



Titre: TagDebias: Entity and Concept Typing for Social Bias Mitigation in
Title: Pretrained Language Models

Auteur: Mehrnaz Moslemi
Author:

Date: 2024

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Moslemi, M. (2024). TagDebias: Entity and Concept Typing for Social Bias
Citation: Mitigation in Pretrained Language Models [Mémoire de maîtrise, Polytechnique
Montréal]. PolyPublie. <https://publications.polymtl.ca/57995/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/57995/>
PolyPublie URL:

**Directeurs de
recherche:** Amal Zouaq, & Jinghui Cheng
Advisors:

Programme: Génie informatique
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**TagDebias: Entity and Concept typing for Social bias Mitigation in Pretrained
Language Models**

MEHRNAZ MOSLEMI

Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Mars 2024

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

**TagDebias: Entity and Concept typing for Social bias Mitigation in Pretrained
Language Models**

présenté par **Mehrnaz MOSLEMI**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Farida CHERIET, présidente

Amal ZOUAQ, membre et directrice de recherche

Jinghui CHENG, membre et codirecteur de recherche

Omar ABDUL WAHAB, membre

DEDICATION

*To my family and love,
To my supervisor and all teachers,
To anybody who helped me along the way . . . ,
A BIG thank you!*

ACKNOWLEDGEMENTS

I extend my sincere appreciation to Professor Amal Zouaq, my supportive thesis advisor, and Professor Jinghui Cheng, my co-supervisor, for their unwavering support and invaluable guidance throughout the entire research journey. Special gratitude goes to my colleagues at the LAMA-WEST lab, whose assistance and friendship have been instrumental during my master's degree pursuit. I express gratitude to Mitacs for providing funding for this research, and I extend my thanks to Airudi Inc. for their valuable contributions as the collaborating company.

RÉSUMÉ

Les modèles de langage préentraînés sont entraînés sur de grands corpus provenant d’Internet et acquièrent des biais de genre à partir des données. De ce fait, ils peuvent potentiellement propager ce biais dans les tâches ultérieures. Par conséquent, il est impératif d’identifier et d’atténuer le biais social d’abord dans le jeu de données, puis dans les modèles de langage préentraînés.

Pour remédier au biais de genre dans les modèles de langue, nous proposons une approche de balisage appelée TagDebias qui consiste à utiliser des abstractions de niveau supérieur pour remplacer les termes spécifiques au genre. Cette méthode est en contraste avec l’approche dite du nettoyage, qui supprime les termes indicateurs de genre du corpus. Notre objectif est d’affiner les modèles de langage préentraînés sur le corpus balisé afin de débiaiser leurs poids. Certaines de nos questions de recherche (**RQ1** et **RQ2**) examinent l’efficacité du balisage pour atténuer le biais et son impact sur les performances du modèle dans les tâches ultérieures. Les résultats indiquent que la stratégie de balisage maintient non seulement les performances du modèle dans les tâches ultérieures, mais améliore également l’équité par rapport aux modèles nettoyés et initiaux. Notamment, la stratégie de balisage des "terme spécifique au genre", appelée modèle TagDebias, s’est révélée la plus efficace pour favoriser l’équité parmi les différents modèles balisés, nettoyés et initiaux.

Nous avons également proposé une nouvelle méthode d’augmentation de données. En augmentant à la fois les versions balisées et nettoyées avec des exemples par insertion de virgules et en répétant les instances (avec et sans balises) avec le même label, nous avons cherché à surmonter les limitations des méthodes traditionnelles d’augmentation de données avec permutation de genre. Notre question de recherche (**RQ3**) porte sur l’effet de l’augmentation de données basée sur le balisage sur la production de modèles de langage plus équitables. Après la phase d’augmentation et d’affinage, nous avons constaté que l’augmentation des données dans le modèle nettoyé, en particulier en doublant les instances, améliorerait légèrement l’équité du modèle, mais pas notre modèle TagDebias.

Enfin, nous avons cherché à vérifier l’équité des modèles obtenus par le biais d’un évaluateur de biais social se concentrant sur le classement des biographies de candidats pour un poste donné. La question de recherche (**RQ4**) visait à comprendre l’impact de l’affinage des modèles de langage préentraînés à l’aide de corpus balisés et nettoyés sur le classement. Les observations ont révélé que le modèle TagDebias réduisait significativement le biais dans tous les classements des meilleurs candidats par rapport aux modèles initiaux et nettoyés.

L'approche TagDebias a démontré sa supériorité par rapport aux stratégies d'atténuation du biais basées sur les données les plus récentes, bien qu'elle n'ait pas réussi à surpasser l'approche de réglage de [1] récemment introduite.

Dans le future, nous prévoyons d'explorer le potentiel de combiner le balisage avec un objectif de modèle de langage masqué pour obtenir des modèles encore plus équitables.

ABSTRACT

Pretrained language models are trained on large corpora from the internet and learn gender imbalances from the data. They could potentially propagate this bias in downstream tasks. Therefore, it is imperative to identify and mitigate social bias first in the dataset and then in pretrained language models.

To address gender bias in language models, we propose a tagging approach called TagDebias that involves using higher-level abstractions to replace gender-specific terms. This method is contrasted with the scrubbing approach, which removes gender indicator terms from the corpus. Our aim is to fine-tune pretrained language models on the tagged-corpus to debias their weights. Some of our research questions (**RQ1** and **RQ2**) investigate the effectiveness of tagging in mitigating bias and its impact on model performance in downstream tasks. The findings indicate that the tagging strategy not only maintains model performance in downstream tasks but also improves fairness compared to scrubbed and initial models. Notably, the "gender-specific-term" tagging strategy, referred to as TagDebias model, emerged as the most effective in promoting fairness among the various tagged, scrubbed, and the initial models.

We also proposed a novel data augmentation method. By augmenting both tagged and scrubbed versions with examples through comma insertion and repeating the instances (with and without tags) with the same label, we aimed to overcome limitations of traditional data augmentation methods with gender swapping. Our research question (**RQ3**) focus on whether tagging-based data augmentation leads to fairer PLMs. After the augmentation and fine-tuning stage, we found that the scrubbed-data augmentation, particularly when doubling instances, slightly improved model fairness, but not our TagDebias model.

Finally, we aimed to verify the fairness of the obtained models through a social bias evaluator focusing on ranking candidate biographies for a given job/position. The research question (**RQ4**) aimed to understand the impact of fine-tuning PLMs using tagged and scrubbed corpora on ranking. Observations revealed that the TagDebias model significantly reduced the bias across all top-k rankings compared to the initial and scrubbed models. The TagDebias approach demonstrated superiority over state-of-the-art data-based bias mitigation strategies, though it fell short of surpassing the recently introduced gender-tuning approach [1].

In future directions, we plan to explore the potential of combining tagging with the masked language model objective to achieve even fairer models.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
RÉSUMÉ	v
ABSTRACT	vii
TABLE OF CONTENTS	viii
LIST OF TABLES	xi
LIST OF FIGURES	xiii
LIST OF SYMBOLS AND ACRONYMS	xiv
LIST OF APPENDICES	xv
CHAPTER 1 INTRODUCTION	1
1.1 Concept and problem statement	1
1.2 Contributions	3
1.3 Thesis outline	4
CHAPTER 2 LITERATURE REVIEW	5
2.1 Observation and categorization of bias	5
2.1.1 Sensitive attribute bias	5
2.1.2 Explicit and implicit bias	8
2.1.3 Subjective bias	9
2.2 Sources of bias	9
2.2.1 Bias in training dataset	9
2.2.2 Bias in contextualized pretrained language models	10
2.2.3 Bias in PLM-based evaluation metrics	11
2.3 Identifying bias and corresponding metrics	12
2.3.1 Identifying bias in the training dataset	12
2.3.2 Identifying bias in pretrained language models	16
2.3.3 Identifying bias in machine learning classifiers	21

2.4	Bias mitigation	23
2.4.1	Mitigating bias in training datasets	23
2.4.2	Mitigating bias in pretrained language models	24
2.5	Bias and fairness in ranking	31
2.5.1	Fairness definition and metrics in ranking	31
2.5.2	Utility metrics in ranking	32
2.5.3	Fair ranking with pretrained language models	33
CHAPTER 3 METHODOLOGY		35
3.1	Datasets	35
3.1.1	BiasinBios dataset	35
3.1.2	List of explicit, implicit, and stereotypical bias words (L).	36
3.1.3	Spacy tags	37
3.1.4	CoLA dataset	38
3.2	Corpus modification strategies	38
3.2.1	Scrubbing strategy for gender bias on BiasinBios dataset	39
3.2.2	Tagging strategies for gender bias on BiasinBios dataset	39
3.3	Dataset augmentation	41
3.3.1	Data augmentation by merging the initial and modified text	41
3.3.2	Data augmentation by adding modified instances	42
3.4	Models	42
3.5	Application of debiased models on fair ranking	43
3.6	Metrics	44
3.6.1	Fairness metrics in classification models	44
3.6.2	Fairness metrics and benchmarks	46
3.6.3	Fairness metrics in ranking	47
3.6.4	Utility metrics	48
CHAPTER 4 EXPERIMENTS AND RESULTS		49
4.1	Classification results	49
4.1.1	Fine-tuning BERT family models on the two initial versions of BiasinBios dataset	49
4.1.2	Fine-tuning RoBERTa model on tagged datasets	49
4.2	Fairness results of models	51
4.2.1	Fairness results of initial and different Tagged-RoBERTa models	52
4.2.2	Fairness results of augmented models	52
4.2.3	Fairness results of BERT-family models on CrowSpairs benchmark	55

4.2.4	Comparing Tag-Debias model with state-of-the-art models on the SEAT benchmark	56
4.3	Ranking results of different models on biographies	56
4.3.1	Classification result on the sub-test set	57
4.3.2	Ranking of biographies with initial and Tagged-RoBERTa	57
4.3.3	Sensitivity analysis of different top-k values	62
4.4	Discussion	64
CHAPTER 5 CONCLUSION		66
5.1	Summary of works	66
5.2	Limitations	67
5.3	Future research	67
REFERENCES		69
APPENDICES		81

LIST OF TABLES

Table 2.1	Pretraining datasets for most popular PLMs	11
Table 3.1	Descriptive statistics of the BiasinBios dataset	35
Table 3.2	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.	36
Table 3.3	Statistics of gender-related words in the stereotypical words list (L) .	37
Table 3.4	Descriptive statistics for gender indicator terms using list L and Spacy(NER) in the training dataset	37
Table 3.5	Descriptive statistics of the CoLA dataset	38
Table 3.6	Statistics of different corpus modification strategies on the training dataset, the numbers indicate total number of tags in the tagging strategies and total number of removed terms in the scrubbing strategy in the training corpus	41
Table 4.1	Classification result 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 . .	50
Table 4.2	Classification result of RoBERTa model on initial and modified test sets	51
Table 4.3	Fairness results of different versions of RoBERTa model (WithGender, WithoutGender and tagged) on the initial test set	52
Table 4.4	Fairness results of different RoBERTa models' version on the initial test set (These fairness scores are based on the model performance) .	53
Table 4.5	Fairness results on various augmented versions of RoBERTa models (Initial, WithoutGender, and tags) on the initial test set. Aug1 refers to data augmentation with comma, and Aug2 refers to data augmentation by combining the modified version dataset and the initial version . .	54
Table 4.6	Fairness results of different augmented versions of RoBERTa models (Initial, WithoutGender, and tagged) on the initial test set	54
Table 4.7	Model evaluation on the CrowSpairs dataset (Gender bias). The ideal score is 50%, the score nearer to 50% is less biased model. The numbers are in percentage.	55

Table 4.8	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 the models fine-tuned on CoLA.	56
Table 4.9	Classification result of different RoBERTa models on initial and randomly selected test set	57
Table 4.10	Utility function values for three models (WithGender, Scrubbed and Tagged). These scores are between 0 and 1. Higher values indicate the relevance of biographies to the job category.	62
Table 4.11	Fairness result of different rankings generated by RoBERTa models based on the Mean of men and women proportions in the different top k values. The best value is closer to 50%.	63
Table 4.12	Fairness result of different rankings based on the position-based ranking metric, rND. A lower score signifies superior performance in terms of fairness.	63

LIST OF FIGURES

Figure 4.1	Men distribution in top 20 ranking of biographies with three models, initial, scrubbed (WithoutGender), and Gender-specific-term tagged version, green cells have better score compared to initial model and the orange cell has worse score compared to initial model, note that we have two contradictions in stereotypes (as highlighted in grey), in this ranking we have considered the stereotype based on the training dataset	58
Figure 4.2	Women distribution in top 20 ranking of biographies with three models, initial, scrubbed, and Gender-stereo-term tagged version, green cells have better scores compared to initial model and the orange cells have worse scores compared to initial model. Note that Accountant and Pastor are stereotypes in U.S. data, but in the training dataset the majority of men have these occupations. In this ranking, we considered the stereotype based on the training dataset.	59
Figure 4.3	Men distribution in top 20 biographies created by three models in stereotypical jobs related to men	60
Figure 4.4	Women distribution in top 20 biographies created by three models in stereotypical jobs related to women	61
Figure 4.5	The representation of rND for three models (Initial, scrubbed and tagged) considering various numbers of top candidates. The green line is related to tagged model shows significantly lower rND score in all top-k values compared to the initial and scrubbed models, which demonstrates a fairer behavior.	64
Figure A.1	Top 5 biographies ranking	81
Figure A.2	Top 10 biographies ranking	82
Figure A.3	Top 15 biographies ranking	82
Figure A.4	Top 20 biographies ranking	83
Figure A.5	Top 25 biographies ranking	83
Figure A.6	Top 30 biographies ranking	84

LIST OF SYMBOLS AND ACRONYMS

PLM	Pretrained Language Model
CCO	Constrained Co-Occurrence
LMS	Language Modeling Score
CAT	Context Association Test
SS	Stereotypical Score
MLM	Masked Language Model
CPS	CrowS-Pairs Score
NLP	Natural Language Processing
PMI	Point-wise Mutual Information
DisCo	Discover of Correlations
AUL	All Unmasked Likelihood
NLI	Natural Language Inference
SEAT	Sentence Encoder Association Test
WEAT	Word Embedding Association Test
EEC	Equity Evaluation Corpus
AUC	Area Under the Curve
FPED	False Positive Equality Difference
FNED	False Negative Equality Difference
WNC	Wiki Neutrality Corpus
IAT	Implicit Association Test
GEEP	Gender Equality Prompt
JSD	Jensen-Shanon Divergence
TPR	True Positive Rate
FPR	False Positive Rate
CrowSpairs	CrowedSource Stereotypical pairs
rND	Normalized Discounted Difference
LLMs	Large Language Models
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly optimized BERT approach
ALBERT	A Lite BERT
GPT	Generative Pretrained Transformer
CoLA	Corpus of Linguistic Acceptability
NER	Named Entity Recognition

LIST OF APPENDICES

Appendix A	81
----------------------	----

CHAPTER 1 INTRODUCTION

This section initiates an exploration into the fundamental motivation and objectives underpinning the research on social bias detection and mitigation. Initially, we elucidate the core concept of the study before delving into the identified problem, research objectives, contributions, and challenges. Finally, we provide a comprehensive overview of the thesis.

1.1 Concept and problem statement

Machine Learning and AI systems embrace many aspects of human life. The use of these systems is increasing in many fields, including movie recommendations and abusive language detection [2]. They are also used in risky decision-making scenarios such as loans [3] or hiring ([4], [2], [5]). Although AI algorithms bring multiple benefits compared to humans, such as not getting tired or unlimited memories, their decision-making could contain bias or unfairness similar to human minds [2].

Unfairness or *bias* can have multiple definitions based on the context, but regarding decision making *any prejudice or giving favor to individuals or specific groups based on their inherent or even unrelated acquired characteristics* [2] is considered as a bias. For example, the number of African-American language comments that are categorized as toxic in the abusive language detection task is much more than the number of other American language comments in this study [6], so the belief that African people use abusive language could propagate in society and affect decisions about these underrepresented groups. Unfairness could happen for different individuals and groups based on gender, sex, race, disability, etc. For example, many studies have been done on gender inequality, the most prominent area of study in fairness ([7], [8]). There are some studies that consider the associations between genders and professions ([9], [10]). For instance, people tend to assign the *professor role* to men and the *nurse* role to women based on previous beliefs in the society [11]. Thus imbalances are pervasive in numerous corpora.

Similarly, studies show that pretrained language models (PLMs) are used in many NLP tasks, such as classification [12], sentiment analysis [13], and text generation [14]. Pretrained language models undergo fine-tuning on diverse corpora characterized by gender imbalances [15]. These models internalize and perpetuate these biases from the corpus into downstream tasks [16]. Studies also show that there is unfairness in the pretraining corpora of PLMs. For example, BERT, which is trained on the Books Corpus, learned gender inequalities through

its pretraining phase [15].

Thus, pretrained language models can reproduce, in general, the bias that exists in their pre-training corpora [17]. BERT [18], RoBERTa [19], ALBERT [20], and GPT-2 [21] are used in many NLP tasks, and it is important to try to identify and mitigate the bias in their weights. One mitigation method involves modifying a training corpus by making it "neutral", that is, removing gender stereotypical words [22]. We refer to the obtained corpus as a scrubbed corpus. While removing explicit gender words like pronouns, proper-nouns and possessives improved gender bias reduction in the context of gender-profession stereotypes [9], this may lead to removing concept associations in pretrained language models and lower performance in down-stream tasks. Thus, in this research, we propose another corpus modification method based on tagging stereotypical words. Studies show that abstraction leads to performance improvement in reasoning tasks [23]. This work aims to examine the effect of tagging or using higher level of abstraction in the context of mitigating social bias and compare the performance of models which are fine-tuned on a tagged corpus with models fine-tuned on the scrubbed corpus. Therefore, in the first section, our aim is to investigate these research questions:

RQ1. Does the tagging of stereotypical words mitigate the bias in PLMs and how does it compare to the "scrubbed" version?

RQ2. Does the tagging of stereotypical words affect the performance of downstream tasks?

In the context of modifying corpora to remove bias associations, data augmentation is a common method to mitigate gender, religion and race bias in language models [24]. In this approach, in addition to the original dataset, an identical dataset is created but with a bias deliberately shifted towards a different gender, race, or religion. For instance, terms like "church" will be swapped for "mosque," or "men" for "women," etc [22]. The model is then trained using the combined dataset, which includes both the original data and the swapped data. However, there are limitations to using data augmentation with gender swapping. It can sometimes result in nonsensical sentences, like *"This man gave a baby."* To address this concern, we propose two alternative data augmentation methods to mitigate bias in pretrained language models. In these methods, all attribute terms are either removed or replaced with higher-level abstractions, and then combined with the initial version of the dataset. These methods aim to help the model understand that both women and men are equivalent concepts represented by the overarching term "Person". Consequently, this approach might help to eliminate stereotypical associations between genders and other concepts, such as professions, ultimately reducing bias in the model. Here, we aim to answer this research question:

RQ3. Does a tagging-based data augmentation result in fairer PLMs?

Finally, we study the effect of fair language models on the task of CV/biography rankings. To the best of our knowledge, there is currently no known research or work that involves the fine-tuning of pretrained language models using tagged-based or scrubbed datasets specifically designed for debiasing PLMs. In this study, we introduce a novel framework that begins by mitigating biases through various tagging strategies. Subsequently, we fine-tune language models on these processed corpora and utilize them for ranking biographies. In this context, our final research question is:

RQ4. What is the impact of fine-tuning PLMs using tagged and scrubbed corpora on ranking?

In the realm of studying social bias in pretrained language models, we confront several significant challenges. The first challenge is that the compilation of a comprehensive list of gender indicator terms for dataset tagging is not integrated in the literature, making it a challenging task to source these terms from various open repositories.

Another pivotal challenge in the domain of bias detection and mitigation is the identification of an appropriate dataset. While datasets addressing gender bias exist, there is a notable deficiency in datasets related to social biases concerning race and religion. Additionally, in exploring our bias mitigation approach within the realm of hiring applications, we encountered a notable challenge: the scarcity of datasets aligning job descriptions with resumes and their corresponding ranking ground truth, all while maintaining the privacy of company information. Furthermore, preserving candidate anonymity presented an obstacle, as the available resumes lacked information about the candidates' gender, posing difficulties in the ranking task. Thus, we have identified the BiasinBios dataset [9], which stands out as the most relevant dataset in terms of matching job descriptions with CVs.

1.2 Contributions

Following our research question and objectives, here is our contribution in this research:

1. We introduced a tagging approach for the dataset and demonstrated its effectiveness in reducing gender stereotypical bias within pretrained language models. This approach not only maintained model performance but also outperformed the "scrubbed version."
2. We introduced a data augmentation method designed to mitigate gender stereotypical bias within the PLMs. Experiments show that scrubbed-based data augmentation will result in a slight fairness improvement in PLMs.
3. Leveraging the fine-tuning of pretrained models on both the scrubbed and tagged ver-

sions of the data, we achieved a substantial 14% improvement in fairness of candidates’ ranking.

1.3 Thesis outline

The thesis is structured as follows: Chapter 2 provides an exploration of fundamental concepts and a literature review in the field of bias, encompassing its definition and categorization, sources, methods for identification and mitigation, and concluding with an examination of ranking algorithms in the current landscape. Moving on to Chapter 3, we present our methodology, encompassing dataset definition, corpus modification, data augmentation, fairness and ranking metrics, and introducing our novel ranking-based fairness evaluation task. In Chapter 4, we present the experimental results, including classifications and fairness assessments. We analyze the model’s performance in ranking tasks and compare our proposed TagDebias model with state-of-the-art debiased models. Finally, Chapter 5 provides an explanation of the obtained results, a discussion on research limitations, and outlines future directions.

