TagDebias: Entity and Concept Tagging for Social Bias Mitigation in Pretrained Language Models

Mehrnaz Moslemi and **Amal Zouaq**
Department of Computer and Software Engineering
LAMA-WeST Lab, Polytechnique Montréal
{mehrnaz.moslemi, amal.zouaq}@polymtl.ca

Abstract

Pre-trained language models (PLMs) play a crucial role in various applications, including sensitive domains such as the hiring process. However, extensive research has unveiled that these models tend to replicate social biases present in their pre-training data, raising ethical concerns. In this study, we propose the TagDebias method, which proposes debiasing a dataset using type tags. It then proceeds to fine-tune PLMs on this debiased dataset. Experiments show that our proposed TagDebias model, when applied to a ranking task, exhibits significant improvements in bias scores.

1 Introduction

Pre-trained language models (PLMs) are extensively utilized in various natural language processing tasks, acquiring a significant amount of knowledge during their pre-training phase. Research has highlighted that these models often inherit substantial social biases present in their pre-training corpora, which may subsequently emerge in the outcomes of downstream tasks (May et al., 2019; Zhao et al., 2018b). So, it is crucial to identify and mitigate social bias in these models.

There are different ways to mitigate social bias both in datasets and pre-trained models. State-of-the-art approaches show effective debiasing methods in PLMs such as increasing dropout regularization (Webster et al., 2020), projection-based debiasing (Liang et al., 2020), and self-debias (Schick et al., 2021) as a post-hoc method to discourage models from generating toxic sentences. Other data-based bias mitigation methods such as counterfactual data augmentation (CDA) (Zhao et al., 2018a) or biased terms removal (scrubbing) (De-Arteaga et al., 2019) have been proposed but exhibit some limitations. Producing counterfactual data and fine-tuning pre-trained language models (PLMs) on an augmented dataset is resource-consuming and, in some cases, impossible. For example, generating a counterfactual example for the sentence "women gave birth" is impossible. Scrubbing biased words removes contextual associations within the PLMs and can decrease model performance in downstream tasks.

In this paper, we propose a framework for mitigating bias in datasets and pre-trained language models by tagging the BiasinBios dataset (De-Arteaga et al., 2019), which is designed to examine gender-profession social biases. More specifically, we propose an approach named "TagDebias" to debias datasets by tagging gender indicator terms. The idea is to replace gender terms with semantic types that represent neutral terms for binary genders (female and male). We then fine-tune pre-trained language models on the debiased dataset, teaching them that each gender term corresponds to the same neutral tag. The proposed method, "TagDebias," has the advantage of not requiring counterfactual data while maintaining model performance compared to the scrubbing method. Furthermore, it outperforms data-based bias mitigation methods, specifically scrubbing and counterfactual data augmentation. To assess the fairness of the debiased models, we test our TagDebias model on a ranking task in the domain of biographies' ranking given a target job title.

In this study, we will answer the following research questions:

Q1 Does tagging stereotypical gender terms mitigate social bias in PLMs?

Q2 Does tagging stereotypical gender terms worsen PLMs' performance?

Q3 Does our proposed TagDebias model have a fairer ranking compared to base and scrubbed PLMs?

In response to Q1, we assess models with various tagging subsets using fairness classification metrics. Our findings reveal that the "Gender-specific-term" model surpasses both the initial and scrubbed

Findings of the Association for Computational Linguistics: NAACL 2024, pages 1553–1567

models. Evaluating model performance on the BiasinBios dataset (Q2), we observe that the tagging approach does not adversely affect model performance. Finally, the TagDebias model demonstrates a substantial enhancement in fairness rankings, exhibiting an improvement compared to the initial and scrubbed models, respectively (Q3).

2 Related works

2.1 Data debiasing techniques

There are different approaches to mitigate social bias in a dataset. One of these approaches is scrubbing gender indicator terms from the corpus (De-Arteaga et al., 2019). While this method mitigates social bias in pre-trained models, it removes context association within both the dataset and PLMs. Another data-based debiasing approach is counterfactual data augmentation (Zhao et al., 2018a; Zmigrod et al., 2019) along with gender swapping. This approach could mitigate bias in downstream tasks; however, it is resource-consuming. Another limitation is that gender swapping is not always possible in different contexts. For example, replacing "the woman is pregnant" with "the man is pregnant" may not be feasible. Due to these limitations in data-based bias mitigation methods, we propose a novel approach for debiasing the dataset by replacing gender indicator terms with a higher level of abstraction (types or tags). The closest work to ours is the Gender-tuning method (Ghanbarzadeh et al., 2023), which proposes the use of a mask language modeling task on gender terms and a modification of the loss function to include the examples generated by the MLM objective in the fine-tuning task. In this work, our goal is to examine the influence of dataset tagging on pre-trained language models, without making any modification to the loss function.

2.2 Fairness evaluation methods

Studies revealed that ranking systems have the potential to exacerbate stereotypical biases present within datasets (Perego et al., 2016). BERT-based ranking models are commonly employed to rank passages in response to a query, using relevance scores calculated through methods like confidence scores, cosine similarity, dot product, or other similarity metrics applied to BERT embeddings. Given that pre-trained models can inadvertently learn biases during their pre-training phase, various techniques have emerged to mitigate bias. In (Rekabsaz et al., 2021), BERT rankers, in conjunction with adversarial learning, are applied to rank passages based on the query using two loss functions. These loss functions ensure that the most related passages with fewer biased identity terms in the representations will be returned. While these dual loss functions strive to find an optimal trade-off between relevance and fairness, there remains a concern that these results could lead to a local optimum (Seyedsalehi et al., 2022). To address this concern, the Bias-aware Fair Ranker model (Seyedsalehi et al., 2022) proposes the penalization of passages when biased terms are found in the passages. This strategy empowers the model to excel in retrieving the most relevant passages at the top of the ranking while relegating the most biased, irrelevant passages to the bottom of the ranking.

Overall, most of the studies are based on modifying the loss function of pre-trained rankers to debias and rank passages; however, this work simply debias biographies by tagging and uses the confidence score associated to job-related biographies for ranking. Interestingly, this job-biography ranking application can be considered as a new task to evaluate the fairness of pre-trained language models.

3 Methodology

3.1 Dataset

The **BiasinBios dataset** is a well-known dataset to study bias, that contains biographies related to people's professional life and jobs. Overall, this dataset contains 393,423 biographies in 28 job categories. There are two versions of this dataset; the first version, called the "WithGender" dataset, is the main version and considers the female and male genders. It includes all explicit gender words (she, he, her, his, John, etc.). The scrubbed version, the "WithoutGender" dataset, eliminates all gender-explicit words in the biographies. These datasets are used to find stereotypical bias related to genders with a categorization task. An unbiased model must categorize biographies according to their job-titles without considering the gender of the person (De-Arteaga et al., 2019). The descriptive statistics of this dataset is shown in Table 1.

