Fairness in Machine Learning: Detecting and Removing Gender Bias in Language Models

Carson Sue, Adam Miyauchi, Kunal S Kasodekar,
Sai Prathik Mandyala, Priyal Padheriya, Aesha Shah

6 December 2022

Abstract

With the increasing presence of AI models in our lives, it has surfaced that the biases that humans may hold can be learned and replicated by these models. Bias in AI can negatively affect already marginalized minorities in a community, and thus efforts have been made to mitigate bias in AI. For our work, we focus on debiasing gender discrepancies in language models. Using previous work on this topic as a base, we developed and tested two different methods: debiasing the dataset and debiasing the corpus embeddings of the model (hard debiasing). We implemented and tested both methods on the BERT model using the Affect in Tweets dataset [14]. Using the Equity Evaluation Corpus, we calculate the total bias for a model without any debiasing and for models using our methods. From our testing, we found that both methods were effective in reducing bias in the model. Dataset debiasing reduced the bias by 29.41% and hard debiasing by 11.76% with a classification accuracy loss of 2.47% and 3.61% respectively.

1 Introduction

Machine learning models are becoming more and more prevalent in our lives and are responsible for making crucial judgments and decisions that have an impact on many people. Some examples include approving loans, various uses in the law, and facial recognition. ProPublica found that the COMPAS tool [1], an algorithm used in some US courts to determine which defendants are most likely to become repeat offenders, was biased towards a certain race. This was due to the facial recognition models having divergent error rates [2] across demographic groups due to their under-representation in the dataset. This was most pronounced for dark-skinned females, whose error rates were 34% higher than those of their light-skinned male counterparts. Surveillance models were trained using datasets with a high majority of images from dark-skinned people, further increasing bias towards them. Thus, inherent bias in the dataset and model can further amplify pre-existing biases towards these demographics [2].

Another example of human bias is in credit. It may be surprising to know that women with better credit scores and similar income and expenses as their male counterparts have

a smaller credit limit. AI systems used by such firms lack explainability and the bias from the data (gender imbalance in data points) seeps into their models. There are many such examples and relevant studies that explore the topic of gender bias across AI. We are looking to detect and mitigate gender bias in Langauge Models for a particular language task.

2 Background

Debiasing in the context of NLP can be done initially in the dataset and then at the Model level. To debias the dataset, a variety of methods have been put forward. For gender debiasing, Diversifying / Augmenting a dataset is currently the most employed method. This method involves generating auxiliary datasets where the gendered entities are swapped and employing training methodologies by combining this with the original dataset. This method would ensure that both the genders are fairly represented.

Debiasing methods at a model level mainly involve modification of the vector representations of the word embeddings. Some ways remove multiple gender dimensions from the vectors to attenuate gender bias, or by shifting the vector to be equally male and female.Algorithms like Hard/Double hard debiasing show promising results in removing bias while not eliminating information that contains necessary information about genders. Some previous studies on mitigating and detecting gender bias are given below:

- **Coreference resolution and gender bias**: Coreference resolution is the process of identifying all the expressions in a text that all refer to the same entity. When coreference resolution was evaluated on the WinoBias dataset [9] for the same contextual text it was observed that in case of anti-stereotypical roles the correct linking prediction was not made and there was bias amplification by the model.

- **Gender Bias in vSRL Tasks:** Visual Semantic Role Labelling tasks have an inherent bias in the dataset with 33% of Cooking Images being Male and 67% being Female. Even when images consist of a Male Agent cooking food in most cases the Agent is detected as a Female [10] by the model. When a similarly distributed dataset is evaluated by the model the results are further skewed with a 16:84 ratio of Male: Female cooking images resulting in bias amplification.

