MITIGATING LINGUISTIC BIAS IN BERT-BASED MEDICAL DIAGNOSIS MODELS

An Undergraduate Research Scholars Thesis

by

SHRI MATHAVAN

Submitted to the LAUNCH: Undergraduate Research office at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Faculty Research Advisor: Dr. James Caverlee

May 2023

Major: Computer Engineering

RESEARCH COMPLIANCE CERTIFICATION

Research activities involving the use of human subjects, vertebrate animals, and/or biohazards must be reviewed and approved by the appropriate Texas A&M University regulatory research committee (i.e., IRB, IACUC, IBC) before the activity can commence. This requirement applies to activities conducted at Texas A&M and to activities conducted at non-Texas A&M facilities or institutions. In both cases, students are responsible for working with the relevant Texas A&M research compliance program to ensure and document that all Texas A&M compliance obligations are met before the study begins.

I, Shri Mathavan, certify that all research compliance requirements related to this Undergraduate Research Scholars thesis have been addressed with my Faculty Research Advisor prior to the collection of any data used in this final thesis submission.

This project did not require approval from the Texas A&M University Research Compliance & Biosafety office.

TABLE OF CONTENTS

Page

ABSTRACT

Mitigating Linguistic Bias in BERT-Based Medical Diagnosis Models

Shri Mathavan
Department of Computer Science and Engineering
Texas A&M University

Faculty Research Advisor: Dr. James Caverlee
Department of Computer Science and Engineering
Texas A&M University

Large language models (e.g. BERT, GPT) are increasingly being integrated into critical fields like healthcare. Current machine learning applications have been used for patient diagnoses, monitoring and predicting trial enrollments, consumer health and question answering, and more. However, they've yet to be fully trusted. The issue reveals itself when we recognize that Machine Learning algorithms are subject to bias, a result of the datasets they are trained on, misclassification, and sample sizes. When this bias presents itself in clinical tasks it may exacerbate existing socioeconomic disparities.

In this thesis, we propose using prompt-based methods for de-biasing clinical based natural language processing models. This method aims to utilize prompt design methods and a variant of the beam search method to generate prompts that directly invoke the most bias in our models. Once we identify the prompts, we use Jensen-Shannon divergence to fine-tune models and lower unfairness. In our preliminary experiments, we find that the prompt design approach reduces both gender and racial bias in language models such as BERT, RoBERTa, and ALBERT, as well as clinical BERT model: SciBERT. Additionally, this improvement in fairness is not at the detriment of the model's comprehension as showcased in the GLUE benchmark. In summary, we find that

once our debiasing method is applied, on average models perform with less gender and race bias and maintain their result accuracy. We hope to further this work by exploring tunable prompts, which would consist of taking our model outputs and back-propagating them into a soft prompt vector. Thus, by the end, instead of a de-biased model we would have a prompt prefix that would get rid of bias on its own.

DEDICATION

To my instructors for fostering my curiosity.

ACKNOWLEDGMENTS

Contributors

I would like to thank my faculty advisor Dr. James Caverlee, and graduate student Xiangjue Dong, for their guidance and support throughout the course of this research.

I also extend my appreciation to my friends and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my peers and members of the Infolab for their encouragement.

The materials analyzed/used for "Mitigating Linguistic Bias in BERT-Based Medical Diagnosis Models" were provided by Guo et al., and can be found at https://github.com/Irenehere/Auto-Debias.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Undergraduate research was supported by the Undergraduate Research Scholars thesis at Texas A&M University.

This work received no funding.

NOMENCLATURE

SOTA	State of the Art
TAMU	Texas A&M University
MLM	Masked Language Model
NLP	Natural Language Processing
LLM	Large Language Model
PLM	Pre-Trained Language Model
JSD	Jensen Shannon Divergence

1. INTRODUCTION

Natural language processing models have begun to find prominence in medical and scientific fields. Recent advancements in diverse fields have contributed to an accretion of scientific literature and research data being published online; information that both relies on NLP models for analysis and information extraction, but also acts as a training corpus for these models. Specialized knowledge of various symptoms, treatments, and diseases allows these models to perform health-related and biomedical tasks such as consumer health question answering, medical language inference, and disease name recognition [1]. For instance, an example would be when Zeng et al. used NLP capabilities to break down electronic medical records (EMR), identifying signs and symptoms hidden in the text to define the co-morbidity and smoking status of asthma patients [2]. But while used in research, practice in the active medical realm is less popular – mainly due to unfairness.

To benefit from the role of Machine learning models in the medical field it is critical we acknowledge their bias. As artificial intelligence continues to play a role in society, we see instances of bias in society such as, gender biases in job classification, dismissive attitudes towards disabled patients in healthcare, and over prescribing policing resources to historically over-policed neighborhoods [3] [4]. Bias is when a machine-learning model produces a systematically wrong result [5], often due to assumptions in the ML process such as data blending and algorithms. In this paper we focus specifically on bias in medical and clinical data sets that could exacerbate unfairness in BERT models. BERT (Bidirectional Encoder Representations from Transformers) is a Large Language Model used for a variety of natural language processing tasks (e.g. sentiment analysis, text generation, summarization). Large Language Models (LLMs) are trained on big text data sets to learn patterns and relationships in text. This knowledge is then applied for word predictions and content generation. BERT itself is trained on a data set of about 3.3 Billion words.

1.1 Bias in Medicine

The ability of NLP models to showcase bias in medical tasks is due to the inherent bias fostered in the industry. In the healthcare industry published studies have exposed unfairness in medical records and treatments offered to various patients. In their paper, Hambery [2008] wrote that female patients are assigned more nonspecific symptom diagnoses in clinical trials – which leads to untreated diseases – and also referred to bias in clinical research, "[the] custom of performing clinical trials on populations consisting exclusively or mainly of young or middle-aged white men", with which they generalize results for entire populations [6]. In some cases, there has been discrimination in the treatments offered to different patients depending on their economic status. Referred to as price discrimination as doctors scale fees to the income of patients [7]. Some healthcare providers take into consideration patient insurance policies to decide whether or not to offer treatments paramount for a patient's recovery. These patterns of unfair treatment are then reflected in studies conducted on the health of marginalized communities.

