXiv:1904.08783

Bengio, Y., Courville, A., & Vincent, e.a. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence

Bhargava, P., Drozd, A., & Rogers, A. (2021). Generalization in nli: Ways (not) to go beyond simple heuristics. arXiv preprint arXiv:2110.01518

Bigdeli, A., Arabzadeh, N., SeyedSalehi, S., Zihayat, M., & Bagheri, E. (2022). A light-weight strategy for restraining gender biases in neural rankers. In: (ECIR 2022)

Bigdeli, A., Arabzadeh, N., Zihayat, M., & Bagheri, E. (2021). Exploring gender biases in information retrieval relevance judgement datasets. In: ECIR 2021, Springer

Bigdeli, A., Arabzadeh, N., Seyedsalehi, S., Zihayat, M., & Bagheri, E. (2021). On the orthogonality of bias and utility in ad hoc retrieval. In: SIGIR 2021, pp. 1748–1752
Bigdeli, A., Arabzadeh, N., SeyedSalehi, S., Zihayat, M., & Bagheri, E. (2022). Gender fairness in information retrieval systems. In: SIGIR 2022, pp. 3436–3439
Bigdeli, A., Arabzadeh, N., Seyedsalehi, S., Zihayat, M., & Bagheri, E. (2022). A light-weight strategy for restraining gender biases in neural rankers. In: ECIR
Bigdeli, A., Arabzadeh, N., Seyedsalehi, S., Mitra, B., Zihayat, M., & Bagheri, E. (2023). De-biasing relevance judgements for fair ranking. In: ECOR Springer
Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings
Burges, C., Shaked, T., Renshaw, E., & Lazier, e.a. (2005). Learning to rank using gradient descent. In: (ICML '05)
Cabrera, Á.A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., & Chau, D.H. (2019). Fairvis: Visual analytics for discovering intersectional bias in machine learning. In: 2019 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 46–56 IEEE
Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.
Du, Y., Guo, X., Shehu, A., & Zhao, L. (2020). Interpretable molecule generation via disentanglement learning. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 1–8
Feldman, T., & Peake, A. (2021). End-to-end bias mitigation: Removing gender bias in deep learning. arXiv preprint arXiv:2104.02532
Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them
Hsu, W.-N., Zhang, Y., & Glass, J. (2017). Unsupervised learning of disentangled and interpretable representations from sequential data. Advances in neural information processing systems **30**
John, V., Mou, L., Bahuleyan, H., & Vechtomova, O. Disentangled representation learning for non-parallel text style transfer. Association for Computational Linguistics
Kopeinik, S., Mara, M., Ratz, L., Krieg, K., Schedl, M., & Rekabsaz, N. (2023). Show me a "male nurse"! how gender bias is reflected in the query formulation of search engine users. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23. Association for Computing Machinery, ???
Krieg, K., Parada-Cabaleiro, E., Schedl, M., & Rekabsaz, N. (2022). Do perceived gender biases in retrieval results affect relevance judgements? In: International Workshop on Algorithmic Bias in Search and Recommendation Springer
Krieg, K., Parada-Cabaleiro, E., & Medicus, e.a. (2023). Grep-biasir: A dataset for investigating gender representation bias in information retrieval results. CHIIR '23
Latif, E., Zhai, X., & Liu, L. (2023). Ai gender bias, disparities, and fairness: Does training data matter? arXiv preprint arXiv:2312.10833
Liu, Z., Zhang, K., Xiong, C., & Liu, e.a. (2021). Openmatch: An open source library for neu-ir research. In: SIGIR 2021, pp. 2531–2535
May, C., Wang, A., Bordia, S., Bowman, S.R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 622–628. Association for Computational Linguistics, Minneapolis, Minnesota https://doi.org/10.18653/v1/N19-1063 . https://aclanthology.org/N19-1063
Nguyen, T., Rosenberg, M., & Song, e.a. (2016). Ms marco: A human-generated machine reading comprehension dataset
Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). MS MARCO: A human generated machine reading comprehension dataset. CoRR arXiv:1611.09268
Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count*. Mahwah, NJ: Lawerence Erlbaum Associates.
Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In: Conference on Empirical Methods in Natural Language Processing https://api.semanticscholar.org/CorpusID:201646309
Rekabsaz, N., Kopeinik, S., & Schedl, M. (2021). Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In: SIGIR 2021, pp. 306–316
Rekabsaz, N., & Schedl, M. (2020). Do neural ranking models intensify gender bias? In: Proceedings of the 43rd International ACM SIGIR Conference, pp. 2065–2068
Rekabsaz, N., Kopeinik, S., & Schedl, M. (2021). Societal biases in retrieved contents: Measurement framework and adversarial mitigation for BERT rankers. CoRR arXiv:2104.13640

