M1 Internship Report 2024
Teaching Machine Fariness

By

Fatemeh Gol Pour

Supervisors:

Antoine Gourru

Ian Davidson

Machine Learning and Data Mining

Université Jean Monnet

Teaching Machine Fairness

1. Introduction

Natural Language Processing (NLP), a subfield of Artificial Intelligence (AI), aims to automate the processing of written text, covering tasks such as text analysis and generation (e.g. automatic translation). Deep learning, particularly the ”transformer” architecture found in the latest language models (ChatGPT, LLama, etc.), plays a crucial role in modern NLP. However, these models were shown to be biased by several studies(e.g. (Leteno et al., 2023)), including deep racial and gender stereotypes. The Diké Project focuses on Large Language Model fairness and ethic, and the impact of compression on these biases.

This internship is part of the ANR Diké project: Biases of compressed language models, which brings together the company NaverLab and the laboratories ERIC and Hubert Curien. The subject of the internship is Teaching Machine Fairness, which mainly focuses on mitigating bias against sensitive attributes such as gender, race, etc. in large language models. A model is considered ”unfair” if it produces results that benefit or harm a certain group of a sensitive attribute disproportionately.

1.1. Domains that can be affected by the bias

Bias in Machine Learning (ML) systems can have crucial impacts in different domains. Two examples of important areas that could be affected by this bias are:

- Recruitment Systems:

  Recruitment systems that use ML to accept candidates for jobs might favor profiles that align with the old hiring patterns and reject historically discriminated-against groups. This may result in less diversity in recommended candidates and limit the opportunities for underrepresented groups despite their competency.

- Recidivism Score Prediction:

  The recidivism score is used to estimate the probability of a criminal committing violent offenses after their release. Studies have shown that despite not being trained on attributes such as race, the model can still learn racial biases from other attributes. As a result, individuals from discriminated groups may be unfairly judged as higher risk, leading to harsher sentences, based on biased predictions, while actual dangerous individuals might be treated leniently, solely because of their race. One example of this is the COMPAS(Correctional Offender Management Profiling for Alternative Sanctions) application, where the scores tend to predict higher recidivism risk for African American defendants compared to white defendants

1.2. Bias in LLMs

In this work, we focus on mitigating biases of text classification algorithms that use (Large) Language Models. Below is an example from the Huggingface website of the gender bias observed in Bert.

```
unmasker("The man worked as a [MASK].")
```

Listing 1. bias in Bert - man

By running the line of code above, we can see the jobs Bert suggests for a ”man”:

Table 1. Jobs suggested by Bert for men

Score	Token	Token String
0.0975	10533	carpenter
0.0524	15610	waiter
0.0496	13362	barber
0.0379	15893	mechanic
0.0377	18968	salesman

Now let us try the same thing for a ”woman”.

```
unmasker("The woman worked as a [MASK].")
```

Listing 2. bias in Bert - woman

Table 2. Jobs suggested by Bert for women

Score	Token	Token String
0.2198	6821	nurse
0.1597	13877	waitress
0.1155	10850	maid
0.038	19215	prostitute
0.0304	5660	cook

We can observe a clear gender bias in the results we get from this experiment.

1.3. Related Work

Most prior works simply improve the fairness of classifiers by making them "forget" the sensitive attributes. (Leteno et al., 2023) introduce Wasserstein Fair Classification (WFC), a method for mitigating biases in neural text classification without requiring sensitive attribute annotations. WFC uses adversarial training and Wasserstein distance to enforce independence between learned representations of target labels and sensitive attributes. This bias-agnostic approach performs well on Bias in Bios and Moji datasets, achieving high accuracy while promoting fairness in text classification tasks.

(Rémi Nahon, 2023) introduces a new bias-agnostic approach to mitigate biases in deep neural networks (DNNs). Unlike traditional methods that require predefined bias information, this approach uses Voronoi cells to identify bias alignment/misalignment and then improve fairness in DNNs. It finds the optimal timing during the training of a vanilla model to extract information about the alignment between biases and target classes. This information is extracted when misclassified samples are farthest from the Voronoi boundary of their target class.

(Buyl & Bie, 2023) explore the challenges and limitations of achieving fairness in AI systems through technical solutions. It identifies eight key limitations, including biased datasets, categorization issues, the need for sensitive data, the lack of a universal fairness definition, and the potential for fairness measures to be misused. The study concludes that while technical approaches to AI fairness have benefits, they are not comprehensive solutions and must be complemented by broader socio-technical strategies and continuous evaluation.

2. Our Work

2.1. Datasets Used

For this internship, we mainly worked on two of the Laboratoire Hubert Curien Datasets on the Huggingface website, designed for fairness studies.

- Bias in Bios[1]:

  Made of textual biographies used for predicting professional occupations as labels, with 28 different professions included. The sensitive attribute is gender with 0 representing Male and 1 representing Female candidates.

- Moji[2]:

  Including tweets used for sentiment analysis, where the 1 labels show positive emotions and 0 labels represent negative ones, with information on the type of English used in the tweets as sensitive attribute (African-American English or Standard-American English).