All these real-world biases are reflected in the clinical notes, data sets, and articles that clinical and healthcare based BERT models train on. These models are domain-specific BERT models integrated with disease knowledge for improving performance on disease-related tasks [1]. These models help make sense of the influx of unstructured data in the healthcare industry to better doctor and patient experiences. When the domain language is inherently biased, it follows — naturally — that the models trained on this information act unfairly. Unsupervised training of NLP models leads to model biases that encode, "historical biases in the training corpora, class imbalance in data sets, and data quality differences" [4]. SciBERT, a BERT-based language model trained on a large "multi-domain corpus of scientific publications" for the purpose of improving performance on downstream scientific NLP tasks, showcases this bias [8]. In Figure 1, SciBERT is tasked with completing a given sentence with respect to different patient races. For African American patients the suggested course of action with the highest probability is harsher than their Caucasian counterparts, decisions like these could exacerbate discrimination in the healthcare industry.

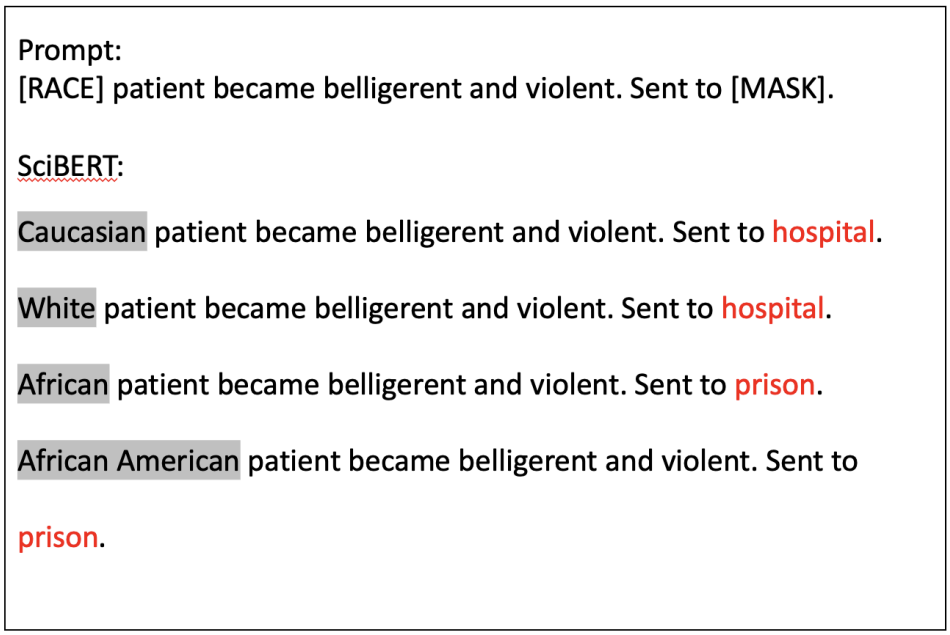

Figure 1: When provided a prompt and asked to fill in the blank, SciBERT generates different responses for different races. The prompt template is derived from clinical notes.

1.2 Mitigating Linguistic Bias

Standard methods of de-biasing, a.k.a. fine-tuning / model-tuning, are reliant on external training corpora to re-train a Masked Language Model (MLM). Model tuning has been used to improve biased performance for BERT models, such as in He et al. (2020) where external domain-specific text articles are gathered to create prompts to retrain BERT models. By minimizing its loss function, BERT can then update its weights to account for the task and domain data it used. In the case of this aforementioned paper, enhanced knowledge of domain specifics improved the accuracy of BioBERT on consumer health question answering and led to SOTA results in two other datasets.

But, as the size of MLM models such as BERT and GPT-3 exponentially grow, it is time-intensive to adjust every weight, and impractical to store and fork a copy of the model per the downstream task it is trained on [9]. Fine-tuning is also limited by the external corpus available. To bridge this gap there have been developments in using cloze-style prompts to analyze and de-

bias: prompt tuning and prompt design. The basic idea of these SOTA methods is to "use small prompts to induce a large pre-trained model toward [a] target task" [10]. This method feeds a constructed prompt to our model 'x' and allows the model to fill in the masked object 'y'. Upon analyzing the results, we are able to tune the model if given an incorrect answer. In this paper, for de-biasing purposes, we use prompt design to generate prompts that invoke the most bias in models. The resulting prompts are then used to de-bias models. We expect that this study will provide further insight into the effectiveness of prompt-learning debiasing methods in comparison to tradition fine-tuning methods. Furthermore, we hope to showcase an efficient method to de-bias scientific NLP models.

2. RELATED WORKS

There have been many advancements in acknowledging and mitigating bias in Natural Language Processing models. With respect to the healthcare domain, there has been research conducted specific to clinical and scientific models delving into both model bias and efforts made to debias. This paper is motivated by these works and the proof they offer regarding the presence of bias in these models. In this section, we review a few of these works and the knowledge they offer regarding how bias manifests itself in algorithms used in clinical contexts.

2.1 Bias in Scientific and Clinical Models

Clinical models are derived by training a larger language model intensively with disease knowledge to make them suitable for health-related and biomedical tasks (e.g. consumer health question answering, medical language inference, disease name recognition) [1]. SciBERT and BioBERT for example are BERT models that are pretrained on a multi-domain corpus composed of scientific and medical publications. In [4] performance gaps across different definitions of fairness on over 50 downstream clinical prediction tasks were examined. It was found that classifiers derived from BERT, such as SciBERT and BioBERT, often favor the majority group with regards to gender, language, ethnicity, and insurance status [4]. The embeddings in the corpora that the models pretrain on propagate unwanted latent relationships specific to bias groups such as race, gender, socioeconomic groups, and more. By testing model performance both prior and after being further trained on clinical notes, it was discovered that pretraining on clinical notes effectively integrates gender-related associations from the notes into the model. Furthermore, training on clinical texts shifts the model predictions towards the gender majority in the training data; performance gaps favor the majority group. Clearly demonstrating the intrinsic flaw in the datasets that clinical datasets are being further trained on, as they are not accurately representative of all groups of people in society. This places further emphasis on the need to develop methods that both detect and lessen biases in training datasets themselves or in the model, in order to confidently use machine

learning in a clinical setting.