As pre-trained language models learn bias from their pre-training corpus, we aim to eliminate this bias association between gender and profession. To do so, we need to know which job in this dataset is more associated with a given gender. We cat-

BiasinBios' dataset statistics			
Title	Train	Validation	Test
Number of biographies	255,710	39,369	98,344
Number of female-related biographies	117,589	18,804	45,710
Number of male-related biographies	138,121	20,565	52,634
Number of job-titles (classes)	28	28	28

Table 1: Descriptive statistics of the BiasinBios dataset

egorize job-titles utilizing data from the U.S. Bureau of Statistics [1] in 2022, where men and women are represented in different proportions for each profession. A higher proportion of men in one job determines that it is stereotypical job for men and if the women proportion is higher, then it is a stereotypical job for women (de Vassimon Manela et al., 2021). Men stereotypical jobs are anti-stereotypical jobs for women and vice versa. Taking this into account, we categorized the professions in the BiasinBios dataset into stereotype and anti-stereotype jobs. For example, in this categorization, nurse, accountant, model, and pastor are categorized as anti-stereotypes for men, while professor, rapper, poet, and software engineer are categorized as anti-stereotypes for women. The full list of this categorization is shown in table 9 in appendix A.

List of gender stereotypical terms. To identify gender-related terms, state-of-the art models rely on a list of terms (Ghanbarzadeh et al., 2023; Webster et al., 2020). In this work, we gather a list of gender-related terms from two sources (Gaucher et al., 2011; Zhao et al., 2018b). These include explicit gender terms [2](women, men, grandmother, grandfather etc.), possessive and pronouns related to genders (she, he, his, her) (Zhao et al., 2018b), implicit gender stereotype adjectives (active, cooperative, considerate, emotional etc.) (Gaucher et al., 2011), and women and men stereotypical jobs (nurse, doctor, etc.) (Zhao et al., 2018b). Table 2 shows statistics of gender-related terms in the BiasinBios dataset along with their corresponding tag.

Spacy named entity recognition tool (NER). We also used the Spacy NER API to identify proper nouns, which often contain information about gender types. The statistics of the tagged outputs using Spacy is provided in Table 2.

[1] U.S. Bureau of Statistics website

[2] Explicit gender terms and profession stereotypes

3.2 Corpus modification strategies

In this section, we define a baseline strategy, specifically the scrubbing approach proposed in (De-Arteaga et al., 2019), and subsequently introduce our TagDebias method. Table 3 shows a comparison of the corpus modification strategies employed in this work.

3.2.1 Scrubbing Strategy

The scrubbing strategy is the act of removing gender indicator terms from the original "WithGender" dataset to eliminate stereotypical associations in pre-trained language models. This strategy is introduced in the BiasinBios dataset (De-Arteaga et al., 2019) and corresponds to the "WithoutGender" version. In this dataset, all proper nouns, men and women related possessives, pronouns, and titles (Ms, Mr, Mrs, etc.) are eliminated from the dataset and replaced by "_".

3.2.2 TagDebias Methodology

TagDebias replaces different gender indicator terms with abstract, neutral types identified in table 2. Different tagging strategies are presented below.

Proper nouns, pronouns and possessives. In this tagging approach, Spacy is used to identify proper nouns mentioned in the text, such as "John" or "Emily". Proper nouns are tagged as "Person". We also tag other gender indicator words from the stereotypical terms list. This list includes pronouns ("she" and "he") and possessives ("his" and "her"), which are replaced respectively by the tag "Person", and the possessive "Their". We aim to examine if tagging these explicit gender indicators would result in fairer models.

Gender-specific terms. Here, our objective is to investigate whether assigning the tag "Person" to all "Gender-specific terms" from the gender stereotypical terms list contributes to enhanced fairness scores. On top of proper nouns, pronouns, and

Gender Indicators	Example	Tag	Number of tags
Proper noun (Spacy NER)	John, Emily	Person	467,046
Pronouns	She, He	Person	441,895
Possessives	Her, His	Their	301,701
Titles	Mr, Mrs, Ms	Mx	28,576
Gender-specific terms	Mama, Papa, Lady, Gentleman	Person	103,439
Gender stereotype adjective	Muscular, Gentle, Active, Kind	Adjective	23,555
Stereotypical jobs	Nurse, Doctor	Job-title	47,637

Table 2: Descriptive statistics for gender indicator terms and their tags in the training dataset.

WithGender Example
He is rated 5.0 stars out of 5 by **his** patients. **She** received **her** M.Sc. and Ph.D. in geological sciences from Brown University.
WithoutGender (Scrubbed) Example
_ is rated 5.0 stars out of 5 by _ patients. _ received _ M.Sc. and Ph.D. in geological sciences from Brown University.
TagDebias Example
PERSON is rated 5.0 stars out of 5 by THEIR patients. PERSON received THEIR M.Sc. and Ph.D. in geological sciences from Brown University.

Table 3: Examples of WithGender, Scrubbed, and TagDebias biographies on the BiasinBios dataset

possessives which are tagged in the previous strategy, all gender-specific terms from the stereotypical terms list such as "girl," "boy," and so on, are subjected to tagging.

Gender stereotype adjectives. Having observed the positive impact of tagging gender-specific terms on the fairness of pre-trained language models, our next step is to explore stereotypical adjectives associated with genders. These adjectives such as "sensitive", "competitive", "hostile" are often associated with specific gender-related behaviors. In this experiment, our aim is to investigate whether replacing these adjectives with the tag "Adjective" results in a fairer model.

Stereotypical jobs. Building upon the "Gender-specific terms tagging" approach, we decided to additionally tag stereotypical professions related to genders in the biographies. The aim of this extension is to investigate whether tagging stereotypical professions with the tag Job-title could result in a fairer model.

Scrubbed-tag. To compare the tagging approach with the scrubbing approach, we tagged all the terms that were scrubbed from the original dataset as described in (De-Arteaga et al., 2019) using proper nouns, pronouns, possessives, and titles (e.g., Mr., Mrs., Ms.). The objective was to compare the impact of scrubbing terms versus tagging them for bias mitigation.