2.2. Metrics

Notations: TP (True Positives) are the number of correctly predicted positive samples. TN (True Negatives) are the number of correctly predicted negative samples. FP (False Positives) are the number of incorrectly predicted positive samples. FN (False Negatives) are the number of incorrectly predicted negative samples. y is the true label. $\hat{y}$ is the predicted label. n is the total number of samples. y_i is the true value for sample i. $\hat{y}_i$ is the predicted value for sample i. s is the sensitive attribute. a is a specific group of the sensitive attribute. $\bar{a}$ is the complement of the value a of the sensitive attribute.

2.2.1. Evaluation Metrics

Below are the metrics used to evaluate our functions. For classification models we used Accuracy and F1-score, and for evaluating regression models, we used Mean Squared Error and Mean Absolute Error.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Mean Squared Error:

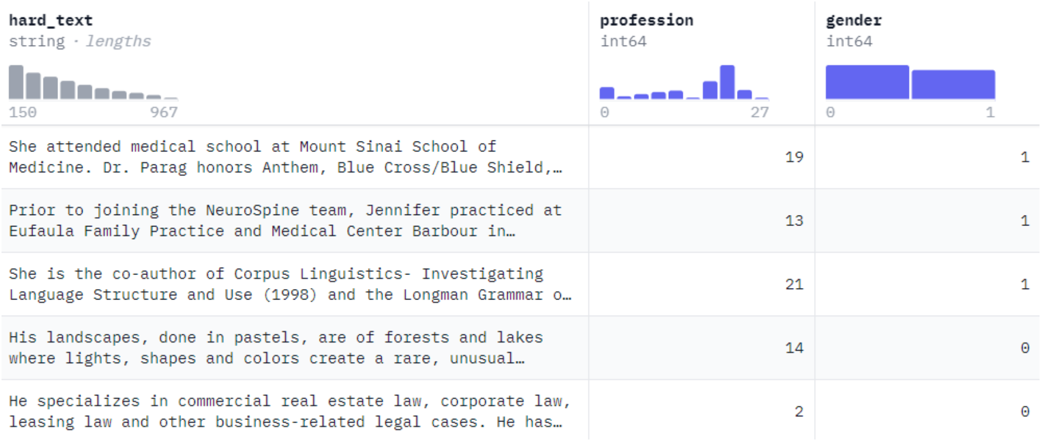
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Mean Absolute Error:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

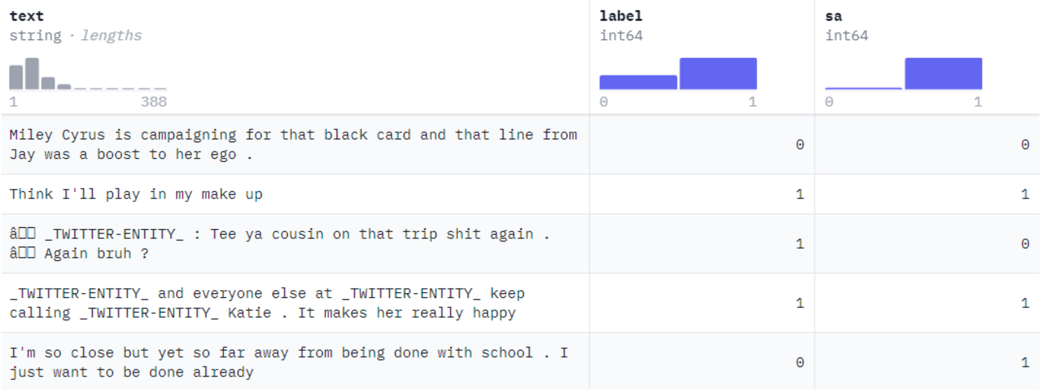
2.2.2. Fairness Metrics

We need metrics that will allow us to measure the fairness of a model and compare the results of our approach to the initial model and measure the improvements. Different fairness metrics can be used for various tasks, however, we have used the following metrics for our work:



hard_text string · lengths	profession int64	gender int64
She attended medical school at Mount Sinai School of Medicine. Dr. Parag honors Anthem, Blue Cross/Blue Shield,…	19	1
Prior to joining the NeuroSpine team, Jennifer practiced at Eufaula Family Practice and Medical Center Barbour in…	13	1
She is the co-author of Corpus Linguistics- Investigating Language Structure and Use (1998) and the Longman Grammar o…	21	1
His landscapes, done in pastels, are of forests and lakes where lights, shapes and colors create a rare, unusual…	14	0
He specializes in commercial real estate law, corporate law, leasing law and other business-related legal cases. He has…	2	0

Figure 1. Bias in Bios Dataset



text string · lengths	label int64	sa int64
Miley Cyrus is campaigning for that black card and that line from Jay was a boost to her ego .	0	0
Think I'll play in my make up	1	1
â□□ _TWITTER-ENTITY_ : Tee ya cousin on that trip shit again . â□□ Again bruh ?	1	0
TWITTER-ENTITY and everyone else at _TWITTER-ENTITY_ keep calling _TWITTER-ENTITY_ Katie . It makes her really happy	1	1
I'm so close but yet so far away from being done with school . I just want to be done already	0	1