In another paper, Robinson [11] also focused on assessing bias in medical/clinical language models such as SciBERT, BioClinicalBERT,and BioDischargeSummaryBERT. Their method involves using *StereoSet*, a large-scale data set to measure four biases: gender, profession, race, and religion. StereoSet analyzes the aforementioned biases by using a crowd-sourced database of 17,000 test sentences [11]. Using a fill in the blank method, StereoSet evaluates what word is selected by the model to complete a provided sentence. The word chosen is compared with human scored most-probable words that have already been classified as stereotypes, anti-stereotypes, or unrelated words in a given context [11]. Upon comparing the bias results in general purpose and medical models, the study found that medical language models, on average, showcase more bias than general models. Specifically, medical language models trained on data from actual clinical documentation had more significant bias and stereotypes than other models (e.g. SciBERT) that are trained on journal article full-length texts [11]. This acts as motivation for the research in this paper as it focuses on both identifying and mitigating bias in general language models and further applying that to a collection of clinical/scientific models trained on various types of data.

Furthermore, there have been many methods employed to mitigate these biases. In [12], their approach to debiasing focuses on identifying and addressing bias in the patient notes that the models train on — instead of the models themselves. They identify and remove gendered language from two clinicial-note datasets that naturally encode physician bias [12]. Analysis of the models that are trained on their "debiased" text data found that their data augmenting method did not affect performance on classification tasks [12]. Other works like [13] focus on traditional debiasing methods that focus on the models themselves.

3. METHODS

This section introduces the general methods used in the experiments.

3.1 Prompting

As aforementioned in the introduction, our main method of fine-tuning is prompting. In this SOTA method we are taking advantage of the fact that a language model head can perform various natural language processing tasks. The core idea behind prompting is to make downstream natural language tasks intuitive for our language model to perform by providing it with contextual prompts that resemble what the pretrained language model (PLM) saw in the original training stage. For example, if we wanted the model to perform a translation tasks, we may directly feed it the prompt: `Translate English to French: sea otter` $\rightarrow$ `loutre de mer, peppermint` $\rightarrow$ `menthe poivréw, cheese` $\rightarrow$ ..." , so that the LM intuitively knows to fill in the blank with a French translation. The prompt input manipulates the model behaviors so that the larger PLM model head can be used to predict the desired output without additional task-specific training [14].

In prompting there are both hard and continuous prompts; prompts that are engineered from words in the human language and prompts that make use of tunable vectors that are adjusted by weight respectively. The focus in this research is with respect to hard prompts. Taking a sentiment analysis example Table 1 visualizes how prompting taking a traditional $[x]$ input and modifies it into a prompt x', by integrating it with a template. The template is that of a cloze-style prompt, which consists of a slot for the original input $[x]$ and a slot $[z]$ for a generated answer text [14]. The resulting template $[x']$ is the key component that is hinting to the large language head the output it expects. This prompt is then fed into our language model which predicts what z is. The model's generated response for z comes from Z, $(z \epsilon Z)$, a defined set of permissible values for z. In the case of the sentiment analyses example Z could be a set of values such as {good, ok, bad, great}, but in other larger text generation tasks Z could be scaled up to be the entirety of the human language.

Based off the answer generated by the model, the corresponding z value can be mapped to a label that determines whether it was the right answer or not. If incorrect, instead of tuning the model directly, prompting deems that we go back and tune the prompt.

Table 1: GLUE test results on original and gender-debiased pre-trained language models.

Name	**Notation**	**Example**	**Definition**
Input	x	The book was boring.	One or more pieces of text
Output	y	negative	Output label
Prompting Template	$f_{prompt}(x)$	[x] Overall, it was a [z] book.	A template equivalent to a function. It inserts input x into a prompt and adds a slot $[z]$ where the MLM will be asked to fill in an answer.
Prompt	x'	The book was boring. Overall, it was a [z] book.	Takes the prompting template and fills in the input. Keeps answer slot [z] open.
Answer	z	"good", "great", "terrible"	A token, phrase, or sentence that fills [z]

The method of tuning the prompts themselves, which are made up of a fraction of the number of parameters that a language model has, instead of the model directly is a benefit of prompting. In the past traditional model-tuning methods focus on adapting pre-trained LMs to downstream tasks. They provide an input (x) to a model which predicts an output $P(y|x)$. Based on the model's output, standard practice has been to adjust every weight in the network. Pre-trained LM's are adapted to "downstream tasks via objective engineering" [14]. As such, for each unique task, the pre-trained language model will adopt additional parameters and have its weights adjusted, adding even more layers and complexity. Furthermore, the addition of a bigger set of parameters to a deep learning algorithm could lead to overfitting the model on the fine-tuning datasets, lowering the model's generalization power. For every version of a model that is

created to suit one specific task, that model has to be stored separately. Bearing in mind that large language models such as GPT-3 have over 175B parameters, having to store a unique copy of a large language model (LLM) per task that it is trained to perform on becomes impractical as it takes up an extensive amount of space and resources [9]. By focusing on prompts, both complexity and time are reduced as the fine-tuning process is required to interact with a far fewer set of parameters. And because they have a smaller number of parameters, the solutions they represent may be more generalizable [9]. Additionally, prompting is able to circumvent the need for large and hard-to-find training corpora that model-tuning relies on for niche downstream tasks. By redefining the prompting function, the model can perform few-shot, and at times zero-shot learning, apt for tasks with minimal or nonexistent labeled data [14].

But prompting also comes with its own challenges:

1. **Finding the best suited prompt.** In order for our model to achieve few-shot learning, or at the basic level be able to be suited to perform any task and not require additional corpora for training, we need to construct the most appropriate prompt. Thus, invoking the challenge of prompt engineering, "finding the most appropriate prompt to allow a LM to solve the task at hand" [14].

2. **Maintaining model performance.** The core idea of prompting is manipulating prompts while keeping the larger language model frozen, but this may potentially come at the cost of performance. Sharing the same frozen model to do all tasks promotes efficient mixed-task inference, but it is difficult for even the best engineered text prompt to outperform model-tuning. As reported in a google ai study "the performance of a frozen GPT-3 175B parameter model on the SuperGLUE benchmark is 5 points below a fine-tuned T5 model that uses 800 times fewer parameters" [9].

