Mitigating Gender Bias in Job Ranking Systems Using Job Advertisement Neutrality

Deepak Kumar*, **Shahed Masoudian***
Alessandro B. Melchiorre, **Markus Schedl**
Johannes Kepler University Linz, Austria
Linz Institute of Technology, AI Lab
{deepak.kumar,shahed.masoudian,alessandro.melchiorre,markus.schedl}@jku.at

Abstract

Transformer-based Job Ranking Systems (JRSs) are vulnerable to societal biases inherited in unbalanced datasets. These biases often manifest as unjust job rankings, particularly disadvantaging candidates of different genders. Most bias mitigation techniques leverage candidates' gender and align gender distributions within the embeddings of JRSs to mitigate bias. While such methods effectively align distributional properties and make JRSs agnostic to gender, they frequently fall short in addressing empirical fairness metrics, such as the performance gap across genders. In this study, we shift our attention from candidate gender to mitigate bias based on gendered language in job advertisements. We propose a novel neutrality score based on automatically discovered biased words in job ads and use it to re-rank the model's decisions. We evaluate our method by comparing it with different bias mitigation strategies and empirically demonstrate that our proposed method not only improves fairness but can also enhance the model's performance.

1 Introduction

Nowadays, transformer-based language models (LMs) are being used for a variety of tasks such as document classification (Adhikari et al., 2019; Kong et al., 2022), information retrieval (Rekabsaz et al., 2021), text generation (Raffel et al., 2020), and recommender systems (RecSys) (Sun et al., 2019). Despite their effectiveness, these models tend to inherit societal biases (e.g., gender bias) present in their training data. Recent studies have concentrated on analyzing the impact of these biases on model decision-making and developing strategies to mitigate them, through preprocessing (Park et al., 2018), in-processing (Kumar et al., 2023b), or post-processing (Pour et al., 2023). Among various applications, the usage of LMs as RecSys, particularly as Job Ranking Systems (JRSs), is of significant importance. Minor alterations in the ranking of JRSs with the contribution of the sensitive attributes can lead to discrimination against certain demographic groups (e.g., females or older individuals). Research in this domain has focused on leveraging encoder or decoder LMs to reduce bias in the job advertisement recommendations for various demographic groups (Rus et al., 2022). Common mitigation strategies often utilize candidates' sensitive attributes (e.g., gender, age), as labeled data to render the model's embeddings agnostic to target attributes (Bhardwaj et al., 2021). In this study, we propose a novel approach that leverages implicit bias within job advertisements to mitigate gender bias on encoder LMs. Instead of relying on candidates' gender, we introduce a new neutrality score, calculated based on implicit biased terms that are automatically derived from job advertisements. We implement our method on two encoder LMs namely BERT-Base and DistillRoBERTa following previous works and due to their strong contextual understanding and representational power to encode natural language. We evaluate our proposed re-ranking strategy and compare it with other successful bias mitigation techniques. Our findings demonstrate that our proposed method not only enhances the model's fairness but can also yield improvements in performance on the primary task—an outcome not achieved by other methods. In summary, our contributions are as follows: (1) We introduce a novel neutrality score derived from implicit biased terms present in job advertisements. (2) We demonstrate that re-ranking jobs according to our neutrality score enhances both fairness and task performance. The code for our study is available at the following link: GitHub.

*These authors contributed equally to this work

Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI), pages 264–271

2 Related Work

JRSs, similar to LMs, suffer from various societal biases (Amer-Yahia et al., 2020) and have been investigated in the past on various popular platforms (Tang et al., 2017; Zhang, 2021; Amer-Yahia et al., 2020). The mitigation approaches for these biases are mostly focused on pre-processing approaches (Kumar et al.), such as replacing gendered pronouns with gender-neutral pronouns (Rus et al., 2022) or directing candidates to dedicated JRSs for particular attributes (Shishehchi and Banihashem, 2019; Ntioudis et al., 2022). Rus et al. (2022) also try in-processing bias mitigation using adversarial debiasing. They tried to make hidden representation agnostic to the candidate's gender adversarially. A post-processing debiasing of JRS is investigated by Li et al. (2023) through reranking the model output based on the candidate's gender to achieve a fairness constraint over the whole dataset.