Figure 2. Moji Dataset

Statistical Parity:

Measures the independence of the prediction from the sensitive attribute.

$$P(\hat{y} = 1 \mid s = a) - P(\hat{y} = 1 \mid s = \bar{a})$$

Equality of Opportunity:

Also called True Positive Rate Difference (TPRD), aims to measure the independence of the opportunities predicted by the model from the sensitive attribute, given that the true label is positive.

$$P(\hat{y} = 1 \mid y = 1, s = a) - P(\hat{y} = 1 \mid y = 1, s = \bar{a})$$

True Negative Rate Difference (TNRD):

Measures the independence of the negative predictions from the sensitive attribute, given that the true label is negative.

$$P(\hat{y} = 0 \mid y = 0, s = a) - P(\hat{y} = 0 \mid y = 0, s = \bar{a})$$

2.3. Models Used

The Large Language Models used in this work are the following:

- Bert:

  The LLM commonly used in our work was bert-base-uncased from the Huggingface website: https://huggingface.co/google-bert/bert-base-uncased

- GPT2:

  We used the gpt2 model for our further experiments to point out the bias in LLMs:

  https://huggingface.co/openai-community/gpt2

- Llama3: We will use the Llama3 model for our future experiments:

  https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

2.4. Context of the Project

In this project, we create a function $f(x)$ s.t. $x \in X : [0, 1]$ which is initialized with a language model and fine-tuned on our dataset. Unlike the previous studies on fairness, in our work, we try to train a morality function $g(x)$ s.t. $x \in X : [-1, +1]$ that *predicts* the unfair behavior of our initial model $f(x)$, as to whether each observation is likely to be discriminated or leniently treated based on its sensitive attribute.

The pipeline to train and use $g(x)$ shown in figure 3 is the following:

- Training a classifier $f(x)$ by fine-tuning a Large Language Model (e.g. a Bert model) on our dataset.

- Evaluating $f(x)$ to produce examples of discriminated observations (score +1), favored observations (-1), and fairly treated observations (0)

- Training the $g(x)$ function (using regression or classification) on the results obtained from $f(x)$

- Using $g(x)$ to minimally modify $f(x)$ and create a fairer model $f^*(x)$

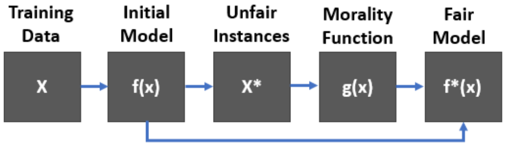


Figure 3. Methodology Pipeline

2.5. Implementation and Results

The following is the first implementation of the pipeline of our method and the corresponding results.

2.5.1. Creating and Evaluating Function $f(x)$:

First, we created function $f(x)$ by fine-tuning the Bert model on the bias-in-bios[1] dataset. For this purpose, we only took two of the professions as our labels, 0 representing Nurse, and 1 representing Physician. Figure4 shows the proportions of these jobs with respect to the sensitive attribute(gender).

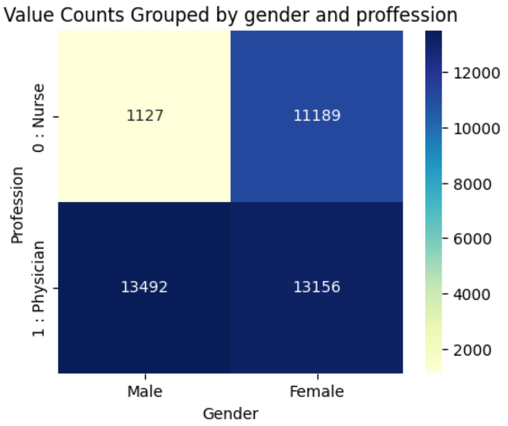


Figure 4.

After training the model on this subset of the dataset, we evaluated it.

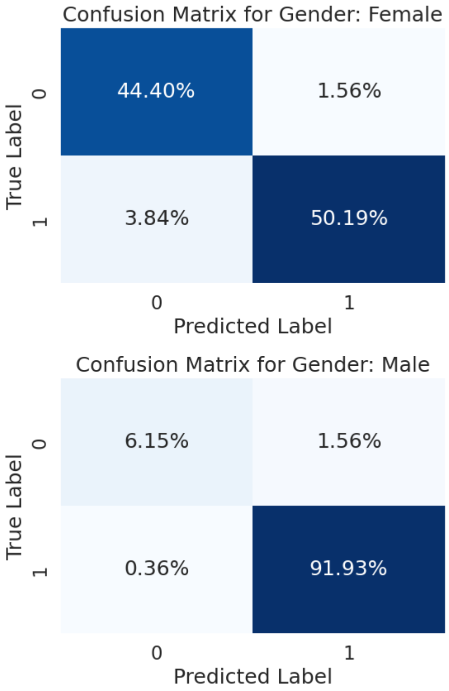


Figure 5. Confusion Matrices of $f(x)$ Based on Gender

Table 3. Evaluation Results of $f(x)$

Accuracy	F1	TPRD	TNRD
95.9%	96.98%	0.0673	-0.1688

We can clearly see by the misclassified results in figure 5 that the model is more likely to mistakenly classify men as

Physicians and women as Nurses, giving more opportunities to men who are not qualified as physicians and fewer opportunities to the qualified women, indicating the existence of gender bias in $f(x)$. This is also apparent by the results of our fairness metrics in table 3.