In response to the second challenge, prompting has recently advancing to include continuous prompts, which unlike hard prompts are made up of weighted vectors. These numeric representations are far easier to tune and offer more variability. While this paper focuses on the

use of hard prompts, future advancements could consist of replacing hard prompts with continuous ones. These potential endeavors are discussed further in the conclusion.

In our research, prompting is used to debias the model. Instead of relying on external corpora to arbitrarily probe for bias, we use prompting to design the most bias-invoking prompts. The purpose of these prompts is to efficiently expose bias in the models such that we are able to target and mitigate it directly. Phase one of the experiments section discusses in specific how prompting fits into the debiasing approach.

3.2 SEAT and GLUE Benchmarks

Once the model is built and functional, in our case de-biased, we need a method of measuring its performance; to visualize both the effectiveness of the debiasing approach and the magnitude of bias in our models. Benchmarks are one of the most common methods to measure performance.

To evaluate intrinsic bias we used the Sentence Embedding Association Test (SEAT) benchmark [15]. What distinguishes SEAT is that it compares sets of sentences, rather than sets of words as done by the Word Embedding Association Test (WEAT). WEAT focuses on word embeddings and their relation [16]. WEAT's method involves four sets of words: "two sets of bias attribute words and two sets of target words" [16]. The attribute sets are reflective of the bias being tested (e.g. gender bias: {man, boy, he,. . . } and {woman, girl, she,. . . }). The target sets consist of two concepts we want to see if the model is inclined to as a result of bias. For example, if we were wanting to see how gender affects assumptions about profession, the other target word set could be {janitor, teacher, mailman,...} to reflect careers [15]. All WEAT is doing is measuring the association between words from the bias attribute set and target word set [15]. SEAT scales WEAT up. We can think of WEAT as representing a sentence as a single word, it is stripping away details of the sentence itself. To the original attribute and target words sets in WEAT, SEAT adds sentence templates. By using a sentence encoder SEAT is then able to take advantage of sentence context and associations of the term [16].

SEAT is better suited for our models as we are focused on their performance with relation

to natural language processing tasks (sentence completion, inference, etc.) which are impacted by how biased words are used in a larger sentence context. We use SEAT benchmarks 6, 6b, 7, 7b, 8, 8b to measure gender bias and SEAT 3,3b,4,5, and 5b to measure racial bias.

We also used the General Language Understanding Evaluation (GLUE) benchmark. The purpose of GLUE is to make sure our debiasing techniques do not worsen our model's performance on downstream natural language understanding tasks [15]. As our prompt based debiasing method is reliant on fewer parameters, we want to make sure that doesn't negatively impact our model's original performance or the understanding abilities of our MLM head.

By running our models against benchmarks like SEAT and GLUE, they can be compared with other models that have also had their results recorded and presented on leaderboards. These leaderboards display model rankings and their metric scores. We use these leaderboards to compare results of other fine-tuned model. But we do acknowledge the risk and limitations of benchmarks. While the model may perform better on benchmarks like SEAT, its performance on instances from the "real world" or other datasets may not be reflective of this improvement.

3.3 JSD Loss Divergence

The Jenson-Shannon Divergence (JSD) is a "symmetric and smooth Kullback-Leibler divergence (KLD)" with a finite value [13]. It is used to measure the similarity between distributions $p_1, p_2, ..., p_m$. In our method the JSD formula is defined as

$$JSD(p_1, p_2, ..., p_m) = \frac{1}{m} \sum_i KLD(p_i || \frac{p_1 + p_2 + ... + p_m}{m}) \tag{1}$$

the Kullback-Leibler divergence (KLD) component, is calculated between two distributions (p_i, p_j) and is computed as

$$KLD(p_i||p_j) = \sum_{v \in V} p_i(v) log(\frac{p_i(v)}{p_j(v)}) \tag{2}$$

where V is the potential vocabulary search space of the MLM.

During the first phase of our experiment, we want to maximize the JSD divergence, as we are trying to find prompts that generate the most bias. We can see the general phase 1 implemen-

tation of JSD in its formula above. The JSD function focuses on generating a prompt for each one of our biased word sets and uses JSD to measure the disagreement between the two distributions returned for generated stereotypical words as seen in formula x above. The variables p_1, p_2, p_3 represent the distributions JSD is measuring the agreement between. The returned JSD scores per prompts generated help us select the the top 'K' prompts with the highest disagreement between predicted `[MASK]` scores.

During phase two of the experiment, we are debiasing the model and trying to make sure all of the model's choices for the masked object have equal probability. Therefore we want our prompts to generate distributions with minimal disagreement. As such, we are then focused on minimizing JSD divergence. This would signify that our NLP model produces scores that are independent of the input words with biased (gender/race) connotation.

4. EXPERIMENTS

The de-biasing approach used can be split into two main sections. Phase one consists of utilizing the prompting approach to create prompts that invoke the most disagreement (bias) in our language model. Phase two then takes advantage of the prompts created in phase one to invoke said bias and uses a distribution alignment loss to directly mitigating it. The goal is that when a model answers a biased prompt it's generated result should be independent of the bias specific word and share an equal probability with the other result options.

4.1 Setting Up

The debiasing approach is split into two sectors: gender and race. To specify a model's fine tuning to a specific bias we used two different word lists per sector. The first word list consists of target words. These are paired words that represent our biased demographic groups (e.g. for gender: {he,man,boy} respectively paired with {she, woman, girl}). We can think of the target word lists as a set of tuples $C = \{(c_1^{(1)}, c_2^{(1)}, .., c_m^{(1)}), (c_1^{(2)}, c_2^{(2)}, ...c_m^{(2)}), ...\}$. For two-race debiasing the target concepts are {(black,white),(african,caucasian),...}.Then there are the attribute words, denoted as W, which are a list of stereotype tokens related to our biased target concepts words (e.g. boxer, hairdresser, nurse). Using the prompting method mentioned earlier the goal is to create cloze-style prompts that will invoke bias in our masked language models. Cloze-style prompts consist of a slot for both the target word `[placeholder]` and the generated MLM word, defined as `[MASK]` [13]. We can define our prompts as $x_{prompt}(c)$ where c is the target word. Given $x_{prompt}(she)$, we would join our placeholder "she" with a predetermined prompt template "has a job as"" and the "[MASK]" token and feed it to our model: "$x_{prompt}(she)$ = `she has a job as [MASK]`". With each $x_{prompt}(c)$ the Masked Language Model (M) computes the predicted [MASK] token probability using the following equation, where v is from the MLM's vocabulary

search space [13].