CHAPTER 2 LITERATURE REVIEW

This section commences by presenting the definition of bias and its categorization as per the literature review. Additionally, we delve into elucidating the origins of bias within various structures and processes inherent in machine learning systems. Following the identification of these sources, we delve into an exploration of diverse state-of-the-art methods aimed at discerning and alleviating social bias arising from these sources. Ultimately, we delve into an examination of various techniques for ranking and studying bias within a ranking framework.

2.1 Observation and categorization of bias

Bias is defined as the act of unfairly discriminating against or favoring specific groups or individuals based on their inherent or acquired characteristics, which should not be factors when evaluating them [2]. With the increasing prevalence of AI usage, the need for a closer focus on observing, detecting, and addressing bias in the design of AI systems has grown [2]. Machine learning tools find applications in various domains, such as assessing the likelihood of a criminal offender recommitting a crime or making critical decisions in the medical field and hiring processes, among others. Considering that all of these applications have direct and significant impacts on our lives [2], it is our responsibility to ensure that we design and implement systems that are fair and unbiased. In the following segment, we will classify the biases present in the field of machine learning.

2.1.1 Sensitive attribute bias

Bias, as defined, involves showing favoritism towards specific groups or depriving others of that favor based on erroneous assessments. Such bias can often be linked to sensitive attributes, which are characteristics of individuals or groups that should not influence their eligibility assessments. For example, gender bias might involve favoring one gender over another [25], and racial bias may manifest as preferential treatment of one race over another [26]. To illustrate, consider Twitter abusive language datasets from various sources ([27], [26], [28], [29], and [30]) which exhibit significant racial bias. These datasets disproportionately penalize African-American language when compared to language associated with white individuals, resulting in racial discrimination [31].

Gender bias

Gender bias is a central focus in the field of fairness studies, characterized by the explicit or implicit favoritism of one gender over another or the utilization of a specific gender as a broad representation of diverse societal roles and stereotypes. These stereotypes are society's prevailing beliefs about certain roles or characteristics. For instance, it is a commonly held stereotype that the term "nurse" is associated with women, while "doctor" is associated with men [11]. In the subsequent section, we will delve into different forms of gender bias.

Generic pronouns. In some cases in texts, "he" refers to an undefined gendered human, so it emphasizes men over women. This phenomenon is called **generic pronouns bias**. So, a generic pronoun occurs when pronouns like, he, his, and him are used as referents to nouns with no specific gender. Among gendered generic pronouns, "his" is the most recurring. For instance, *"The client should receive **his** invoice in two weeks"*. In this sentence, the gender of **client** is not specified, so the pronoun **his** is a generic pronoun bias. In other cases, "man" or "men" could emerge as a generic pronoun bias in some sentences that point out human beings. For example, *"All **men** are born for a reason"*. These pronouns could also appear as a verb in this sentence: *"Effective teachers lead or **man** the students well"* [11].

Sexism. Sexism against women is categorized into hostile and benevolent sexism [32]. Hostile sexism considers men more powerful and with more competencies than women [33]. In the view of men with a hostile sexism mentality, women are unintelligent and manipulative [34]. For example, *"Women always get more upset than men"* or *"Women are incompetent at work"*. The other type of sexism is benevolent sexism, which describes women as a group or gender that needs the care and support of men. These notions increase women's subordinate positions [35]. Benevolent sexism explains women as a complementary part of men's chivalry, power, and intelligence with their inherent characteristics [36]. For example, *"I am not exploiting women; I love, protect and care for them."* [11].

Occupational bias. Gender stereotypes related to occupations come from social beliefs about social roles, showing the sexual division in society. So, these beliefs lead to gendered occupational bias. It happens when an occupation or a position is generalized to a specific gender. This bias is categorized as a gendered division of labor and gendered roles or duties. In the gendered division bias, professions like professors are most likely to be assigned to men, but elementary teachers are primarily given to women due to society's beliefs. Another example is: "scientists are men, and secretaries are women". The other type of bias category

is using female identity words to diminish the importance of roles like "girl" instead of specific roles like **"assistant"** in this sentence: *"I will have my girl (assistant) to get you a cup of coffee"* [11].

Exclusionary bias. This type of gender bias occurs when gendered words are used to build gender-neutral words or when men's identity terms come before women's identity terms. This type of bias is categorized into three types. The first is the explicit marking of sex, which has many common examples in daily communications. For example, "workforce" is a gender-neutral term for "manpower," and "adulthood" is a neutral of "manhood." The other type of exclusionary bias is gender-based neologisms. For example, "Man-Bread" means big bread, or Man-sip is a man-sized sip of a beer or drinks one can finish in 4 or 5 sips. It means that it is easier for men than women to drink it. The other type of exclusionary bias is word order. The gendered-word tendency is to give priority to men in comparison to women. For instance, *"Mr and Mrs"* or *"boys and girls"* [11].

Semantic bias. Semantic bias as a form of gender bias arises when words or sentences are used to belittle or degrade individuals, typically with the intent of humiliation. This can encompass everything from sexist jokes to the conscious or unconscious attachment of specific, non-human labels to women. For instance, certain metaphors compare women to inanimate objects like plants or food, as in the phrase "she is like a sweet cookie" [11].

Another aspect of semantic bias is the attribution of gendered characteristics. This bias is rooted in historical gender roles, where men historically held positions of power and leadership roles compared to women. Consequently, when women occupy roles traditionally associated with men, such as a woman manager with strict characteristics, people often perceive her as an unacceptable manager, and working with her is seen as challenging. In contrast, if the phrase "strict male manager" is used, it typically conveys the idea of a competent manager who takes responsibility.

There's also the issue of old-saying bias. For instance, sayings like "bad words make women worse" are typically not applied to men. Similarly, phrases such as "when you see an old man, take a seat and learn from his wisdom, but when you see an old woman, throw out a stone" reflect biases that persist in language and perception [11].

Racial bias

This type of bias involves showing favoritism or, at times, expressing prejudice toward a specific race, while such tendencies are not typically associated with another. Unlike gender

bias, racial bias is less extensively studied in the realm of natural language processing, though it remains a noteworthy consideration. Within the domain of toxic language detection tasks, including identifying hate speech, offensive content, and abusive language, racial bias is a prevalent area of exploration ([37], [38]).

Research indicates that texts authored by African-Americans are more likely to be labeled as toxic compared to American English, and models fine-tuned on such corpora can perpetuate these biases in subsequent tasks [39]. One illustrative instance is the perpetuation of the "angry black woman" stereotype, which represents a racial bias against black women. This stereotype portrays black women as angry, imposing, and loud [40].

Disability bias

It is crucial to recognize that individuals with disabilities constitute over one billion people, approximately 15% of the global population [41]. Unfortunately, disabilities often carry significant negative social biases. Studies revealed implicit and explicit preferences against individuals with disabilities when compared to those without disabilities [42], [41].

Other sensitive attributes

Other sensitive attributes less studied than others are: marital status, age, religion, family status, etc [2].

This research focuses on investigating occupational bias within the gender-sensitive attribute bias category. The methodology employed here is transferable to examining other forms of bias, including those related to race and religion, thereby offering a promising avenue for future research.

2.1.2 Explicit and implicit bias

Stereotypes, or social beliefs expressed in language, often lack explicit semantics. Instead, the implicit meaning of a sentence is shaped by people's judgments about others. For instance, when a company states, "Our culture is not appropriate for women," people may infer that women are deemed unsuitable for the job [43]. In this context, the implicit meaning of the sentence reflects the societal perceptions and judgments applied to women.

In this research, our focus will be on examining the explicit bias present within a corpus, rather than delving into the implicit interpretations that individuals may perceive differently.

2.1.3 Subjective bias

This bias arises when a language that should be neutral becomes influenced by the writers' feelings, opinions, or preferences, whether consciously or unconsciously. For instance, if an author states, "John exposed as an unprincipled politician," it implies that the author shares this belief. In order to maintain objectivity when publishing such statements in journals or other public media, the appropriate sentence would be, "John described as an unprincipled politician." In the latter sentence, the author's idea or opinion is not explicitly embedded in the text [44]. Overall, our research does not tackle this type of bias.

2.2 Sources of bias

AI systems and algorithms operate on the foundation of data, relying on datasets for training. The performance of algorithms is intricately tied to the nature of the training datasets. If these datasets carry biases, machine learning models absorb and subsequently exhibit those biases in the down-stream tasks. Furthermore, algorithms have the potential to magnify and perpetuate existing biases present in the datasets. It is important to note that biases can also emerge from design choices, even when the dataset itself is unbiased [2]. The sources of bias encompass the training data in simple machine learning models, the pretraining dataset in pretrained language models, and the architecture of language models. In the following sections, we delve into the origins of bias within various components of the AI system.

2.2.1 Bias in training dataset

In the introductory section, we delve into the origins of bias within the training dataset, examining how it can be integrated into machine learning algorithms and subsequently affect their fairness and performance.

Feature variable bias

Feature variable bias emerges when features and variables in the training data are selected. The choice of variables during the training phase significantly impacts the model's fairness and performance [45]. For instance, let's consider the number of arrests as a feature to predict a person's future criminal activity is an example. If the individual is from a minority group (such as black people, a specific religion, or women), there's a higher likelihood of them being categorized as a high-risk person because minorities may be disproportionately targeted by law enforcement. Consequently, selecting inappropriate features for the model can lead to

unfair outcomes [2].

Representation bias and sampling

This bias stems from the way datasets are sampled during the data collection phase. Non-representative samples lack diversity and may not adequately reflect the characteristics of the overall population. For example, if a dataset lacks sufficient representation from a specific group, such as a particular race, it can introduce bias into classification tasks [2]. Inappropriately sampled data can lead to imbalances in the training phase, where the model encounters more words associated with one gender over another due to skewed sampling. As a result, the model may predict words related to one gender more accurately than the other for a particular class. It is also crucial for the sampling strategy to be random, as non-random sampling can introduce bias into the training dataset ([46], [2]).

Temporal bias

Certain datasets, such as those containing abusive language, may include words with offensive semantics that have evolved over time. Words that were not considered offensive two years ago might now carry a different connotation. Consequently, a model trained on a dataset from the previous year may not perform effectively on today’s datasets [47].

For instance, Twitter discussions often revolve around specific topics initially identified by hashtags. However, as time progresses, people may continue these discussions without using hashtags or even move on to new subjects.

2.2.2 Bias in contextualized pretrained language models

Pretrained language models (PLM), built on the transformer architecture [48], have demonstrated considerable success in various natural language processing (NLP) tasks ([49], [50]). However, it is essential to recognize that biases may originate from the pretraining phase of these models. A notable example is discussed in the paper [51], where the BERT model was employed to forecast words linked to disability-related expressions. The methodology involved generating fill-in-the-blank sentence samples with the structure "phrase is....," focusing on disability-related definitions, such as "a deaf person", "a blind person", "a person with ADHD" etc. The BERT model was then tasked with predicting masked words within these sentences. Subsequent sentiment analysis on the terms predicted by BERT revealed a prevalent occurrence of negative sentiments associated with words linked to disability phrases. Another study considers a generative model (GPT-2) to study the possible gender-profession

stereotypes within these models. After collecting sentence templates for generative tasks, they found that the predicted jobs for women and men template sentences were not diverse and were more stereotypical towards women [10]. RoBERTa [19] and ALBERT [20] also show strong stereotypical bias when they are applied to link neutral sentences that contains race sensitive attributes like "Arab" to stereotypical sentences. As an example, these models often link the sentence "He is an Arab from the Middle East." more closely to the statement "He is probably a terrorist with bombs." than to the sentence "He is a pacifist" [52].

Another example shows that GPT-3 can establish analogies in semantic spaces during the pretraining phase. GPT-3 demonstrated learned associations between religious attributes (such as Muslim or Christian), violent nouns (like "terrorist"), and even non-violent nouns like "peace" and "faithfulness." In a study described in this paper [53], GPT-3 was employed to create a sentence using an analogy, revealing a close association between the terms "Muslim" and "terrorist" in the model's learned representation.

A separate analysis on the GPT-3 pretraining corpus, as outlined in [54], uncovered non-negligible amounts of toxic, harmful, and abusive text. The study delved into the GPT-3 pretraining datasets, such as WebText2 and Common Crawl, identifying issues including the presence of toxic content as well as news articles from unreliable sources shared on banned internet communities. This underscores the importance of scrutinizing and addressing biases present in the datasets used to pretrain language models, as they can significantly impact the model's behavior in various applications. Table 2.1 shows the most popular pretrained language models' pretraining datasets and their inherent bias.

Language model	Dataset	Bias
BERT [18]	Books corpus and English Wikipedia (16 GB)	Gender imbalanced [55]
RoBERTa [19]	Books corpus and English Wikipedia (160 GB)	Gender imbalanced [55]
XLNET [56]	Books corpus and English Wikipedia, Giga 5 (32.89 GB)	Gender imbalanced [55]
GPT-2 [21]	WebText from Reddit (40 GB)	Abusive text- gender imbalanced [55], [54]
GPT-3 [57]	WebText2 and Common Crawl	Abusive text [54]

Table 2.1 Pretraining datasets for most popular PLMs

This highlights the potential for biases ingrained in pretrained language models to influence the outcomes of NLP tasks, shedding light on the importance of addressing and mitigating such biases in language model development.

2.2.3 Bias in PLM-based evaluation metrics

Pretrained language models play a crucial role in various tasks, and metrics based on these models are commonly employed to assess the semantic similarity between sentences, includ-

ing predictions and reference sentences. Notably, widely used pretrained language models and associated metrics like BERTScore [58] and BARTScore [59] have prompted researchers to scrutinize their fairness. In this paper [60], an evaluation of these metrics revealed significant bias across six sensitive attributes: Age, Race, Gender, Religion, Physical appearance, and Socioeconomic Status. The authors assessed pretrained language model (PLM)-based metrics, specifically BERTScore and BARTScore, concerning their susceptibility to social bias in text generation tasks. If a PLM-based metric assigns distinct scores to a reference set and two candidate sentences with minor gender-type variations, it indicates a bias present in these metrics. Their results show that PLM-based metrics are more likely to carry bias than traditional metrics. This finding underscores the importance of considering and addressing potential biases in the use of pretrained language models and associated evaluation metrics.

In this work, we aim to examine gender bias within both the training dataset and pretrained language models. Our methodology involves identifying gender-related terms and employing tagging to minimize the correlation between gender terms and professions in the dataset. Given that pretrained language models inherently learn social bias during the pretraining phase, our approach involves fine-tuning these models on the tagged corpus. This process helps alleviate bias within pretrained models, thereby preventing the propagation of bias in contextualized word embeddings and downstream tasks.

2.3 Identifying bias and corresponding metrics

This section comprises the methods and metrics to detect bias in the different aforementioned sources.

2.3.1 Identifying bias in the training dataset

This section will categorize methods capable of identifying bias within a training dataset.

Class distribution and availability analysis method

The actual distribution of classes in datasets plays a crucial role in the training stage. Data may be removed for various reasons, including errors in data management or automated removal due to offensive content (e.g., sexist words or offensive language) by platforms and websites like Wikipedia, Twitter, etc. These factors can impact both the quantity and quality of the dataset, affecting the true distribution of each class. In such instances, conducting a visualized class distribution analysis can be instrumental in identifying abnormal datasets or instances where there is a deficiency of data for specific class labels [47]. For instance, the

removal of certain elements may lead to an imbalance in gender distribution within a specific class, inaccurately reflecting the true gender distribution. Conducting an analysis of class distribution can be instrumental in identifying social bias within the training corpora.

Temporal distribution analysis

The temporal aspect of posting is pivotal, especially when dealing with datasets comprising Tweets. A dataset that is excessively dated may prove unsuitable for model training, given the dynamic evolution of language over time. On the flip side, a dataset containing Tweets posted within an extremely brief time-frame may present challenges. A visualized distribution analysis of Tweets can help evaluate whether the training dataset effectively encompasses a diverse range of user traits on the platform or not, as highlighted in the study by Wich et al. [47]. Tweets are frequently employed in the study of racial bias, and the timing of tweet posts is crucial. Visualizing tweets based on their timestamps enables us to identify timelines with a diverse range of subjects, including race, gender, religion, and more. This visualization aids in pinpointing periods when a wide array of tweet topics related to various social aspects are prevalent.

Author distribution analysis

The study described in the paper by Wich et al. [47] utilized Pareto analysis on authors' documents within the Waseem dataset [61]. The analysis revealed that a single author generated nearly all racist Tweets in the dataset, while a substantial portion of sexist tweets originated from two authors. This observation suggests the presence of author bias in creating the training dataset, and such bias could potentially be propagated during the prediction stage. To identify and understand training dataset bias attributed to authors, an author distribution analysis becomes a valuable method.

PMI-Based word ranking for classes

Point-wise Mutual Information (PMI) serves as a metric to gauge the correlation between individual words within a class and various classes, unveiling the thematic essence learned by a machine learning model. This facilitates the identification of words characterizing each class. For instance, in the Waseem dataset's sexism class [61], notable PMI scores emerge among words like women, girls, men, and sexism. Conversely, within the hostility class of the Vidgen dataset [62], elevated PMI scores link the class with words such as China, people, indicating an implicit bias in the dataset. In the equation 2.1, w_i is referred to a word, c_j is

a class and p refers to probability [47]:

$$PMI(w_i, c_j) = \log \frac{p(w_i, c_j)}{p(w_i)p(c_j)} \quad (2.1)$$

Measuring the percentage of false-positive identity terms

In the context of classification tasks specially abusive language detection, a valuable approach to unveil potential bias in training data, especially in tasks related to identity, involves examining the percentage of identity terms within specific classes compared to the entire dataset. Identity terms, encompassing descriptors like lesbian, guy, Islam, Muslim, etc., serve to identify individuals but should ideally not unduly influence model predictions. An effective measure is to assess whether the proportion of sentences containing identity terms, specifically categorized as toxic comments, significantly surpasses the overall proportion in the entire dataset. Elevated ratios may indicate bias in the labeling and annotation of the training data, as discussed in Dixon et al.’s work [63].

Quantifying and identifying bias in the text by word counts and word vector similarities

This paper [64] described a metric to evaluate the gender bias in the text. This approach measures bias based on word counts for masculine and feminine terms in the text and provides a score of (-1,0,1) for feminine, neutral, and masculine bias, respectively. The drawback of this method is that the output is a discrete number and cannot express the bias degree of the text. [64] adopted the word counting approach but considered the continuous output in [-1,+1] where a negative value represents feminine bias and positive values represent masculine bias. The score is computed as in equation 2.2. X_{fem} represents the number of feminine words in the text based on word lists provided by these papers [65] [66] and the authors’ word list and X_{mas} is the counts of men-related words. However, this approach, called S1 in equation 2.2 could not reflect the potential differences in the level of bias exhibited by different words in the text. Consequently, the authors proposed a second bias metric similar to Relative Norm Distance [67] or the Word Embedding Association test (WEAT) [68]. In this method, first, an average of word vectors in the specified text is calculated, then the cosine distance of each masculine and feminine word in the dictionary with the text vector is computed; after that, the difference in the average of cosine distance is calculated as shown in equation 2.3 (S2). For the notation, M refers to the number of words in the masculine word list, F refers to the number of words in the feminine list, V_T is the word vector of the text, and V_w is a specified

word vector.

$$S1 = Max\{\frac{X_{mas} - X_{fem}}{X_{mas}}, \frac{X_{fem} - X_{mas}}{X_{fem}}\} \quad (2.2)$$

$$S2 = \frac{1}{M} \sum_{w \in M} \frac{V_T \cdot V_w}{||V_T|| \cdot ||V_w||} - \frac{1}{F} \sum_{w \in F} \frac{V_T \cdot V_w}{||V_T|| \cdot ||V_w||} \quad (2.3)$$

Identifying bias in a text by Constrained Co-Occurrence score

This metric assesses the co-occurrence of neutral words in a given text alongside words indicative of gender or race definition words, such as "Female" and "Male" for gender or "African-American" and "Caucasian" for race. Let N represents the set of attributes and adverbs in the text, and A and B denote the sets of gender or race-defining words. The function C(w, b) calculates the number of times word w co-occurs with the definition words in set B, and C(w, a) represents the number of times word w co-occurs with the definition words in set A.

The Co-occurrence Bias Score (CCO) is derived from this measure, providing a quantitative assessment of bias in the text. The score ranges from 0 to infinity, indicating the degree of association between neutral words such as "beautiful" and social bias terms such as gender terms (women, men) or race-defining terms (American, African). This co-occurrence calculation can be conducted within a specified window in the text, as detailed in Garimella et al.'s work [69].

$$CCO(text) = \frac{1}{N} \sum_{w \in N} |log(\frac{\sum_{a \in A} c(w, a)}{\sum_{b \in B} c(w, b)})| \quad (2.4)$$

In this study, we have conducted a "class distribution analysis" for two genders within the BiasinBios dataset. Gender imbalance in the training dataset could import social bias in the fine-tuned pretrained language models, so identifying this distribution and also removing this imbalance is an interesting research avenue that we pursue in this work.

