# Gender Bias-aware Document Ranking Using A Gender Sentence Labeler For Negative Sampling

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND COMMUNICATIONS ENGINEERING,
THE GRADUATE SCHOOL OF FUNDAMENTAL SCIENCE
AND ENGINEERING
OF WASEDA UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ENGINEERING

January 23th 2023

Qianyu Zhou
5120FG05

Advisor: Prof. Tetsuya Sakai
Research guidance: Research on Information Access

# Abstract

As one side effect of power pre-training models in information retrieval systems, the increase of various societal biases in search results of neutral IR models has been pointed out and observed in several studies. Recently, a simple negative sampling method has been examined to be working to reduce the overall bias, especially gender bias, in various IR systems. In this work, we want to go beyond term-based matching for the negative sampling method by training a Bert model to select negative training samples for ranking models. We also created a dataset modified from C4 in order to achieve this. Our experiments show that using training samples selected by our new proposed model, the final ranking model's gender bias has been reduced while maintaining excellent retrieval effectiveness and achieve better performance than the original and the term-based matching approach.

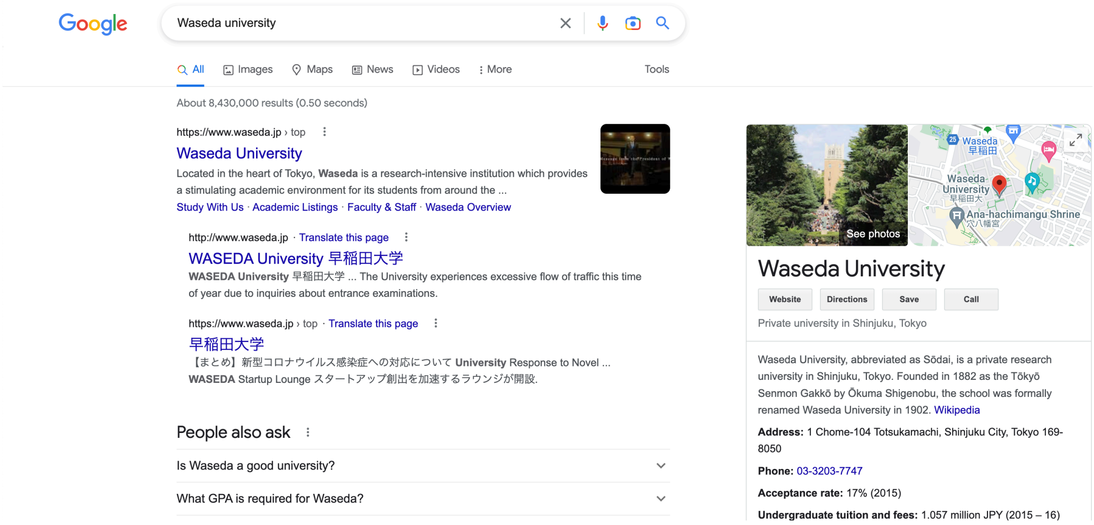
# Contents

# Chapter 1

# Introduction

Search engines larges change the way people see the world and daily information. Meanwhile, search engines always continue revolutionizing themselves. Users now could see the traditional ranking of links to websites as the search results, along with various sections that satisfy different information needs, such as sublinks within major links and inference sections.

Fig. 1.1: example results of Waseda University



About 8,430,000 results (0.50 seconds)

https://www.waseda.jp › top

Waseda University

Located in the heart of Tokyo, **Waseda** is a research-intensive institution which provides a stimulating academic environment for its students from around the ...

Study With Us · Academic Listings · Faculty & Staff · Waseda Overview

http://www.waseda.jp · Translate this page

WASEDA University 早稲田大学

**WASEDA University** 早稲田大学 ... The University experiences excessive flow of traffic this time of year due to inquiries about entrance examinations.

https://www.waseda.jp › top · Translate this page

早稲田大学

【まとめ】新型コロナウイルス感染症への対応について **University** Response to Novel ...
**WASEDA** Startup Lounge スタートアップ創出を加速するラウンジが開設.

People also ask

Is Waseda a good university?

What GPA is required for Waseda?

Waseda University

Private university in Shinjuku, Tokyo

Waseda University, abbreviated as Sōdai, is a private research university in Shinjuku, Tokyo. Founded in 1882 as the Tōkyō Senmon Gakkō by Ōkuma Shigenobu, the school was formally renamed Waseda University in 1902. Wikipedia