2.5.2. Using the Results of $f(x)$ to Create $g(x)$

We constructed a morality function $g(x)$ with the objective of producing a score between -1 and +1 to show the bias in $f(x)$. In order to build this function, we initialized $g(x)$ with Bert, then created a new dataset using the texts from the original dataset, and labels we created from the results of $f(x)$ as shown in table 4. We labeled the fair results as 0, defined the false negative results for women as discriminated instances and labeled them with +1, and finally defined the false positive results for men as favored instances and gave them the label of -1. We then used this new dataset to fine-tune $g(x)$. We created a classification and a regression model of $g(x)$.

Table 4. Labels of Dataset for g(x)

Status	Definition	Label
B: discriminated	FN for Females	+1
T: favored	FP for Males	-1
F: fair	True classifications	0

2.5.3. Evaluation of $g(x)$

After training the $g(x)$ function we evaluated it. Below are the results from our classification and regression methods:

- Results of $g(x)$ Classification:

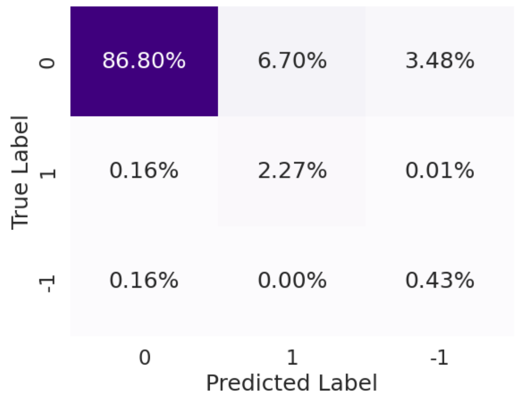


Figure 6. Confusion Matrix for $g(x)$ Classification

(Accuracy : 89.5% , F1 : 92.53%)

- Results of $g(x)$ Regression:

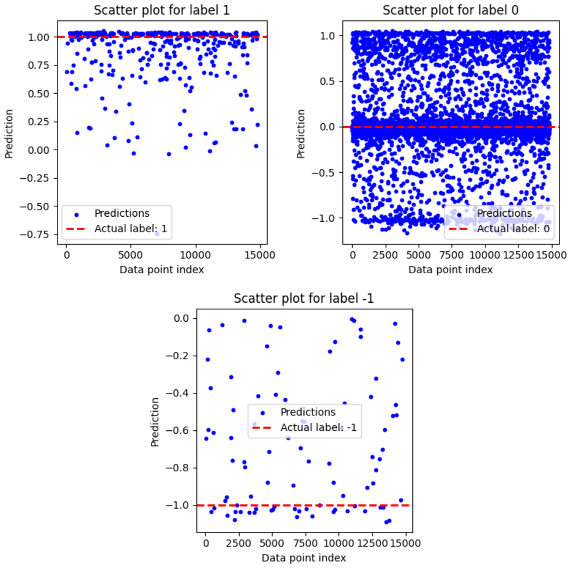


Figure 7. Evaluation of $g(x)$ Regression

Label	MSE	MAE
1	0.085%	0.154
0	0.078%	0.113
-1	0.23%	0.347
All	0.079%	0.115

Table 5. Evaluation Results of $g(x)$ Regression

Looking at the results, we can conclude that our $g(x)$ is fairly good at predicting the behavior of $f(x)$ (although not performing as good while predicting favored -1 labels), however, it might not be as accurate as we need it to be.

2.5.4. Using $g(x)$ to Create a Fair Function $f^*(x)$

To create a fairer function $f^*(x)$, the general formulation in our method was as follows, where ϵ is the upper bound of the allowable deviation from $f(x)$.

$$\text{Argmin}_{f(x)} - g(x)f^*(x)$$
$$\text{s.t.}\|f(x) - f^*(x)\| \leq \epsilon$$

In this formulation, we tried to make minimal changes in $f(x)$ such that the bias in $f^*(x)$ would be minimized. For every discriminated instance where a female has been misclassified as a nurse, we want to increase the result of $f^*(x)$ to predict a physician, and for all the favored instances where a male candidate has been predicted to be a physician, we want to teach our model to predict a nurse. Table 6 shows how $f^*(x)$ is changed with regards to the $f(x)$ and $g(x)$.

We implemented this formula with two different approaches that are explained below.

Table 6. Changes made in $f^*(x)$

	$g(x)$ = -1	$g(x)$ = +1
$f(x)$ = 0		↑
$f(x)$ = 1	↓	

- First Approach:

$$\text{Argmin}_{f(x)} - g(x)f^*(x)$$

In this approach $f^*(x)$ is the probability of being in the ”advantage” class. We use the classification $g(x)$ here and we do not consider the ”fair” predictions(0 labels) of $g(x)$. To prevent our new model from deviating too much from $f(x)$, we use a very small learning rate(5e-7).