2.3.2 Identifying bias in pretrained language models

Here we explore some of the most prominent methods and metrics designed to detect and quantify bias in PLMs and word embeddings. We employed some of these metrics in this work and indicate that in the appropriate sections.

Difference in probability distributions over stereotypical word or sentence

Local biases occur when a generated prediction has an undesirable association with the context. For example, some language models assign a higher probability to the doctor in the first sentence "he works as a [doctor]" in comparison to the second sentence "she works as a [doctor]" [70]. Thus, the difference in these probabilities shows bias in these language models.

Discovery of correlations (DisCo)

This metric is instrumental in evaluating bias within masked language models. It utilizes templates with two slots, exemplified by $T=X \text{ likes to } [MASK]$, where the first slot (X) incorporates gender-related terms such as gender names, women, men, or related professions. The second slot is then populated by the masked language model, and the metric examines the top three predictions. If these predictions differ for two genders, the metric identifies bias. The quantification of this bias involves averaging the count of differing predictions across the two genders, as detailed in Delobelle et al.'s work [71].

A modification introduced by the paper [72] replaces the top three predictions with a probability filter denoted as $p(x_m, T) > 0.1$ which is the probability of the masked token X_m in the template sentence T. In this way, all predicted words with a probability higher than 0.1 (and not only the top three) are compared between the two genders and this provides a comprehensive assessment between two genders.

Context association test (CAT score)

The paper by Nadeem et al. [52] introduced the StereoSet dataset along with three metrics designed to assess social bias within pretrained language models. This dataset is tailored for evaluating language models' ability to predict the next sentence or word in a prompt, with three options: "unrelated," "stereotype," and "anti-stereotype." The goal is to determine which option is more likely to fill in the blank.

The proposed metrics include the **Language Modeling Score (LMS)**, **Stereotypical Bias**

Score (SS), and **CAT Score (CAT)**. The LMS quantifies the percentage of examples in the dataset where the language model favors meaningful options (stereotypical or anti-stereotypical words) over meaningless ones. A stereotype refers to a societal belief regarding a specific group of individuals. As an illustration, a common stereotype is the assumption that the majority of nurses are women, as mentioned in a study by Nadeem et al. (2020) [52]. On the contrary, an anti-stereotype challenges such preconceptions, proposing an opposite viewpoint, such as the idea that most nurses are men. The SS measures the percentage of examples in which the language model prefers stereotypical words over anti-stereotypes. A fair model is characterized by an SS score of 50% and an LMS score of 100%.

To balance accuracy and fairness, the authors introduced the CAT score. This score is determined by combining the LMS and SS scores. According to equation 2.5, the best CAT score (100%) occurs when the LMS score is 100% and the SS score is 50%, indicating a language model that excels in both accuracy and fairness.

$$CAT = lms * \frac{\min(ss, 100 - ss)}{50} \quad (2.5)$$

Bias detection in Masked-Language model (MLM) with CrowS-Pairs Score (CPS)

In the assessment of bias, this paper [73] introduced the CrowS-Pairs dataset and employed the adapted pseudo-log-likelihood metric, as proposed by [74] [75]. The CrowS-Pairs dataset consists of sentence pairs with the sole difference being sensitive attributes, such as gender pronouns (e.g., "He" or "She"). An example includes "John ran into his old football friend" vs. "Shaniqua ran into her old football friend," where [John and his] are the modified parts, while the rest remains unaltered.

In contrast to the approach introduced in this paper [52], where the sensitive attribute is masked, this method [73] conditions the masking of other tokens based on the sensitive attributes. The model's advantage lies in assigning a higher probability to "John" not merely due to its frequent occurrence in the pretraining corpus but rather because of the social bias ingrained in the language model.

The scoring involves evaluating sets of unmodified tokens (U) and modified tokens (M) for both stereotype and anti-stereotype sentences. The unbiased model predicts an equal distribution of stereotypes and anti-stereotypes and achieves a bias score of 50%. CrowSpairs score (CPS) is defined as below:

$$CPS = \sum_{i=0}^{|C|} \log P(u_i \in U | U_{\setminus u_i}, M, \theta) \quad (2.6)$$

All Unmasked Likelihood score (AUL)

All Unmasked Likelihood score [76] is another metric to measure bias in masked language models. In contrast to other metrics such as CPS and CAT scores, it predicts all tokens in a test sentence given the MLM embedding of its unmasked input. AUL creates an embedding for a test sentence without masking any of its tokens, using information related to all of the tokens in that sentence. By requiring the MLM to simultaneously predict all of the unmasked tokens in a sentence, it avoids any biases due to masking a subset of the input tokens, such as highly frequent words. In fact the CPS metric has some drawbacks. Firstly, when one token is masked at a time from context, the context is changed due to the lack of this token, resulting in a loss of information that could help predict a masked token. Therefore the accuracy of the MLM can decrease. Secondly, the remaining tokens can also be biased. The solution is that instead of masking tokens, the whole sentence is provided to the MLM, and all tokens between the first and the last tokens are predicted without masking (Byte pair encoding). In this method, the number of predicted tokens is the same as the initial sentence tokens. AUL is calculated as in equation (2.7) in which S refers to the test sentences, θ refers to masked language parameter, w_i is the current token to be predicted; equation AULA (2.8) considers the importance of tokens in a sentence with the attention parameter, α_i . The bias score is the proportion of the number of examples where the language model prefers stereotype sentence S^{st} over anti-stereotypical sentence S^{at} over the whole examples (2.9), where the highest "biasscore" is 100.

$$AUL(S) = \frac{1}{|S|} \sum_{i=1}^{|S|} \log P_{MLM}(w_i|S; \theta) \quad (2.7)$$

$$AULA(S) = \frac{1}{|S|} \sum_{i=1}^{|S|} \alpha_i \log P_{MLM}(w_i|S; \theta) \quad (2.8)$$

$$Biasscore = \frac{100}{N} \sum_{S_{st}, S_{at}} I(f(s^{st}) > f(s^{at})) \quad (2.9)$$

Log probability bias score

BERT is a Masked language model that can predict an association between target words and attributes for example male and programmer. To identify the association between man and programmer in BERT, first, a masked sentence is fed into the BERT language model; for example, [Mask] is a programmer, then the probability of this sentence "he is a programmer"

is assigned as (p_{im}) , and then the probability is re-weighted by the prior probability of this sentence "he is a [Mask]"(pm_{prior}). However, the aim is to determine how much the model prefers male to female for a programming career; so, this procedure is repeated for the "She" pronoun. The log probability bias is calculated as below [77]. In equation 2.10, p_{im} is the probability of a stereotype sentence like the one described above, pm_{prior} is the prior probability for the sentence "im", p_{ig} is the probability of the anti-stereotype sentence, and pg_{prior} is the prior probability for the sentence "ig" :

$$Logscore = \log \frac{p_{im}}{pm_{prior}} - \log \frac{p_{ig}}{pg_{prior}} \quad (2.10)$$

Identifying bias in PLM-based metrics

As stated in section 1.1, pretrained language models (PLMs) are widely employed in various tasks, including the construction of PLM-based metrics. These metrics can be categorized based on the tasks they perform. Matching-based metrics, such as BERTScore and Mover-Score [78], assess the semantic similarity between two sentences, typically a reference and a predicted sentence, to evaluate the model's performance. Regression-based metrics [79], exemplified by BLUERT [79], predict a numerical score for the language model's output. Another category involves generation-based metrics, like BARTScore [59], which calculates precision and recall using the conditioned probability of reference and predicted sentences.

In evaluating bias in text-generation metrics, a study [60] utilized template-based sentences with and without sensitive attributes. The reference sentence, devoid of any sensitive attributes, was compared to two candidate sentences, each containing a sensitive attribute. For instance, the reference sentence "The carpenter made a desk for the clerk because the clerk needed a desk" was compared to Candidate 1: "The carpenter made a desk for the clerk because she needed a desk" and Candidate 2: "The carpenter made a desk for the clerk because he needed a desk."

To quantify bias in PLM-based metrics, the paper introduced an equation involving similarity scores. Let's assume $\hat{S}_{i,1}$ represents the similarity score between a candidate sentence one and the reference, and $\hat{S}_{i,2}$ represents the similarity score between candidate sentence two and the reference, and N represents the number of reference sentences. The bias in the PLM-based metric can be calculated using the following equation:

$$Bias = \frac{1}{N} \sum_{i=1}^N |\hat{S}_{i,1} - \hat{S}_{i,2}| \quad (2.11)$$

Word embedding association test (WEAT)

Word embeddings are representations of words as points in a vector space. In fixed word embeddings, for example, GloVe, by considering two sets of target words (programmer, engineer, scientist, etc.) and (nurse, teacher, librarian, etc.) and two sets of attribute words (man, male, him, his, etc.), and (woman, female, her, ...), the hypothesis is that there is not any difference between two sets of target words in terms of similarity to the attribute words. Bias in word embeddings is calculated by the difference between the mean of cosine similarity of the target words with two sets of stereotype and anti-stereotype attributes [68]. Let A and B be men-related and women-related attributes, respectively, and X is the set of target words that are primarily stereotypes for men, and Y is the stereotypical target terms for women. In equation 2.12, denote $w(x_1, A, B)$ is a function that measures the association of a specific word x_1 with the attribute sets A and B. It is defined as the difference between the average cosine similarity of x_1 with words in A and the average cosine similarity of x_1 with words in B. WEAT score is calculated as 2.13:

$$w(x_1, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(x_1, a) - \frac{1}{|B|} \sum_{b \in B} \cos(x_1, b) \quad (2.12)$$

$$WEAT(X, Y, A, B) = \sum_{x \in X} w(x, A, B) - \sum_{y \in Y} w(y, A, B) \quad (2.13)$$

The sentence Encoder Association Test (SEAT)

The Sentence Encoder Association Test (SEAT) serves as another benchmark dataset and metric for evaluation of social bias within contextualized pretrained models [40]. SEAT is an extension of the Static Word Embedding Association Test (WEAT) [68], designed specifically for contextualized word embedding settings. In SEAT, the evaluation involves leveraging simple sentences 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. Let us denote A and B as the sets of attribute terms such as, she, he, man and woman. X and Y as the sets of target terms such as (family, profession, career). Based on the WEAT score described in [68], for the specific word w, $S(w, A, B)$ is described as a difference between the mean of w's cosine similarity with the words from A word list and w's mean cosine similarity with the word list

B. The effect size is then calculated by the below equation. Let us denote μ and σ as a mean and standard deviation, respectively. The score closer to zero is the best:

$$S(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \quad (2.14)$$

$$d = \frac{\mu([s(x, A, B)]_{x \in X}) - \mu([s(y, A, B)]_{y \in Y})}{\sigma([s(t, X, Y)]_{t \in A \cup B})} \quad (2.15)$$

In this study, we employed various metrics to assess bias in pretrained language models before and after mitigation. To gauge the model’s inclination towards stereotypes, anti-stereotypes, or unbiased representations, we utilized the CrowSpairs score. Additionally, the SEAT score enabled us to evaluate the bias in sentence representations within pretrained language models. This score served as a valuable tool for comparing our model with the current state-of-the-art in bias mitigation strategies.

2.3.3 Identifying bias in machine learning classifiers

AUC difference score

The evaluation of gender bias in language models becomes challenging when tested on biased datasets that mirror the same distribution. In a study [80], an attempt is made to address this issue by creating an unbiased test set. The underlying principle is that a fair machine learning model should exhibit comparable accuracy in predicting the classification task of abusive language on both the original biased test set and the newly generated unbiased test set.

The equitable test set comprised sentences containing identity terms associated with both men and women, categorized into neutral and abusive language classes. The evaluation metric employed for assessing model performance on both test sets is AUC (Area under the Curve). The disparity in AUC scores between the original biased test set and the unbiased test set is used as a measure of the degree of bias present in the model.

$$\text{AUC difference} = \text{AUC}_{\text{biased}} - \text{AUC}_{\text{unbiased}} \quad (2.16)$$

Pinned AUC equality difference

This is a method to identify and quantify bias in the classification models. First, the dataset is separated into subgroups such as gay, Muslim, black, white, etc. The task is a classification

of sentences into two toxic and non-toxic classes. To evaluate whether the machine learning model contains bias, the first task involves sampling from the whole dataset and joining the sampled dataset to the sample from each subgroup. Pinned AUC (pAUC) is the Area under the curve or, in other words, the model’s accuracy on the dataset’s subset. The difference between the model’s AUC on the whole test data and pinned AUC defines bias [63]. Note that "t" refers to a given subgroup and T is total number of subgroups.

$$\text{Pinned AUC Equality Difference} = \sum_{t \in T} |AUC - pAUC_t| \quad (2.17)$$

False Positive/Negative Equality Difference Scores

This method, first, calculates the FPR (False positive rate) and FNR (False negative rate) for the test set of the dataset (see equations 2.18 and 2.19). 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 t. 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 [63]. The formulas are described below:

$$FPR = \frac{FP}{FP + TN} \quad (2.18)$$

$$FNR = \frac{FN}{FN + TP} \quad (2.19)$$

$$FPED = \sum_{t \in T} |FPR - FPR_t| \quad (2.20)$$

$$FNED = \sum_{t \in T} |FNR - FNR_t| \quad (2.21)$$

While the aforementioned metrics are not exhaustive, they are among the most prominent ones in the literature.

In this study, in addition to employing standard classification metrics such as precision (P), recall (R), F1, and accuracy (Acc) scores, we have incorporated FPED and FNED scores for assessing the fairness of the PLMs’ classifiers. In the context of a classification task,

FNED and FPED scores provide insights into how much the model distinguishes between two genders when predicting true labels. These metrics are widely utilized in examining social bias in classification tasks, as highlighted in previous research [81] [82].

To evaluate the effectiveness of our proposed model’s sentence representation in mitigating bias and to compare it with other debiasing approaches from the literature, we have also employed the SEAT score. Additionally, we have utilized the CrowSpairs metric as a benchmark. This metric allows us to gauge whether our debiased model behaves neutrally in selecting between stereotypical and anti-stereotypical sentences.

2.4 Bias mitigation

Upon identifying social bias within various sources, our focus shifts towards examining strategies for mitigating these biases in both the training dataset and pretrained language models.

2.4.1 Mitigating bias in training datasets

This section outlines the methods to mitigate social bias in the training dataset.

Counterfactual data augmentation (CDA)

Sometimes datasets have disproportional gender types. To mitigate this, [24] proposed data augmentation. In this method, besides the original data, an identical dataset to the original dataset is created. Still, it is biased toward another gender, and the model is trained on the union of original and data-swapped sets. Data augmentation is widely used, and it could mitigate gender bias as well as other types of bias like religion by swapping and augmenting attributes, for example, **church** instead of **mosque** [22]. This method is used in a wide range of tasks; for instance, in hate speech detection, data augmentation substantially reduced False positive equality difference (FPED) and False negative equality difference (FNED) using Convolutional Neural Networks by a wide margin [80].

To address data imbalance, which can lead to unintended bias, other strategy involves augmenting the training dataset with additional data. For instance, in the paper by Dixon et al. [63], the identity term "gay" constitutes only 0.5% of the overall dataset, yet three percent of the toxic class comprises sentences related to this identity term. This imbalance arises due to the inappropriate distribution of identity terms across classes. To rectify this, additional data corresponding to each identity term or sensitive attribute can be added to the non-toxic class dataset, aiming to balance the data with respect to the initial distribution of these

attributes throughout the entire dataset. In the context of a classification task, if the model is influenced by sentence length, additional data points for each sentence length can also be introduced [63].

There are also limitations to data augmentation; while it is easy to use, creating an annotated list is expensive and time-consuming; these are also large datasets (original and gender-swapped), which can increase training time. Lastly, sometimes swapping can lead to nonsensical sentences, for example, *this man gave birth to a child* [83].

Scrubbing

Scrubbing represents one approach to mitigate social bias in the training dataset. Initially introduced by this paper [9], the method involves removing a list of gender indicator terms, including proper nouns (e.g., John, Emily), pronouns (she and he), and possessives (her, his, himself, etc.), from the text.

Gender-tagging

In some cases, like machine translation (MT), gender tagging can be used in order to help the model predict more accurately when the destination language is sensitive to gender. For example, English to French. This method tags gender, M, or F at the beginning of each data point, so the model can realize the gender of the speaker ([84], [85]). However, the gender tagging method can be expensive in time and memory usage: knowing the gender of the source requires meta-information, and obtaining this can be costly [16].

2.4.2 Mitigating bias in pretrained language models

In this section, we will explain the common methods to mitigate social bias in the pretrained language models.

Projection-based debiasing

Hard debiasing. Hard debiasing is a method used to reduce gender bias in word embeddings while preserving the useful properties of the embedding. It involves the application of a debiasing algorithm to the word embedding to effectively reduce gender stereotypes present in the data. The process begins by identifying a gender subspace in the embedding that captures the bias. To define the gender subspace, the study uses the principal components of gender pair difference vectors and computes its principal components (PCs). The top

PC, denoted by the unit vector g , captures the gender subspace. This direction is identified by analyzing the variance in the vectors and observing the gradual and uniform decrease in eigenvalues. The next step is to neutralize and equalize or soften the gender bias. In the case of hard debiasing, the algorithm ensures that gender-neutral words are zero in the gender subspace and perfectly equalizes sets of words outside the subspace. This means that any neutral word is equidistant to all words in each equality set. The gender-specific and gender-neutral terms were derived through a combination of manual selection, crowdworker input, and the use of a linear classifier to identify and classify words based on their gender associations. The algorithm effectively removes gender stereotypes from the word embedding while maintaining the useful properties and quality of the embeddings. It has been shown to significantly reduce both direct and indirect gender bias in the word embeddings [86].

SentenceDebias. This approach builds upon the hard debiasing technique for word embeddings introduced in [86], extending it to address bias in sentence representations. The method is applicable to various bias types, such as gender, race, and more. In the initial phase, gender attributes ("She" and "He") are incorporated into straightforward sentence templates, which are then inputted into a pretrained language model. The resulting sentence representations are derived from this language model.

The subsequent step involves identifying the bias subspace, encompassing aspects like gender, religion, and race. Subsequently, the sentence representation is projected onto this bias subspace. By subtracting this resulting projection from the original sentence representation, the method effectively neutralizes bias, rendering the sentence unbiased [87] [22].

Debiasing pretrained models by prompting

Distracting word associations. The approach presented in this paper [53] outlines a technique for mitigating biases in text generated by the GPT-3 language model. In order to eliminate the implicit association between the terms "Muslim" and "Terrorist," the author introduced a brief phrase at the outset of the main text prompt. This additional language serves to positively link "Muslim" with other adjectives like "trusted" and "hard-working." The experimental outcomes demonstrate a significant reduction in the occurrence of violent language in sentence completion tasks across multiple iterations.

Self debiasing. Self debiasing method is a type of bias mitigation that does not need any change through a model parameter or training dataset. In this case, first, the model is encouraged to produce toxic comments, by prompting potential bias-related sentences, for

example: "Nurse is the job of ...". Language models (GPT-2, T5) tend to predict women more than men. After that, the approach uses the prompt "this is a discrimination toward women," in the second sentence generation. The language model tends to assign a lower probability to the next word as "women" [88].

Dropout regularization

Dropout regularization serves as a preventive measure against overfitting in large machine learning models. Its utility extends to addressing erroneous associations between words and sentences. For instance, in tasks like coreference resolution, it is crucial that models like BERT do not disproportionately predict "she" for a "nurse" compared to "he." Dropout regularization intervenes in the attention mechanism of BERT, disrupting the potential for inaccurate gender associations. By preventing overfitting, dropout helps the model avoid learning noise and idiosyncrasies present in the training data that may lead to biased representations. It encourages the model to learn more general and less data-specific patterns, potentially mitigating biases associated with specific social characteristics. This regularization method is versatile and can be applied to mitigate biases related to race and religion, as discussed in [22].

In the case of the publicly available ALBERT model, the dropout regularization is initially set to zero. However, increasing this parameter to 0.5 has been shown to decrease incorrect associations without a substantial reduction in the model's overall accuracy, as evidenced by findings in [89].

Counterfactual role reversal mitigation method in distilled models

In the current landscape, language models are growing in size and resource requirements. To address this challenge, distilled models are employed to replicate the functionalities of large language models in performing natural language processing (NLP) tasks. In this context, the distilled model serves as the student, while the original extensive language model acts as the teacher. However, the process of knowledge distillation can inadvertently transfer social biases from the teacher model to the student.

One approach to mitigate this bias, known as counterfactual role reversal, involves generating two sentences with different genders, such as "He works in the hospital as a..." and "She works in a hospital as a...". Large language models, like GPT-2, may exhibit a tendency to assign a higher probability to a doctor for the pronoun "He" than for "She." To prevent the transfer of gender disparities into the student model, the combined likelihood of the original statement

and its counterfactual counterpart is fed into the student loss function instead of the original probability.

To achieve this combination, the study outlined in [90] employed various methods, including the formulation $p_{com} = \log(\frac{e^{z^t} + e^{z^s}}{2})$, where z^t represents the probability of the next token in the original sentence, and z^s corresponds to the probability of the next token in the counterfactual sentence [90].

Debiasing pretrained language models by modifying loss function

Using gender equalizing loss function in word-level language models. Word-level language models, such as RNN or LSTM, may exhibit gender bias in tasks like next word prediction, particularly in sentences associating gender with professions. To mitigate this bias, a modification is introduced to the entropy loss function of these models, aiming to equalize the probability distribution of words related to men and women.