3 Problem Description

We aim to develop methods that can debias gender discrepancies in language models. We have worked on two different approaches: debiasing the dataset and debiasing the corpus embeddings in the model. Debiasing the dataset aims to prevent the model from learning biases in the first place, while debiasing the embeddings aims to fix the bias that the model has leaned. We debias the dataset by generating new samples of gendered words with their gendered counterpart and obscuring names of people. For the corpus embeddings, we debias by neutralization and equalization of the bias. Both methods will be explained in further detail in their respective sections. In addition, there must be a distinction between gendered words with appropriate bias and non-gendered words with inappropriate bias. For

example, the words “king” and “queen” have appropriate gender bias as gender is part of their definition. However, words such as “scientist” and “housekeeper” can have gender bias towards that isn’t inherent to their definition, thus being inappropriate. It is imperative that gendered words are not debiased while non-gendered words are, and thus we need to be able to distinguish the two.

4 Dataset Description

For this project we have used an emotion classsification data known as The Affects in Tweets Dataset [14]. The dataset is a compilation of around 5847 unique tweets. The collection of corpus of the tweet dataset was done by polling the Twitter API. The tweets containing words related to emotion like annoyed, happy, angry, surprised were collected. Below figure represents a word cloud generated from the tweets:

Figure 1: Word Cloud

We have used this dataset as it aligns with the need of the problem statement we are trying to solve. It can be observed that in the raw dataset each tweet specifically falls under four emotional categories - Anger, Sadness, Fear and Joy. The graph below shows the distribution of these emotional categories in the raw data:

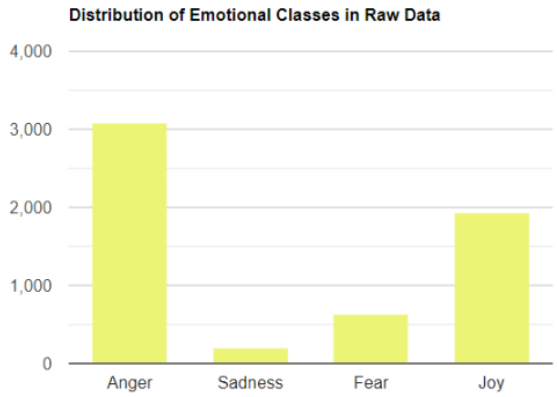


Figure 2: Distribution of Emotional Categories

5 Dataset Debiasing

We first explore removing gender bias by altering the underlying training dataset. The first step to gender debias our dataset involved balancing the occurrences of gendered terms. We gathered a collection of gendered word pairs from [15]. Figure 3 provides some example gendered word pairs included in the collection. For every occurrence of a gender word in the

Male	Female
man	women
father	mother
daughter	son

Figure 3: Example Gendered Word Pairs

original Tweet training set, we generated a new sample where the gendered word is replaced with the gendered word counterpart. Figure 4 provides an example of this augmentation process.

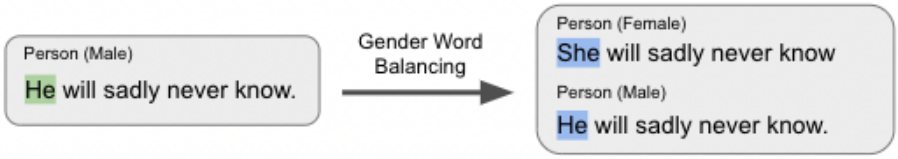


Figure 4: Example Gender Word Balancing

In addition to gender balancing, we also gender-neutralized people's names. We identified names in the training corpus using a BERT named entity recognition model. We replaced all person entities identified by the NER model with the gender-neutral word 'NAME'. Figure 5 provides an example of this process. The resulting training dataset contains an equal

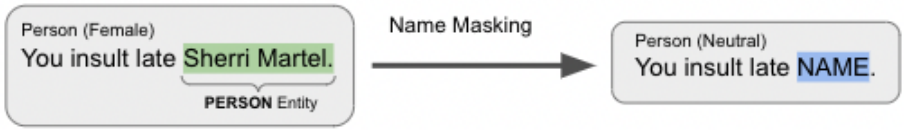


Figure 5: Example Name Masking

representation of both male and female genders.