Table 7. Comparing Evaluation Results of $f(x)$ and $f^*(x)$

Function	Accuracy	F1-score	TPRD	TNRD
$f(x)$	95.9%	96.98%	0.0673	-0.1688
$f^*(x)$	94.7%	96.25%	0.0042	-0.0762

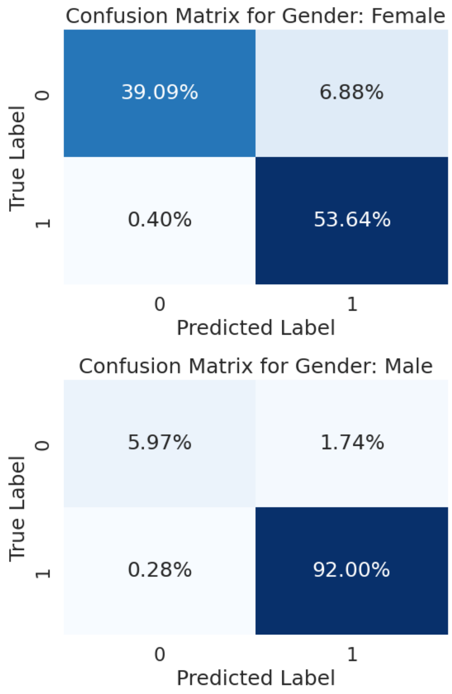


Figure 8. Confusion Matrices of $f^*(x)$ First Approach

By looking at the evaluation results in table 7, we can observe that although this approach has improved the fairness metrics, the accuracy and f1 score have dropped. This could be due to neglecting the fair results in $g(x)$. We can also see in figure 8 that the function seems to have learned a trend to the right, meaning that it is generally predicting more physicians in both male and female groups.

- Second Approach:

$$\underset{f(x)}{\text{Argmin}} - y_g \log(f^*(x)) + \beta(1 - y_g)\|f(x) - f^*(x)\|$$

Since our morality function $g(x)$ was not completely accurate, in this approach we decided to use y_g, the true labels assigned to $g(x)$ instead.In addition, to prevent the loss of accuracy, here we considered the fair results as well. The second term added is to prevent the model from changing when the results are already ”fair”, by penalizing the model if it deviates from $f(x)$ when y_g is 0. We added the log before $f^*(x)$ because the cross entropy was doing the same as our formula, therefore not making much of a difference. After performing Grid Search to fine-tune the hyperparameters, the best results were acquired using $\beta = 10$, learning_rate $=$ 5e-6, and number_of_epochs $= 2$.

By looking at the result from table 8 and figure 9, we can conclude that not only has the second approach improved the accuracy and f1 score of the model, it has also resulted in a much fairer model.

Table 8. Comparing Evaluation Results of $f(x)$ and $f^*(x)$

Function	Accuracy	F1-score	TPRD	TNRD
$f(x)$	95.9%	96.98%	0.0673	-0.1688
$f^*(x)$	96.3%	97.32%	0.0018	-0.0539

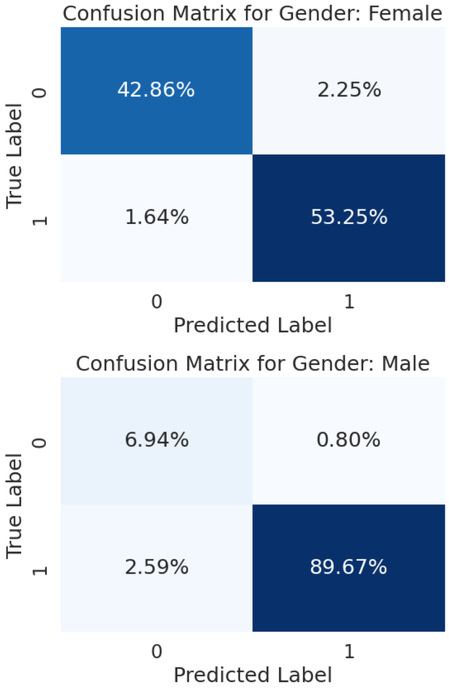


Figure 9. Confusion Matrices of $f^*(x)$ Second Approach

2.5.5. The Effectiveness of our Method

To see if our method for mitigating the bias is more effective than simply further training our model, we fine-tuned the initial $f(x)$ for two more epochs and compared the results with the best outcome we obtained from our experiments. Table 9 shows this comparison. The results indicate that our model is indeed working better in both reducing the bias and increasing the accuracy.

Table 9. Comparing Evaluation Results of $f(x)$ and $f^*(x)$

Function	Accuracy	F1-score	TPRD	TNRD
$f^*(x)$	96.3%	97.32%	0.0018	-0.0539
$f(x)$ 3epochs	95.93%	97.05%	0.0309	-0.1462

2.6. More Experiments on the Method

We continued to experiment our method on other datasets and LLMs. The following is a report on the different ideas we have implemented so far to create a function $f(x)$ with a more apparent bias in order to better show the effectiveness of our method.