The cross-entropy loss at time step t is conventionally defined as follows, where V represents the vocabulary, y is the ground truth label, and $\hat{y}$ is the model’s output:

$$L^{CE}(t) = - \sum_{w \in V} y_{w,t} \log(\hat{y}_{w,t}) \quad (2.22)$$

The augmented component of the loss function, denoted as L^B , is designed to equalize the predicted probabilities of gender-related words. Here, G denotes the size of the gender pairs set (she-he, her-his, etc), while F and M are sets encompassing gender-related terms, as described in [91].

$$L^B(t) = \frac{1}{G} \sum_i^G \left| \log \frac{\hat{y}_{f_i,t}}{\hat{y}_{m_i,t}} \right| \quad (2.23)$$

DEBIASBERT by further pretraining. This method used BERT as a pretrained language model and pretrain it further with bias mitigation objectives and its loss function. This method consists of two stages, Equalizing and Declustering. In the Equalizing stage, the model attempts to equalize the association between every gender-neutral word with two sensitive attributes, for the gender (Men and women-related words) and for the race (African-American and Caucasian-related words) [91]. To achieve this equalization, in addition to cross-entropy loss, the below equation (2.24) is used. K is the number of gender or race word pairs (men-women for gender and African-American for race). Group A and Group B are sensitive attributes (women-related and men-related words for gender). The second

stage is declustering. Some neutral words are implicitly associated with gendered or racial words. For example, "delicate" is gender-neutral but strongly associated with women. These attributes for race could be "blackness," "under-world," and "oversized," and they are closer to African identity terms. In equation 2.25, these attributes are referred as socialgroupA, such as "beautiful" for women related attributes and socialgroupB, such as "powerful" for men related attributes. Upon obtaining representations of these words from the Brown corpus, where the chosen words exhibit a high projection value in the she-he direction for gender bias or the slave-manager direction for racial bias [92], and identifying neutral words (those not part of gender-indicator terms), the following equation (2.25) is incorporated into both the cross-entropy and EqLoss functions. This addition aims to mitigate both implicit and explicit associations, as highlighted in the work by Garimella et al. [69]. In these equations (2.24, 2.25), $\lambda > 0$ defines is an equalizing weight, and p is the probability of the masked token.

$$EqLoss = \lambda \frac{1}{k} \sum_{i=1}^k \left| \log \left(\frac{P[groupA_i]}{P[groupB_i]} \right) \right| \quad (2.24)$$

$$DeclustLoss = \lambda \left| \log \left(\frac{\sum_{i=1}^{|A|} P([socialgroupA_i])}{\sum_{i=1}^{|B|} P([socialgroupB_i])} \right) \right| \quad (2.25)$$

Mitigating bias in decoder with BiasPenalizing Loss. This method mitigates bias in the decoder while the model generates a summarization task. In the summarization task, bias is defined as discriminating against sensitive groups (gender, race) due to language choices. While debiasing methods in the encoder could mitigate bias, the input sequence could also carry bias, resulting in a bias in the output sequence. In the summarization task, to alleviate this type of bias in the decoder, in addition to negative log-likelihood loss, a penalizing loss is added to the loss function as shown below (2.27). W is defined as all sets of adjectives and adverbs; these words come before and after sensitive group terms such as "beautiful women" and makes implicit association between gender and some adjectives and adverbs, $b_i(W_i)$ is a bias score of W_i computed with equation 2.26, and $P(W_i)$ is a probability of W_i . In the equation (2.26), k represents the number of gender words, Group A and B represent the definition words for two underrepresented groups (Female and Male), $P[groupA_j, W_i]$ represents the probability of co-occurrence of the word W_i with the group A_j in the input context (with the specified context window) [69]. The resulting "BiasPenalizingLoss" will be added to the standard loss function of the model.

$$b_i(W_i) = \frac{1}{k} \sum_{j=1}^k \left| \log \left(\frac{P[groupA_j, W_i]}{P[groupB_j, W_i]} \right) \right| \quad (2.26)$$

$$BiasPenalizingLoss = \sum_{i=1}^{|W|} (e^{b_i} * P(W_i)) \quad (2.27)$$

Debiasing pretrained language models by efficient fine-tuning

Numerous approaches exist for mitigating gender bias in pre-trained language models, with one common strategy being re-training these models on a neutral corpus [93] and unfreezing the whole parameters of the model in the fine-tuning stage [93]. However, a drawback of this method is the risk of catastrophic forgetting, where the model loses knowledge acquired during the initial pre-training phase which leads to a decrease of the language model performance in downstream tasks [94], [95]. Thus a common strategy is to target only a subset of the models hyper-parameters.

Unfreezing a small set of model parameters. This paper [94] applied efficient fine-tuning by only changing only 1 percent of the model parameters in the fine-tuning stage to reduce social bias and simultaneously increase model performance. First, the entire model is frozen, then they apply one or more of the below options. 1. unfreezing the norm layer parameter 2. unfreezing the word embedding 3. unfreezing the word positioning embeddings 4. adding a linear input transformation 5. adding a linear output transformation. After applying these options, this paper [94] concludes that, on average, the performance of the model increased by 0.29 by the LMS score [52] and the StereoSet Score (SS) [52] decrease by 0.92 and the best performing option is applying the combination of option 1 and option 3 together.

Adaptor-Based debiasing (ADELE). Adaptors are lightweight neural model components injected into neural models' layers for efficient fine-tuning. All parameters are frozen in the fine-tuning step of the pretrained language model. This keeps all the knowledge obtained from the pretraining phase. The methods only tunes the adaptor parameters in the fine-tuning stage and injects one adaptor module into each layer of pretrained transformer [96]. Debiasing adaptors can also be used with other debiasing methods, such as modifying loss functions or counterfactual data augmentation [24] [72].

Improving gender fairness with Gender Equality Prompt (GEEP). In the context of addressing gender bias, the paper by [97] proposes an innovative technique that involves freezing the initial parameters of the model to retain previously learned information.

Given that much of the gender bias is associated with professions, the study focuses on mitigating this bias. The first step involves employing data augmentation to create a neutral corpus. In the case of [97], this is achieved by using the English Wikipedia corpus and filtering out sentences containing professions commonly linked with stereotypical bias, such as doctor and nurse. Gender swapping and data augmentation methods are then applied to generate a gender-neutral training dataset.

In the second step, new parameters specific to profession tokens are introduced and injected into the pretrained language model. During the re-training process on the neutral corpus, only these generated parameters are updated, while the initial parameters remain frozen. This method is notably efficient compared to other re-training approaches, offering an effective means of improving gender fairness in pretrained language models [97].

Auto-Debias method without external corpus. This automated method aims to mitigate bias ingrained in Masked pretrained language models (MLM) without relying on an external fine-tuning corpus [8]. The necessary corpus is automatically generated through beam search from the MLM, referred to as prompts. These prompts consist of masked-style sentences, exemplified by phrases like "[PLACE HOLDER] has a job as MASKED." During the beam search, the Masked token is predicted by the MLM with target terms such as (Men, Women, girl, boy, he, she, etc.), and the placeholder slot is filled with profession stereotypes like nurse, doctor, receptionist, and so on.

In the first step, the method identifies prompts with high stereotypical bias using the beam search technique. Essentially, it pinpoints the most biased prompts in the MLM across various demographic groups for a specific profession.

Moving to the second step, fine-tuning is performed. For a particular profession (attribute term), the objective is for the fair MLM model to generate equal probabilities for all targets in all the top k masked prompts related to that masked profession.

Gender-tuning. Gender-tuning is a novel debiasing approach that integrates Masked Language Modeling (MLM) training objectives into fine-tuning on downstream tasks' datasets to simultaneously debias and fine-tune Pre-trained Language Models (PLMs). The process involves perturbing training examples by masking existing gender-related words based on a given gender-word list and replacing them with new words to disrupt the association between gender words and other words in the training examples. The perturbed examples are then classified according to the original examples' ground-truth labels to compute a joint loss from perturbation and classification for training the Gender-tuning model. This joint loss,

denoted as L_{joint} , is an aggregation of the perturbation loss ($L_{perturb}$) and fine-tuning loss ($L_{fine-tuning}$), with a weighting factor α to adjust the contribution of the two training losses in computing the joint loss. The objective of Gender-tuning is to minimize the joint loss to ensure that the label of the perturbed example is the same as the label of the original training example. The key advantage of Gender-tuning is its integration of the debiasing process into the fine-tuning, allowing for simultaneous debiasing and fine-tuning without requiring separate pre-training or additional training data.

In this work, although we refrained from fixing the model parameters, we conducted fine-tuning on the tagged dataset version. This approach proves to be more effective than employing data augmentation and has the added advantage of preserving contextual associations in Pretrained Language Models (PLMs), as opposed to the alternative of removing terms from the text, which can strip away meaning from the corpus and subsequently from the PLMs during fine-tuning. Our approach differs from gender-tuning in that we refrain from altering the loss function of Pretrained Language Models (PLMs) during the fine-tuning process. Instead, we simply tag the corpus and fine-tune PLMs based on that. This approach is more efficient compared to gender-tuning.

2.5 Bias and fairness in ranking

Since our work aims to leverage pretrained language models for job-candidate rankings, this section explores how bias and fairness are approached in ranking tasks.

Ranking algorithms are extensively employed in diverse online platforms, including LinkedIn, TaskRabbit, and Fevrr. Past research indicates a notable bias in these algorithms, particularly concerning gender and race [98]. There is a pressing need for further investigation to comprehensively understand how resumes (CVs) and job descriptions impact the ranking of candidates. Additionally, evaluating existing fair ranking algorithms is crucial to determine whether they contribute to improved hiring decisions or not, as highlighted in a study by Suhr et al. [98]. In the following discussion, we provide an overview of the definition of fairness in ranking and explore the role of pretrained models in recent papers within the realm of ranking tasks.

2.5.1 Fairness definition and metrics in ranking

In the realm of fair ranking, the predominant focus lies in ensuring equitable treatment of the items undergoing ranking. Typically, fairness in ranking entails that comparable items or groups of similar items receive comparable visibility, meaning they occupy similar positions in

the ranking [99]. There are different approaches for fair ranking. One of these methodologies is ranking with fairness constraints [100], which focuses on the representation of different groups (for example in gender, men and women) in the different top k values. In this method, they specify the lower and upper bound for the number of candidates in top k. For instance, $L(\text{women},4)=2$ means that at least 2 out of 4 top k candidates must be women.

Another approach implemented in this study ensures parity in the representation of protected groups, such as women in men stereotypical professions or men in women stereotypical roles, within different top k values. This method aims to align the distribution of these protected groups in the top k values with their proportional representation in the entire population, as referenced in prior studies [99] [101].

In determining job stereotypes, a profession is considered biased toward women if the proportion of women in that profession exceeds that of men, and vice versa [102]. This is what we call a stereotypical job for women (or men). These proportions are extracted from US Bureau statistics, as elaborated in the methodology section 3.1.1 of the study.

Normalized discounted difference (rND). The Normalized Discounted Difference measures the disparity between the proportion of protected attributes (G_1) in the top-k rankings (k starts at 5 in our ranking) and the total input data in the ranking. The protected attribute refers to a characteristic or feature of an individual that is considered sensitive or subject to protection against discrimination, such as "black people" in the hate speech texts. 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 $|G_1|$ 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.

$$rND = \frac{1}{Z} \sum_{k=5}^n \frac{1}{\log_2^k} \left| \frac{|G_1^{1,\dots,k}|}{k} - \frac{|G_1|}{n} \right| \quad (2.28)$$

2.5.2 Utility metrics in ranking

In the context of ranking tasks, a group of candidates is assessed based on various qualification metrics, including skill levels, GPA, grades, and other pertinent indicators of suitability. Typically, the straightforward approach involves a utility function that calculates the total scores of candidates occupying the highest positions. The average total of scores of top-k candidates is named as "**Average total Utility**". This metric ensures that the re-ranking of top-k candidates maintains their qualification for top positions. The corresponding formulas

are below [103]:

$$U_1^k(T) = \frac{1}{k} \sum_{i=1}^k Y_T(i) \quad (2.29)$$

In this formula, U^k determines the utility of top k rankings, T determines the type of ranking, $T(i)$ corresponds to the candidate at position i , and $Y_T(i)$ is the score of candidate at position i in ranking T . Simply, the average summation of all candidates' scores at top k positions is calculated.

The other metric of utility deals with minority groups and it is the average of the sum over all scores of minority groups in top rankings. For example, the minority group of stereotypical jobs related to men is women and the minority group of stereotypical jobs related to women is men. Therefore, we calculate "**Average protected group Utility**" as follows.

$$U_2^k(T) = \frac{1}{k} \sum_{i=1}^k Y_T(i) * \mathbb{I}[T(i) \in G] \quad (2.30)$$

Here, $\mathbb{I}$ is an indicator variable that turns to 1 when the $T(i) \in G$ which means that the candidate corresponds to the minority group and otherwise is zero. Overall, this score ensures that not only minority groups are well present in the top-k candidates but also that they are present with high ranking scores which means that the minority group candidates are qualified.

2.5.3 Fair ranking with pretrained language models

Research has revealed that ranking systems have the potential to exacerbate stereotypical biases present within datasets [104]. 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 pretrained models can inadvertently learn biases during their pretraining phase, various techniques have emerged to mitigate bias.

One such approach involves the application of adversarial learning in conjunction with BERT language models [105]. The primary objective of employing adversarial training with BERT rankers is to maximize the relevance of passages to a query while minimizing the presence of protected attributes (attributes that are considered sensitive and should not be used in a way that results in bias or discrimination) within the resulting passages. Training BERT

involves two key loss functions. The first is the maximum marginal loss, which guides the model in learning the relevance between passages and queries. The second is a cross-entropy loss, which determines whether the vector representation of a passage-query pair contains any bias-related information. The overarching goal is to strike a balance between the effectiveness of the retrieval system in identifying the most relevant passages and ensuring less biased results [105].

While these dual loss functions strive to find an optimal trade-off between relevance and fairness, there remains a concern that this result may be a local optimum, potentially diminishing the relevance of documents at the top of the ranking. To address this concern, the Bias-aware Fair Ranker model is introduced [106]. Within this model, the relevance score of irrelevant passages in the loss function is penalized if bias terms are present in the passages. In contrast, relevant passages do not receive any penalizing score if bias terms are present within them. 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.

These studies primarily concentrate on modifying the loss functions of pretrained rankers to achieve fairness in passage-query ranking.

To the best of our knowledge, there has been no prior research that explores the impact of fine-tuning these models on corpora through tagging and scrubbing. In this work, we introduce a novel framework that begins by rendering biographies gender-neutral through the process of tagging and scrubbing specific terms. Subsequently, we rank these biographies using pretrained models. Our research demonstrates that replacing gender indicator terms in the corpus with higher-level abstractions and fine-tuning pretrained language models leads to fairer rankings.

CHAPTER 3 METHODOLOGY

This segment commences with an introduction to the datasets utilized, followed by an examination of diverse techniques for altering the corpus to alleviate gender bias. Additionally, our proposed tag-based data augmentation methods are detailed in the "Data Augmentation" section 3.3. Subsequently, we describe the models employed for fine-tuning on various versions of datasets. We also expand upon our application-based social bias evaluation method, specifically biography ranking. Finally, comprehensive explanations of all classification and fairness metrics are provided in section 3.6.

3.1 Datasets

3.1.1 BiasinBios dataset

BiasinBios is a well-known dataset to study bias in the NLP domain [9]. This dataset contains biographies of people which are related to their professional life and jobs. Overall, this dataset has 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 of the dataset. 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 [9]. We follow the same dataset splitting as the original paper, with 255,710 biographies in the train set, 39,369 in the validation set and 98,344 in the test set. Table 3.1 shows descriptive statistics of the dataset.

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 3.1 Descriptive statistics of the BiasinBios dataset

As pretrained language models learn bias from their pre-training corpus, we aim to eliminate this bias association between gender and profession. To do that, first we need to know which

job in this dataset is more associated with a given gender (the dataset considers the female and male genders). To do so, we categorize job-titles utilizing data from the U.S. Bureau of Statistics ¹, as we observed that men and women are represented in different proportions within each profession in 2022. A higher proportion of men in one job determines that it is a stereotypical job for men and if the women proportion is higher, then it is a stereotypical job for women [102]. 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, and (professor, rapper, poet, and software engineer) are categorized as anti-stereotypes for women. The full list of this categorization is shown in table 3.2.

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 3.2 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.

3.1.2 List of explicit, implicit, and stereotypical bias words (L).

To identify gender-related terms, state-of-the art models rely on list of terms ([1], [89]). In this work, we accumulate various lists from these sources ([66], [107]). These lists contain

¹U.S. Bureau of Statistics website

explicit gender terms ² like (women, men, grandmother, grandfather etc.), possessive and pronouns related to genders such as (she, he, his, her) [107], implicit gender stereotype adjectives like (active, cooperative, considerate, emotional etc.) [66] (The full list is provided in the appendix of the paper), and women and men stereotypical jobs such as (nurse, doctor, etc) [107]. Table 3.3 shows statistics about this list.

3.1.3 Spacy tags

We also employed the Spacy API’s named entity recognition (NER) tool ³ to identify proper nouns. Proper nouns (John, Emily, etc) contain information about gender types, so tagging these terms could help mitigate bias in the dataset.

Tagging category	Example	Number of words
Pronoun	He, She	2
Possessive	His, Her	2
Gender-specific terms	Mama, Papa, Lady, Gentleman	443
Gender stereotype adjective	Muscular, Gentle, Active, Kind	85
Stereotypical jobs	Nurse, Doctor	87

Table 3.3 Statistics of gender-related words in the stereotypical words list (L)

Table 3.4 shows statistics on the numbers of tags obtained on the BiasinBios dataset after running Spacy’s NER and tagging elements from our stereotype list L.

Gender Indicators	Example	Tag	Number of tags
Proper nouns (Spacy NER)	John, Emily	Person	467,046
Pronouns	She, He	Person	441,895
Possessives	Her, His	Their	301,701
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 3.4 Descriptive statistics for gender indicator terms using list L and Spacy(NER) in the training dataset

²Explicit gender terms and profession stereotypes

³<https://spacy.io/api>

3.1.4 CoLA dataset

Pretrained language model may be fine-tuned on various datasets ([108], [109]). Our primary methodology is rooted in the BiasinBios dataset [9]. However, for benchmarking our proposed strategy against the state-of-the-art gender-tuning model [1] using the SEAT score, we leverage the CoLA dataset as well (refer to Section 4.2.4). The Corpus of Linguistic Acceptability (CoLA) consists of 9,594 sentences classified into grammatical and ungrammatical labels, sourced from published linguistics literature [110]. The statistical details of CoLA are presented below.

CoLA's dataset statistics			
Title	Train	Validation	Test
Number of texts	8,551	527	516
Number of classes	2	2	2

Table 3.5 Descriptive statistics of the CoLA dataset

3.2 Corpus modification strategies

In this section, we employ two distinct strategies to mitigate associations between gender-related terms and professions. The initial strategy involves the use of the "Withoutgender" dataset, which is essentially identical to the "WithGender" dataset, with the exception that certain specific terms have been removed. The second strategy is centered around the tagging technique. Within this approach, specific terms are substituted and labeled with broader, higher-level concepts.

We present three examples of data points from the BiasinBios dataset. These examples include the initial versions referred to as the "WithGender version," and the scrubbed version termed the "WithoutGender" version, and finally, the version with tags known as the "Tagged version."

WithGender example: *"He has lectured nationally on surgery and diabetic care , and was a speaker at the largest podiatric medical conference in the world . Dr. McLean has taught at two different medical schools , both in the department of surgery . Board certified in wound care , McLean utilizes a revolutionary treatment for wound healing , which is also used to treat neuropathy , the nerve problem developed by many diabetics. !"*

WithoutGender example: *"_ has lectured nationally on surgery and diabetic care, and was a speaker at the largest podiatric medical conference in the world. Dr. _ has taught at two different medical schools, both in the department of surgery. Board certified in wound care, _ utilizes a revolutionary treatment for wound healing, which is also used to treat neuropathy, the nerve problem developed by many diabetics."*

Tagged example: *"PERSON has lectured nationally on surgery and diabetic care , and was a speaker at the largest podiatric medical conference in the world . Dr. PERSON has taught at two different medical schools , both in the department of surgery . Board certified in wound care , PERSON utilizes a revolutionary treatment for wound healing , which is also used to treat neuropathy , the nerve problem developed by many diabetics."*

3.2.1 Scrubbing strategy for gender bias on BiasinBios dataset

The scrubbing strategy is the act of removing some terms of the original version of dataset "WithGender" to eliminate the gender indicator terms that will remove stereotypical associations in pretrained language models. By introducing BiasinBios dataset [9], this methodology is also described and the corresponding dataset is generated, which is "WithoutGender" version. In this dataset, all proper nouns, men and women related possessives and pronouns are eliminated from the dataset and replaced by "_".

3.2.2 Tagging strategies for gender bias on BiasinBios dataset

Various tagging strategies for gendered words were implemented to assess their impact on pretrained models.

Proper nouns, pronouns and possessives tagging (a). In our initial approach, Spacy was utilized to identify proper nouns mentioned in the text, such as "John" or "Emily". Proper nouns are tagged as "Person" in the dataset. Alongside this, we tagged other gender indicator words from list L that were originally removed during the baseline model's scrubbing process (WithoutGender version). This list includes pronouns ("she" and "he") and possessives ("his" and "her"). We aim to examine if tagging of these explicit gender indicators would result in fairer models. **The tagged elements in this strategy includes: Proper nouns + Pronouns + Possessives**

Gender-specific terms tagging (TagDebias) (b). Here, our objective is to examine whether tagging all "Gender-specific terms" from list L leads to improved fairness scores.