3.3 Models and Experimental Setup

Our primary task is multi-class classification, involving 28 job categories, with biographies as inputs. We applied our various tagging strategies on the BiasinBios dataset. To identify our baseline classification performance, we chose to employ sequence classification models from the BERT family. We fine-tuned three pre-trained language models: BERT, ALBERT, and RoBERTa from Hugging face libraries with the base models "bert-base-uncased", "albert-base-v2" and "roberta-base". We started with the "WithGender" and "WithoutGender" baselines to determine the best performing model for further experiments. Our training setup consisted of training the models for 3 epochs, using a learning rate of 2e-6, and employing a batch size of 16.

Once we identified the best performing model based on evaluation metrics like F1 score, accuracy, precision, and recall, our next step involved fine-tuning the selected model on the modified datasets. Subsequently, we assessed fairness using metrics to be discussed in the following sections.

3.4 TagDebias Evaluation on the Fair Ranking Task

After identifying the most equitable TagDebias model, our goal is to evaluate this model in a fair ranking task, specifically the ranking of biographies for a given profession compared to a base model. In our methodology, we consider each biography in the BiasinBios dataset as a CV, with each

job category class serving as a "job description". To create rankings based on the similarity between each biography and job category, a scoring mechanism is essential. We chose the confidence score as our ranking metric for ordering all biographies within each job category.

To ensure unbiased ranking, we took steps to mitigate the impact of gender distribution. We randomly selected an equal number of biographies associated with men and women from each job category in the initial test set. Consequently, each job category comprises 50% men's and 50% women's biographies. A fair model ensures the same distribution of input candidates in the top ranking which is 50% (Zehlike et al., 2022; Yang and Stoyanovich, 2017). Subsequently, we employed three distinct models for ranking purposes: **Base Model**: fine-tuning a base PLM on the classification of the With-Gender dataset; **Scrubbed Model:** Fine-tuning a base PLM on the scrubbed version of the dataset; and **TagDebias Model:** We utilized our most effective tagged model, which is the gender-specific terms model (see section 4.1).

Once the biographies were ranked, we selected the top-k biographies from each job category and assessed the gender distribution in this subset. When analyzing gender distribution, we took into account stereotypes associated with jobs based on the distribution of the training dataset. For instance, if we observed a lower percentage of men in the top 20 rankings generated by a debiased model for a job typically associated to men, compared to the distribution generated by the initial model, that would indicate that our debiased model exhibits reduced bias, while the majority of biographies related to that job in the training dataset are men's biographies.

Finally, we performed a sensitivity analysis experiment, examining various top-k values to ascertain whether our proposed TagDebias model consistently delivers fairer outcomes across all these top-k values.

3.5 Metrics and Test Datasets

This section is segmented into two parts: the initial part concentrates on fairness metrics, while the latter delves into metrics linked with ranking.

3.5.1 Fairness Evaluation Metrics

Besides standard classification metrics to assess the ability of models to classify biographies into their related job title, we employed a few fairness evaluation metrics.

False Positive/Negative Equality Difference Scores. This method, first, calculates FPR (False positive rate) and FNR (False negative rate) for the test set of BiasinBios. The test set is categorized into two subgroups men and women. A subgroup is designated by t. FPR_t/ FNR_t determines the False positive/negative rate of the subgroup biographies. The FPED and FNED evaluate how balanced or equitable the model's predictions are in terms of false positives and false negatives across different groups. A lower score indicates a fairer classifier (Dixon et al., 2018). The formulas are described below:

$$FPED = \sum_{t \in T} |FPR - FPR_t| \quad (1)$$

$$FNED = \sum_{t \in T} |FNR - FNR_t| \quad (2)$$

3.5.2 CrowSpairs

We also employed test sets and metrics that are specifically used for evaluating social bias in Pretrained Language Models (PLMs) (Nangia et al., 2020) such as the CrowdSource Stereotypical pairs (CrowS-Pairs) benchmark dataset and metric. This dataset consists of sentence pairs categorized as leaning towards less or more stereotypes. CrowSpairs introduces three key metrics. The Stereotype Score (SS) gauges the percentage of instances where language models show a preference for more stereotypical examples over less stereotypical ones. The Anti-Stereotype Score (Anti-SS) quantifies the percentage of instances where language models favor more anti-stereotypical examples over less anti-stereotypical ones. The CrowS Pairs Score (CPS) assesses the percentage of all examples, encompassing both stereotypical and anti-stereotypical pairs, where language models prefer the higher stereotypical or anti-stereotypical option over the less stereotypical or anti-stereotypical counterpart. For each of these metrics, an ideal score is 50%.

3.5.3 SEAT

The Sentence Encoder Association Test (SEAT) serves as another benchmark dataset and metric (May et al., 2019). SEAT is an extension of the Static Word Embedding Association Test (WEAT) (Caliskan et al., 2017), designed specifically for contextualized word embedding settings. In SEAT, the evaluation involves leveraging simple sentences

Model	P%	R%	F1%	Acc%
Initial Model				
BERT	**81.49**	76.68	78.29	84.72
ALBERT	79.18	73.99	75.50	83.00
RoBERTa	81.27	**77.95**	**79.16**	**84.97**
Scrubbed Model				
BERT	**81.35**	75.94	77.63	84.33
ALBERT	78.59	73.74	75.15	83.00
RoBERTa	81.22	**77.42**	**78.79**	**84.65**

Table 4: Classification results of BERT-family models fine-tuned on two datasets' versions (WithGender and WithoutGender (scrubbed version)) - the classification results are computed on the WithGender version test set.

with placeholders, such as "This is a/an [word]." The placeholder is then replaced with demographic groups (e.g., man, woman) and stereotypical words (such as attributes, careers, etc.). A fair model is expected to not exhibit significant differences between demographic groups and their respective similarities with stereotypical words. The evaluation metric is based on the effect size, often reported as a bias score in PLMs. The Absolute Average effect size, closer to zero, indicates a fairer model.

3.5.4 Fairness metrics in ranking

In this study, we explore ranking scenarios that incorporate sets of examples with binary sensitive attributes, designated as protected (G1) and favored (G2) attributes, respectively. Protected attributes are characteristics of individuals that require safeguards to prevent discrimination and bias, often legally protected, while favored attributes are characteristics that should not confer unjust advantages and should be treated neutrally in various contexts to ensure fairness. In our context, protected attributes refer to anti-stereotypical jobs and favored refer to stereotypical jobs. For instance, G1 might represent females in software engineering roles, while G2 could denote females in model positions.