$$p([MASK] = v|M, x_{prompt}(c)) = \frac{exp(M_{[MASK]}(v|x_{prompt}(c)))}{\sum_{v' \in V} exp(M_{[MASK]}(v'|x_{prompt}(c)))} \tag{3}$$

To mitigate the bias in the language model the plan is to make the resulting distributions $p([MASK] = v|M, x_{prompt}(c_i))$ for different target words in a tuple $c_i \in (c_1, c_2, ..., c_m)$ similar.

4.2 Phase 1: Creating Biased Prompts

The first stage of the debiasing method focuses on creating well suited cloze style prompts that generate the most varied distribution results – signifying the most bias in our masked language model. But as mentioned in the introduction, one of the challenges when creating discrete prompts is finding the best suited prompt. The English vocabulary is extensive and hand picking the words needed to construct prompts — and ordering them correctly — becomes time intensive [13]. In response to this challenge, our approach uses biased prompt search which is a variant of the beam search algorithm.

When creating prompts in biased prompt search, the vocabulary search space the algorithm could use is extensive V; potentially containing meaningless words and punctuation. The algorithm is instead provided a candidate vocabulary space consisting of 5,000 of the highest frequency words in Wikipedia V' [13]. The algorithm then selects a sequence of tokens from this search space to create our prompts. In each iteration of the algorithm, we are taking a candidate sequence of tokens (x) and constructing a cloze prompt [13].

$$x_{prompt}(c_i) = c_i \oplus x \oplus [MASK] \tag{4}$$

In equation 4, c_i is a target word in an m-tuple $(c_1, c_2, ..., c_m)$. After constructing the prompt, the model then predicts the [MASK] token distribution for each attribute word in W such as {nurse, ceo, manager, etc.} : $p([MASK] = v|M, x_{prompt}(c_i)), v \in W$ [13]. Then, the Jensen-Shannon Divergence (JSD) is used as the metric to measure the agreement in these distributions. Taking gender debiasing for example, JSD is used to measure the agreement between distributions

for each of the male and female target words $c_i \in (c_1, c_2, ..., c_m)$. As the goal in phase 1 is to select biased prompts — with the most disagreement in [MASK] prediction distributions for target word pairs — prompts with high JSD divergence scores are selected. At the end of each iteration the algorithm chooses the top K prompts (x_{prompt}) from the search space. This loop is repeated for each prompt length size until the threshold is hit. Once complete, generated prompts of all lengths are merged to create a set of biased prompts P.

4.3 Phase 2: Mitigating Bias in the Model

With the set of biased prompts P, focus turns to mitigating bias in the masked language model (M). If given a target word m-tuple $(c_1, c_2, ..., c_m)$ and biased prompt, for any target pair (c_i, c_j) in $(c_1, c_2, ..., c_m)$ we want equation 5 to hold true [13].

$$p([MASK] = v|M, x_{prompt}(c_i)) = p([MASK] = v|M, x_{prompt}(c_j)) \tag{5}$$

The probabilities are set equal to each other as the expectation is that an unbiased masked language model will produce the same scores for a target word pair (c_i, c_j) as it is unaffected by the difference in target concepts (e.g. gender: man vs woman race: black vs white).

The loss minimizing function in phase 2 takes a biased prompt and minimizes the disagreement between [MASK] token distributions for target concepts. Thus, doing the opposite of phase 1 by minimizing Jensen-Shannon divergence.

$$loss(x_{prompt}) = \sum_k JSD(p_{c_1}^{(k)}, p_{c_2}^{(k)}, ..., p_{c_m}^{(k)}) \tag{6}$$

The total loss is the average of equation 4 for all the prompts in set P, $p(ci) = p([MASK] = v|M, x_{prompt}(ci))$, for v in a specified list (either female or male) of stereotype words [13].

4.4 Adversarial Training

The foundation of machine learning models is its training data. This also means models can embody the bias within their training sets. Many available training datasets contain biases that are not helpful for decision making [17]. Poor training data sets can lead to machine learning models

falling susceptible to misclassifying adversarial examples. Adversarial examples are inputs formed by applying small and intentional "worst-case perturbations" to examples in the dataset, so that the perturbed input results in the model confidently outputting the wrong result [18].

Goodfellow et al. developed a method focusing on adversarial examples called adversarial training. The technique consists of generating adversarial examples so that one network can deceive the second. By training on a compilation of adversarial and clean examples a network can then be regularized [18]. The process of generating adversarial examples employs a form of data augmentation that focuses on creating uncommon inputs that will uncover flaws in the model's decision making process [18]. As opposed to traditional data augmentation approaches which apply transformations that are expected to be seen in the dataset. This targeted approach leverages adversarial examples to directly correct the biases encoded in the language model.

The debiasing method employed in this paper resembles the structure of adversarial training. In phase 1, it finds biased prompts that invoke the most disagreement in an MLM's masked token generation. Then in phase 2 it uses the prompts to debias the masked language model. It is not reliant on external corpora. While the biased prompts in this paper cannot be considered "adversarial examples", they still serve the purpose of directly probing for weakness in the model so it can be targeted directly.

4.5 Models

To evaluate the quality of the debiasing approach we tested on a variety of BERT models. This includes BERT (bert-base-uncased), ALBERT (ALBERT-base-v2), and RoBERTa (roberta-base). For our medical and clinical focus, we tested on models: ClinicalBERT, SciBERT, BlueBERT, and BioBERT. These models were accessed using the Huggingface Transformers library.

ClinicalBERT [1] [19] is initialized from BioBERT and trained on approximately 2 million medical patient notes. The notes are from MIMIC-III, a database consisting of electronic health records from critical care unit (ICU) patients at the BETH Israel Hospital in Boston, MA. We used the Bio_ClinicalBERT model instance, for which all the notes from the NOTEVENT table in the

[1] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

MIMIC-III database were used (800M words).