The work most closely related to ours is that of Rekabsaz et al. (2021), who introduced neutrality score based on explicit bias words derived from a pre-defined dictionary to enforce neutrality in information retrieval. Our approach diverges from theirs in several key aspects. Firstly, our focus is implicit gendered language in job advertisements, building on the methodology established by Kumar et al. (2023a) for candidate ranking systems. Additionally, we formulate our neutrality score based on the biasedness of words rather than relying solely on binary gendered terms. Lastly, we apply our neutrality score directly to the ranking process of the model, enhancing both its performance and neutrality.

3 Methodology

To find better representation between genders, we introduce a three-stage approach: (1) we acquire the biased words in job advertisements and their biasedness(Section 3.1). (2) we use specific words assigned for each class of job and introduce a neutrality score based on their biasedness and frequency of usage in job advertisements(Section 3.2). (3) we utilize the new job advertisement neutrality score to re-rank the jobs.

3.1 Acquiring Biased Words

In order to extract the implicit biased words, we follow the footsteps of Kumar et al. (2023a) on candidate ranking system. We introduce gender counterfactual of the CVs and unitize integrated gradient (Sundararajan et al., 2017) to find the contribution of words in job advertisement towards the ranking score of candidates and their gender counterfactual. Then we normalize and scale the ranking scores according to the rank of the candidate. Finally we average over all job advertisements belonging to the same job class. We call these values the biasedness of the words, and the words with biasedness above a certain threshold are bias words.

3.2 Neutrality Score

In order to obtain the neutrality score, we collect a bag of the top 20 bias words* for each job class with normalized biasedness score. Then, we calculate the neutrality score (N) for each document ($D = d_1, d_2, ...d_n$) based on the frequency of occurrence ($f_w^{d_i}$) of each of the bias words (w) in the respective document (d_i) and the biasedness (b_w) of words (w) following equation 1.

$$\mathrm{N}_{d_i} = \begin{cases} 1, & \text{if } \sum_{w \in Top_{20}} f_w^{d_i} \leq 1 \\ 1 - \frac{\sum_{w \in Top_{20}} b_w f_w^{d_i}}{\sum_{w \in Top_{20}} f_w^{d_i}}, & \text{otherwise} \end{cases} \quad (1)$$

The neutrality score, ranging from 0 to 1, reflects the level of bias in a job ad. Considering gender was used as an indirect bias indicator, we expect increasing neutrality in recommended documents to help make the model fairer toward gender subgroups.

3.3 Re-ranking

Re-ranking of documents serves as an effective post-processing technique to enhance the neutrality of the model. In this approach, the model initially ranks job advertisements based on their relevance scores. Subsequently, we take top-ranked advertisements (the top 10 advertisements based on relevance), and re-order according to a neutrality score, thus improving the overall neutrality of the recommendations.

4 Experiment Setup

4.1 Dataset

The dataset is based on job advertisements from UK portals and candidates are biographies from

*We used BERT-base for finding biased words. Given the context length of the model to be 512, we put threshold on the biasedness of individual word to be above 10/512. This choice led to 20 words found on average per job advertisement. We tried 1/512 and 100/512 threshold too, this led to low neutrality for all ads and neutrality being almost binary respectively.

the BIOS dataset (De-Arteaga et al., 2019). First, we match the labels in the job advertisements and the labels in biographies to create ground truth relevance. We only keep job classes with at least 10 job advertisements. Then, we replaced all names with Bob for male candidates and Alice for female candidates. This helps us to mitigate the effect of the degree of genderdness that different names have. As another pre-processing step, we remove the mention of the current profession from biographies to make the task more difficult. Subsequently, biographies are sampled to ensure equal distribution across all job classes, i.e., 200 candidates per job class. Furthermore, we try to mimic the real-world gender distribution of the UK job landscape for each job class. For each job class in our dataset, we collect the most recent gender distribution from different sources (See Appendix). The resulting dataset contains 2085 job advertisements for 14 job classes and 200 biographies for each job class. The biographies are split into train, test, and validation splits of 70, 20 and 10 percent. We load the training set with 4 negative samples for each positive sample.