In this study, we will utilize the position-based ranking that is proposed by (Yang and Stoyanovich, 2017). The fundamental premise behind this metric lies in the significance attributed to higher ranks in candidate assessments. Essentially, these metrics aim to ensure that the gender distribution among top-ranked candidates closely mirrors the gender distribution observed in the input data. To assess and quantify bias within these rankings, we employ a position-based ranking metric which is described below.

Normalized discounted difference (rND). The Normalized Discounted Difference measures the disparity between the proportion of protected attributes (G1) in the top-k rankings (k starts at 5 in our ranking) and the total input data in the ranking. In this study, the protected attribute includes women in men stereotypical jobs and men in women stereotypical jobs. Z serves as a normalizer and denotes the highest possible value of rND, calculated using the specified total input data in the ranking (n) and the size of |G1| in the input data. After normalization, the score ranges between 0 and 1, where zero indicates no bias, and 1 represents a fully biased ranking.

$$rND = \frac{1}{Z} \sum_{k=5}^{n} \frac{1}{\log_2(k)} \left| \frac{|G_1^{1,\ldots,k}|}{k} - \frac{|G_1|}{n} \right| \quad (3)$$

4 Results

4.1 Model performance and fairness results

Table 4 presents the classification report for the three language models on the initial and scrubbed datasets. Based on these results, the RoBERTa model fine-tuned on both the initial and scrubbed versions outperforms other PLMs in terms of F1 score, accuracy, and recall. Therefore, we will conduct the remaining experiments using the RoBERTa model.

As we can see in Table 5, the scrubbed model and all the tagging models have a lower FNED score compared to the initial model. Among these models, the "Gender-specific-term" model has the lowest FNED score and the scrubbed version has the second lowest FNED score compared to other models (lower FPED/FNED means less bias). The scrubbed-tag model exhibits a lower FPED score in contrast to the scrubbed model. Moreover, it demonstrates a reduced FPR and an increased TPR when compared to the scrubbed model. This outcome underscores the potential of tagging identical

Model	↓FNED%	↓FPED%	↑TPR$_{antistereo}$%	↓FPR$_{stereo}$%	F1%	Acc%
Initial model	4.81	0.107	72.14	0.729	79.16	84.97
Scrubbed	3.64	0.109	73.21	0.723	78.79	84.65
Possessive and pronouns	3.97	0.105	73.29	0.715	**78.86**	**84.82**
Gender-specific terms	**3.56**	0.107	**73.89**	0.715	78.71	84.75
Gender-stereo-ADJ	4.04	0.112	73.46	0.722	**78.86**	84.79
Stereotypical jobs	4.46	**0.007**	71.71	0.834	76.54	83.07
Scrubbed-tag	3.69	0.100	73.43	**0.709**	78.75	84.79

Table 5: Fairness and performance of RoBERTa models on the BiasinBios WithGender version test set.

Model	SEAT6	SEAT6b	SEAT7	SEAT7b	SEAT8	SEAT8b	Avg. esize
RoBERTa-base	0.92	0.20	0.98	1.46	0.81	1.26	0.938
RoBERTa-InitialModel	1.55	0.69	1.40	1.42	1.13	0.77	1.160
RoBERTa-Scrubbed	1.23	0.19	-0.54	0.68	-0.09	1.13	0.643
RoBERTa-TagDebias-BiasinBios	1.42	0.04	-0.54	-0.37	-0.03	0.10	0.417
RoBERTa-TagDebias-CoLA	0.93	-0.41	-0.06	0.24	-0.15	0.20	0.338
RoBERTa-TagDebias-CoLA-NoSpacy	0.83	-0.23	-0.03	0.21	-0.19	0.25	0.294
RoBERTa-CDA	0.98	0.01	0.08	1.29	0.99	1.16	0.752
RoBERTa-gendertuning-CoLA	0.05	0.00	0.15	0.07	0.70	0.03	**0.166**

Table 6: Model evaluation on the SEAT benchmark dataset. Scores closer to zero are better. The Initial Model is fine-tuned on the WithGender dataset of BiasinBios, while the base model is not fine-tuned. Light grey represents the models fine-tuned on BiasinBios, while dark grey represents our models fine-tuned on CoLA.

terms, which are scrubbed in the scrubbed version, to foster a fairer model. In terms of FPED score, the "Stereotypical jobs" model obtains the lowest FPED score. However, among tagged models, the gender-stereo-ADJ model increased the FPED score compared to the initial model, which indicates a higher bias compared to the initial model. Tagged (except Stereotypical jobs) and scrubbed models improved the true positive rate of anti-stereotypical jobs compared to the initial model, which means that these models could better categorize anti-stereotypical jobs than the initial model. The decrease in the FPR metric shows that the scrubbed and tagged models (except Stereotypical jobs) do not incorrectly assign anti-stereotypical biographies, such as "she works as a doctor," to stereotypical jobs (e.g., nurse). Among all tagged models, tagging stereotypical jobs appears to be the less interesting approach and could negatively affect the fairness of the models.

It is crucial to evaluate model performance even when employing bias mitigation strategies. As it is shown in Table 5, compared to the initial model, both the accuracy and F1 scores of tagged models remain very close to the initial model. This implies that the "TagDebias" approach not only enhances the model's fairness but does so without detriment to its overall performance. Overall, based on these metrics, the "Gender-specific term" model has the lowest bias scores compared to other models. Henceforth, we refer to this model as "TagDebias" for all subsequent experiments.

SEAT benchmark. Table 6 shows the result of different data bias mitigation methods on the RoBERTa model, based on the SEAT benchmark dataset and evaluation metric (Average Absolute value-esize)(lower is fairer). To test the models' behavior on different datasets, some of our models are fine-tuned on BiasinBios (-BiasinBios) while others are fine-tuned on CoLA (-CoLA) (Warstadt et al., 2019). Some of our experiments remove the tagging of proper nouns (NoSpacy). The results show that our proposed TagDebias model outperforms the base model, as well as the scrubbed and counterfactual data augmentation (CDA) methods, but it does not outperform

Model	CPS	SS	Anti-SS
RoBERTa-base	54.96	59.12	**48.54**
RoBERTa-InitialModel	45.42	41.51	**51.46**
RoBERTa-Scrubbed	44.46	**42.77**	47.57
RoBERTa-TagDebias	**47.33**	37.11	63.11
ALBERT-base	54.20	47.17	65.05
ALBERT-InitialModel	54.58	**48.43**	64.08
ALBERT-Scrubbed	**50.38**	47.17	**55.34**
ALBERT-TagDebias	48.85	40.25	62.14
BERT-base	58.02	55.35	62.14
BERT-InitialModel	57.25	54.72	61.17
BERT-Scrubbed	54.58	54.09	55.34
BERT-TagDebias	**53.05**	**52.20**	**54.37**

Table 7: Model evaluation on the CrowSpairs dataset (Gender bias). The ideal score is 50%, scores nearer to 50% indicate a less biased model. The numbers are in percentage.

the gender-tuning approach (Ghanbarzadeh et al., 2023). When our TagDebias model is fine-tuned on the CoLA dataset, it enhances the SEAT score compared to the TagDebias-BiasinBios model. We also observed that the TagDebias model fine-tuned on CoLA instead of BiasinBios and without employing Spacy (thus, not tagging proper nouns), emerged as our best-performing model among the other TagDebias models. RoBERTa-CDA results are taken from (Meade et al., 2022), and RoBERTa-gendertuning-CoLA results are taken from (Ghanbarzadeh et al., 2023).