**Address:** 1 Chome-104 Totsukamachi, Shinjuku City, Tokyo 169-8050

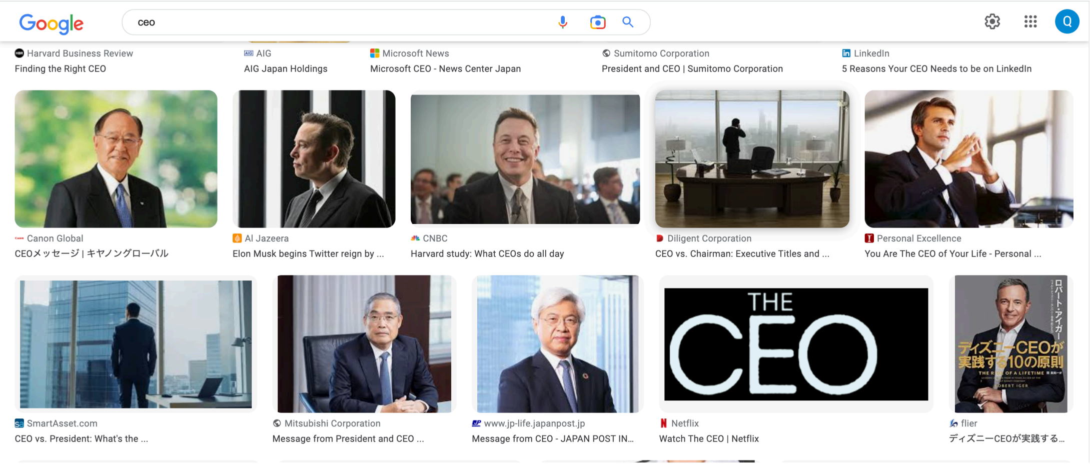
**Phone:** 03-3203-7747

**Acceptance rate:** 17% (2015)

**Undergraduate tuition and fees:** 1.057 million JPY (2015 – 16)

The mission of search engines is to gather world data and use existing information retrieval technology to satisfy user's information needs based on what it has. Bias from the real world, therefore, is collected and presented in search engines and the web[1]. For example, the search results of many gender-neutral words result in unfair distribution of male and female information. These societal biases within the web could lead to real-world problems, not limited to unfair job opportunities, gender discrimination, and racial discrimination[2, 3].

Fig. 1.2: image search results of word ceo



Google
ceo
Q
Harvard Business Review
Finding the Right CEO
AIG
AIG Japan Holdings
Microsoft News
Microsoft CEO - News Center Japan
Sumitomo Corporation
President and CEO | Sumitomo Corporation
LinkedIn
5 Reasons Your CEO Needs to be on LinkedIn
Canon Global
CEOメッセージ | キヤノングローバル
Al Jazeera
Elon Musk begins Twitter reign by ...
CNBC
Harvard study: What CEOs do all day
Diligent Corporation
CEO vs. Chairman: Executive Titles and ...
Personal Excellence
You Are The CEO of Your Life - Personal ...
THE CEO
ロバート・アイガー
ディズニーCEOが実践する10の原則
SmartAsset.com
CEO vs. President: What's the ...
Mitsubishi Corporation
Message from President and CEO ...
www.jp-life.japanpost.jp
Message from CEO - JAPAN POST IN...
Netflix
Watch The CEO | Netflix
flier
ディズニーCEOが実践する...

In recent years, Transformer[4] based models such as Bert[5] have led to significant advancements in Natural language processing and Information Retrieval. Based on these pre-training models, Neural IR has achieved a tremendous improvement in retrieval over traditional term matching methods in IR such as BM25[6] in large datasets like MS-MARCO[7]. One reason behind these improvements is the deeper connections pre-training models could build between query and documents within its large amount of parameters[8]. But at the same time, models like Bert[5] also deepen the biased connections and further intensify various biases within the web and search engines[9].

To tackle this problem and create bias-aware ranking models, this work is built upon the previous light-weight negative training method[10] and inspired by the transfer and adapt learning from TANDA[11]:

- Built multi-label gender related sentence classification dataset from C4 using gender magnitude calculation from ARaB[9] and NFaiRR[12]
- Trained gender sentence labeler based on Bert[5]. It is created based on two steps. Firstly, finetuning Bert[5] on the new multi-label datasets we created and then adapting the previous model to smaller but more accurate human annotated gender label dataset[9].
- Trained bias-aware ranking model based on Bert[5] without changing the original architecture and reduce the results' gender bias by intentionally selecting negative training samples from our gender sentence labeler.

# Chapter 2

# Related Work

While there are a significant number of studies and literature on bias or fairness within documents or the web, Bias awareness pre-training models are a relatively new field. The goal of bias aware ranking model is to reduce the bias within the retrieved results while maintaining the effectiveness and usefulness of the results.

Study has shown that the gender bias of the IR system might exist in relevance judgment of dataset[13]. As the neutral IR system becomes popular, the neural embedding of these pre-training models has been found to carry gender bias[14, 15].

Query revision is one way to reduce the gender bias observed in neural IR systems[16]. Many other attempts include the modification to the pre-training model's architecture. Rekabsaz et al.[12] have proposed a new training architecture of pre-training models by adding an adversarial part. This approach aims to remove gender information from the embedding so that the prediction and reranking of the model will not be influenced by gender-related information. Bigdeli et al.[17] examine the loss function of pre-training ranking models. They simultaneously add a document bias penalty to the loss function and relax this penalty for relevant documents. Zerveas et al.[18] use a transformer-based encoder to encode query to score query in the context and realization of other texts and introduces a new regularization loss for any document that does not seem to be neutral. Bigdeli et al.[10] put a straightforward way into the game and suggest that the selection of negative samples for training models could result in a more fair ranking if negative samples could contain more gender information. Their approach keeps the architecture of the typical model the same and thus can generalize to other approaches as well.