4.2 Models

We use CrossEncoder (Reimers and Gurevych, 2019) as our JRs, and we run CrossEncoder with BERT-Base (Devlin et al., 2018) and DistilRoBERTa (Liu et al., 2019). Both models are transformer-based encoder language models used for various natural language processing, such as document classification and information retrieval. The models are based on a self-attention mechanism, which allows them to focus on specific parts of the sequence that the model deems to be informative about the task. We use BM25 (Lin et al., 2021) as our initial ranker and CrossEncoder as our final ranker for both training and evaluation. This helps us achieve high performance.

4.3 Debiasing methods

Data Augmentation: A baseline pre-processing approach to mitigate bias in language models is to balance the presence of females and males before training. We used balancing with weighted sampling between males and females of each job class. For weights we calculate the proportion of female to male for each class and multiply it by the total proportion of female to male appearance in the dataset. (e.g., $Weight^{doctor}_{female} = Weight_{\frac{male}{female}} \frac{\#\text{male doctor}}{\#\text{female doctor}}$)

Regularization: In this method the task head which is responsible to estimate the relevancy of the document is used to estimate the neutrality as well adding a new optimization loss to the model. In other words we are forcing the network to rank documents not just by relevancy but neutrality as well. The overall objective loss is binary cross-entropy loss at its core for relevancy, as shown in Eq. 3 where z_i is the logits of the language models. For regularization, we use L1 distance between neutrality scores and logits of the language model. λ is the regularization coefficient which determines the power of regularization. Equation 2 shows the overall loss of the proposed regularization method.

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda \mathcal{L}_{regularization} \quad (2)$$

$$\mathcal{L}_{task} = y_i \log \sigma(z_i) + (1 - y_i) \log(1 - \sigma(z_i)) \quad (3)$$

4.4 Training and Evaluation

Training: We train models for 15 epochs with a learning rate 1×10^{-5}. Training use *AdamW* optimizer and $\lambda = 2$ which proved best in our experiments. We avoid using any early stopping as for multi-optimization objectives there is no clearly defined method to stop model training. Instead, we slow down training by using 3 epochs warm-up and linear decay of learning rate until the end of training which helps the model settle down toward the end of training. We report the mean and standard deviation of the results over 3 independent runs to account for variations.

Evaluation: For the evaluation we used Normalized Discounted Cumulative Gain (NDCG) of top 10 scored job advertisements averaged over all users as the main ranking task. We also evaluate our model on several fairness metrics as follow:

Neutrality. As baseline for evaluation we use our own introduced neutrality score and check the average top 10 job ad neutrality after re-ranking to compare with other bias mitigation methods.

Performance Gap.(Deldjoo et al., 2024; Wang et al., 2023) Performance gap between males and females is an indicator of empirical fairness of the model. Ideally, the performance gap between different demographic groups should be zero. For this metric we calculate the NDCG of the top 10 ranked job ads for our target attribute $\rho = male, female$ and consider the difference as Gap: $Gap = |NDCG@10_{male} - NDCG@10_{female}|$

Table 1: Task and fairness performance result of BERT-Base and DistilRoBERTa trained on job advertisement ranking dataset with different debiasing methods like data-augmentation, regularization, and re-ranking.