CrowSpairs. Table 7 displays the CrowSpairs scores of TagDebias-BiasinBios using BERT-family models. Notably, the TagDebias model attains the lowest CrowSpairs (CPS) score among the models examined in the BERT and RoBERTa models, with a score closer to 50% indicating an ideal unbiased model. TagDebias on the BERT model achieved the lowest CPS, SS, and Anti-SS scores compared to other BERT models.

The results on the RoBERTa and ALBERT models indicate that, in contrast to other models, TagDebias tends to attribute a higher probability to anti-stereotypical sentences rather than stereotypical ones. Consequently, this leads to a higher anti-stereotype score and a lower stereotypical score when compared to other models.

4.2 Fairness in Biographies' ranking

To identify the impact of TagDebias on the ranking task, we first compute the mean of each gender distribution in their corresponding stereotypical jobs in different top-k rankings as illustrated in Table 8. By considering all top-k values, we see that in the majority of cases (except top 25), the scrubbed version reduced the representation of men and women in their stereotypical roles compared to the initial model, bringing it closer to 50%. Most importantly, our proposed TagDebias version consistently enhanced this balance across all top-k values, showcasing a superior performance compared to the scrubbed version (see appendix A.5).

For instance, in the initial model, the mean of men and women distribution in all stereotypical jobs in the top 20 is 60.71%. This percentage decreases to 58.75% for the scrubbed model and further drops to 51.25% for our proposed TagDebias model. This suggests that both the Scrubbed and TagDebias versions display reduced proportions of men and women in the top 20 candidates compared to the initial model in their respective stereotypical jobs. These percentages approach 50%, reflecting a fairer ranking, as illustrated in the corresponding figure in the appendix (see appendix A.5, figure 8)

Sensitivity analysis for different top-K ranking. Following the ranking of various top-k values, our objective is to employ a position-based fairness metric to measure the bias in each of the top-k rankings generated by each model.

As we can see in Figure 1, the TagDebias model outperformed the initial model in the total rND score compared to the scrubbed version. This result is consistent among all top-k rankings, where the tagged model consistently demonstrates the lowest bias score compared to both the initial and scrubbed models. Another noteworthy observation is the tagged model's stable fairness performance when compared to the initial and scrubbed models.

5 Conclusion

We introduced a novel data-based approach for mitigating social bias in Pre-trained Language Models (PLMs) known as TagDebias. Our experiments demonstrate the successful mitigation of gender bias in several models. A comparative analysis

Model	Top5	Top10	Top15	Top20	Top25	Top30	AVG
Mean of men and women proportions in top k rankings %							
Initial model	63.57	65.71	62.50	60.71	59.57	59.54	61.93
Scrubbed-version	60.36	57.86	59.43	58.75	61.14	58.79	59.39
TagDebias-model	**50.00**	**52.50**	**50.50**	**51.25**	**52.29**	**51.86**	**51.40**

Table 8: Fairness results of different rankings generated by RoBERTa models based on the mean of women and men distribution in their corresponding stereotypical jobs (closer to 50% is fairer).

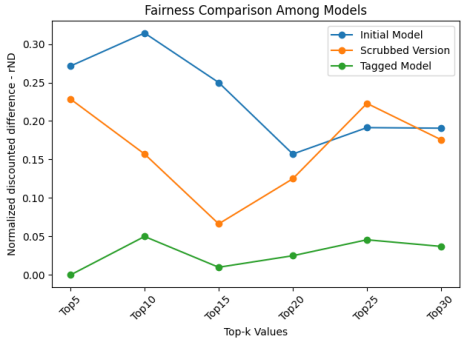


Figure 1: rND score for different Top k rankings

indicates that TagDebias surpasses the fairness performance of two other data-based bias mitigation methods—counterfactual data augmentation and scrubbing—while preserving performance in downstream tasks. Furthermore, we proposed a ranking-based gender bias evaluation, revealing a significant enhancement in fair ranking by the TagDebias model compared to the initial and scrubbed models.

6 Limitations

While our proposed TagDebias method effectively addressed some of the gender bias in several models, it does have some limitations. One aspect is that the choice of dataset and model impacts the results. We conducted experiments on the CoLA dataset and we found significant improvement in SEAT score compared to tagging on BiasinBios. Consequently, the dataset used for tagging and fine-tuning could impact the amount of bias mitigation in PLMs. Similarly, some models appear to benefit more from our tagging strategy. Future work should identify the reason for these differences.

Additionally, the tagging of gender indicator terms relies primarily on a limited list derived from the literature review. Expanding this list to encompass a broader range of stereotypical terms would be beneficial.

More importantly, the tagging strategy demonstrated in this study could be extended to address various forms of social bias, including race and religion. Finally, it does not seem sufficient to debias the datasets with tags, since the gender-tuning approach (Ghanbarzadeh et al., 2023) outperforms TagDebias in terms of SEAT scores. Our future experiments should consider whether a MLM objective based on our tagging strategy and/or a fine-tuning with the MLM generated examples contribute to further reducing bias. Ultimately, we aim to conduct experiments using large pre-trained language models to evaluate the effectiveness of TagDebias in mitigating gender bias within these models.

Acknowledgements

This research was enabled in part by support provided by Calcul Québec (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca). This work was supported by Mitacs through the Mitacs Accelerate program. We would also like to acknowledge that we employed ChatGPT to improve the paper's writing.

References

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

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*, pages 120–128.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242.

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*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. *arXiv preprint arXiv:2307.10522*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

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.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Raffaele Perego, F Sebastiani, JA Aslam, I Ruthven, and J Zobel. 2016. Proceedings of the 39th international acm sigir conference on research and development in information retrieval, sigir 2016, pisa, italy, july 17-21, 2016. ACM.