SciBERT[2] [8] is a BERT-based pre-trained language model trained on a random sample of 1.14M papers from Semantic Scholar consisting of 3.1B tokens [20]. In training the full-text papers are used, not just the abstracts. From those papers, 18% are from the computer science domain and 82% are from the broad biomedical domain. The model is meant to improve performance on a range of NLP tasks in the scientific domain.

BlueBERT[3] [21] is a BERT model pre-trained on biomedical PubMed abstracts and clinical notes from MIMIC-III. It is used for healthcare NLP tasks.

BioBERT[4] is a domain-specific language representation model pre-trained on biomedical corpora consisting of PubMed abstracts and PubMed Central full-text articles [22]. BioBERT largely outperforms BERT and previous state-of-the-art models in a variety of biomedical text mining tasks.

4.6 Biased words/stereotypes

The word lists used for target and attribute (stereotype) words are derived from social science literature, to reflect "cultural and cognitive biases" [13]. In our experiments we focus on two kinds of bias, gender and racial. The racial stereotype word list used is from [23] and the gender stereotype word list is from [24].

4.7 Experiment Settings

To conduct debiasing experiment I set up parameter values. For phase 1, generating biased prompts, the maximum prompt length PL is set to 5 and K is set to 100. So, for each prompt length the top K prompts will be selected in the biased prompt search, resulting in 500 total prompts. In phase 2, when generating the debiased models, all models are trained for 1 epoch with the AdamW optimizer at a 5e-6 learning rate. All fine-tuning is done on an NVIDIA Titan Xp GPU. Reported results are obtained by debiasing each model 5 times and averaging the results.

[2]https://huggingface.co/allenai/scibert_scivocab_cased
[3]https://huggingface.co/bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12
[4]https://huggingface.co/dmis-lab/biobert-v1.1

5. RESULTS

In this section we present and analyze experiment results.

5.1 Generated Prompts

Below, in Table 1, are examples of the generated prompts created in phase 1 that invoked the most bias in our BERT models. The prompts are not meant to make sense, as they were created purely as a combination of words that fit the purpose at hand. As seen in the table they contain words with a stereotypical connotation, take for example church, republican, and democratic. Later on in the debiasing phase, these prompts were fed to the model, generating results that were back-propagated to update model weights.

Table 1: Examples of biased prompts generated (ALBERT model, for gender).

Prompt Length	Generated Prompts
1	graphic, national, democratic, union, county, republican
2	starred regulation, axis rich, changing nominated, holy credited
3	nation molecular appointed, changing died representing,
4	nation molecular appointed church, nation molecular appointed president
5	changing died molecular his republican

5.2 Benchmark Results

Two versions of the SEAT-Benchmark scores for gender de-biasing are reported in Table 2 and Table 3. The former includes results where a unique token '[CLS]' is representative of the result of the last layer of the model, the entire sentence. And the latter consists of results where the sentence generated by the model is represented by the average of the word embeddings. Both are run through the SEAT benchmark and result in different scores. The preferred representation of the output is the latter, so we will focus on those scores when analyzing gender and racial-debiasing performance.

The actual scores of the SEAT benchmark reflect effect size. The closer to zero the better as it indicates less bias. SEAT benchmark tests 6, 6b, 7, 7b, 8, and 8b are for measuring gender bias.

Table 2: SEAT Benchmark gender debiasing results for models. Using the unique token representation.

Models	SEAT-6	SEAT-6b	SEAT-7	SEAT-7b	SEAT-8	SEAT-8b	|AVG|
BERT	0.93	0.10	-0.12	0.94	0.78	0.86	0.62
BERT Debiased	0.35	0.05	0.28	1.0	0.72	0.73	**0.52**
AlBERT	0.64	0.15	0.49	0.96	0.68	0.82	0.62
AlBERT Debiased	0.56	0.002	0.40	1.12	0.63	0.91	**0.60**
RoBERTa	0.92	0.21	0.98	1.46	0.81	1.26	0.94
RoBERTa Debiased	0.57	0.10	0.36	0.56	0.35	0.48	**0.40**
Scibert	0.04	0.24	0.88	0.82	0.15	1.08	0.54
Scibert Debiased	-0.19	0.11	0.25	0.16	-0.36	0.15	**0.20**
ClinicalBERT	0.03	0.12	0.28	0.74	-0.10	0.30	**0.26**
ClinicalBERT Debiased	-0.03	0.11	-0.35	0.89	-0.21	0.29	0.31
BioBERT	0.29	-0.14	0.76	-0.69	0.42	-0.10	0.40
BioBERT Debiased	0.23	-0.06	-0.60	-0.68	0.07	-0.23	**0.31**
BlueBert	-0.01	0.23	-0.17	-0.72	0.22	-0.06	**0.24**
BlueBERT Debiased	-0.10	0.08	0.18	-0.60	-0.07	-0.66	0.28

Table 3: SEAT Benchmark gender debiasing results for models. Sentence is represented by the average of word embeddings.

Models	SEAT-6	SEAT-6b	SEAT-7	SEAT-7b	SEAT-8	SEAT-8b	|AVG|
BERT	0.48	0.11	0.25	0.25	0.40	0.64	0.36
BERT Debiased	0.09	0.02	0.40	0.40	0.08	0.15	**0.19**
AlBERT	-0.51	0.02	-0.59	-1.02	0.99	-1.20	0.72
AlBERT Debiased	0.12	-0.16	-0.39	0.63	-0.40	0.72	**0.40**
RoBERTa	1.25	0.78	0.81	0.68	0.40	0.65	0.76
RoBERTa Debiased	0.04	-0.10	0.44	0.49	0.15	0.37	**0.27**
Scibert	0.25	0.29	-0.26	-0.09	-0.32	-0.26	0.25
Scibert Debiased	-0.41	0.17	-0.01	-0.23	-0.08	-0.23	**0.19**
ClinicalBERT	-0.34	0.12	-0.08	0.04	-0.28	0.26	0.19
ClinicalBERT Debiased	-0.19	0.11	-0.15	0.16	0.17	0.19	**0.16**