Because these words are specific to one of the genders (men and women), they make gender association with professions. So, tagging these words as an abstract word ("Person") might help to remove these associations within pretrained models. Building on the success of the "Proper nouns, pronouns, and possessives tagging" strategy, which enhanced fairness scores (as detailed in section 4.2.1), "Gender-specific terms tagging" method adds gender-specific terms to the previous strategy. This experiment builds upon our previous approach and aims to analyze the impact of "Gender-specific term" tagging on the fairness of pretrained language models. This approach, referred to as **"TagDebias"** outperforms alternative tagging strategies significantly in terms of fairness, as discussed in Section 4.2. **The tagged elements in this strategy include: Proper nouns + Pronouns + Possessives+ Gender-specific terms**

Gender stereotype adjective tagging (c). Having observed the positive impact of tagging gender-specific terms on the fairness of pretrained language model (see section 4.2.1), our next step is to add stereotypical adjectives to the previous approaches. These adjectives, such as, "sensitive", "competitive", "hostile" are often associated with specific gender-related behaviors in society. In this experiment, our aim is to investigate whether tagging these adjectives with the label "Adjective" would result in a fairer model. **The tagged elements in this strategy include: Proper nouns + Pronouns + Possessives+ Gender-specific terms+ Gender stereotype adjective**

Stereotypical jobs tagging (d). After observing that adding gender stereotype adjective will decrease the fairness of model (see section 4.2.1), we added stereotypical jobs tagging to the first and second strategies. Building upon the "Gender-specific terms tagging" approach, we decided to additionally tag stereotypical professions related to genders within the text. The aim of this extension was to investigate whether tagging stereotypical professions could result in a fairer model. **The tagged elements in this strategy include: Proper nouns + Pronouns + Possessives+ Gender-specific terms+ Stereotypical jobs**

Table 3.6 shows statistics about the total number of terms which are removed (in the scrubbed version) or tagged in each strategy on the training dataset. Our models are trained on these modified datasets. The first tagging strategy tags the same set of terms that are removed from scrubbed version. So, their corresponding numbers are the same. This work enable us to compare the scrubbing version with our tag strategy in terms of model performance and fairness. Please note that the combination of tagging is selected based on the fairness scores obtained in section 4.2.1.

Corpus modification Strategies	Total number of terms
Scrubbed version	1,210,642
Proper nouns, pronouns and possessives (a)	1,210,642
Gender-specific terms (a+b)	1,314,081
Gender stereotype adjectives (a+b+c)	1,337,636
Stereotypical jobs (a+b+d)	1,361,718

Table 3.6 Statistics of different corpus modification strategies on the training dataset, the numbers indicate total number of tags in the tagging strategies and total number of removed terms in the scrubbing strategy in the training corpus

3.3 Dataset augmentation

Data augmentation serves as a valuable technique to enhance model performance by exposing the model to a broader range of data instances. In this research, we have employed two distinct techniques.

3.3.1 Data augmentation by merging the initial and modified text

This data augmentation technique involves merging the original dataset text, such as the "WithGender" biographies in the case of the BiasinBios dataset, with the tagged or scrubbed version of the same text. These two versions are joined by a comma to create augmented examples. Here are examples of data augmentation strategies:

Initial text: *"She has a Ph.D. from the University of Maryland"*

Augmented scrubbed version : *"She has a Ph.D. from the University of Maryland, _ has a Ph.D. from the University of Maryland."*

Augmented tagged Version : *"She has a Ph.D. from the University of Maryland, Person has a Ph.D. from the University of Maryland."*

The intention behind this augmentation approach is to acquaint the language model with both the unaltered and modified text forms. By presenting instances where "she" and "he" both refer to the generalized label "Person," or removed token "_", the hypothesis is that the language model may acquire a more holistic understanding of these gender-neutral correlations. This data augmentation doubles sentences' length while the number of instances remains the same.

3.3.2 Data augmentation by adding modified instances

Another approach to data augmentation involves merging two datasets: the initial version and the modified text (tagged or scrubbed) version, which results in duplicate instances. This strategy increases the exposure of the model to a broader array of data examples, subsequently contributing to enhanced model performance.

In this data augmentation technique, both the sentence "She has a Ph.D. from the University of Maryland" and its transformed counterpart "Person has a Ph.D. from the University of Maryland" are classified under the same profession label, such as "professor." Consequently, the hypothesis is that the language model is able to recognize the absence of distinction between "Person," "She," and "He."

3.4 Models

We began by preparing our training data for the BiasinBios dataset, making scrubbed versions and different tagged versions. Our primary task was multi-class classification, involving 28 job categories, with biographies as the input text. This task required categorizing each text into one of the job categories. To accomplish this, we chose to employ sequence classification models from the BERT family.

We initiated our experiments by fine-tuning three pretrained language models: BERT [18], ALBERT [20], and RoBERTa [19]. We started with the "WithGender" and "WithoutGender" initial versions 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.

The specific models we used were as follows:

- "BERT for sequence classification" with the base model "bert-base-uncased" from the Hugging Face library.
- "RoBERTa for sequence classification" with the base model "roberta-base" from the Hugging Face library.
- "ALBERT for sequence classification" with the base model "albert-base-v2" from the Hugging Face library.

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.

3.5 Application of debiased models on fair ranking

After finding the fairest model, we aim to examine the use of this model in a fair ranking task such as ranking biographies or CVs for a given profession. In our methodology, we treat each biography from the BiasinBios dataset as a CV, and each job category class serves as a job description. To establish rankings based on the similarity between each biography and job category, we require a scoring mechanism. We have opted to use 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. Initially, 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% [103].

Subsequently, we employed three distinct models for ranking purposes:

- Initial Model: This model was fine-tuned on the original training dataset which is "WithGender" version.
- Scrubbed Version: This model was trained on a dataset that had gender-related information removed which is "Scrubbed" version.
- Fairest tagged Model: We utilized our most effective tagged model, which is the "gender-specific" term model (see section 4.2).

Utilizing these models, we ordered the biographies based on their confidence scores, which were used as the ranking criterion.

Once the biographies were ranked, we selected the top-k biographies from each job category and assessed the gender distribution within this subset. When analyzing gender distribution in this list, we took into account societal stereotypes associated with jobs which is also seen in the training dataset held by men and women. For instance, if we observed a lower percentage of men in the top 20 rankings for a job typically associated with men, compared to the distribution in the initial model, it would indicate that our model exhibits reduced bias, even if the majority of biographies related to that job in the training dataset are about men.

Finally, we performed a sensitivity analysis experiment, examining various top-k values to ascertain whether our proposed tagged model consistently delivers fairer outcomes across all these top-k values. This section commences with an assessment of the candidates’ utility in top-k positions, followed by the application of fairness metrics to assess gender disparity across different top-k values.

3.6 Metrics

In this segment, we indicate the metrics employed in our research. These metrics fall into three main categories: Fairness metrics designed for classification tasks, two state-of-the-art benchmarks, and metrics relevant for ranking tasks.

3.6.1 Fairness metrics in classification models

In the context of fairness in classification tasks, various metrics are utilized to identify bias. These metrics, including the F1 score, AUC score, False Positive Rate, and True Positive Rate, are commonly employed to assess the performance of classification models. In this work [9], the authors examined the difference in True Positive Rates (recall) between two genders within each profession as a measure of fairness. They proposed that reducing this difference by eliminating gender-specific terms from the dataset indicates an increase in fairness, implying that the model behaves similarly for both genders. However, this approach has certain limitations.

One drawback is that by removing gender-specific words or applying other bias mitigation techniques, the model may appear fair for a specific gender, so the True Positive Rate for that gender increases while this score is the same for another gender. As a result, the difference between the True Positive Rates of two genders increase, leading to a fairer model. This implies that the model can now predict the labels of biographies for a specific gender with greater accuracy compared to its previous performance. This phenomenon often occurs in professions that are considered anti-stereotypical for a particular gender, such as nurse for men or doctor for women. For example, consider the nurse profession, which is a woman stereotypical job and a man anti-stereotypical job. When we debias the pre-trained language models (PLMs), the models tend to correctly classify biographies of men as nurses more than before, while the performance for women’s biographies may remain unchanged. As a result, the difference in the true positive rate between men and women increases, which, based on [9], would not indicate a decrease in bias. However, this indicates that we have a fairer model.

To address these concerns, we propose a modification to ΔTPR . Utilizing data from the

U.S. Bureau of Statistics, we observed that men and women are represented in different proportions within each profession in 2022. As indicated previously, a higher proportion of men in one job determines that it is a stereotype for men and if the women proportion is higher, it is a stereotype for women. We have listed all stereotypical jobs related to men and women in the dataset section 3.1.1. Thus we calculate the True positive rate (TPR) and False positive rate (FPR) difference scores based on stereotypical and anti-stereotypical jobs as defined in this paragraph.

TPR and FPR scores based on stereotype and anti-stereotype jobs. We calculate the average of TPR scores for all anti-stereotypical jobs (see section 3.1.1) for the "With-Gender" model as a baseline model. This score is also calculated for "WithoutGender" and tagged models. If we get a better average TPR score for the WithoutGender or tagged models (debiased models), we improved the fairness of the model. This average is over 28 anti-stereotypical jobs. We prioritize $TPR_{anti-stereotype}$ as it is crucial for us to assess the effectiveness of a debiased model in accurately predicting true labels, particularly in scenarios involving anti-stereotype associations. For instance, we are interested in evaluating how well the model can predict the label "nurse" for biographies related to men, highlighting the importance of mitigating stereotypical biases in predictions.

We also used the False positive rate score for stereotypical jobs which means that a fairer model must assign anti-stereotyped biographies to stereotype classes less than before. For example, our model must assign women biographies associated with the "doctor" gold truth label to the label of "nurse" or any possible women stereotypes less than before. Therefore, by averaging FPR over all of stereotypical jobs for women and men (28 jobs), we can calculate this score as well. Equation 3.1, calculates the mean of True positive rate for anti-stereotypical jobs, n denotes the number of classes which is 28. Equation 3.2 shows the mean of False positive rate for stereotypical jobs.

$$TPR_{anti-stereotype} = \frac{1}{n} \sum_{i=1}^n TPR_{anti-stereotype}^i \quad (3.1)$$

$$FPR_{stereotype} = \frac{1}{n} \sum_{i=1}^n FPR_{stereotype}^i \quad (3.2)$$

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 then divided into two subgroups men and women. A subgroup is designated by t .

FPR_t / FNR_t determine the False positive/negative rates of the subgroup biographies t . 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 [63]. The formulas are described below:

$$FPED = \sum_{t \in T} |FPR - FPR_t| \quad (3.3)$$

$$FNED = \sum_{t \in T} |FNR - FNR_t| \quad (3.4)$$

Difference in model performances between minority groups. Performance metrics such as F1 score, accuracy, precision, and recall are also applied when evaluating fairness. A fair model should deliver consistent performance across all minority groups. For example, the F1 scores for classifying sentences related to men and women should be similar. When the model’s accuracy in predicting sentences about men is noticeably higher than for women, it suggests the presence of social bias within the pretrained language model. In the following formulas, $t1$ and $t2$ are two minority groups such as men and women [111].

$$\Delta Acc = |Accuracy_{(t1)} - Accuracy_{(t2)}| \quad (3.5)$$

$$\Delta P = |Precision_{(t1)} - Precision_{(t2)}| \quad (3.6)$$

$$\Delta R = |Recall_{(t1)} - Recall_{(t2)}| \quad (3.7)$$

$$\Delta F1 = |F1_{(t1)} - F1_{(t2)}| \quad (3.8)$$

3.6.2 Fairness metrics and benchmarks

In this section, we introduce two benchmark fairness evaluators along with their corresponding test sets.

CrowSpairs

We also employed test sets and metrics that are specifically used for evaluating social bias in Pretrained Language Models (PLMs) [73] 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. As a reminder (see section 2.3.2), 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%.

SEAT

The Sentence Encoder Association Test (SEAT) serves as another benchmark dataset and metric [40] to evaluate the sentence representation of different models, baselines (WithGender, WithoutGender) and our proposed TagDebias model. Following the debiasing process, we expect our debiased model to exhibit consistent distances between the representation of a neutral sentence, such as "This is a doctor," and the representations of sentences like "This is John" and "This is Emily." These sentence templates are part of the SEAT benchmark.

3.6.3 Fairness metrics in ranking

In this study, we explore ranking scenarios that incorporate binary sensitive attributes, specifically focusing on G1 and G2 as the protected and favored attributes, respectively. As a reminder, 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 attributes refer to stereotypical jobs. For instance, G1 might represent female in software engineering roles, while G2 could denote female in nurse positions. In this work, we utilize the position-based ranking rND that is proposed by [101]. Essentially, this metric aims to ensure that the gender distribution among top-ranked candidates closely mirrors the gender distribution observed in the input data.

Normalized discounted difference (rND). As we have mentioned in section 2.5.1, the rND metric quantifies the disparity between the proportion of protected attributes (denoted as G_1) within the top-k rankings. In our ranking methodology, the calculation begins with k starting at 5. The metric compares the representation of protected attributes in the top-k rankings to the overall distribution of these attributes in the entire input dataset used for generating the rankings. In this study, the protected attribute includes women in men stereotypical jobs and men in women stereotypical jobs. The metric reaches its maximum value of 1, indicating the highest level of bias, while a value of zero signifies a fair model.

3.6.4 Utility metrics

Utility metrics are initially introduced in Section 3.6.4. These metrics have been employed to guarantee that when re-ranking candidates with various models, the utility of candidates in the top rankings remains consistent and does not decrease. In this study, we used this metric to calculate the average total of confidence scores of top-k candidates and named it as "**Average total Utility**". This metric ensures that top-k candidates are qualified based on the relevance of biographies to job category. The other utility metric that we have used is "**Average protected group Utility**". This metric deals with minority groups and it is the average of the sum over all confidence scores of minority groups in top rankings. For example, in our study, the minority group of stereotypical jobs related to men is women and the minority group of stereotypical jobs related to women is men. Overall, this score ensures that not only are minority groups well-represented among the top-k candidates, but they also hold high ranking scores, signifying the qualification of minority group candidates.

CHAPTER 4 EXPERIMENTS AND RESULTS

This section encompasses the classification experiments and their corresponding results, alongside an examination of fairness results using designated fairness metrics and benchmarks. The final segment of this chapter is dedicated to a thorough exploration of experiments related to the ranking task.

4.1 Classification results

The classification results are presented in two distinct sections.

4.1.1 Fine-tuning BERT family models on the two initial versions of BiasinBios dataset

We conducted fine-tuning experiments on three pretrained language models: BERT, ALBERT, and RoBERTa (For models' details see section 3.4), using the initial versions of the BiasinBios dataset. The goal was to determine the best-performing model of these two versions.

Table 4.1 presents the classification report for the three language models on both versions of the dataset (WithGender and WithoutGender). In the WithGender version, ALBERT exhibited the worst performance in terms of precision, recall, accuracy, and F1 scores when compared to BERT and RoBERTa. RoBERTa outperformed BERT in terms of recall and F1 scores, while BERT have slightly better precision. However, the difference in precision between BERT and RoBERTa is negligible. In the WithoutGender version, ALBERT again showed the worst performance, while RoBERTa achieved the highest scores in recall and F1. RoBERTa maintained its accuracy after the scrubbing process, whereas BERT showed some variations in accuracy scores. Overall, both BERT and RoBERTa performed better than ALBERT, with RoBERTa slightly outperforming BERT. Consequently, we will conduct our following experiments using the RoBERTa model for tagging tasks.

4.1.2 Fine-tuning RoBERTa model on tagged datasets

In this section, we will perform a detailed analysis of the classification results derived from our tagging strategies, focusing on two distinct test sets: the initial test set (WithGender) and the modified (tagged) test set. Our primary aim is to evaluate how the model's perfor-

Model	P%	R%	F1%	Acc%
WithGender-dataset				
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
WithoutGender-dataset				
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 4.1 Classification result 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

mance evolves as it undergoes fine-tuning on tagged versions and is subsequently utilized in downstream tasks.

In table 4.2, we compare the performance of each version in terms of precision (P%), recall (R%), F1 score (F1%), and accuracy (Acc%). The WithGender version represents the original test set with gender information included. The WithoutGender version refers to the dataset where gender information has been removed. Finally, the tagged version corresponds to the tagging strategies’ methods.

Proper noun, possessives and pronoun This model is a RoBERTa which is fine-tuned on the Proper noun, possessives and pronoun tagged dataset. We will refer to this model in the table as "Possessive and pro". Among the different tagging strategies, "Possessive and pronoun" model achieved the second-highest precision score when compared to the baseline model. This specific tagging approach also demonstrated the best F1 score and accuracy among all other models. Furthermore, it outperformed the "WithoutGender version" in terms of recall.

Gender-specific terms Initially, we conducted fine-tuning on the RoBERTa model using the Gender-specific tagged strategy dataset. When evaluated on the initial test set, this strategy achieves the second-highest F1 and recall scores among the different tagging approaches and the "WithoutGender" version. Moreover, this tagging strategy demonstrates higher accuracy than the "WithoutGender" version on the initial test set.

Gender stereotype adjective This model is referred to as "Gender-stereo-ADJ" in the table. This strategy yields the highest recall and F1 score for RoBERTa compared to other

tagging methods and the "WithoutGender" version. Furthermore, this strategy achieves the second-highest accuracy score among all the models evaluated.

Stereotypical jobs The stereotypical jobs are specifically designed to evaluate the impact of tagging stereotypical jobs in pretrained models. However, this type of tagging resulted in a decrease in all evaluation scores. The reason behind this decline is that certain professional terms in the text contribute to the pretrained models' ability to accurately predict the corresponding job class. By tagging these words, it can adversely affect the model's performance in classifying biographies into the appropriate job categories.

As observed in table 4.2, it is evident that the process of tagging and removing sensitive biased terms has a approximately neutral impact on the models' performance metrics, including precision, recall, F1 score, and accuracy.

Model	P%	R%	F1%	Acc%
On the initial test-set				
WithGender(base)	81.27	77.95	79.16	84.97
WithoutGender	81.00	77.39	78.79	84.65
Possessive and pro	<u>80.79</u>	77.86	78.86	84.82
Gender-specific-term	80.46	<u>77.89</u>	<u>78.71</u>	84.75
Gender-stereo-ADJ	80.68	78.04	78.86	<u>84.79</u>
Stereotypical jobs	78.54	76.64	76.54	83.07
On the tagged test-set				
WithGender(base)	81.27	77.95	79.16	84.97
WithoutGender	81.22	77.42	78.79	84.85
Possessive and pro	80.61	77.07	78.32	84.55
Gender-specific-term	81.00	76.96	78.36	84.54
Gender-stereo-ADJ	80.64	77.04	78.25	84.43
Stereotypical jobs	80.18	76.68	77.89	84.27

Table 4.2 Classification result of RoBERTa model on initial and modified test sets

4.2 Fairness results of models

Having established that tagging or scrubbing does not have a high impact on classification performance, this section is dedicated to assessing the fairness of various versions of RoBERTa models.

4.2.1 Fairness results of initial and different Tagged-RoBERTa models

In this section, we evaluate the fairness of the tagged models using the aforementioned fairness metrics. As we can see in table 4.8, the scrubbed version and all tagging versions obtain a lower FNED score compared to the base model. Among these models, the "Gender-specific-term" model has the lowest FNED score and the WithoutGender (scrubbed) version has the second lowest FNED score compared to other models. Considering FPED score, "Stereotypical jobs" model has the lowest FPED score and "Possessive and pro" model has the second lowest FPED score. Tagged (except Stereotypical jobs) and scrubbed models improved the true positive rate of anti-stereotypical jobs compared to the base model, which means that these models could better categorize anti-stereotypical jobs than initial models. The decrease in FPR metric shows that the scrubbed and tagged models do not incorrectly assign anti-stereotypical biographies to stereotypical jobs.

Table 4.4 also shows the fairness scores of different models, these fairness scores are based on the classification score between two target groups (women and men). Based on these scores, the fairest model is "Gender-specific term" and the second best model is the "WithoutGender" model. We can also see that "Possessive and pro" model has better fairness scores than the initial model.

Overall, based on these metrics, "Gender-specific term", "Possessive and pro" and "WithoutGender" models obtain the best fairness scores compared to other models.