Model	NDCG@10 ↑	Neut ↑	Gap ↓	p_{value} ↑	LDR ↓	CF_{Gap} ↓
BERT-Base	$0.812_{0.005}$	$0.738_{0.001}$	$0.117_{0.002}$	$< 10^{-3}$	$0.738_{0.027}$	$0.027_{0.005}$
+data augmentaion	$0.798_{0.002}$	$0.737_{0.002}$	$0.126_{0.007}$	$< 10^{-3}$	$0.704_{0.036}$	$0.031_{0.004}$
+regularization	$0.744_{0.009}$	$\mathbf{0.821}_{0.008}$	$0.124_{0.002}$	$< 10^{-3}$	$\mathbf{0.441}_{0.053}$	$\mathbf{0.025}_{0.000}$
+re-ranking	$\mathbf{0.870}_{0.007}$	$0.738_{0.001}$	$\mathbf{0.065}_{0.001}$	0.005	$0.496_{0.039}$	$\mathbf{0.025}_{0.005}$
DistilRoBERTa	$0.779_{0.014}$	$0.735_{0.004}$	$0.138_{0.008}$	$< 10^{-3}$	$0.635_{0.052}$	$0.021_{0.005}$
+data augmentaion	$0.734_{0.021}$	$0.735_{0.002}$	$0.145_{0.011}$	$< 10^{-3}$	$0.669_{0.071}$	$0.031_{0.009}$
+regularization	$0.670_{0.011}$	$\mathbf{0.809}_{0.010}$	$0.127_{0.098}$	$< 10^{-3}$	$0.494_{0.042}$	$0.037_{0.002}$
+re-ranking	$\mathbf{0.843}_{0.013}$	$0.735_{0.004}$	$\mathbf{0.087}_{0.007}$	$< 10^{-3}$	$\mathbf{0.399}_{0.056}$	$\mathbf{0.019}_{0.005}$

We perform T-test between male and female NDCG with a threshold of 10^{-3} as significance test and report p-values.

Counterfactual Gap. We use the counterfactual dataset explained in 3.1 to calculate a new fairness metric. First, a counterfactual candidate ($\hat{c}$) is created based on the gender of the original candidate (c) for each candidate in the test set (C). Then, for each candidate, a gap is calculated between the model performance over the original and its counterfactual input:

$$CF_{Gap} = \frac{\sum_{c \in C} |NDCG@10_c - NDCG@10_{\hat{c}}|}{|C|}$$

List Difference Rate (LDR).(Zhang, 2021) We take a list-wise approach for our next fairness metric. Instead of calculating the NDCG difference between the ranked list (Q) for the original candidate (Q^c) and counterfactual candidates ($Q^{\hat{c}}$), we calculate the normalized Hamming distance between the two lists:

$$LDR@10 = \frac{\sum_{c \in C} \sum_{i=1}^{10} \mathbb{1}(Q@10_i^c = Q@10_i^{\hat{c}})}{|C|}$$

This metric measures the impact of altering the gender pronoun on the ranked list.

We compare the results of re-ranking a post-processing method with balancing a pre-processing method and regularization an in-processing method in section 5.

5 Results

As it can be seen from table 1 for both BERT-Base and DistilRoBERTa models, the baseline has a decent NDCG@10 performance with a high performance gap between genders. We can see that by applying data augmentation, the model's performance decreases while not affecting neutrality. Interestingly, data augmentation causes an increase in both Gap and CF Gap metrics but reduces the LDR. As for the regularization, we can observe that while also reducing performance on the main task, regularization manages to increase the Neutrality score

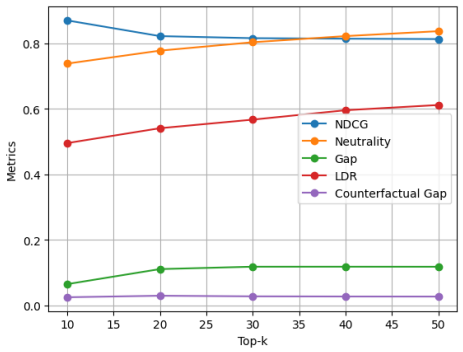


Figure 1: Changes in fairness metrics for BERT-Base as we increase the k in the re-ranking of top-k retrieved documents