With the gender-balanced dataset created, we trained a BERT Base model for our emotion classification task. BERT or Bidirectional Encoder Representations from Transformers, is a transformer-based machine learning model developed by Google in 2018 that is widely used for multiple natural language processing tasks. In this study, we utilized a publicly available pre-trained BERT Base model to classify the emotion of Tweets. We accessed the model through the BERT-Sklearn[1] library, which provides an easy-to-use interface for pre-

[1]BERT-Sklearn: `https://github.com/charles9n/bert-sklearn`

trained models hosted on HuggingFace[2]. The pre-trained model consists of 12 layers, 768 hidden, 12 heads, 110 million parameters, and a vocabulary size of 30,522. [16] We adapted the pre-trained model to our emotion classification task by adding a single fully connected layer. We trained the model for three epochs using a learning rate of 2e-5 and batch size of 32. We present the results of the outlined dataset debiasing methods in Section 8.

6 Hard Debiasing

Word Embeddings are learned numerical vector representations of words in the vocabulary. These embeddings are learned by training neural network models such as word2vec with a skip-gram of CBOW scheme and picking the hidden representations as our embeddings. Word embeddings can capture simple relations between corpus words using vector arithmetic. More conceptually, similar word embedding vectors have larger cosine similarities than unsimilar words.

Human-generated corpora have an inherent bias in the data. This bias comes in varied formats from gender, racial bias and ethnic bias. Word embeddings created using these corpora's data consequently carry considerable bias, particularly strong gender bias. This bias is further amplified through downstream models and tasks. When a word is gender-neutral by definition but has a learnt embedding that is more closely associated with a particular gender, the word has a gender bias [12]. More technically we define the gender bias of the word w by its projection on the gender direction as follows:

$\vec{w} = \vec{he} - \vec{she}$ *(assuming normalized vectors)*

The larger the projection the more biased the word is. We define gender direction as the difference between the she embedding vector and the he embedding vector. Direct bias is defined as simply summing up the cosine of the angle between neutral words and the gender direction. Generally speaking embeddings without any bias would have zero projection in the gender direction. Thus gender neutral terms ought to be equidistant from the he-she gender embedding vectors [12].

Our goal in this project is to identify the gender bias in our dataset and debias the data initially and then debias the word embeddings from our dataset corpus. We explore several debiasing approaches after debiasing our data to debias the word embeddings for our dataset corpus. In order to reduce bias amplification in downstream tasks and maintain all the semantic linkages between the embeddings, we experiment with hard debiasing strategies here.

In this project we employ **Hard Debiasing** that is the method of ***Neutralizing and Equalizing*** the gender-neutral vectors in this gender subspace. Conceptually, our goal is to make sure that the gender-neutral terms are equidistant from the "he" and "she" equality pairs as this implies no gender bias.

In order to debias the word embeddings initially we need to identify the gender subspace [11]. We do this by taking the difference of some of the pre-known gender pairs that define

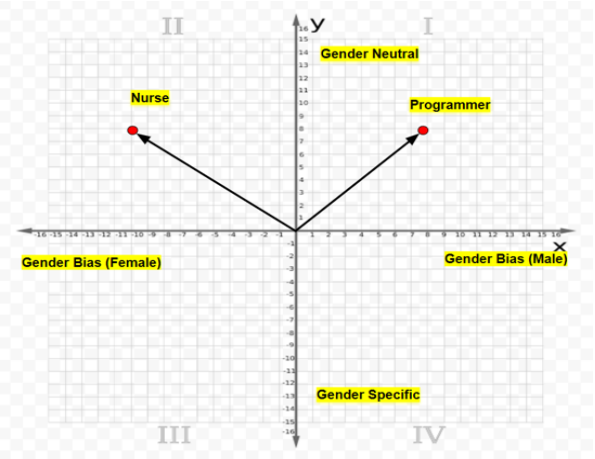
[2]HuggingFace: `https://huggingface.co/`

the concepts of gender. Some of these sets are:

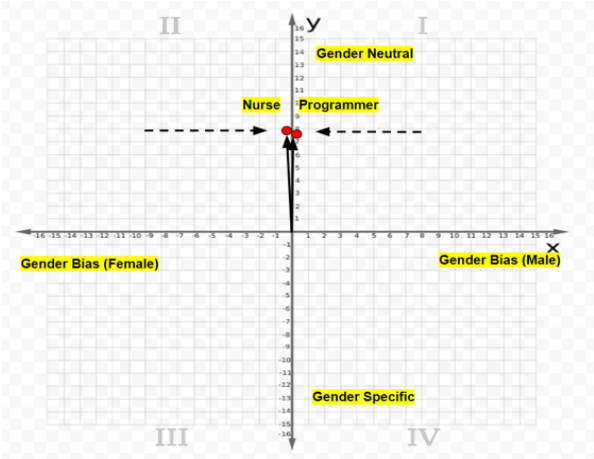
$\vec{girl} - \vec{boy}$
$\vec{female} - \vec{male}$
$\vec{mother} - \vec{father}$
$\vec{daughter} - \vec{son}$
$\vec{gal} - \vec{guy}$

We obtain these subsets of equality pairs and perform SVD on these opposite gender pairs to obtain the direction of the gender subspace. Then we obtain the word embeddings (size=300) of our dataset corpus by training a Word2Vec model using nltk. After defining the vocab of our corpus we do the following:

- We employ a combination of neutralisation followed by equalization on our corpus embeddings. We define and calculate direct bias by taking a dot product of gender direction ($\vec{man} - \vec{woman}$) of gender-specific words like *"he"* and *"she"* with gender-neutral words like "programmer" and "nurse" [11]. We created a JSON file with a large list of such gender-specific and gender-neutral words.
- For neutralisation initially we identify the gender subspace by utilizing our gender-specific pairs.

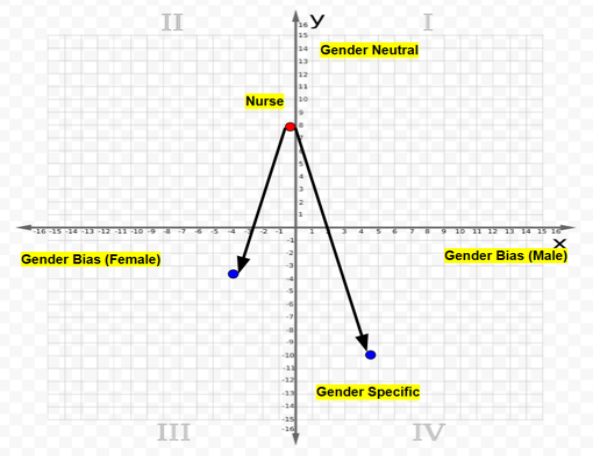


(a) Before Neutralisation

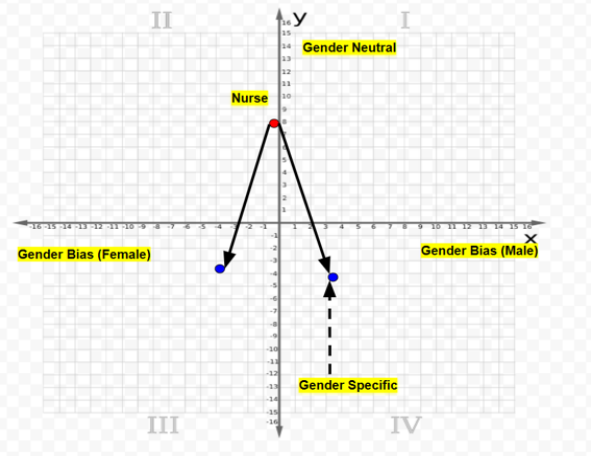


(b) After Neutralisation

- Then we nullify the magnitude of components of gender-neutral components to a minimum value through subtraction of their specific components in this subspace [11][13].
- For equalization, we equalize the vector distance of all gender-neutral words outside the subspace from all our gender-specific subspaces using vector algebra on their projections [11][13]. Finally, we would hard-debias the corpus embeddings using a combination of these *Neutralization + Equalization*. These debiased word embeddings generated after this process would be stored by us and then used for further downstream tasks and models thus minimizing bias amplification and propagation.