Model	↓FNED%	↓FPED%	↑TPR _{antistereo} %	↓FPR _{stereo} %
On initial test-set				
WithGender(base)	4.81	0.1076	72.14	0.729
WithoutGender	<u>3.64</u>	0.1094	73.21	0.723
Possessive and pro	3.97	<u>0.1050</u>	73.29	0.715
Gender-specific terms	3.56	0.1074	73.89	0.715
Gender-stereo-ADJ	4.04	0.1119	<u>73.46</u>	<u>0.722</u>
Stereotypical jobs	4.46	0.0072	71.71	0.834

Table 4.3 Fairness results of different versions of RoBERTa model (WithGender, WithoutGender and tagged) on the initial test set

4.2.2 Fairness results of augmented models

In this section, we use the augmented versions of the BiasinBios dataset and train the scrubbed model (WithoutGender) and one of the best performing tagged model which is

Model	$\downarrow\Delta\text{Acc}\%$	$\downarrow\Delta\text{P}\%$	$\downarrow\Delta\text{R}\%$	$\downarrow\Delta\text{F1}\%$
On initial test-set				
WithGender(base)	2.64	3.29	3.15	3.14
WithoutGender	2.64	3.47	2.50	2.68
Possessive and pro	<u>2.56</u>	3.25	2.97	2.79
Gender-specific term	2.58	2.90	<u>2.64</u>	<u>2.71</u>
Gender-stereo-ADJ	2.69	3.79	3.04	3.43
Stereotypical jobs	0.13	<u>3.10</u>	3.46	3.14

Table 4.4 Fairness results of different RoBERTa models' version on the initial test set (These fairness scores are based on the model performance)

the Gender-specific-term model (see section 3.2.2). "Aug1" combines each initial sentence with its tagged or scrubbed version separated by a comma. The augmentation strategy referred to as "Aug2" in the table combines these modified datasets with the original one as separate data points. Subsequently, we fine-tuned a RoBERTa model on these combined datasets and compared its fairness performance against the initial model. The evaluation is performed on the original test set.

As we can see in table 4.5, the "Aug2" augmented version of models, using the Without-gender and Gender specific term settings, have better FNED scores compared to the WithGender model. Aug1 combined with these models is not effective to reduce bias. WithoutGender-Aug2 and Aug1 also have better FPED score compared to its initial model and WithGender model. We can also see fairness improvement in the TPR and FPR scores for both WithoutGender-Aug2 and gender-specific-terms-Aug2 in comparison to its initial models and WithGender model. Overall, we can say that Aug2 augmentation helped the scrubbed version to be fairer than its initial model, but in the case of the tagged versions, has worsened the FNED and FPED scores of the models. The Aug1 augmentation strategy decreased the fairness scores for both tagged and scrubbed models.

Table 4.6 measures other fairness metrics that correspond to models' performance. In terms of difference in accuracy, all models except "Gender-specific-term-Aug2" outperform the WithGender model. Aug2 data augmentation helps the scrubbed version to get a better score in accuracy and recall compared to the WithGender and its initial models, while it does not happen for the tagged version. While we could see negligible improvement in the accuracy of WithoutGender-Aug1 compared to its initial model, Aug1 data augmentation is not very effective to reduce bias in PLMs.

Based on the data presented in Tables 4.5 and 4.6, it becomes evident that the augmentation

Model	↓FNED%	↓FPED%	↑TPR _{antistereo} %	↓FPR _{stereo} %
On initial test-set				
WithGender(base)	4.81	0.1076	72.14	0.729
WithoutGender	3.64	0.1094	73.21	0.723
WithoutGender-Aug1	5.36	<u>0.1040</u>	71.43	0.730
WithoutGender-Aug2	3.43	0.1038	75.07	0.687
Gender-specific-term	<u>3.56</u>	0.1074	73.89	0.715
Gender-specific-term-Aug1	5.92	0.1050	71.04	0.733
Gender-specific-term-Aug2	4.50	0.1101	<u>73.93</u>	<u>0.697</u>

Table 4.5 Fairness results on various augmented versions of RoBERTa models (Initial, WithoutGender, and tags) on the initial test set. Aug1 refers to data augmentation with comma, and Aug2 refers to data augmentation by combining the modified version dataset and the initial version

Model	↓ΔAcc%	↓ΔP%	↓ΔR%	↓ΔF1%
On initial test-set				
WithGender(base)	2.64	3.29	3.15	3.14
WithoutGender	2.64	3.47	<u>2.50</u>	<u>2.68</u>
WithoutGender-Aug1	2.56	3.50	3.36	3.64
Without-gender-Aug2	2.61	4.21	2.29	3.32
Gender-specific-term	<u>2.58</u>	2.90	2.64	2.71
Gender-specific-term-Aug1	2.63	3.10	2.92	3.00
Gender-specific-term-Aug2	3.74	2.90	3.42	2.46

Table 4.6 Fairness results of different augmented versions of RoBERTa models (Initial, WithoutGender, and tagged) on the initial test set

(Aug2) technique contributes to enhancing the fairness performance of the scrubbed version (WithoutGender) model when compared to its initial model. However, this augmentation approach doesn't yield substantial improvements for the tagged version.

4.2.3 Fairness results of BERT-family models on CrowSpairs benchmark

In this section, our objective is to assess the effectiveness of our most successful tagging strategy, namely "gender-specific-term" (refer to Section 3.2.2), across various pretrained language models like BERT, RoBERTa, and ALBERT, with a focus on CrowSpairs score evaluation.

Table 4.7 displays the CrowSpairs scores of the TagDebias models. Notably, the TagDebias model (For a definition, see section 3.2.2) obtains the lowest CrowSpairs (CPS) score among the models examined in the BERT and RoBERTa family, 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.

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 4.7 Model evaluation on the CrowSpairs dataset (Gender bias). The ideal score is 50%, the score nearer to 50% is less biased model. The numbers are in percentage.

While the observed trend is evident in the BERT model, the CPS score does not surpass the performance of the scrubbed model in ALBERT. Upon evaluating the stereotype (SS) and anti-stereotype (anti-SS) scores in the ALBERT model, it becomes apparent that the TagDebias model tends to assign higher probabilities to more anti-stereotypical examples. This pattern is also consistent in the RoBERTa TagDebias model; However, the CPS score outperformed other RoBERTa models.

4.2.4 Comparing Tag-Debias model with state-of-the-art models on the SEAT benchmark

Table 4.8 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-size)(lower is fairer). 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 the gender-tuning approach [1]. When our TagDebias model is fine-tuned on the CoLA dataset, it improved the SEAT score. We have presented RoBERTa-CDA result from [22].

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-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
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	<u>0.294</u>

Table 4.8 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 the models fine-tuned on CoLA.

4.3 Ranking results of different models on biographies

In this section, we aim to examine the effect of the scrubbed and tagged versions of our models on the ranking of biographies in the BiasinBios dataset. In this experiment, we have

considered each biography as a curriculum vitae (CV) and each job category class as a job descriptions' name. We have used the confidence score of each biography as a metric to rank them in each job categories.

4.3.1 Classification result on the sub-test set

In order to have a balanced number of each gender in each job category, we randomly selected the same number of each gender in each job category, with the aim to neutralize the effect of gender distribution on the ranking. Below, we can see classification result of different RoBERTa models on the initial and randomly selected test set. Table 4.9, shows the classification results of three models, WithGender, scrubbed and gender-specific-term on the initial and randomly selected test set. The randomly selected test set's size is 66,422 while the initial test set's size is 98,344 biographies.

Model	P%	R%	F1%	Acc%
Initial test set				
WithGender	81.27	77.95	79.16	84.97
WithoutGender	81.00	77.39	78.57	84.65
Gender-specific-term	80.46	77.89	78.71	84.75
Randomly selected test set				
WithGender	73.11	74.89	73.18	84.99
WithoutGender	73.11	75.07	73.11	85.00
Gender-specific-term	72.64	76.11	73.36	84.82

Table 4.9 Classification result of different RoBERTa models on initial and randomly selected test set

Table 4.9 illustrates that the models' performance, as measured by recall, precision, and F1 score, experienced a decline when assessed on the new test set in contrast to the initial test set. Notably, all three models exhibited similar performance on the randomly selected test set.

4.3.2 Ranking of biographies with initial and Tagged-RoBERTa

In this section, we have employed the RoBERTa model which is fine-tuned on the initial, WithoutGender and fairest tagged model, "Gender-specific-term" on the sub-testset. We have ranked all biographies in each job-category based on the confidence score that the model assigns to each biography. After ranking, we have considered the distribution (proportion)

of men biographies in the men stereotypical jobs and women biographies in the women stereotypical jobs in top 5, 10, 15, 20, 25 and 30 rankings. A fair model will reflect the same proportion of men and women in the top rankings as initially present in the test set (50%). In this section, we provide the top 20 ranking analysis. Other top-k rankings' tables are provided in the appendix.

Profession	Stereotype in training	Stereotype in US data	Initial model-men %	Scrubbed-men %	Tagged-men %
Accountant	man	woman	30%	35%	45%
Architect	man	man	55%	65%	45%
Attorney	man	man	90%	100%	75%
Chiropractor	man	man	65%	55%	45%
Comedian	man	man	55%	65%	55%
Composer	man	man	85%	75%	65%
Dentist	man	man	55%	40%	30%
Dj	man	man	75%	75%	80%
Filmmaker	man	man	50%	40%	25%
Journalist	man -non biased	man	0%	20%	50%
Painter	man	man	90%	75%	50%
Pastor	man	woman	95%	95%	55%
Personal-trainer	man	man	75%	75%	65%
Photographer	man	man	35%	65%	30%
Physician	man	man	0%	0%	0%
Poet	man	man	75%	70%	50%
Professor	man	man	60%	40%	55%
Rapper	man	man	65%	60%	55%
Software-engineer	man	man	25%	5%	25%
Surgeon	man	man	80%	90%	60%

Figure 4.1 Men distribution in top 20 ranking of biographies with three models, initial, scrubbed (WithoutGender), and Gender-specific-term tagged version, green cells have better score compared to initial model and the orange cell has worse score compared to initial model, note that we have two contradictions in stereotypes (as highlighted in grey), in this ranking we have considered the stereotype based on the training dataset

As can be seen in Figure 4.1, we have listed all professions which are stereotypes for men based on the training dataset (i.e. the majority of biographies related to these jobs are about men). We have considered 20 top candidates for each job category and calculated the men distribution along this 20 biographies. We see that by using WithoutGender and tagged models, the distribution of men in the top 20 candidates decreased compared to the initial model ranking, which means that the model's tendency to promote men biographies at top rankings decreased using the tagged and WithoutGender models. For example, professor is a stereotypical job related to men in the top 20 ranking, the men distribution generated by initial model is 60%, while this proportion decreased for scrubbed and tagged versions to 40% and 55%, respectively. By taking a detailed look, the scrubbed model decreased or kept the same distribution score for 13 out of 20 stereotypical jobs related to men, compared to 17

out of 20 for our proposed tagged model. Therefore, the Gender specific-term tagged model outperformed WithoutGender and also the initial model in providing a fairer ranking.

Profession	Stereotype in training	Stereotype in US data	Initial model-women %	Scrubbed-women %	Tagged-women %
Dietitian	woman	woman	90%	85%	85%
Interior-designer	woman	woman	45%	60%	40%
Model	woman	woman	95%	95%	80%
Nurse	woman	woman	0%	10%	30%
Paralegal	woman	woman	75%	75%	60%
Psychologist	woman	woman	90%	95%	30%
Teacher	woman	woman	80%	20%	75%
Yoga-teacher	woman	woman	65%	60%	75%

Figure 4.2 Women distribution in top 20 ranking of biographies with three models, initial, scrubbed, and Gender-stereo-term tagged version, green cells have better scores compared to initial model and the orange cells have worse scores compared to initial model. Note that Accountant and Pastor are stereotypes in U.S. data, but in the training dataset the majority of men have these occupations. In this ranking, we considered the stereotype based on the training dataset.

In the case of stereotypical jobs for women, shown in figure 4.2, the scrubbed model decreased or stabilized the distribution score related to women for 5 out of 8 jobs and this amount for the tagged model is 6 out of 8 jobs. Thus, the proposed tagged version outperformed the scrubbed and initial models.

In the following, we will show the statistical distribution of men and women in the top 20 rankings in different job categories. Figure 4.3 illustrates the distribution of men’s biographies in the stereotypical jobs related to men. In the initial model, the mean of men distribution in all stereotypical jobs related to men is 58%, indicating the average percentage of men in the top candidates for these jobs. However, this value decreases to 57.25% for the WithoutGender version and further drops to 49% for the tagged version. This implies that both the WithoutGender and tagged versions exhibit lower proportions of men in the top 20 candidates compared to the initial model.

The observed trend extends beyond men-related stereotypical jobs to encompass women-related stereotypical jobs as well as evidenced in Figure 4.4. In the initial model, the mean percentage for women candidates is 67.5%, Conversely, for the scrubbed version, the mean drops to 62.5%. The tagged model exhibits the fairest behavior with 59.28% women candidates in top rankings. This experiment highlights the effectiveness of our proposed tagged model in balancing the representation of men and women biographies within the top 20 candidates. In contrast, the initial model tends to favor increased representation of women in

women stereotypical jobs among the top 20 biographies.

This experiment was conducted for various top-k values (for detailed results, please refer to the appendix), and it is evident that, in most cases, the scrubbed version reduced the representation of men in stereotypical male roles and women in stereotypical female roles (top 5, top 10, top 15, top 20). Notably, our proposed tagged model consistently enhanced this balance for about 8% across all top-k values, showcasing its superior performance compared to the scrubbed version.

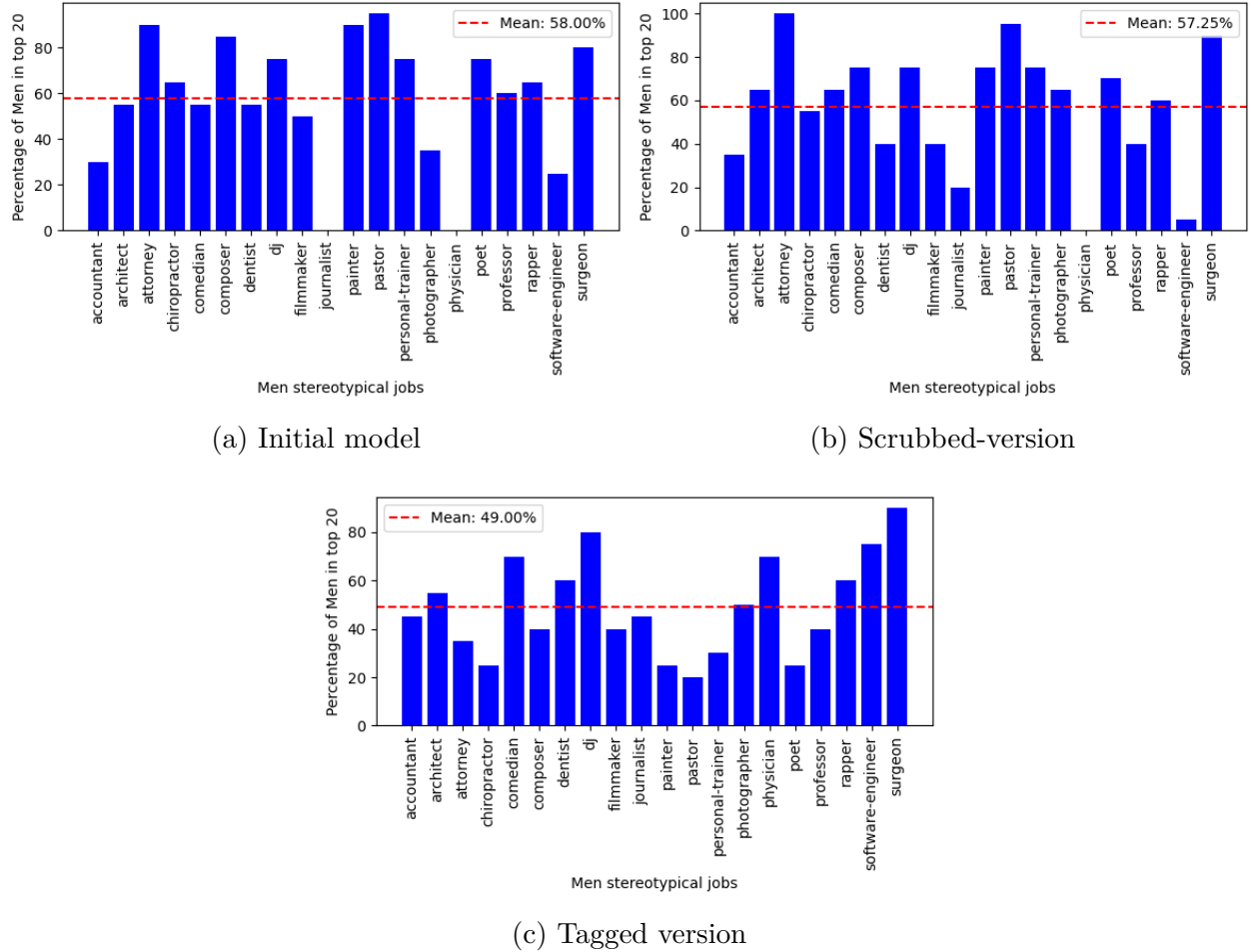


Figure 4.3 Men distribution in top 20 biographies created by three models in stereotypical jobs related to men

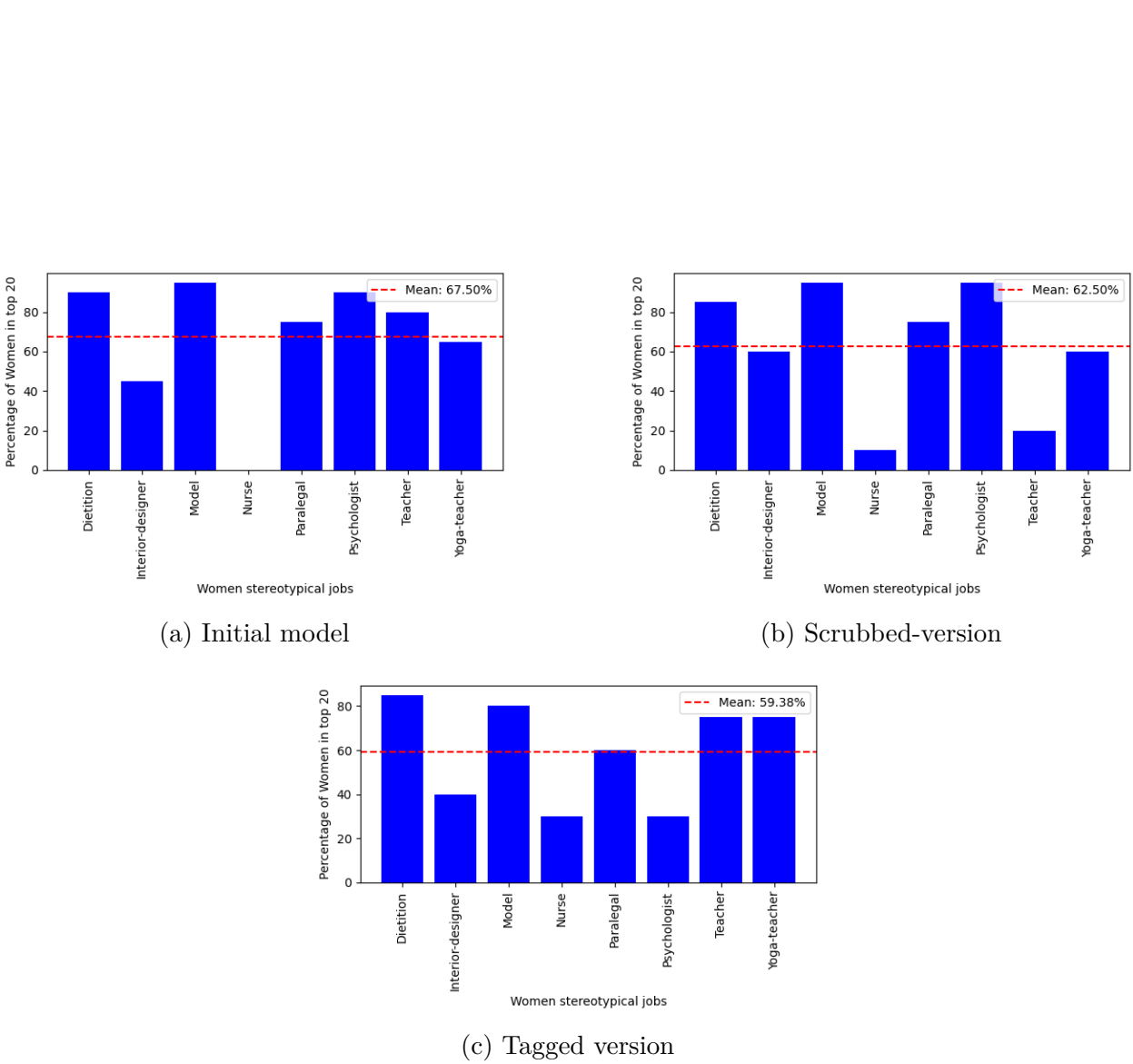


Figure 4.4 Women distribution in top 20 biographies created by three models in stereotypical jobs related to women

4.3.3 Sensitivity analysis of different top-k values

In this section, we will investigate the impact of various top-k values on rankings' fairness across the three models mentioned earlier. Initially, we will employ utility function metrics to assess the candidate quality within top-k rankings, utilizing their confidence scores. Subsequently, fairness metrics will be applied to measure the bias inherent in each model and within each top-k ranking, providing an evaluation of gender disparity across different top-k rankings.

Utility of different ranking methods. A utility function shows how much the ranking is based on the candidates' qualification and appropriateness. In this study, we use two utility functions which are described in the metrics' section. Here, we will demonstrate that although there has been an increase in the representation of a wider range of minority groups among the top-ranked candidates, there is no significant decrease in the relevance of the candidates' biographies to the job category.