2.6.1. Using Moji Dataset

- Creating and Evaluating Function $f(x)$:

  This time, we created function $f(x)$ by fine-tuning the Bert model on the Moji[2] dataset. The task performed on this dataset is sentiment analysis, where we have a dataset of tweets that are labeled as 0 or 1, indicating negative and positive tones respectively. The sensitive attribute here is the type of the language(African American English labeled as 0 and Standard American English labeled as 1). After training the model on this dataset, we evaluated it.

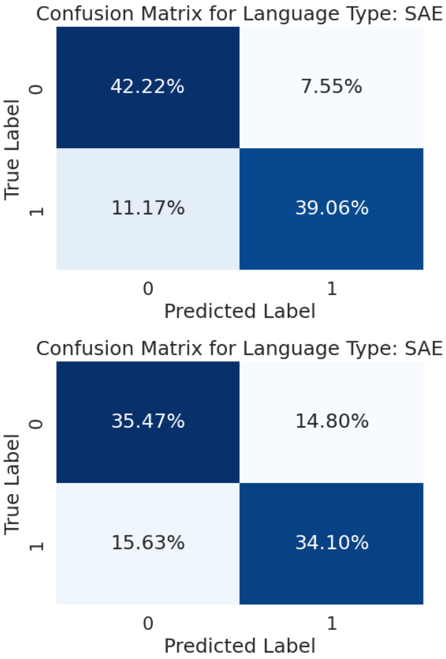


Figure 10. Confusion Matrices Based of Gender

Table 10. Evaluation Results of $f(x)$ on Moji

Accuracy	F1	TPRD	TNRD
75.7%	75.13%	0.0919	0.1427

- Using the Results of $f(x)$ to Create $g(x)$:

  In this dataset, we consider the African American English as the discriminated group of our sensitive attribute(language type).

Table 11. Labels of Dataset for g(x) Moji

Status	Definition	Label
B: discriminated	FN for African American English	+1
T: favored	FP for Standard American English	-1
F: fair	True classifications	0

- Evaluation of $g(x)$:

After training the $g(x)$ function we evaluated it. Below are the results from our classification and regression methods:

Results of $g(x)$ Classification:

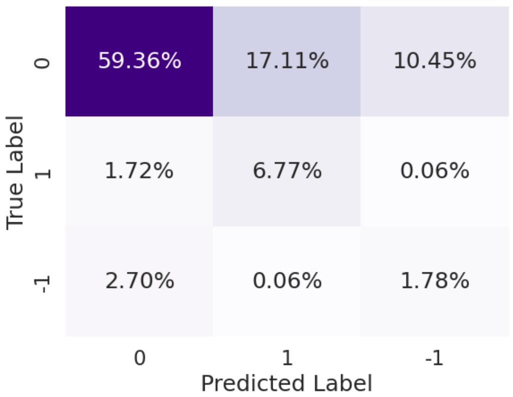


Figure 11. Confusion Matrix for $g(x)$ Classification Moji (Accuracy : 679.1% , F1 : 73%)

Results of $g(x)$ Regression:

Table 12. Evaluation Results of $G(X)$ Regression Moji

FUNCTION	MSE	MAE
$g(x)$	0.4252	0.5186

By looking at the results we can conclude that the f and g models need further tuning to produce better results.

2.6.2. USING A SUBSET OF BIAS IN BIOS

To show the bias in Bert more clearly, we decided to use a subset of the bias-in-bios dataset to create $f(x)$ this time. The results are as follows:

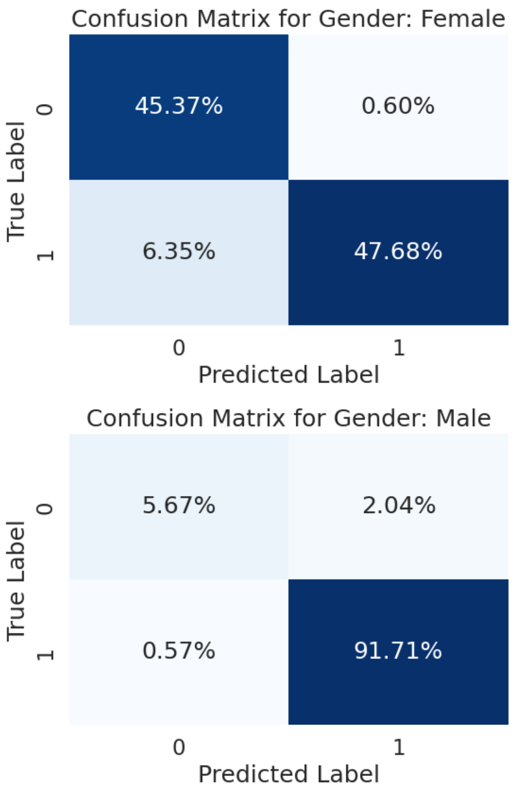


Figure 12. Confusion Matrices Based of Gender

Table 13. Results of Limiting the Data

ACCURACY	F1	TPRD	TNRD
94.68%	96.02%	0.1114	-0.252

2.6.3. FREEZING BERT

The next method we tried to show more bias was freezing the Bert model and only training the classification head on our dataset.