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

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Bhaskar Mitra, Morteza Zihayat, and Ebrahim Bagheri. 2022. Bias-aware fair neural ranking for addressing stereotypical gender biases. In *EDBT*, pages 2–435.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*, pages 1–6.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in ranking, part i: Score-based ranking. *ACM Computing Surveys*, 55(6):1–36.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Stereotypical jobs categorization

Here is the list of men and women stereotypical jobs based on the U.S. Bureau Statistics [3].

Women Stereotypical Job	Men Stereotypical Job
Accountant	Architect
Dietitian	Attorney
Interior-designer	Chiropractor
Model	Comedian
Nurse	Composer
Paralegal	Dentist
Pastor	DJ
Psychologist	Filmmaker
Teacher	Journalist
Yoga-teacher	Painter
	Personal-trainer
	Photographer
	Physician
	Poet
	Professor
	Rapper
	Software-engineer
	Surgeon

Table 9: BiasinBios job categorization into stereotypical categories for both genders based on the U.S. Bureau of Statistics. Note that anti-stereotypical jobs for men are stereotypical jobs for women and vice versa.

A.2 SEAT score

This section describes the SEAT score. Let A and B be sets of attribute terms such as she, he, man and woman. X and Y are sets of target terms such as (family, profession, career). Based on the WEAT score which is described in the (Caliskan et al., 2017):

$$S(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \tag{4}$$

$$d = \frac{\mu([s(x, A, B)]_{x \in X} - \mu([s(y, A, B)]_{y \in Y}}{\sigma([s(t, X, Y)]_{t \in A \cup B}} \tag{5}$$

S(w, A, B) is described as the difference between the mean of w's cosine similarity with the words from attribute terms A and w's mean cosine similarity with the attributes terms B. The effect size is calculated by the below equation. We denote μ and σ as means and standard deviation, respectively. The score closer to zero is the best.

A.3 CoLA dataset

Corpus of Linguistic Acceptability is a set of 10,657 sentences that are labeled as grammatical and ungrammatical, from published linguistics literature (Warstadt et al., 2019).

[3] U.S. Bureau of Statistics website

A.4 PLMs architectures and experimental setup

In this study, we have used BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020) pre-trained models from Hugging face libraries. BERT-base with 110M parameters, RoBERTa-base with 125M and ALBERT-base-v2 with 11M parameters. For experiments, we have used 3 epochs, a learning rate of 2e-6 and a batch size of 16. All experiments have been done with NVIDIA A100 GPUs. The experiments are computed with an average of three runs.

A.5 Different rankings generated by the models

In this section, we present various top-K rankings (5, 10, 15, 20, 25, 30) generated by the initial, scrubbed, and TagDebias models. The bars illustrate the proportion of women and men in stereotypical jobs associated with each gender. Across all rankings, TagDebias consistently demonstrates a fairer ranking, with a mean approaching 50%.