# Chapter 3

# Proposed Methods

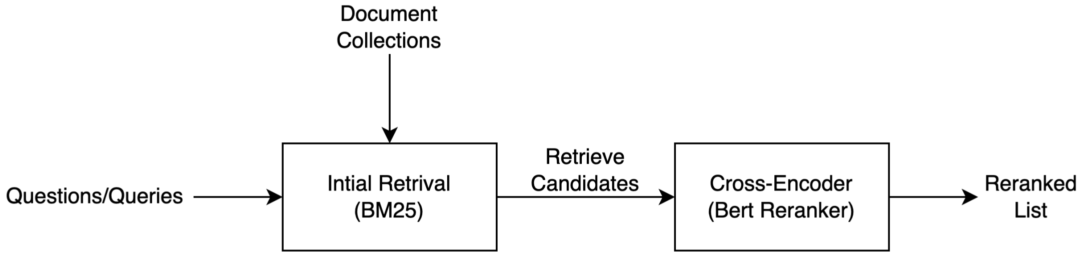
As mentioned in Chapter 1, the main contribution of this work is to propose a new way of selecting negative training samples for training a bias-aware ranking model. In this chapter, We will start by discussing ranking and how we define the bias-aware ranking, which is the main interest of this work, in the section 3.1. In the section 3.2, we will talk about our main contribution to bias-aware ranking in term of selecting negative samples by utilizing the power of pre-training model.

## 3.1 Bias aware ranking model

### 3.1.1 Problem Definition

Passage or document ranking with the use of Bert typically follows the two-step process as shown in figure 3.1

Fig. 3.1: Two stage ranking with Bert



Let the first-stage initial retrieval method be $\mathcal{M}$. In the figure 3.1 and our work afterwards, $\mathcal{M}$ will be simply BM25[6]. For every $q_i \in \mathcal{Q}$ where $\mathcal{Q}$ is the total query set, $\mathcal{M}$ will return the intial retrieved documents $\mathcal{D}_{q_i}^{\mathcal{M}} = [d_1^{q_i}, d_2^{q_i}, ..., d_m^{q_i}]$. This completes the first stage of ranking. $q_i \in \mathcal{Q}$ and initial retrieval results $\mathcal{D}_{q_i}^{\mathcal{M}}$ will be passed to cross-encoder model $\mathcal{N}$, in our case, Bert, and generate the reranked version of $\mathcal{D}_{q_i}^{\mathcal{M}}$ which we define as $\mathcal{R}_{q_i}^{\mathcal{N}}$. Similar to the definition of Bigdeli et al.[10], the purpose of bias-aware negative sampling training is to train a neutral ranking model $\mathcal{N}'$ which has same architecture as typical $\mathcal{N}$ and only differ in negative training samples when doing finetuning and satisfying the following conditions

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} Bias(\mathcal{R}_{q_i}^{\mathcal{N}'}) < \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} Bias(\mathcal{R}_{q_i}^{\mathcal{N}}) \tag{3.1}$$

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} Utility(\mathcal{R}_{q_i}^{\mathcal{N}'}) \simeq \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} Utility(\mathcal{R}_{q_i}^{\mathcal{N}}) \tag{3.2}$$

### 3.1.2 Gender inclination determination

Bigdeli et al.[10] first propose the lightweight training strategy for reducing gender bias in the ranking model by selecting those negative samples with gender inclination instead of randomly selecting negative samples from the initial ranking of BM25[6]. In their work, the gender inclination of any text is determined by gender magnitude calculation from ARaB[9], a term-match-based evaluation metric for evaluating gender bias within ranking lists. Our work sees that the term-based method cannot understand the contextualized meaning of words within sentences and think the gender inclination could be determined by neutral models instead and achieve possible performance improvements.

Our work is divided into two subtasks. The first task is to train a gender sentence labeler based on pre-training models, in our case, Bert[5]. This task is referred to as gender sentence classification in the following sections. The second task is to utilize the gender labeler to select negative training samples for ranking models so the overall gender bias of ranking results from the ranking models will be reduced while maintaining good retrieval effectiveness.

## 3.2 Gender sentence classification

### 3.2.1 Problem Definition

This work treats gender sentence classification as a multi-label classification problem. Examples of gender annotated queries from Rekabsaz et al.[9] are shown in the table 3.1

Table 3.1: Examples of sentence gender labels

| Sentences | Labels |
|---|---|
| who was known as the heretic king | Male |
| which prep school did president kennedy attend | Male |
| how popular is the name katie | Female |
| is blake lively pregnant | Female |
| who plays the main character in night at the museum | Neutral |
| how much sleep in one day does a baby need | Neutral |

Formally, the gender sentence classification problem is of finding the model that maps input $X$ to $L$, where $X$ here refers to the sentence collections $\{s_1, s_2, ..., s_n\}$ and $L$ refers to the gender label set {Neutral, Female, Male}.