but still fails to reduce the Gap between male and female performance of the model. Also, it can be seen that regularization manages to reduce the LDR and CF Gap. This is due to regularization trying to increase neutrality at the cost of relevance. Finally, on both models with re-ranking, we can observe that re-ranking based on the neutrality score significantly increases the model's performance while having the best reduction in Gap and CF Gap. it is noticeable that although LDR of the model is higher than regularization still compared to baseline the LDR metric is reduced. With p-value, we can observe that the male and female NDCG are indistinguishable only for BERT re-ranking. As expected, we can see that re-ranking based on neutrality on the top 10 relevant results has no effect on the overall neutrality score. We also analyzed the re-ranking of the results based on neutrality score on more than the top 10 rankings (Fig.1) and observed that as the number of top candidates increases, the neutrality, LDR, the gap increases while NDCG@10 decreases. This means that the bias mitigation effect decreases with the increase of top-k candidates for re-ranking, and at the same time, ranking performance also decreases. Which is similar to the regularization results.

6 Conclusions

In this study, we address bias in job ranking systems by introducing a novel neutrality score using the biasedness of words present in job advertisements. We employed this neutrality score as a reranking strategy following evaluation and demonstrated its effectiveness in enhancing model performance. Our results show that integrating the neutrality score not only mitigates bias but also improves overall performance metrics, offering an easy and effective approach to job ranking. In the future, we plan to target non-binary gender.

7 Limitations and ethical concerns

Our work has limitations along several dimensions. First, dataset is the most significant issue in the recruitment domain. Due to the sensitive nature of the job candidate's profile, there is an absence of a reliable dataset with CVs. We addressed the dataset issue by using biographies as an alternative. But our curated dataset itself is limited along several axes, such as small dataset, dataset from a specific geography, limited number of occupations, and assigned names. We plan to create an artificial dataset to resolve the problem in the recruitment domain. We use gender pronouns to infer binary gender from biographies, which don't cover the nuanced definition of gender and can be considered both a limitation and an ethical issue of the work at hand. This limits our study to a binary gender setting. We plan to resolve this issue by incorporating non-binary gender candidates into an artificially created dataset. Finally, we narrowed our study from broad existing language models that use different architectures, such as LSTM and RNNs, to transformer-based language models. Specifically, we conducted our experiments with BERT and RoBERTa, which limited the work's findings to transformer-based language models.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

Sihem Amer-Yahia, Shady Elbassuoni, Ahmad Ghizzawi, Ria Mae Borromeo, Emilie Hoareau, and Philippe Mulhem. 2020. Fairness in online jobs: A case study on TaskRabbit and Google. In *International Conference on Extending Database Technologies (EDBT)*, Copenhagen, Denmark.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation*, 13(4):1008–1018.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2024. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction*, 34(1):59–108.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jun Kong, Jin Wang, and Xuejie Zhang. 2022. Hierarchical bert with an adaptive fine-tuning strategy for document classification. *Knowledge-Based Systems*, 238:107872.

Deepak Kumar, Tessa Grosz, Elisabeth Greif, Navid Rekabsaz, and Markus Schedl. 2023a. Identifying words in job advertisements responsible for gender bias in candidate ranking systems via counterfactual learning.

Deepak Kumar, Tessa Grosz, Navid Rekabsaz, Elisabeth Greif, and Markus Schedl. Fairness of recommender systems in the recruitment domain: An analysis from technical and legal perspectives. *Frontiers in Big Data*, 6:1245198.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023b. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia. Association for Computational Linguistics.

Yunqi Li, Michiharu Yamashita, Hanxiong Chen, Dongwon Lee, and Yongfeng Zhang. 2023. Fairness in job recommendation under quantity constraints. In *AAAI-23 Workshop on AI for Web Advertising*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2356–2362.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Dimos Ntioudis, Panagiota Masa, Anastasios Karakostas, Georgios Meditskos, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2022. Ontology-based personalized job recommendation framework for migrants and refugees. *Big Data and Cognitive Computing*, 6(4):120.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Manasa Bharadwaj, Nikhil Verma, Ali Pesaranghader, and Scott Sanner. 2023. Count: Contrastive unlikelihood text style transfer for text detoxification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8658–8666.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Navid Rekabsaz, Robert West, James Henderson, and Allan Hanbury. 2021. Measuring societal biases from text corpora with smoothed first-order co-occurrence. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 549–560. AAAI Press.