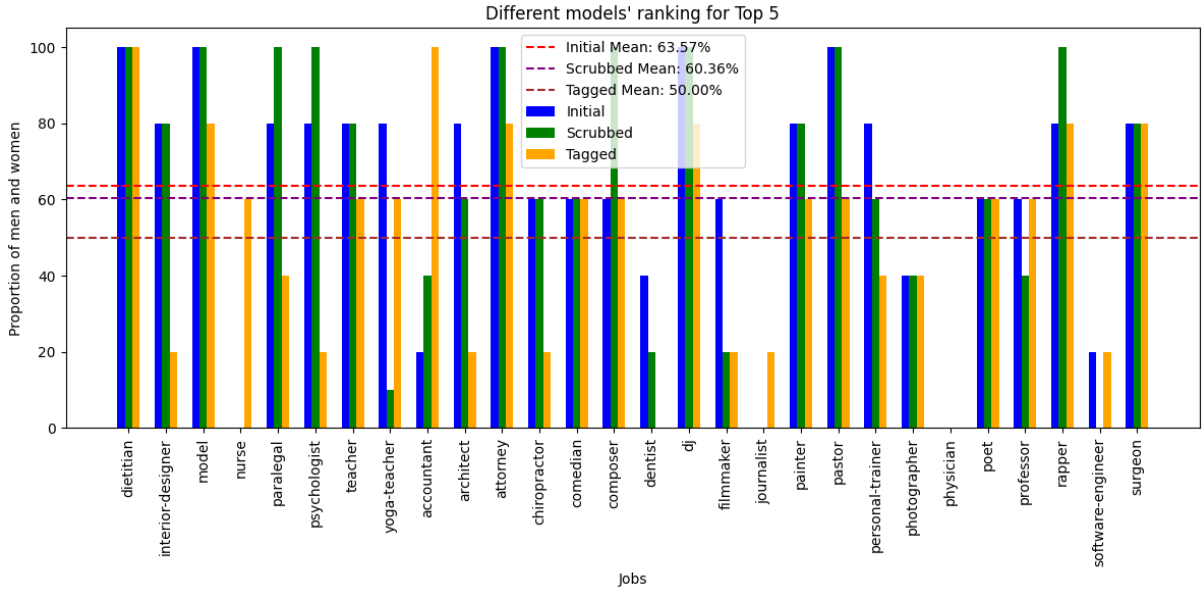


Figure 2: Top 5 biographies ranking

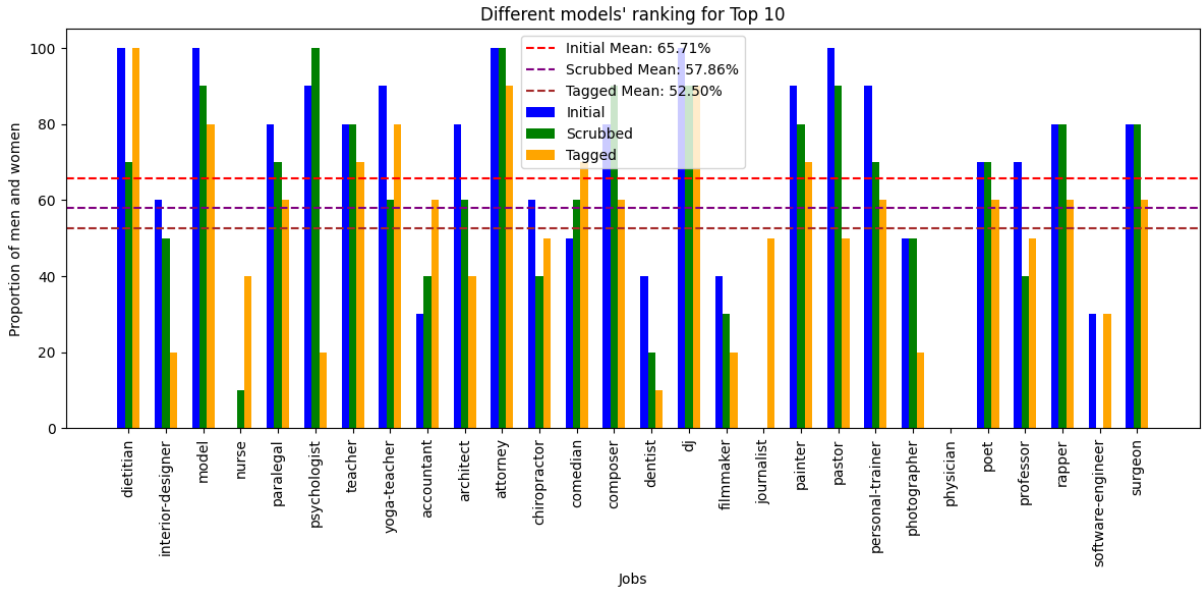


Figure 3: Top 10 biographies ranking

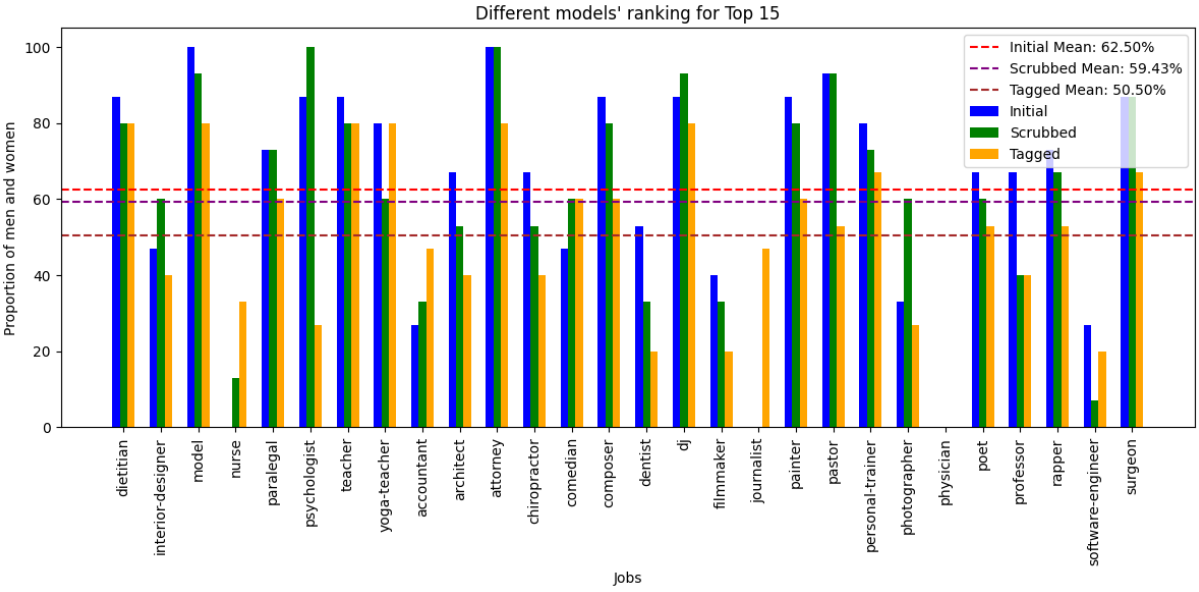


Figure 4: Top 15 biographies ranking

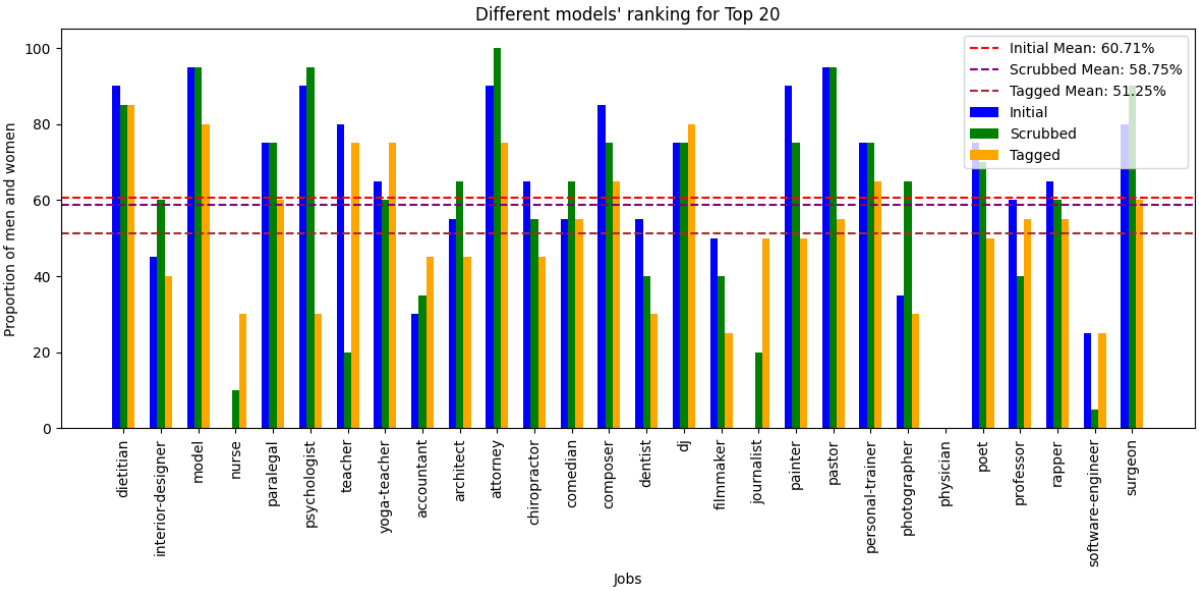


Figure 5: Top 20 biographies ranking

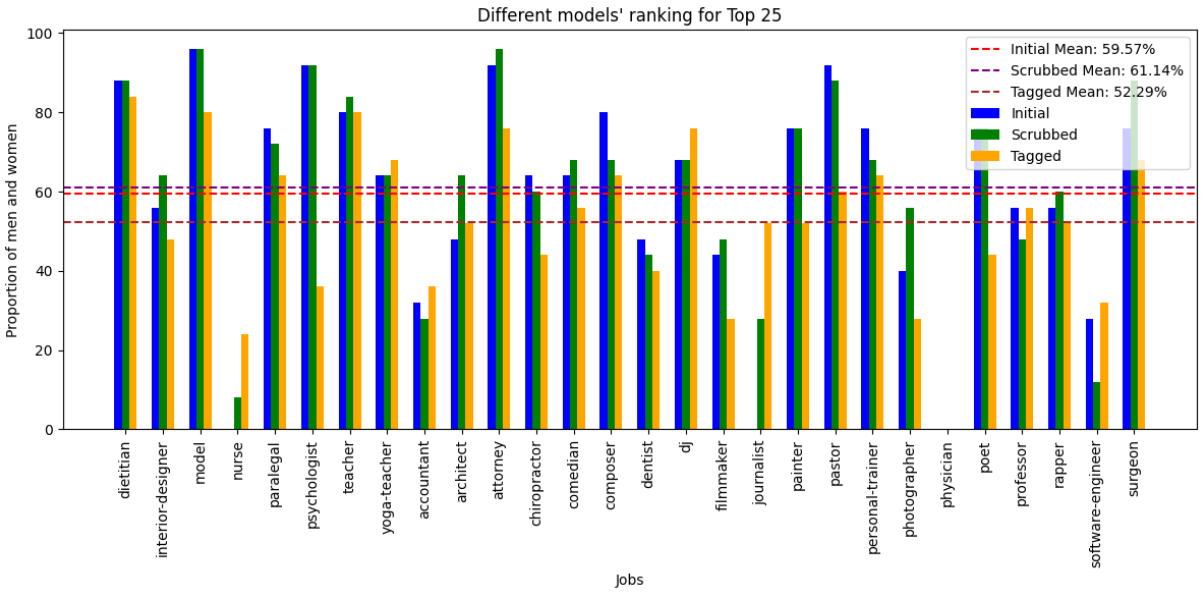