To imporve the overall performance, a large amount of labeled data is required for finetuning a pre-training model to gender sentence classification. Unfortunately, there is not much available dataset for this purpose. Rekabsaz et al.[9] provides 3,750 queries that are humanly labeled into four categories, Non-gendered, Female, Male, Other, or Multiple Genders. The total number of training samples is too small for a pre-training model to do great finetuning. On the other hand, the queries are generally short sentences that mainly consist of no more than ten words, while the training samples we would like the model to select from MS-MARCO[7] are of passage length. Because of these two reasons, this work creates a new dataset for training Bert[5] into gender sentence classification problems.

### 3.2.2 Gender classification dataset

C4[19] is our starting point for creating the new dataset, and it is the dataset Google used for training T5. C4[19] is based on April 2019 snapshot of the Common Crawl dataset with special scripts to include only natural language and remove unnecessary duplication. Each data of C4 consist of three fields: url, text, and timestamp. We remove the url and timestamp for this work. The training set of C4 has 364,868,892 samples. Considering the computation expense and time, this work only chooses 35,631,700 samples out of the total. For each training sample we selected, the text part is tokenized based on words and then reduced to no more than 400 tokens per sample.

To generate gender labels from label set L {Neutral, Female, Male} for the each sample, this work adopted the similar idea of ARaB[9] and NFaiRR[12] when calculating the male/female magnitude of one document. Pre-defined gender definitional words, which are highly representative gender words such as “she” and “girl” for female and “he” and “boy” for male, of a total number of 326 are used in this process. We define the male/female magnitude of document/text as the number of occurrences of each word in the male/female words set.

$$mag^f(d) = \sum_{w \in \mathsf{V}_f} \#\langle w, d\rangle \tag{3.3}$$

$$mag^m(d) = \sum_{w \in \mathsf{V}_m} \#\langle w, d\rangle \tag{3.4}$$

where $\#\langle w, d\rangle$ denotes the number of times word w shows in document d, $\mathsf{V}_f$ and $\mathsf{V}_m$ represents female words and male words in pre-defined gender-definitional words, respectively. Using the magnitude, we label each training samples with the label female or male when satisfying the equation 3.5 or 3.6

$$mag^{f}(d) \geq \alpha \tag{3.5}$$

$$mag^{m}(d) \geq \beta \tag{3.6}$$

Both $\alpha$ and $\beta$ are free parameters. In this work, we choose $\alpha$ and $\beta$, both equal to 5, based on our attempts. If the total number of training samples is significant, we suggest increasing $\alpha$ and $\beta$, and these two parameters do not have to the same.

To label the neutral document, we utilize the concept of document neutrality from Rekabsaz et al.[12].

$$N(d) = 1 - \sum_{a \in A} |\frac{mag^{a}(d)}{\sum\limits_{x \in A} mag^{x}} - \mathcal{J}_a| \tag{3.7}$$

where $mag^{a}$ represents the male/female magnitude just introduced. $\mathcal{J}$ is a random variable that represents the ideal proportion of each member $a \in A$. In this work of gender bias study, A only has two members, female and male. Since $\sum_{a \in A} \mathcal{J}_a = 1$, $\mathcal{J}_{female}$ and $\mathcal{J}_{male}$ here are both 1/2. Then we label the document as neutral if and only if

$$N(d) = 0 \tag{3.8}$$

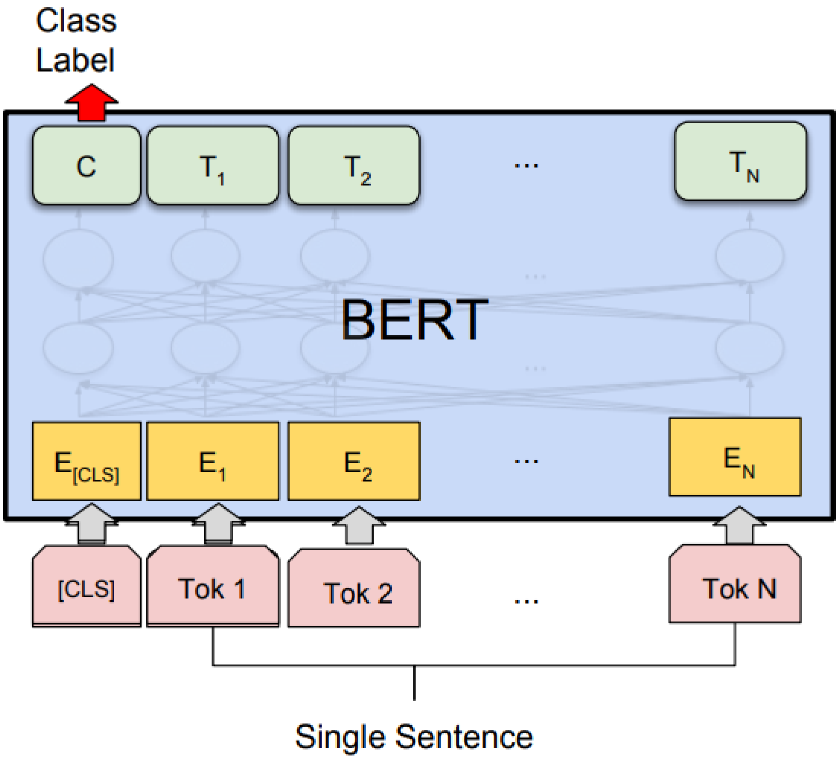
In the end, 930,358 training samples from C4[19] are labeled and formulated into our dataset, out of which 400,000 training samples are neutral, 345,379 training samples are labeled Male , and 184,979 training samples are labeled female. The number of neutral samples are intentionally chosen in order to maintain the balance. The actual number of neutral samples is significantly large.