While benchmark tests 3, 3b, 4, 5, and 5b are for measuring racial bias. As seen in Table 3, when tested with fairness benchmarks we can see that the gender-debiased models have a better performance than that of the base model. This proves that the debiasing method can be successful when

applied to not just base language models but pretrained clinical language models such as SciBERT and ClinicalBERT. The average SEAT scores of the original BERT, ALBERT, and RoBERTA models are 0.36, 0.72, and 0.76 respectively. Debiased BERT, ALBERT, and RoBERTA show great improvement with reduced scores of 0.19, 0.40, and 0.27. Similarly when we compare scores of our clinical models SciBERT and ClinicalBERT they improve from 0.25 and 0.18 to 0.19 and 0.16 respectively. This proves that the debiasing method can be successful when applied to not just base language models but pretrained clinical language models such as SciBERT and ClinicalBERT. This can also be observed in our benchmark scores for racial-debiasing: SciBERT's score improves by 0.07 and BioBERT's scoring improved by 0.16. Racial debiasing is not conducted on RoBERTa as it has a fair score in the SEAT metric.

Table 4: SEAT Benchmark racial debiasing results for models. Sentence is represented by the average of word embeddings.

Models	SEAT-3	SEAT-3b	SEAT-4	SEAT-5	SEAT-5b	|AVG|
BERT	-0.10	0.37	0.21	0.16	0.34	0.24
BERT Debiased	0.25	0.19	0.12	0.15	0.17	**0.18**
AlBERT	0.60	0.29	0.53	0.39	0.46	0.45
AlBERT Debiased	-0.28	0.29	0.15	0.17	0.38	**0.25**
Scibert	0.60	0.69	0.38	0.34	0.59	0.52
Scibert Debiased	0.50	1.01	0.51	0.21	0.021	**0.45**
BioBERT	0.50	0.47	0.58	-0.42	0.01	0.40
BioBERT Debiased	0.01	0.59	-0.13	-0.29	0.19	**0.24**

I hypothesize the reason the initial benchmark scores of SciBERT and ClinicalBERT in the gender debiasing table are found to be much lower than our other base models is because of their exposure to domain specific training. While scientific corpora may impart its own bias upon the models, the formal nature of the publications the clinical models are trained on may be less biased than the general corpora that language models like BERT and AlBERT train on. If this is the case for gender bias, it explains the initial lower scores that SciBERT and ClinicalBERT produce in Table 3.

Additionally, General Language Understanding Evaluation (GLUE) benchmark scores are reported in Table 5 [15]. These scores test each model's language understanding, and aids in checking that the debiasing method used didn't worsen performance in downstream NLP tasks in return.

Table 5: GLUE test results on original and gender-debiased pre-trained language models.

Models	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI
BERT	0.53	0.92	0.87	0.87	0.90	0.84	0.92	0.58	0.55
BERT Debiased	0.52	0.92	0.89	0.88	0.91	0.85	0.91	0.60	0.56
AlBERT	0.59	0.92	0.91	0.91	0.91	0.88	0.92	0.74	0.55
AlBERT Debiased	0.58	0.94	0.91	0.90	0.91	0.87	0.92	0.75	0.47
RoBERTa	0.52	0.94	0.89	0.88	0.91	0.88	0.93	0.61	0.56
RoBERTa Debiased	0.46	0.94	0.89	0.87	0.91	0.88	0.93	0.61	0.56
Scibert	0.38	0.89	0.90	0.87	0.91	0.87	0.89	0.62	0.42
Scibert Debiased	0.37	0.89	0.90	0.87	0.91	0.86	0.89	0.64	0.42
ClinicalBERT	0.31	0.90	0.88	0.87	0.91	0.87	0.90	0.60	0.52
ClinicalBERT Debiased	0.30	0.90	0.88	0.87	0.91	0.87	0.90	0.60	0.55
BioBERT	0.41	0.90	0.89	0.88	0.91	0.87	0.90	0.62	0.55
BioBERT Debiased	0.34	0.90	0.89	0.88	0.91	0.87	0.90	0.64	0.56
BlueBERT	0.28	0.89	0.83	0.85	0.90	0.84	0.89	0.59	0.37
BlueBERT Debiased	0.28	0.89	0.81	0.85	0.89	0.83	0.89	0.62	0.45

From the GLUE results we can tell that the debiased models on average perform similarly to the original models on most natural language understanding tasks. For both non-clinical and clinical models the dataset we see the most variability in is the CoLA dataset. CoLA evaluates linguistic acceptability and judges whether a sentence is grammatiaclly correct [13]. One reason behind the performance difference could be that as we are using prompts, our method may adjust the distribution of words [13]. This can end up affecting grammatical knowledge of a pre-trained language model. But the difference we do see is minor in CoLA. As per our scientific and clinical

models we do not see a large difference in most of the datasets, excluding BioBERT's CoLA score. Demonstrating that the debiasing method does not have a negative affect on downstream task performance on both general and domain-specific models.

5.3 Exploratory Work

The experiment results were not always favorable for all applications of our debiasing method on clinical and scientific models. Tables 6 and 7 below denote stagnant or worse SEAT benchmark performance of debiased models for gender and race debiasing respectively.

Table 6: SEAT Benchmark gender debiasing results for BioBERT and BlueBERT. Performance worsens/stagnant after model has been debiased.

Models	SEAT-6	SEAT-6b	SEAT-7	SEAT-7b	SEAT-8	SEAT-8	|AVG|
BioBERT	0.21	-0.16	0.26	-0.34	0.03	0.02	0.17
BioBERT Debiased	0.18	0.05	-0.24	-0.31	0.14	0.10	**0.17**
BlueBERT	-0.01	- 0.10	0.73	-0.53	-0.25	0.33	**0.33**
BlueBERT Debiased	-0.03	-0.26	0.84	0.60	-0.19	-0.28	0.37

Table 7: SEAT Benchmark racial debiasing results for ClinicalBERT and BlueBERT. Performance worsens/stagnant after model has been debiased.