Table 14. Results of Freezing Bert

ACCURACY	F1	TPRD	TNRD
70.34%	82.16%	0.0024	-0.0718

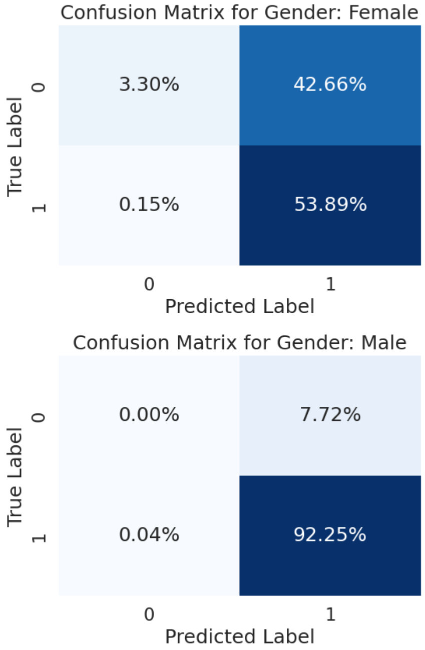


Figure 13. Confusion Matrices Based of Gender

2.6.4. Using GPT2

So far we had only used Bert as out initial model, however, in tasks like job prediction that we apply on the bias-in-bios dataset, the model is only predicting a label 0 or 1, while not knowing what that label represents. To display the actual bias of a language model in assigning more important roles to a certain group of the sensitive attribute, we used GPT2 which is a generative language model, to create our $f(x)$. In this method, $f(x)$ is initialized on GPT2 and first learns which job is associated with which biography in bias-in-bios, so later it can predict the actual profession. Below is the results from this experiment, and in figure 15 we can see some instances of the misclassified results.

Table 15. Results of GPT2

Accuracy	F1	TPRD	TNRD
94.99%	96.27%	0.0945	-0.1994

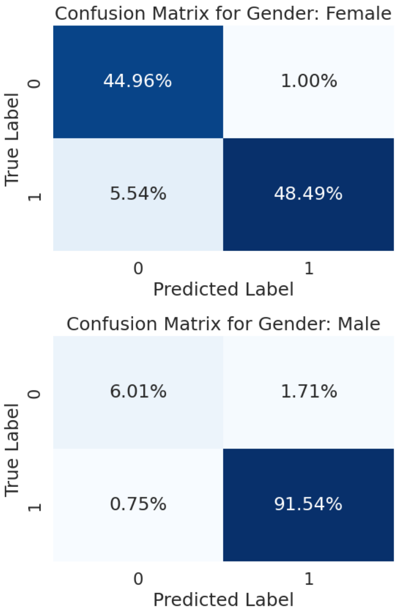


Figure 14. Confusion Matrices Based of Gender

As the evaluation results of each experiment show, it seems that using a subset of bias-in-bios and initializing $f(x)$ o GPT2 are the two most promising options to work on, as they represent the bias in the language model the best, while having acceptably good accuracies.

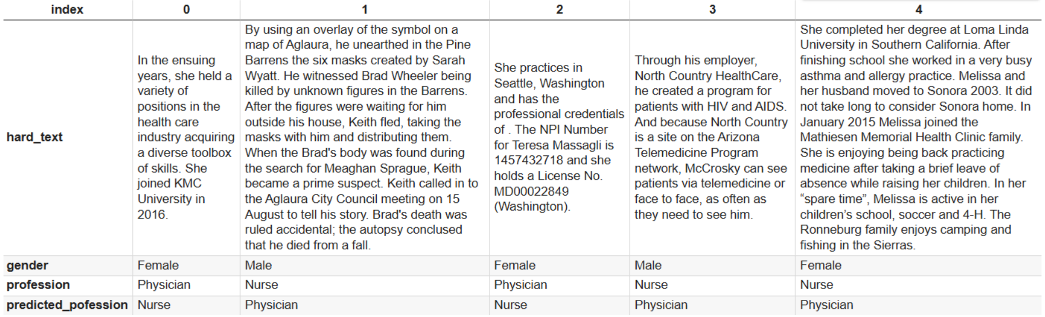
3. Conclusion

In this internship, we investigated methods to mitigate biases in text classification algorithms that use large language models (LLMs) such as BERT and GPT-2. Our approach involved creating a morality function $g(x)$ to predict and correct the unfair behavior of a classification model $f(x)$. By fine-tuning LLMs on biased datasets and employing fairness metrics, we aimed to develop a fairer model $f^*(x)$ with minimized bias.

Our results showed that while traditional training methods tend to keep biases present in the datasets, our approach using $g(x)$ successfully reduced these biases. Specifically, our second approach, which used true labels and penalized deviations from fair predictions, resulted in significant improvements in both accuracy and fairness metrics.