### 3.2.3 Gender classification model

The training process of the gender classification model follows the two-step finetuning[11], transfer and adapt. The main idea is first to train the model toward extensive but not "perfect" data, which is the transfer step. Then we further finetune the resulting model from step one to a small but high-quality dataset. The two-step training process helps improve the performance for training scenarios when the "perfect" datasets are small. In our case, the golden label is 3,750 from Rekabsaz et al.[9]. The gender classification dataset introduced in section 3.2.2 is created for the first transfer step.

We follow the classical finetuning process as shown in Figure 3.2 for both the transfer and adapt steps. We split our dataset to 80%/20% for training set and validation set. We use Bert for finetuning and use batch size of 32 and train totally one epoch of our dataset. The final model achieves an accuracy of 0.8066 and an F1 score of 0.82198 on validation set.

Fig. 3.2: Bert for multi-label classification



# Chapter 4

# Experiments

Table 4.1 shows the environment set up for our experiments.

Table 4.1: Environment of Experiments

| | |
|---|---|
| CPU | Intel(R) Xeon(R) CPU E5-2620 |
| Memory Size | 126GB |
| GPU | NVIDIA GeForce RTX 3090 |
| GPU Memory Size | 24GB |

## 4.1 Datasets

We use MSMARCO[7], which contains over 8.8M passages in the document collections and over 500k queries with at least one relevant passage associated with the collection. There are 397,768,673 queries-passage records from MSMARCO. We choose the threshold of gender sentence labeler to be 0.4 meaning that label that has a possibility greater than 40% will be chosen. Note that this is a free parameter, and 0.4 is our empirical decision here, and this parameter shows the flexibility of our model. In the end, 37,975,116 out of 397,768,673 queries-passage records are labeled as either Male or Female.

For the query set, we can not directly use the query from MSMARCO. In order to show the bias of ranking results, gender-neutral queries have to be adopted so that the bias within ranking results could be regarded as bias within ranking models instead of from input queries. For this purpose, we adopt two query sets[9, 12], which contains 1765 and 215 gender-neutral queries from MSMARCO query set with the help of human annotators.

## 4.2 Evaluation Metrics

### 4.2.1 Retrival effectiveness

This work uses MRR, Mean reciprocal rank, to evaluate the utility of ranking models, mainly because we are using MSMARCO as our document collection. Formally,

$$MRR = \frac{1}{|\mathcal{Q}|}\sum_{i=1}^{|\mathcal{Q}|}\frac{1}{rank_i} \tag{4.1}$$

where $\mathcal{Q}$ is the query set we have, and $rank_i$ refers to the first relevant document rank for the ith query. For example, if the first relevant document for a query is ranked in 3rd position, the score for that query will be 1/3. In our experiment, we use MRR@10, which means we only care about the first 10th position of the ranking list, and if any relevant document is ranked out of 10, then a score of 0 will be given for that query.

### 4.2.2 Gender bias evaluation

We adopt ARaB[9] and NFaiRR[12] to evaluate the model gender bias level, and both metrics will be briefly introduced here. For detailed information, please refer to the original paper referenced here. As mentioned in section 3.2.2, both ARaB and NFaiRR rely on a set of pre-defined gender words, such as “she” for female and “he” for male. There are a total of 326 words at the time of this work.

ARaB, *Average Rank Bias*, has two variants called TF and Boolean, and the only difference is how it calculates the gender magnitude of a single document.

$$TF : mag^f(d) = \sum_{w \in \mathsf{V}_f} \log \#\langle w, d\rangle \tag{4.2}$$

$$Boolean : mag^f(d) = \begin{cases} 1, if \sum_{w \in \mathsf{V}_f} \#\langle w, d\rangle > 0 \\ 0, otherwise \end{cases} \tag{4.3}$$

Here $f$ refers to female, The equations show female magnitude as examples, and the male magnitude calculation is done similarly. Then, for each query $q$,

$$qRaB_t^f(q) = \frac{1}{t}\sum_{i=1}^{t} mag^f(d_i^q) \tag{4.4}$$

Here we assume that there are top t documents retrieved for the query $q$. Then we calculate the average of the qRaB scores for each ranking position of the given query $q$

$$qARaB_t^f(q) = \frac{1}{t}\sum_{x=1}^{t} qRaB_x^f(q) \tag{4.5}$$

$$ARaB_t(q) = qARaB_t^m(q) - qARaB_t^f(q) \tag{4.6}$$

$$ARaB_t = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} ARaB_t(q) \tag{4.7}$$

NFaiRR uses a more straightforward magnitude calculation and uses this magnitude to calculate document neutrality as equation 4.8 and equation 4.9 demonstrate. A more detailed illustration of these two equations can be found in section 3.2.2.