Clara Rus, Jeffrey Luppes, Harrie Oosterhuis, and Gido H Schoenmacker. 2022. Closing the gender wage gap: Adversarial fairness in job recommendation. In *2nd Workshop on Recommender Systems for Human Resources, RecSys-in-HR 2022*. CEUR-WS.

Saman Shishehchi and Seyed Yashar Banihashem. 2019. JRDP: A job recommender system based on ontology for disabled people. *Int. J. Technol. Hum. Interact.*, 15(1):85–99.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Shiliang Tang, Xinyi Zhang, Jenna Cryan, Miriam J. Metzger, Haitao Zheng, and Ben Y. Zhao. 2017. Gender bias in the job market: A longitudinal analysis. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Trans. Inf. Syst.*, 41(3).

Shuo Zhang. 2021. Measuring algorithmic bias in job recommender systems: An audit study approach.

Job	Male,Female
Architect	69, 31
Photographer	85, 15
Psychologist	20, 80
Teacher	25, 75
Nurse	11, 89
Software Engineer	84, 16
Painter	68, 32
Personal Trainer	65, 35
Dietitian	6, 94
Dentist	46, 54
Interior Designer	17, 83
Senior Software Engineer	90, 10
Accountant	55, 45
Paralegal	37, 63

Table 2: UK job's gender distribution sources.

8 Appendix

8.1 A1

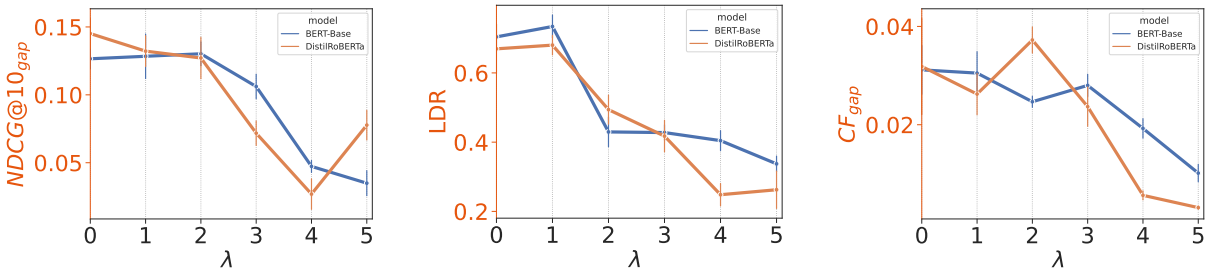
The UK job's gender distribution (Tab. 2) from multiple sources is used for replicating the gender distribution in our dataset.

The examples of words used for neutrality calculation are presented in Tab. 3. These words are not biased words from the human perspective but from the model's perspective. The objective of the work is not to remove these words from job advertisements but to reduce the bias effects caused by the presence of these words.

The effect of lambda over regularization is explored in Fig. 2.

Job	Biased words
senior software engineer	software, senior, engineer, development, team, engineering, experience, design, code, java
software engineer	software, engineer, team, development, experience, technology, engineering, data, code
dentist	dental, dentist, practice, associate, nhs, care, patients, clinical, private, patient
paralegal	legal, para, team, firm, law, litigation, client, property, role, commercial
nurse	nurse, nursing, nurses, residents, home, training, registered, clinical, shifts, team
teacher	school, pupils, teaching, teachers, children, teacher, students, staff, schools, curriculum
architect	architect, projects, design, architectural, practice, residential, team, working, architects
accountant	accountant, accounting, accounts, management, tax, finance, audit, reporting, business
painter	painter, decor, painters, painting, looking, shift, working, refurbishment, email

Table 3: Some examples of words used for neutrality score.



(a) $Gap_{NDCG@10}$ (b) LDR (c) CF_{Gap}

Figure 2: Changes in the different Fairness Metrics (a): performance gap (b): LDR (c): Counterfactual gap as we increase the regularization power by increasing λ