Models	SEAT-3	SEAT-3b	SEAT-4	SEAT-5	SEAT-5b	|AVG|
ClinicalBERT	0.37	0.44	0.67	0.04	0.36	**0.38**
ClinicalBERT Debiased	0.43	0.51	0.71	-0.69	-0.52	0.57
BlueBERT	0.36	-0.79	0.37	-0.37	0.90	**0.56**
BlueBERT Debiased	0.35	-0.78	0.46	-0.46	0.76	**0.56**

Gender debiased models BioBERT and BlueBERT did not show an improvement in benchmark performance. While BioBERT's score did not change, BlueBERT's worsened by an decrement of 0.04. Pertaining to racial debiasing, ClinicalBERT noticeably under performs compared to the other clinical and scientific models tested. Race debiased SciBert improved by 0.070 while

debiased ClinicalBERT performance actually worsened as the average SEAT benchmark score increased by 0.19, indicating more bias. When gender-debiasing as well, while Scibert improved by 0.06, ClinicalBERT fell short and only improved performance by 0.0206. In BlueBERT's case, while the debiased version's benchmark score didn't worsen, it remained stagnant, showcasing that the debiasing method did not have a positive effect on the model.

Further work may consist of looking into what debiasing method is better suited for these scientific models to generate a consistent improvement in benchmark scores for *both* gender and race debiasing. Exploration could also consist of initially looking at ClinicalBert, BioBERT, and BlueBERT's training datasets and mitigating the encoded biases present there.

6. CONCLUSION

In this work we obtained results that demonstrate the existence of bias in both non-specialized large language models and clinical models, and the effectiveness of our prompt design based debiasing method. Without the use of external corpora we were able to create biased prompts to extract bias from our models. Using the model's generated response we were then able to normalize its distribution disagreement and work at mitigating the effects of bias.

But we recognize the limits of hard prompts. The larger a language model is, the more parameters it has, making it difficult to manually design a discrete prompt that out-performs a model copy that is tuned to perform a certain task. To tackle this there are derivations of the prompting method that could lead to further advancing a model's performances. One such example is prompt tuning. Unlike our use of prompt design, which engineered a hard prompt made up of concrete real words, prompt tuning replaces the hard prompt with a continuous one — represented by a collection of tokens. Instead of consisting of 'real' words the tokens are tunable vectors. While the soft prompt isn't comprehensible it performs the same function of the hard prompt. Similar to prompt design we provide the model with an input consisting of our (vector) prompt and embedded input and compare the model's result with the expected target value. But this time, we take the loss we calculate and back-propagate it to generate gradient updates. These gradient updates are then applied to the tunable vectors. This means the continuous prompt can be optimized with more granularity than the hard prompt since it is not limited by words that exist, making it is much easier for the prompts to condense information [9]. Because of prompt tuning's granularity it is able to catch up to model-tuning performance as the size of the language model it's training on increases. It is no longer limited by words in the human vocabulary. Therefore, even as the model gets larger, prompts are able to be as effective. This is intuitive, as the larger the model the more adept it is in performing various tasks.

We can also take this work further by focusing more on the data sets that train our clinical

and scientific models. In this paper we used intrinsic metrics which focus on the up-stream language model [25]. But, we could focus on extrinsic metrics instead, which evaluate for fairness by comparing system predictions on downstream tasks; measuring for fairness by looking at downstream bias [26][25]. For our scientific models that could look like training SciBERT on MedNLI — a data set that provides a natural language inference task based on patient medical history — and then testing on a contructed Bias-MedNLI to check for bias.

REFERENCES

[1] Y. He, Z. Zhu, Y. Zhang, Q. Chen, and J. Caverlee, "Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition," 2020.

[2] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system," *BMC Medical Informatics and Decision Making*, vol. 6, 07 2006.

[3] A. Ali, K. Scior, V. Ratti, A. Strydom, M. King, and A. Hassiotis, "Discrimination and other barriers to accessing health care: Perspectives of patients with mild and moderate intellectual disability and their carers," *PLOS ONE*, vol. 8, pp. 1–13, 08 2013.

[4] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, and M. Ghassemi, "Hurtful words: Quantifying biases in clinical contextual word embeddings," 2020.

[5] I. Straw and C. Callison-Burch, "Artificial intelligence in mental health and the biases of language based models," *PLOS ONE*, vol. 15, pp. 1–19, 12 2020.

[6] K. Hamberg, "Gender bias in medicine," *Women's Health*, vol. 4, no. 3, pp. 237–243, 2008.

[7] R. A. Kessel, "Price discrimination in medicine," *The Journal of Law and Economics*, vol. 1, pp. 20–53, 1958.

[8] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," 2019.

[9] B. Lester and N. Constant, "Guiding frozen language models with learned soft prompts," Feb 2009.

[10] H. Chung and K. H. Park, "Lightweight prompt learning with general representation for rehearsal-free continual learning," in *Sixth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems*.

[11] R. Robinson, "Assessing gender bias in medical and scientific masked language models with stereoset," 2021.

[12] J. R. Minot, N. Cheney, M. Maier, D. C. Elbers, C. M. Danforth, and P. S. Dodds, "Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance," 2021.

[13] 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)*, (Dublin, Ireland), pp. 1012–1023, Association for Computational Linguistics, May 2022.

[14] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," 2021.

[15] N. Meade, E. Poole-Dayan, and S. Reddy, "An empirical survey of the effectiveness of debiasing techniques for pre-trained language models," 2021.

[16] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, "On measuring social biases in sentence encoders," 2019.

[17] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," 2018.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[19] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical bert embeddings," 2019.

[20] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, R. Kinney, S. Kohlmeier, K. Lo, T. Murray, H.-H. Ooi, M. Peters, J. Power, S. Skjonsberg, L. L. Wang, C. Wilhelm, Z. Yuan, M. van Zuylen, and O. Etzioni, "Construction of the literature graph in semantic scholar," 2018.

[21] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets," 2019.

[22] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234–1240, sep 2019.

[23] T. Manzini, Y. C. Lim, Y. Tsvetkov, and A. W. Black, "Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings," 2019.

[24] M. Kaneko and D. Bollegala, "Debiasing pre-trained contextualised embeddings," 2021.

[25] J. He, M. Xia, C. Fellbaum, and D. Chen, "Mabel: Attenuating gender bias using textual entailment data," 2022.

[26] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai, "Bias in bios," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, jan 2019.