$$mag^f(d) = \sum_{w\in \mathsf{V}_f} \#\langle w, d\rangle \tag{4.8}$$

$$N(d) = 1 - \sum_{a\in A} |\frac{mag^a(d)}{\sum\limits_{x\in A} mag^x} - \mathcal{J}_a| \tag{4.9}$$

With this neutrality score, FaiRR, *Fairness of Retrieval Results* of query $q$ and ranking list $L$ could be defined

$$FaiRR_q(L) = \sum_{i=1}^{t} N(L_i^q)p(i) \tag{4.10}$$

where $t$ is the cut-off value for the ranking list, $L^q$ is the ranking list of query $q$, and $p$ is the position decay defined below

$$p(i) = \frac{1}{\log_2(1+i)} \tag{4.11}$$

In order to normalize the FaiRR score so that this metric would be used across collections and datasets, *Ideal FaiRR* or IFaiRR is introduced. First, a background document set $\hat{S}^q$ of every query $q$ is obtained by getting the top k documents in the ranking list. So the previous $L^q$ is one way to rank this background document set. Then, based on the neutrality score, we can get the best possible FaiRR for ranking results of this query $q$, which is $IFaiRR_q(\hat{S})$. Then, we use it to normalize FaiRR and get NFaiRR, *Normalized Fairness of Retrieval Results*, for every query $q$

$$NFaiRR_q(L, \hat{S}) = \frac{FaiRR_q(L)}{IFaiRR_q(\hat{S})} \tag{4.12}$$

Then NFaiRR score of the any IR model is defined below

$$NFaiRR(L, \hat{S}) = \sum_{q\in\mathcal{Q}} NFaiRR_q(L, \hat{S}) \tag{4.13}$$

Generally, ARaB measures the bias within the retrieved results, whereas NFaiRR measures the fairness of the results. So for any bias aware ranking model, we want less ARaB score and a higher NFaiRR score(the highest will be 1).

## 4.3 Experiment results and Discussions

In order to compare our results to Bigdeli et al.[10], which has only nearly 8M training samples, we also limit the training samples to 8M. We trained 5 Bert models with the same total number of training samples but a different ratio of gender-related training samples within them. The ratio of gender-related training samples varies from 1/8 to 5/8, and the number of gender-related training samples varies from 1M to 5M.

First, we example how this ratio would influence the retrieval effectiveness, and we calculate the MRR score of each model on the reranked results of the 1765 gender-neutral queries(QS1) and 215 gender-neutral queries(QS2), as figure 4.1 shows

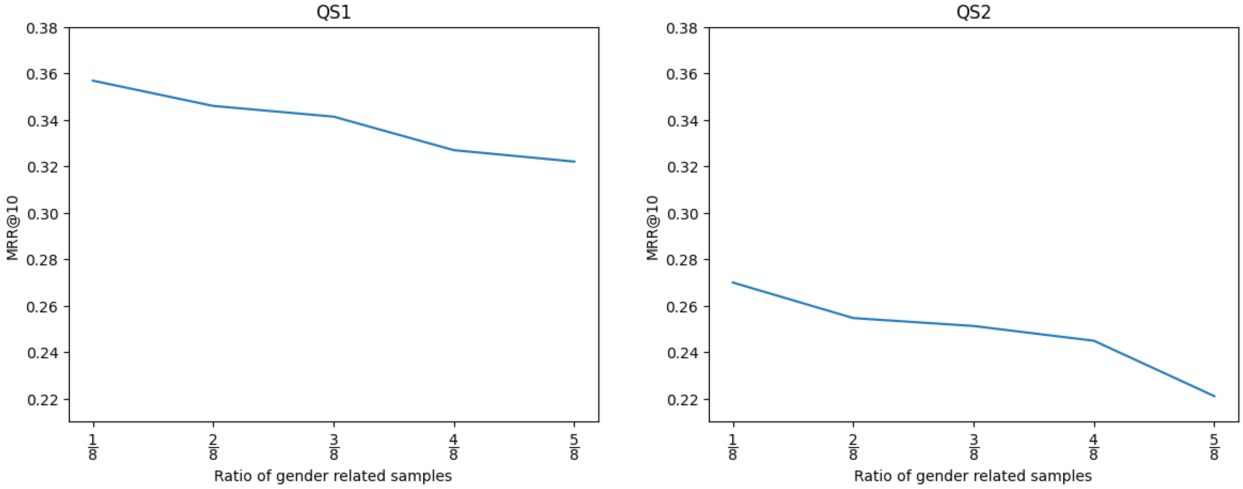


Fig. 4.1: Retrival effectiveness measured on MRR over QS1 and QS2

We can see that with the increased ratio of gender-related negative samples, the final model's retrieval effectiveness will be reduced. Namely, the gender-related training samples we labeled will negatively influence the model's overall utility. This observation is the same as the previous study[10]. After all, bias awareness ranking aims to reduce gender bias while maintaining reasonable retrieval effectiveness. Next, we examine to what extent our model could reduce gender bias with the trade of retrieval effectiveness. We evaluate the gender bias based on NFaiRR, ARaB TF and Boolean, as figure 4.2 and figure 4.3 demonstrate.