Rekabsaz, N., & Schedl, M. (2020). Do neural ranking models intensify gender bias? In: Proceedings of the 43rd International ACM SIGIR Conference

Sarhan, M.H., Eslami, A., Navab, N., & Albarqouni, S. (2019). Learning interpretable disentangled representations using adversarial vaes. In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 1, pp. 37–44 Springer

Seyedsalehi, S., Bigdeli, A., Arabzadeh, N., Mitra, B., Zihayat, M., & Bagheri, E. (2022). Bias-aware fair neural ranking for addressing stereotypical gender biases. In: EDBT, pp. 2–435

Seyedsalehi, S., Bigdeli, A., Arabzadeh, N., Zihayat, M., & Bagheri, E. (2022). Addressing gender-related performance disparities in neural rankers. SIGIR '22

Tsang, M., Liu, H., Purushotham, S., Murali, P., & Liu, Y. (2018). Neural interaction transparency (nit): Disentangling learned interactions for improved interpretability. Advances in neural information processing systems **31**

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems, 33*, 5776–5788.

Wang, T.-H., Xiao, W., Seyde, T., Hasani, R., & Rus, D. (2023). Measuring interpretability of neural policies of robots with disentangled representation. In: Conference on Robot Learning, pp. 602–641 PMLR

Zerveas, G., Rekabsaz, N., Cohen, D., & Eickhoff, C. (2022). Mitigating bias in search results through contextual document reranking and neutrality regularization. In: SIGIR

Zerveas, G., Rekabsaz, N., Cohen, D., & Eickhoff, C. (2021). Coder: An efficient framework for improving retrieval through contextual document embedding reranking. arXiv preprint arXiv:2112.08766

Zhao, W.X., Zhou, K., Li, J., & al. (2023). A survey of large language models. ArXiv arXiv:2303.18223

Zhao, L., & Callan, J. (2010). Term necessity prediction. In: CIKM'10, pp. 259–268

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender Bias in Contextualized Word Embeddings

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:1804.06876

Zhou, Y., Shen, T., Geng, X., Tao, C., Xu, C., Long, G., Jiao, B., & Jiang, D. (2023). Towards robust ranker for text retrieval. In: Findings of the ACL 2023, pp. 5387–5401

Zhu, X., Xu, C., & Tao, D. (2021). Where and what? examining interpretable disentangled representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5861–5870

Zou, J.Y., Hsu, D.J., Parkes, D.C., & Adams, R.P. (2013). Contrastive learning using spectral methods. Advances in Neural Information Processing Systems **26**

Authors and Affiliations

Shirin Seyedsalehi[1,2] · Sara Salamat[2] · Negar Arabzadeh[3] · Sajad Ebrahimi[4] · Morteza Zihayat[2] · Ebrahim Bagheri[1]

✉ Shirin Seyedsalehi
shirin.seyedsalehi@torontomu.ca

Sara Salamat
sara.salamat@torontomu.ca

Negar Arabzadeh
narabzad@uwaterloo.ca

Sajad Ebrahimi
sebrah05@uoguelph.ca

Morteza Zihayat
mzihayat@torontomu.ca

Ebrahim Bagheri
ebrahim.bagheri@utoronto.ca

[1] University of Toronto, Toronto, Canada

[2] Toronto Metropolitan University, Toronto, Canada

[3] University of Waterloo, Waterloo, Canada

[4] University of Guelph, Guelph, Canada