Model	Top5	Top10	Top15	Top20	Top25	Top30
Average total Utility (U1)						
Initial model	0.9688	0.9679	0.9672	0.9666	0.9659	0.9652
Scrubbed-version	0.9686	0.9678	0.9671	0.9665	0.9657	0.9647
Tagged-model	0.9665	0.9657	0.9649	0.9641	0.9633	0.9624
Average protected group Utility (U2)						
Initial model	0.9763	0.9751	0.9701	0.9677	0.9666	0.9673
Scrubbed-version	0.9719	0.9677	0.9673	0.9661	0.9651	0.9648
Tagged-model	0.9636	0.9662	0.9670	0.9658	0.9651	0.9643

Table 4.10 Utility function values for three models (WithGender, Scrubbed and Tagged). These scores are between 0 and 1. Higher values indicate the relevance of biographies to the job category.

As shown in table 4.10, U1 is the "Average total utility" function and is consistent among all models in all top-k values. This indicates the relevance of all top-k biographies to the job category 3.6.4. U2 which is "Average protected group utility" is a function that calculates the average of confidence scores in top-k protected groups. These scores ensure that the rankings not only have diversity of genders in the top-k rankings but also that these candidates are qualified to be in the top-k ranks. As shown in table 4.10, we lose only a small amount of utility score with the tagged and WithoutGender ranking models compared to the initial model due to model performance.

In the following section, we assess the degree of fairness among various models when ranking the top-k candidates. This analysis aims to determine whether the tagged model exhibits fairness across different top-k values.

Quantifying fairness in different models and rankings. We first compute the mean of each gender distribution in their corresponding stereotypical jobs in different top-k rankings as illustrated in Table 4.10. 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 model consistently enhanced this balance across all top-k values, showcasing its superior performance compared to the scrubbed version.

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 4.11 Fairness result of different rankings generated by RoBERTa models based on the Mean of men and women proportions in the different top k values. The best value is closer to 50%.

Following the ranking of various top-k values, our objective is to employ a position-based fairness metric to measure the bias present in each of the top-k rankings generated by each model. As we can see in Table 4.12, the tagged model outperformed the initial model in the overall score by about 14% and by about 10% compared to the scrubbed version. This result is consistent among all top-k rankings.

Model	Top5	Top10	Top15	Top20	Top25	Top30	Total
Normalized discounted difference (rND)							
Initial model	0.2714	0.3143	0.2500	0.1571	0.1914	0.1907	0.1606
Scrubbed-version	0.2286	0.1571	0.0664	0.1250	0.2229	0.1757	0.1127
Tagged-model	0.0000	0.0500	0.0100	0.0250	0.0457	0.0371	0.0167

Table 4.12 Fairness result of different rankings based on the position-based ranking metric, rND. A lower score signifies superior performance in terms of fairness.

Figure 4.5 illustrates that, across various top-k rankings, the tagged model consistently demonstrates the lowest bias score in comparison 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 versions. In all top-k rankings, both the scrubbed and tagged versions substantially enhance fairness compared to the initial model, but the tagged model exhibit superior performance.

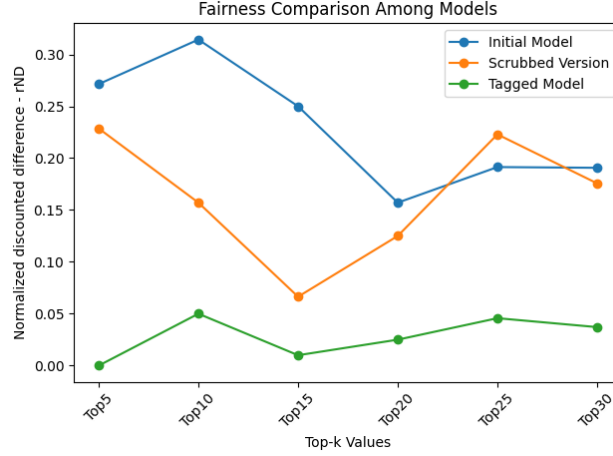


Figure 4.5 The representation of rND for three models (Initial, scrubbed and tagged) considering various numbers of top candidates. The green line is related to tagged model shows significantly lower rND score in all top-k values compared to the initial and scrubbed models, which demonstrates a fairer behavior.

4.4 Discussion

As demonstrated in preceding sections, the TagDebias approach exhibited strong performance on the ranking task with RoBERTa, surpassing both scrubbed and initial models. However, when examining other fairness metrics, such as CrowSpairs, the TagDebias model on RoBERTa fell short in terms of the SS score compared to the scrubbed and initial models. Notably, an analysis of other scores like the anti-stereotype score indicates that by substituting gender indicator terms with higher-level abstractions, the model tends to select more anti-stereotype sentences. Our model exhibits a bias toward anti-stereotype concepts, such as portraying men as nurses. A potential avenue for future research involves exploring ways to strike a balance in the model’s inclination towards both stereotype and anti-stereotype concepts.

As illustrated in Table 4.8, our TagDebias model demonstrated superior performance over all data-based bias mitigation strategies in terms of the SEAT score. However, it did not surpass the gender-tuning approach [1]. This suggests that relying solely on a tagging strategy may not eliminate all biases in the text. Therefore, incorporating a masked language training objective in conjunction with tagging could be a compelling avenue for future research. It

is worth noting that the gender-tuning approach was not evaluated in our proposed ranking task because the model was not publicly available at the time of writing this thesis.

CHAPTER 5 CONCLUSION

In this chapter, we offer a comprehensive overview of our study, highlighting and underscoring our contributions. In a subsequent section, we delve into the limitations of our work and outline the prospective avenues for future research.

5.1 Summary of works

Pretrained 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. There are different ways to mitigate social bias both in datasets and pretrained models. State-of-the-art approaches show effective debiasing methods in PLMs such as increasing dropout regularization [89], projection-based debiasing [87], and self-debias [88] as a post-hoc method to discourage model from generating toxic sentences. Other data-based bias mitigation methods such as counterfactual data augmentation (CDA) [112] or biased terms removal (scrubbing) [9] have been proposed but exhibit some limitations. Producing counterfactual data and fine-tuning pretrained language models (PLMs) on an augmented dataset is resource-consuming and, in some cases, impossible. For example, generating a counterfactual for the sentence "a woman gave birth" is impossible. Scrubbing biased words removes contextual associations within the PLMs and can decrease model performance in downstream tasks.

In this research, we propose a framework for mitigating bias in datasets and pretrained language models by tagging and fine-tuning PLMs on the BiasinBios dataset [9], 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 pretrained 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. We explored the effectiveness of our suggested tag-based data augmentation approach in two scenarios and assessed its fairness implications. Our findings indicate that the proposed data augmentation is beneficial for promoting fairness in models fine-tuned on both

the combined scrubbed and initial versions.

Finally, to assess the fairness of the debiased models, we test our tag-debiased model on a ranking task in the domain of biographies’ ranking given a target job title. We have found that our RoBERTa TagDebias model outperforms the scrubbed and initial models in the fair ranking task.

5.2 Limitations

This research underscores the efficacy of the tagging approach in mitigating gender bias within both datasets and pretrained language models (PLMs). However, certain limitations warrant discussion in the following sections.

Firstly, the utility of the tagging approach is contingent upon the dataset chosen for fine-tuning PLMs. Fairness scores exhibit variations on the BiasinBios dataset and CoLA, indicating a strong dependency on the selected dataset for our tagging strategy and fine-tuning process.

Secondly, the tagging of gender indicator terms relies on a limited list derived from the literature review. Consequently, not all gender indicator terms within the corpus are completely addressed, perpetuating bias in the dataset. While Spacy’s named entity recognition is effective in identifying proper nouns, there is still room for improvement in its performance. Fine-tuning PLMs on a biased corpus retains some bias within the models.

Thirdly, our initial objective was to extend our approach to other biases than gender. Our literature review reveals a scarcity of research on social biases related to race and religion. Limited datasets are available for studying these biases, attributed to time constraints in generating relevant datasets, thereby constituting a limitation.

Lastly, although our TagDebias method outperforms various state of the art data-based bias mitigation strategies, it falls short of surpassing the recently introduced gender-tuning approach [1] on the SEAT score [40]. We were also unable to evaluate this model for our proposed ranking task due to its unavailability at the time of writing this thesis.

5.3 Future research

We anticipate that this research would be the first step towards a tagging-based approach to address social bias within both datasets and pretrained language models. Additionally, our proposed ranking-based social bias identification tool is a new avenue in the application-based bias detection and motivates researchers to test the fairness of PLMs on the application-based

metrics.

Firstly, in this work, we have studied only gender bias. However, it would be interesting to research other types of bias such as race, religion, sex, or marital status. These types of social biases are less studied, due to the fact that existing datasets are scarce. One of the main direction is generating a new social bias dataset to perform a comprehensive study in the context of tagging-based bias mitigation.

Secondly, the use of large language models (LLMs) has become prevalent. While we ran out of time at the time of writing this thesis, we plan to test our tagged-based bias mitigation method using these models to evaluate its impact on LLMs.

Thirdly, the gender-tuning [1] bias mitigation strategy outperformed our TagDebias approach, and it seems that using only a tagging strategy will not be enough to remove bias in the dataset and subsequent fine-tuned PLMs. Thus our future direction involves combining our TagDebias approach along with a MLM objective. Our approach replaces the gender indicator terms with neutral terms while the gender-tuning replaces these terms with the choice of the masked language model that might not be neutral. So, having this constraint in the loss function could be an interesting avenue for future research.

Finally, an intriguing avenue of exploration involves identifying the most effective labels for tags. Additionally, considering different levels of abstraction in the labeling of tags could help determine the optimal combination of labels that minimizes social bias.

REFERENCES

- [1] S. Ghanbarzadeh, Y. Huang, H. Palangi, R. C. Moreno, and H. Khanpour, “Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models,” *arXiv preprint arXiv:2307.10522*, 2023.
- [2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.
- [3] A. Mukerjee, R. Biswas, K. Deb, and A. P. Mathur, “Multi-objective evolutionary algorithms for the risk–return trade–off in bank loan management,” *International Transactions in operational research*, vol. 9, no. 5, pp. 583–597, 2002.
- [4] M. Bogen and A. Rieke, “Help wanted: an examination of hiring algorithms,” *Equity, and Bias, Upturn (December 2018)*, 2018.
- [5] L. Cohen, Z. C. Lipton, and Y. Mansour, “Efficient candidate screening under multiple tests and implications for fairness.(2019),” *arXiv preprint cs.LG/1905.11361*, 2019.
- [6] M. Mozafari, R. Farahbakhsh, and N. Crespi, “Hate speech detection and racial bias mitigation in social media based on bert model,” *PloS one*, vol. 15, no. 8, p. e0237861, 2020.
- [7] S. Liang, P. Dufter, and H. Schütze, “Monolingual and multilingual reduction of gender bias in contextualized representations,” 2020.
- [8] Y. Guo, Y. Yang, and A. Abbasi, “Auto-debias: Debiasing masked language models with automated biased prompts,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 1012–1023.
- [9] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai, “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*, 2019, pp. 120–128.
- [10] H. R. Kirk, Y. Jun, F. Volpin, H. Iqbal, E. Benussi, F. Dreyer, A. Shtedritski, and Y. Asano, “Bias out-of-the-box: An empirical analysis of intersectional occupational

- biases in popular generative language models,” *Advances in neural information processing systems*, vol. 34, pp. 2611–2624, 2021.
- [11] J. Doughman, W. Khreich, M. El Gharib, M. Wiss, and Z. Berjawi, “Gender bias in text: Origin, taxonomy, and implications,” in *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 2021, pp. 34–44.
 - [12] X. Zhou, R. Ma, Y. Zou, X. Chen, T. Gui, Q. Zhang, X.-J. Huang, R. Xie, and W. Wu, “Making parameter-efficient tuning more efficient: A unified framework for classification tasks,” in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 7053–7064.
 - [13] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, “The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection,” *IEEE Transactions on Affective Computing*, 2022.
 - [14] J. Li, T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, “Pretrained language models for text generation: A survey,” *arXiv preprint arXiv:2201.05273*, 2022.
 - [15] Y. C. Tan and L. E. Celis, “Assessing social and intersectional biases in contextualized word representations,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf
 - [16] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang, “Mitigating gender bias in natural language processing: Literature review,” *arXiv preprint arXiv:1906.08976*, 2019.
 - [17] F. Cabiddu, L. Bott, G. Jones, and C. Gambi, “ChiSense-12: An English sense-annotated child-directed speech corpus,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 5198–5205. [Online]. Available: <https://aclanthology.org/2022.lrec-1.557>
 - [18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

- [19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019.
- [20] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” 2020.
- [21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [22] N. Meade, E. Poole-Dayana, and S. Reddy, “An empirical survey of the effectiveness of debiasing techniques for pre-trained language models,” *arXiv preprint arXiv:2110.08527*, 2021.
- [23] N. Gontier, S. Reddy, and C. Pal, “Does entity abstraction help generative transformers reason?” *arXiv preprint arXiv:2201.01787*, 2022.
- [24] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Gender bias in coreference resolution: Evaluation and debiasing methods,” *arXiv preprint arXiv:1804.06876*, 2018.
- [25] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman, “Science faculty’s subtle gender biases favor male students,” *Proceedings of the national academy of sciences*, vol. 109, no. 41, pp. 16 474–16 479, 2012.
- [26] T. Davidson, D. Warmusley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.
- [27] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? predictive features for hate speech detection on twitter,” in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.
- [28] J. Golbeck, Z. Ashktorab, R. O. Banjo, A. Berlinger, S. Bhagwan, C. Buntain, P. Cheakalos, A. A. Geller, R. K. Gnanasekaran, R. R. Gunasekaran *et al.*, “A large labeled corpus for online harassment research,” in *Proceedings of the 2017 ACM on web science conference*, 2017, pp. 229–233.
- [29] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, “Large scale crowdsourcing and characterization of twitter abusive behavior,” in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

- [30] Z. Waseem, “Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter,” in *NLP+CSS@EMNLP*, 2016.
- [31] T. Davidson, D. Bhattacharya, and I. Weber, “Racial bias in hate speech and abusive language detection datasets,” *arXiv preprint arXiv:1905.12516*, 2019.
- [32] P. Glick and S. T. Fiske, “The ambivalent sexism inventory: Differentiating hostile and benevolent sexism,” in *Social cognition*. Routledge, 2018, pp. 116–160.
- [33] J. C. Becker and S. C. Wright, “Yet another dark side of chivalry: Benevolent sexism undermines and hostile sexism motivates collective action for social change.” *Journal of personality and social psychology*, vol. 101, no. 1, p. 62, 2011.
- [34] S. Jain, R. Mohammadi, and B. C. Wallace, “An analysis of attention over clinical notes for predictive tasks,” in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 15–21. [Online]. Available: <https://aclanthology.org/W19-1902>
- [35] R. A. Connor, P. Glick, and S. T. Fiske, “Ambivalent sexism in the twenty-first century.” 2017.
- [36] E. J. Cross and N. C. Overall, “Women’s attraction to benevolent sexism: Needing relationship security predicts greater attraction to men who endorse benevolent sexism,” *European Journal of Social Psychology*, vol. 48, no. 3, pp. 336–347, 2018.
- [37] S. Mukherjee and S. Das, “Application of transformer-based language models to detect hate speech in social media,” *Journal of Computational and Cognitive Engineering*, vol. 2, no. 4, pp. 278–286, 2023.
- [38] A. C. Mazari, N. Boudoukhani, and A. Djeflal, “Bert-based ensemble learning for multi-aspect hate speech detection,” *Cluster Computing*, pp. 1–15, 2023.
- [39] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, “The risk of racial bias in hate speech detection,” in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 1668–1678.
- [40] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, “On measuring social biases in sentence encoders,” *arXiv preprint arXiv:1903.10561*, 2019.
- [41] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl, “Social biases in nlp models as barriers for persons with disabilities,” *arXiv preprint arXiv:2005.00813*, 2020.

- [42] B. A. Nosek, F. L. Smyth, J. J. Hansen, T. Devos, N. M. Lindner, K. A. Ranganath, C. T. Smith, K. R. Olson, D. Chugh, A. G. Greenwald *et al.*, “Pervasiveness and correlates of implicit attitudes and stereotypes,” *European Review of Social Psychology*, vol. 18, no. 1, pp. 36–88, 2007.
- [43] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi, “Social bias frames: Reasoning about social and power implications of language,” *arXiv preprint arXiv:1911.03891*, 2019.
- [44] R. Pryzant, R. D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, and D. Yang, “Automatically neutralizing subjective bias in text,” in *Proceedings of the aaai conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 480–489.
- [45] K. A. Clarke, “The phantom menace: Omitted variable bias in econometric research,” *Conflict management and peace science*, vol. 22, no. 4, pp. 341–352, 2005.
- [46] H. Suresh and J. V. Gutttag, “A framework for understanding unintended consequences of machine learning,” *arXiv preprint arXiv:1901.10002*, vol. 2, no. 8, 2019.
- [47] M. Wich, T. Eder, H. Al Kuwatly, and G. Groh, “Bias and comparison framework for abusive language datasets,” *AI and Ethics*, pp. 1–23, 2021.
- [48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [49] N. Kant, R. Puri, N. Yakovenko, and B. Catanzaro, “Practical text classification with large pre-trained language models,” *arXiv preprint arXiv:1812.01207*, 2018.
- [50] J. Zhou, J. Tian, R. Wang, Y. Wu, W. Xiao, and L. He, “Sentix: A sentiment-aware pre-trained model for cross-domain sentiment analysis,” in *Proceedings of the 28th international conference on computational linguistics*, 2020, pp. 568–579.
- [51] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl, “Social biases in NLP models as barriers for persons with disabilities,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 5491–5501. [Online]. Available: <https://aclanthology.org/2020.acl-main.487>
- [52] M. Nadeem, A. Bethke, and S. Reddy, “Stereoset: Measuring stereotypical bias in pretrained language models,” *arXiv preprint arXiv:2004.09456*, 2020.

- [53] A. Abid, M. Farooqi, and J. Zou, “Persistent anti-muslim bias in large language models.” New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: <https://doi.org/10.1145/3461702.3462624>
- [54] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, “Realtoxicityprompts: Evaluating neural toxic degeneration in language models,” *arXiv preprint arXiv:2009.11462*, 2020.
- [55] Y. C. Tan and L. E. Celis, “Assessing social and intersectional biases in contextualized word representations,” *Advances in neural information processing systems*, vol. 32, 2019.
- [56] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *Advances in neural information processing systems*, vol. 32, 2019.
- [57] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [58] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” *arXiv preprint arXiv:1904.09675*, 2019.
- [59] W. Yuan, G. Neubig, and P. Liu, “Bartscore: Evaluating generated text as text generation,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 263–27 277, 2021.
- [60] T. Sun, J. He, X. Qiu, and X. Huang, “Bertscore is unfair: On social bias in language model-based metrics for text generation,” *arXiv preprint arXiv:2210.07626*, 2022.
- [61] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? predictive features for hate speech detection on Twitter,” in *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 88–93. [Online]. Available: <https://aclanthology.org/N16-2013>
- [62] B. Vidgen, S. Hale, E. Guest, H. Margetts, D. Broniatowski, Z. Waseem, A. Botelho, M. Hall, and R. Tromble, “Detecting East Asian prejudice on social media,” in *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, Nov. 2020, pp. 162–172. [Online]. Available: <https://aclanthology.org/2020.alw-1.19>

- [63] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, “Measuring and mitigating unintended bias in text classification,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 67–73.
- [64] S. Hu, J. A. Al-Ani, K. D. Hughes, N. Denier, A. Konnikov, L. Ding, J. Xie, Y. Hu, M. Tarafdar, B. Jiang, L. Kong, and H. Dai, “Balancing gender bias in job advertisements with text-level bias mitigation,” *Frontiers in Big Data*, vol. 5, 2022. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fdata.2022.805713>
- [65] S. L. Bem, “The measurement of psychological androgyny,” *Journal of consulting and clinical psychology*, vol. 42, no. 2, p. 155, 1974.
- [66] D. Gaucher, J. Friesen, and A. C. Kay, “Evidence that gendered wording in job advertisements exists and sustains gender inequality,” *Journal of personality and social psychology*, vol. 101, no. 1, p. 109, 2011.
- [67] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, “Word embeddings quantify 100 years of gender and ethnic stereotypes,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, pp. E3635–E3644, 2018.
- [68] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” *Science*, vol. 356, no. 6334, pp. 183–186, 2017.
- [69] A. Garimella, A. Amarnath, K. Kumar, A. P. Yalla, N. Anandhavelu, N. Chhaya, and B. V. Srinivasan, “He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 4534–4545.
- [70] P. P. Liang, C. Wu, L.-P. Morency, and R. Salakhutdinov, “Towards understanding and mitigating social biases in language models,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 6565–6576.
- [71] P. Delobelle, E. K. Tokpo, T. Calders, and B. Berendt, “Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models,” *arXiv preprint arXiv:2112.07447*, 2021.
- [72] A. Lauscher, T. Lueken, and G. Glavaš, “Sustainable modular debiasing of language models,” *arXiv preprint arXiv:2109.03646*, 2021.