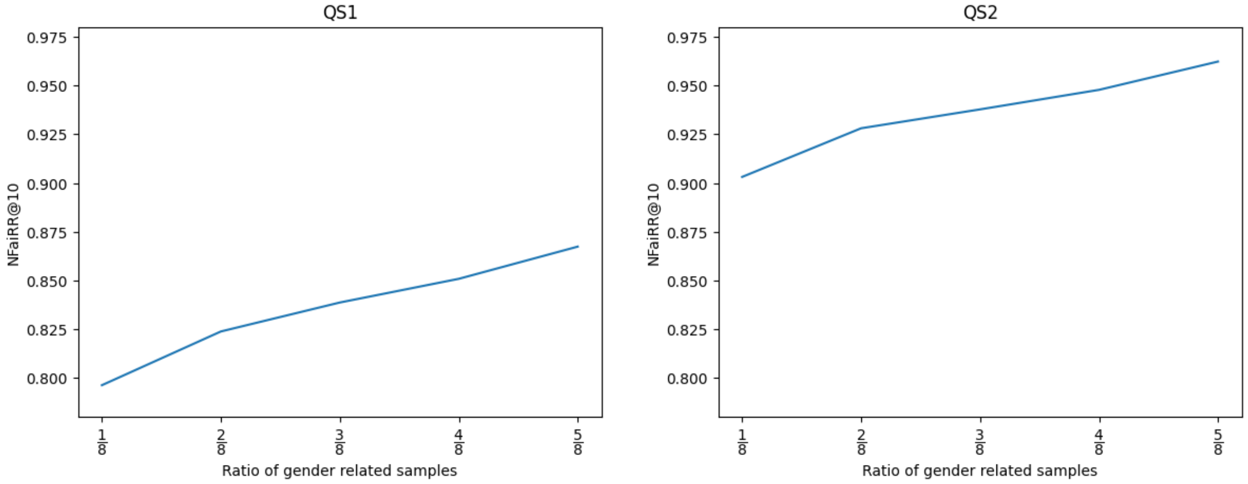


Fig. 4.2: Gender bias of Retrieved results of our models measured on NFaiRR

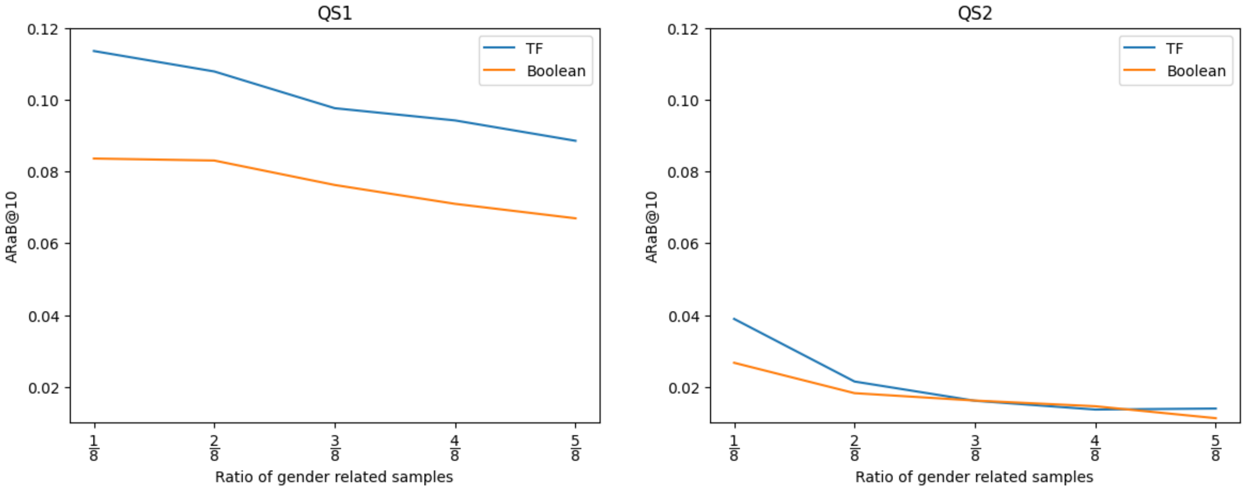


Fig. 4.3: Gender bias of Retrieved results of our models measured on ARab

It is worth noting that the higher NFaiRR score and lower ARaB TF and Boolean are expected for lower gender bias within the reranked list of our models. Figure 4.2 and Figure 4.3 shows a consistent trend of gender bias changes over all three measurements. When we increase the ratio of gender-related samples, the gender bias decrease with no exception. To determine the good ratio to use and see how our models perform compared with other models, we trained two extra models with completely original training samples and with the use of lightweight negative sampling method[10]. Table 4.2 shows our comparisons over retrieval effectiveness and gender bias level.