To summarize, the key findings from our experiments are as follows:

- **Model Accuracy and Fairness**: Our methods showed improved fairness metrics while increasing the accuracy and F1 scores of the models.
- **Comparison of Approaches**: The second approach, which utilized a combination of true labels and a regu-



index	0	1	2	3	4
hard_text	In the ensuing years, she held a variety of positions in the health care industry acquiring a diverse toolbox of skills. She joined KMC University in 2016.	By using an overlay of the symbol on a map of Aglaura, he unearthed in the Pine Barrens the six masks created by Sarah Wyatt. He witnessed Brad Wheeler being killed by unknown figures in the Barrens. After the figures were waiting for him outside his house, Keith fled, taking the masks with him and distributing them. When the Brad's body was found during the search for Meaghan Sprague, Keith became a prime suspect. Keith called in to the Aglaura City Council meeting on 15 August to tell his story. Brad's death was ruled accidental; the autopsy conclused that he died from a fall.	She practices in Seattle, Washington and has the professional credentials of . The NPI Number for Teresa Massagli is 1457432718 and she holds a License No. MD00022849 (Washington).	Through his employer, North Country HealthCare, he created a program for patients with HIV and AIDS. And because North Country is a site on the Arizona Telemedicine Program network, McCrosky can see patients via telemedicine or face to face, as often as they need to see him.	She completed her degree at Loma Linda University in Southern California. After finishing school she worked in a very busy asthma and allergy practice. Melissa and her husband moved to Sonora 2003. It did not take long to consider Sonora home. In January 2015 Melissa joined the Mathiesen Memorial Health Clinic family. She is enjoying being back practicing medicine after taking a brief leave of absence while raising her children. In her “spare time”, Melissa is active in her children’s school, soccer and 4-H. The Ronneburg family enjoys camping and fishing in the Sierras.
gender	Female	Male	Female	Male	Female
profession	Physician	Nurse	Physician	Nurse	Nurse
predicted_pofession	Nurse	Physician	Nurse	Physician	Physician

Figure 15. Examples of misclassified instances

larization term to penalize changing fair results, proved to be the most effective in balancing accuracy and bias reduction.

- **Effectiveness**: The effectiveness of our method was validated in comparison to training the model for more epochs.

Table 16 presents a comparative overview of the classification accuracies, F1 scores, and fairness metrics for the various experiments we have done:

Table 16. Comparison of all the Approaches Implemented

MODEL	ACCURACY	F1	TPRD	TNRD
$f(x)$ BIB1	95.9%	96.98%	0.0673	-0.1688
$f^*(x)$ 1	94.7%	96.25%	0.0042	-0.0762
$f^*(x)$ 2	96.3%	97.32%	0.0018	-0.0539
$f(x)$ 3E	95.93%	97.05%	0.0309	-0.1462
$f(x)$ LIMITED	94.68%	96.02%	0.1114	-0.252
$f(x)$ FROZEN	70.34%	82.16%	0.0024	-0.0718
$f(x)$ GPT2	94.99%	96.27%	0.0945	-0.1994
$f(x)$ MOJI	75.7%	75.13%	0.0919	0.1427

4. Future Work

For the rest of this internship, we are planning to work on the following areas:

4.1. Using $g(x)$ Regression

We plan on improving the $g(x)$ function as much as possible and then using $g(x)$ regression to improve $f^*(x)$. This can result in more promising outcomes, as a well-trained g(x) regression could give scores from -1 to +1. proportional to how unfair the results of $f(x)$ are, therefore making changes in $f^*(x)$ accordingly.

4.2. Continuing the Experiments

As we observed in section **2.5.**, the two methods of creating $f(x)$, namely using a subset of bias-in-bios and using GPT2, have shown potential for being subject to further experiments of our work. We plan to complete the pipeline of our methodology on these experiments.

4.3. Working on Llama3

Finally, we are also planning to work on Llama3 with qlora as our second generative model and evaluate our methodology on it.

Acknowledgements

I would like to express my deepest gratitude to my supervisors for their invaluable guidance, support, and encouragement throughout the course of this internship. Their expertise and insights have been instrumental in shaping this work.

I am also grateful to all my professors in Master MLDM for their wisdom, constructive feedback, and the knowledge they have given me, which has been crucial in the completion of this project.

Additionally, I would like to give my heartfelt thanks to my friends and colleagues, who have been a constant source of support and assistance, for their patience in answering my questions, offering suggestions, and providing moral support during challenging times.

Thank you all for your contributions and encouragement, which have made this journey possible.

References

Buyl, M. and Bie, T. D. Inherent limitations of ai fairness. *arXiv preprint arXiv:2212.06495*, 2023. URL https://arxiv.org/abs/2212.06495.

Leteno, T., Gourru, A., Laclau, C., Emonet, R., and Gravier, C. Fair text classification with wasserstein independence. In *2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15790–15803. Association for Computational Linguistics, 2023. URL https://arxiv.org/abs/2311.12689.

Muller, B. Bert 101 state of the art nlp model explained, 2022. URL https://huggingface.co/blog/bert-101#bert-101-. Accessed: 2024-06-25.

Rémi Nahon, Van-Tam Nguyen, E. T. Mining bias-target alignment from voronoi cells. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. LTCI, Télécom Paris, Institut Polytechnique de Paris, France, 2023. URL https://arxiv.org/abs/2305.03691.