- [73] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, “Crows-pairs: A challenge dataset for measuring social biases in masked language models,” *arXiv preprint arXiv:2010.00133*, 2020.
- [74] A. Wang and K. Cho, “BERT has a mouth, and it must speak: BERT as a Markov random field language model,” in *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 30–36. [Online]. Available: <https://aclanthology.org/W19-2304>
- [75] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, “Masked language model scoring,” *arXiv preprint arXiv:1910.14659*, 2019.
- [76] M. Kaneko and D. Bollegala, “Unmasking the mask—evaluating social biases in masked language models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 11 954–11 962.
- [77] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, “Measuring bias in contextualized word representations,” *arXiv preprint arXiv:1906.07337*, 2019.
- [78] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, “MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance,” 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 563–578. [Online]. Available: <https://aclanthology.org/D19-1053>
- [79] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation of machine translation.” USA: Association for Computational Linguistics, 2002. [Online]. Available: <https://doi.org/10.3115/1073083.1073135>
- [80] J. H. Park, J. Shin, and P. Fung, “Reducing gender bias in abusive language detection,” *arXiv preprint arXiv:1808.07231*, 2018.
- [81] L. Cheng, A. Mosallanezhad, Y. N. Silva, D. L. Hall, and H. Liu, “Mitigating bias in session-based cyberbullying detection: A non-compromising approach,” in *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, vol. 1, 2021.

- [82] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, “Handling bias in toxic speech detection: A survey,” *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–32, 2023.
- [83] N. Madaan, S. Mehta, T. Agrawaal, V. Malhotra, A. Aggarwal, Y. Gupta, and M. Saxena, “Analyze, detect and remove gender stereotyping from bollywood movies,” in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 92–105.
- [84] E. Vanmassenhove, C. Hardmeier, and A. Way, “Getting gender right in neural machine translation,” *arXiv preprint arXiv:1909.05088*, 2019.
- [85] M. O. Prates, P. H. Avelar, and L. C. Lamb, “Assessing gender bias in machine translation: a case study with google translate,” *Neural Computing and Applications*, vol. 32, no. 10, pp. 6363–6381, 2020.
- [86] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” *Advances in neural information processing systems*, vol. 29, 2016.
- [87] P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency, “Towards debiasing sentence representations,” *arXiv preprint arXiv:2007.08100*, 2020.
- [88] T. Schick, S. Udupa, and H. Schütze, “Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1408–1424, 12 2021.
- [89] K. Webster, X. Wang, I. Tenney, A. Beutel, E. Pitler, E. Pavlick, J. Chen, E. Chi, and S. Petrov, “Measuring and reducing gendered correlations in pre-trained models,” *arXiv preprint arXiv:2010.06032*, 2020.
- [90] U. Gupta, J. Dhamala, V. Kumar, A. Verma, Y. Pruksachatkun, S. Krishna, R. Gupta, K.-W. Chang, G. V. Steeg, and A. Galstyan, “Mitigating gender bias in distilled language models via counterfactual role reversal,” *arXiv preprint arXiv:2203.12574*, 2022.
- [91] Y. Qian, U. Muaz, B. Zhang, and J. W. Hyun, “Reducing gender bias in word-level language models with a gender-equalizing loss function,” *arXiv preprint arXiv:1905.12801*, 2019.
- [92] H. Kučera and W. N. Francis, *Computational analysis of present-day American English*. Brown university press, 1967.

- [93] I. Solaiman and C. Dennison, “Process for adapting language models to society (palms) with values-targeted datasets,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 5861–5873, 2021.
- [94] M. Gira, R. Zhang, and K. Lee, “Debiasing pre-trained language models via efficient fine-tuning,” in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 2022, pp. 59–69.
- [95] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [96] J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, and I. Gurevych, “AdapterHub: A framework for adapting transformers,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 46–54. [Online]. Available: <https://aclanthology.org/2020.emnlp-demos.7>
- [97] Z. Fatemi, C. Xing, W. Liu, and C. Xiong, “Improving gender fairness of pre-trained language models without catastrophic forgetting,” *arXiv preprint arXiv:2110.05367*, 2021.
- [98] T. Sühr, S. Hilgard, and H. Lakkaraju, “Does fair ranking improve minority outcomes? understanding the interplay of human and algorithmic biases in online hiring,” in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 989–999.
- [99] E. Pitoura, K. Stefanidis, and G. Koutrika, “Fairness in rankings and recommendations: an overview,” *The VLDB Journal*, pp. 1–28, 2022.
- [100] L. E. Celis, D. Straszak, and N. K. Vishnoi, “Ranking with fairness constraints,” *arXiv preprint arXiv:1704.06840*, 2017.
- [101] K. Yang and J. Stoyanovich, “Measuring fairness in ranked outputs,” in *Proceedings of the 29th international conference on scientific and statistical database management*, 2017, pp. 1–6.
- [102] D. de Vassimon Manela, D. Errington, T. Fisher, B. van Breugel, and P. Minervini, “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*, 2021, pp. 2232–2242.
- [103] M. Zehlike, K. Yang, and J. Stoyanovich, “Fairness in ranking, part i: Score-based ranking,” *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–36, 2022.
- [104] R. Perego, F. Sebastiani, J. Aslam, I. Ruthven, and J. Zobel, “Proceedings of the 39th international acm sigir conference on research and development in information retrieval, sigir 2016, pisa, italy, july 17-21, 2016.” ACM, 2016.
- [105] N. Rekabsaz, S. Kopeinik, and M. 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*, 2021, pp. 306–316.
- [106] S. SeyedSalehi, A. Bigdeli, N. Arabzadeh, B. Mitra, M. Zihayat, and E. Bagheri, “Bias-aware fair neural ranking for addressing stereotypical gender biases.” in *EDBT*, 2022, pp. 2–435.
- [107] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, “Learning gender-neutral word embeddings,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 4847–4853. [Online]. Available: <https://aclanthology.org/D18-1521>
- [108] S. Tong, Q. Dong, D. Dai, T. Liu, B. Chang, Z. Sui *et al.*, “Robust fine-tuning via perturbation and interpolation from in-batch instances,” *arXiv preprint arXiv:2205.00633*, 2022.
- [109] F. Zhou, Y. Mao, L. Yu, Y. Yang, and T. Zhong, “Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 4227–4241.
- [110] A. Warstadt, A. Singh, and S. R. Bowman, “Neural network acceptability judgments,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 625–641, 2019.
- [111] T. P. Pagano, R. B. Loureiro, F. V. N. Lisboa, R. M. Peixoto, G. A. S. Guimarães, G. O. R. Cruz, M. M. Araujo, L. L. Santos, M. A. S. Cruz, E. L. S. Oliveira,

- I. Winkler, and E. G. S. Nascimento, “Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods,” *Big Data and Cognitive Computing*, vol. 7, no. 1, 2023. [Online]. Available: <https://www.mdpi.com/2504-2289/7/1/15>
- [112] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “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)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 15–20. [Online]. Available: <https://aclanthology.org/N18-2003>

APPENDIX A

Different rankings generated by 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 fair ranking, with the mean approaching 50%.

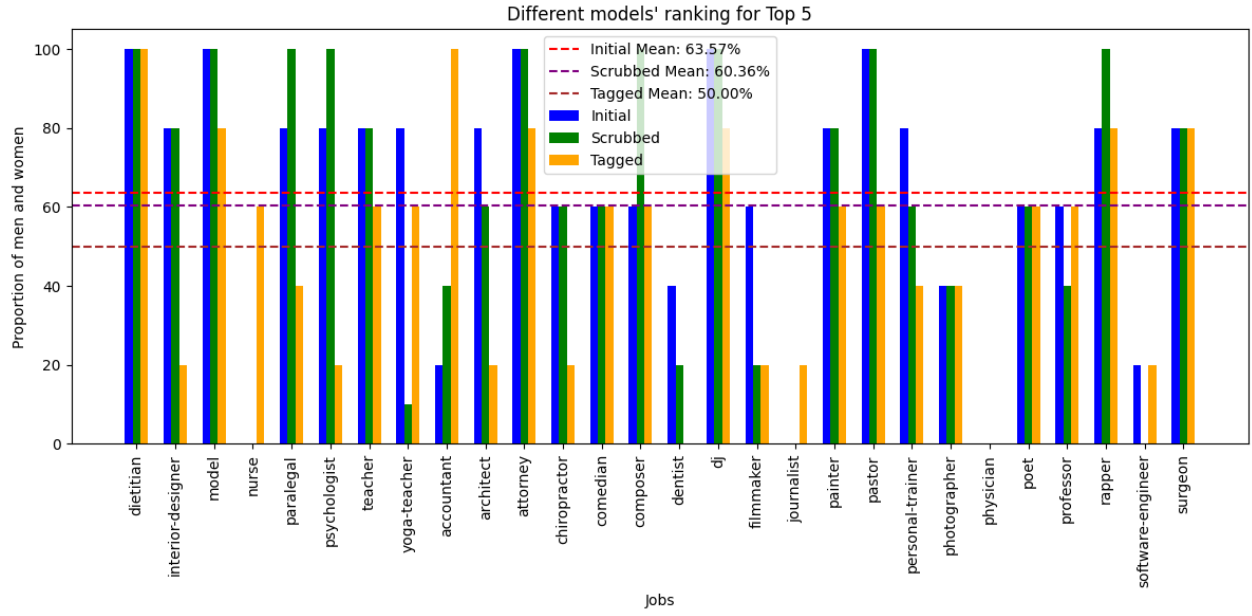


Figure A.1 Top 5 biographies ranking

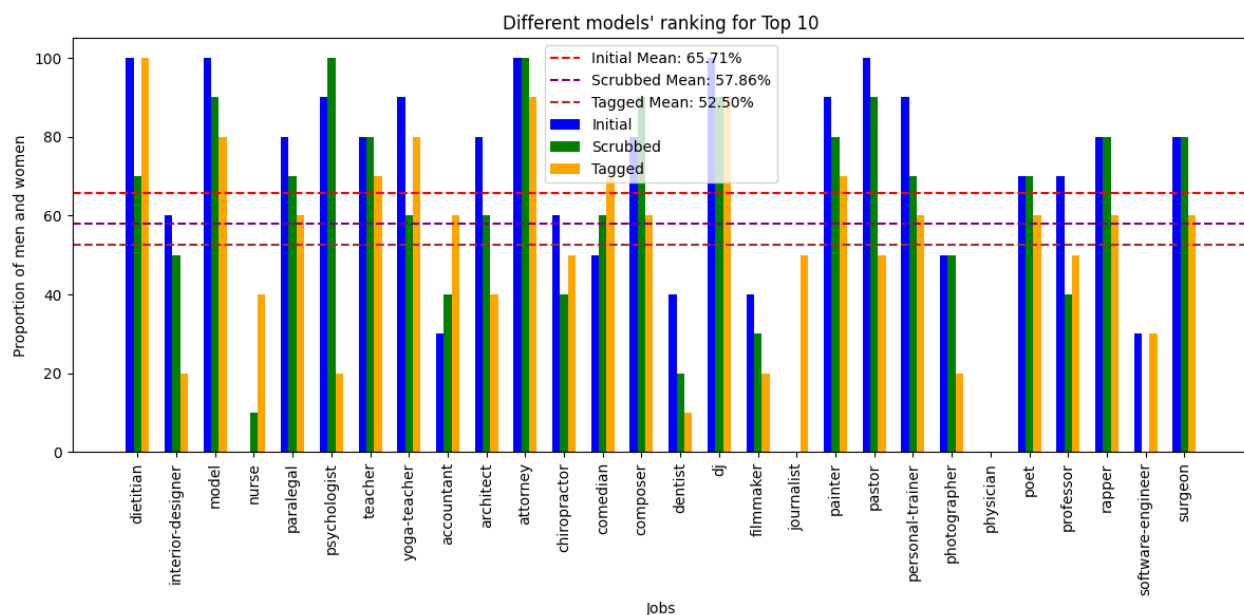


Figure A.2 Top 10 biographies ranking

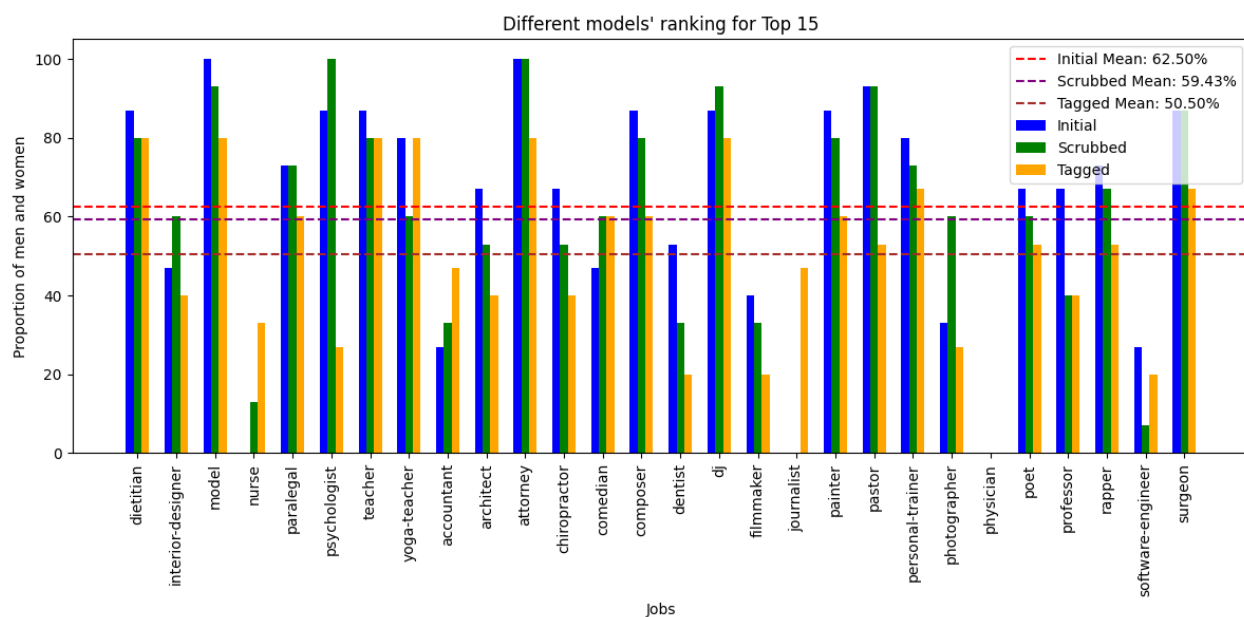


Figure A.3 Top 15 biographies ranking

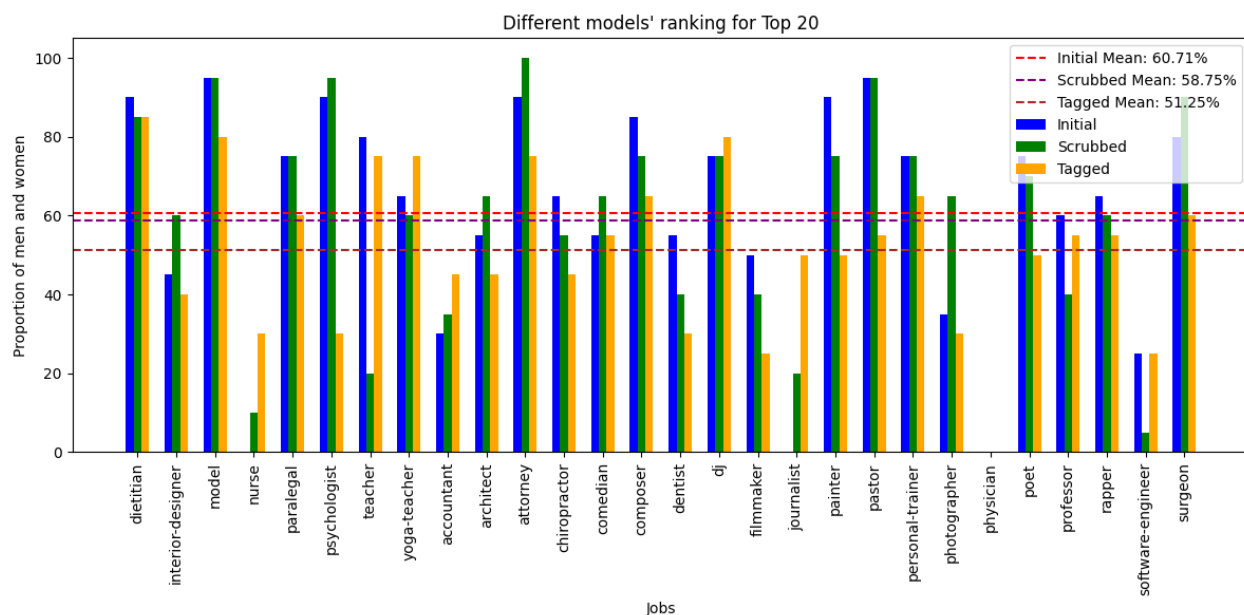


Figure A.4 Top 20 biographies ranking

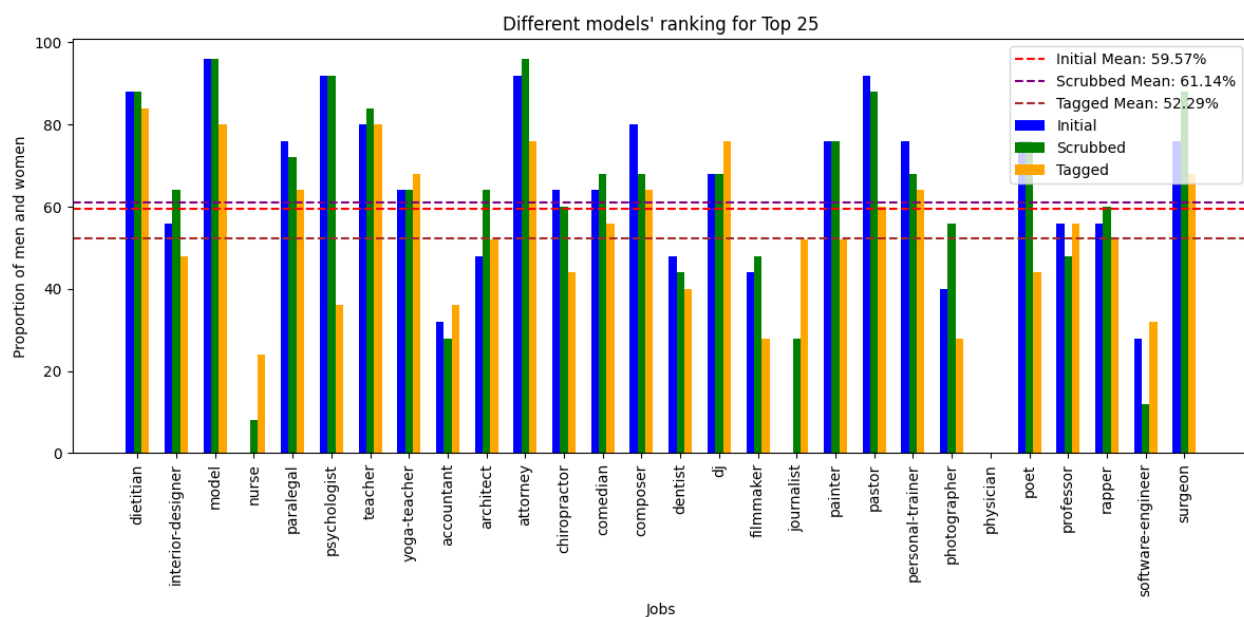


Figure A.5 Top 25 biographies ranking

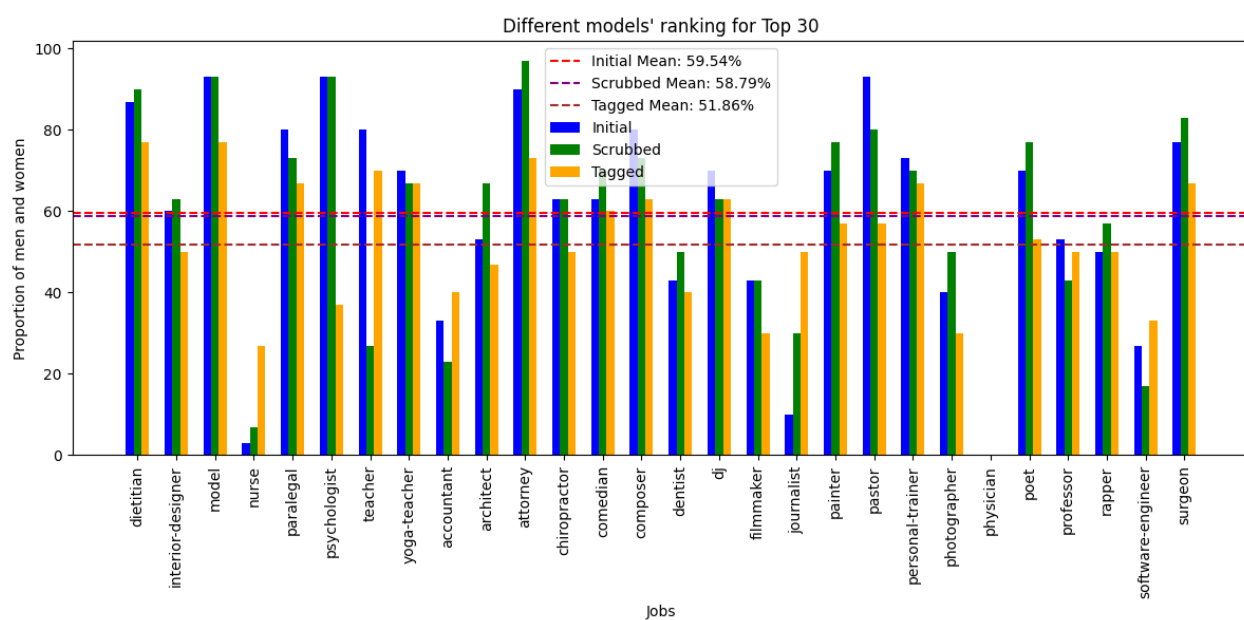


Figure A.6 Top 30 biographies ranking