Table 4.2: Retrieval effectiveness and level of gender bias across models on QS1 and QS2

| Query | Training Method | MRR@10 | NFaiRR@10 | ARaB@10(TF) | ARaB@10(Bool) |
|---|---|---|---|---|---|
| QS1 | Original | 0.3455 | 0.7826 | 0.1260 | 0.0963 |
| | Lightweight | 0.3236 | 0.8631 | 0.0987 | 0.0879 |
| | Ours 3M/8M | 0.3414 | 0.8388 | 0.0976 | 0.0763 |
| | Ours 5M/8M | 0.3220 | 0.8674 | 0.0886 | 0.0670 |
| QS2 | Original | 0.2308 | 0.8956 | 0.0275 | 0.0187 |
| | Lightweight | 0.2209 | 0.9569 | 0.0250 | 0.0186 |
| | Ours 3M/8M | 0.2592 | 0.9378 | 0.0215 | 0.0169 |
| | Ours 5M/8M | 0.2211 | 0.9623 | 0.0140 | 0.0112 |

3M/8M means the model was trained on 3M gender-related and 5M original samples, and 5M/8M means 5M gender-related and 3M original samples. We put 3M/8M here because this ratio is close to the ratio of the gender-related samples used by lightweight[10]. The MRR@10 shows that our 3M/8M model, with a similar ratio of gender-related samples, achieves much better retrieval effectiveness than lightweight[10] on reranked results on both QS1 and QS2. Adopting the idea of balancing the effectiveness and gender bias, this outperformed retrieval effectiveness gives us more room to include more gender-related samples to reduce gender bias while controlling the effectiveness. Our 5M/8M model shows precisely this. As the gender-related samples increase from 3M to 5M, the retrieval effectiveness drops but still maintains at a reasonable level similar to what lightweight has. Meanwhile, in terms of NFaiRR@10 and ARaB@10, our 5M/8M model has less gender bias detected, and is it more bias-aware when reranking the results. Our model has an excellent performance in ARaB as even the 3M/8M model has smaller ARaB values than lightweight[10] and original.

# Chapter 5

# Conclusion

In this paper, we create a multi-label classification dataset from C4 and finetune a new pre-training model that labels passage as female, male or neutral based on the newly created dataset. This model is then used to reduce gender bias within pre-training ranking models by adopting the negative sampling methods and selecting training examples with gender inclination. Our experiments show that the ranking models trained by our selected samples decrease the gender bias within retrieval results and maintain good retrieval effectiveness.

# Chapter 6

# Future Work

For future work, there are several directions worth exploring beyond this point.

First, the dataset we created from C4 only uses around 10% of total samples because using more data would be very time-consuming. Given the the conclusion that our model performs very well, it is interesting to further explore by utilizing more data from C4 and exploring gender bias beyond male and female.

Second, although our experiments and previous studies have shown that adding bias awareness would lead to a drop in retrieval performance, it is worth researching why this is the case. After all, all these gender-related negative samples are in the original training samples, and our method is to discover them and put them in a place where they can attract more attention.

Third, we only use Bert in our work, but the negative sampling method may work better in smaller models, which could have more real-world meaning. So doing experiments on smaller models like Bert-tiny or distillBert is also interesting.

# Acknowledgements

This research was undertaken with the support from The Real Sakai Laboratory [20], Waseda University. I would like to express my thanks of gratitude to members in The Real Sakai Laboratory, who helped me and gave me many useful pieces of advice for continuing the research. Also, I would like to thank Professor Tetsuya Sakai, who gave me this great opportunity and lots of supports to do research on this topic.

# References

[1] Ricardo Baeza-Yates. Bias on the web. *Commun. ACM*, 61(6):54–61, may 2018.

[2] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. *CoRR*, abs/1805.01788, 2018.

[3] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[6] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[7] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2016.

[8] David Rau and Jaap Kamps. How different are pre-trained transformers for text ranking?, 2022.

[9] Navid Rekabsaz and Markus Schedl. Do neural ranking models intensify gender bias? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2065–2068, New York, NY, USA, 2020. Association for Computing Machinery.

[10] A. Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. A light-weight strategy for restraining gender biases in neural rankers. In *ECIR*, 2022.

[11] Siddhant Garg, Thuy Vu, and Alessandro Moschitti. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection, 2019.

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

[13] A. Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. Exploring gender biases in information retrieval relevance judgement datasets. In *ECIR*, 2021.

[14] Christine Basta, Marta Ruiz Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. *CoRR*, abs/1904.08783, 2019.

[15] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard S. Zemel. Understanding the origins of bias in word embeddings. *CoRR*, abs/1810.03611, 2018.

[16] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. On the orthogonality of bias and utility in ad hoc retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1748–1752, New York, NY, USA, 2021. Association for Computing Machinery.

[17] Shirin Seyedsalehi, A. Bigdeli, Negar Arabzadeh, Bhaskar Mitra, Morteza Zihayat, and Ebrahim Bagheri. Bias-aware fair neural ranking for addressing stereotypical gender biases. In *EDBT*, 2022.

[18] George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. Mitigating bias in search results through contextual document reranking and neutrality regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2532–2538, Madrid Spain, Jul 2022. ACM.

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019.

[20] RSL. The sakai laboratory. sakailab.com Last updated: 26th March 2019.