Figure 6: Top 25 biographies ranking

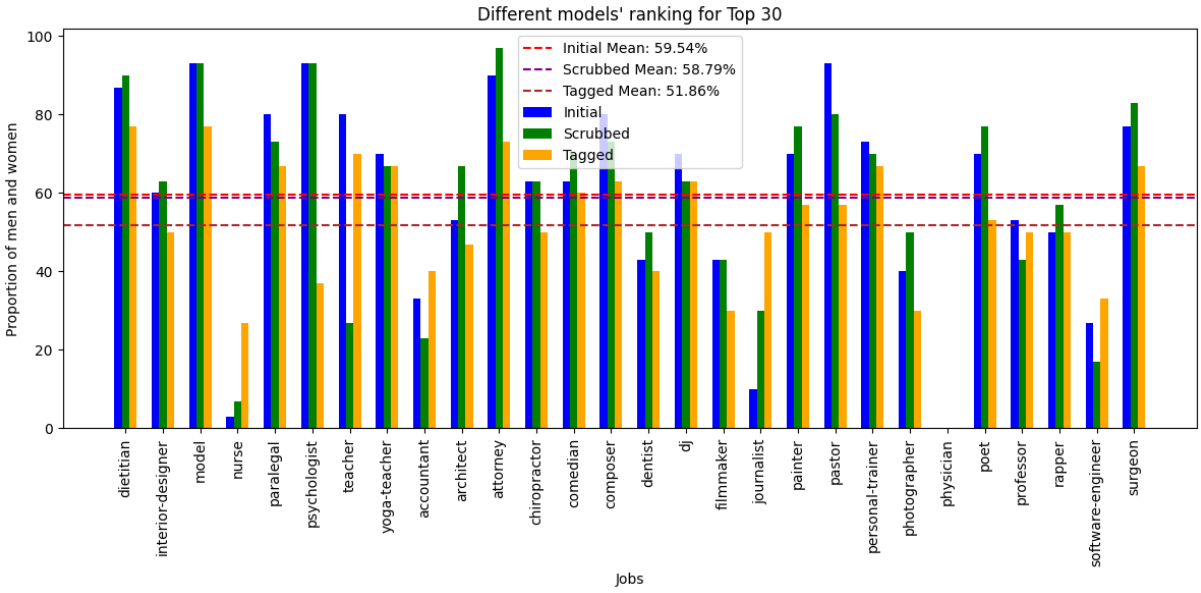
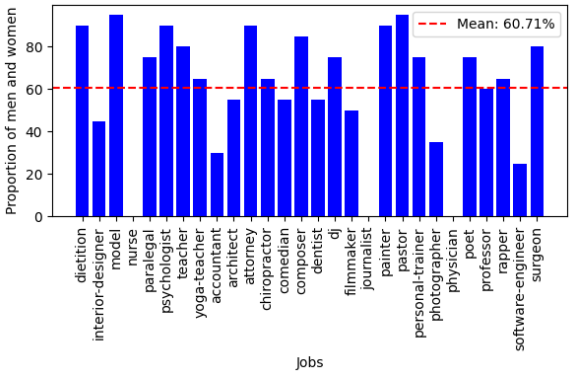
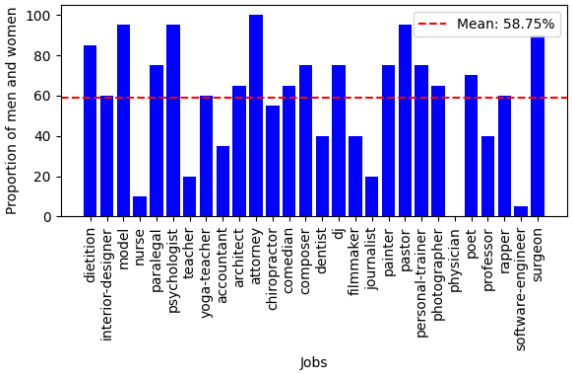


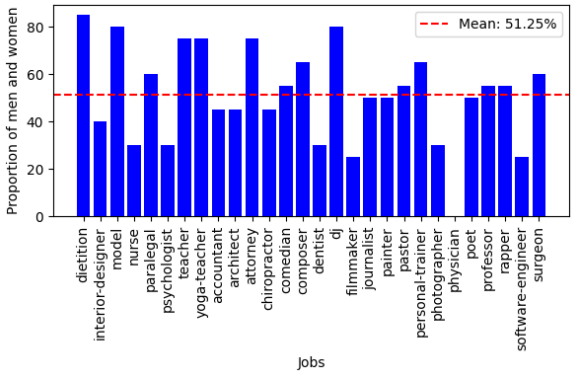
Figure 7: Top 30 biographies ranking



(a) Initial model



(b) Scrubbed model



(c) TagDebias model

Figure 8: Top 20 rankings generated by different RoBERTa models