(a) Before Equalisation



(b) After Neutralisation

7 Measuring Gender Bias

To measure gender bias, we subject our models to the Equity Evaluation Corpus (EEC) [17]. The EEC consists of sentence pairs that only differ by a gendered term. For example, the sentence pair "He feels irritated" and "She feels irritated" only differ by the gendered words "he" and "she". Altering only the gendered part of the sentence should not change the model's prediction. Using this idea, we defined two quantitative metrics to measure gender bias.

$$Bias_{M,F}(s) = \begin{cases} 0 & classification(s_M) = classification(s_F) \\ 1 & otherwise \end{cases} \tag{1}$$

Equation 1 measures the bias of a single input pair s. If the model produces the same classification for the male and female versions of the test sample, then we say there is no bias. Otherwise, we say there is a bias of 1. To measure the total bias of the EEC, we defined Equation 2.

$$Total\ Bias_{M,F}(sc) = \frac{1}{N} \sum_{i=0}^{N} Bias_{M,F}(s_i) \tag{2}$$

Equation 2 represents the arithmetic mean of all test pairs in the Equity Evaluation Corpus.

8 Results and Evaluation

The main objective was to perform Gender Debiasing while also preserving the level of performance of the original model. To measure this we performed Gender Debiasing and measured both the pre and post-Debiasing classification accuracies of the model.

We performed both Gender Debiasing techniques - Dataset Debiasing and Hard Debiasing. Then we calculated the respective Gender Biases as per the formulation proposed in Section 7. We also made note of the Classification Accuracies. By calculating the bias in the baseline model, we could perform a comparative analysis of both methods.

8.1 Baseline Performance and Bias

Each tweet in the Dataset belonged to one of four distinct classes of emotions -Fear, Anger, Sadness, and Joy. We used a baseline BERT Classifier on the original dataset to classify each tweet based on its emotional Class.We also calculated the baseline levels of Gender Bias in each class.The results are tabulated in Table 1.

Classification Accuracy - Baseline			0.83
Quantitative measure of Bias in each class			
Fear	Anger	Sadness	Joy
19	10	117	1

Table 1: Baseline Metrics

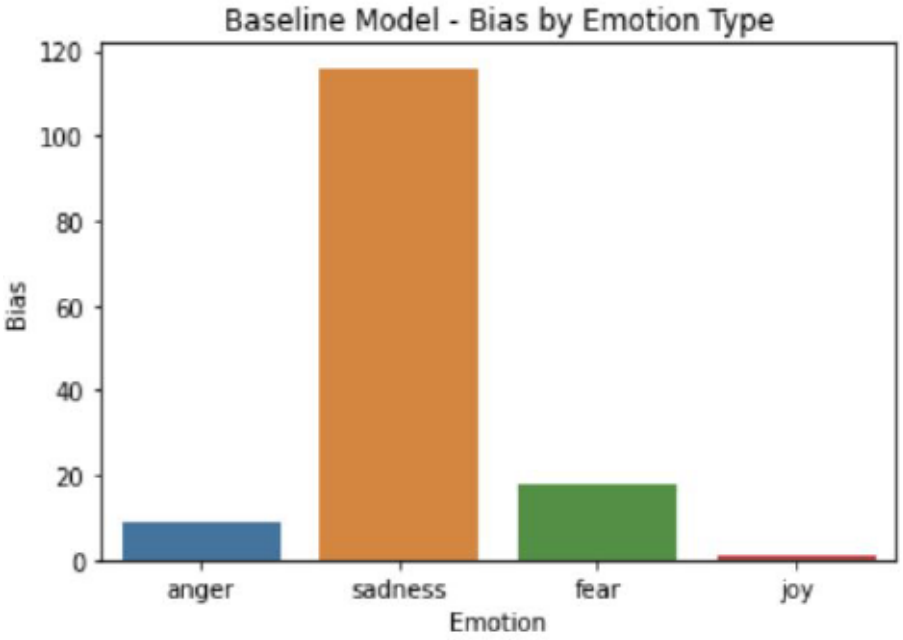


Figure 8: Baseline Bias Histogram

8.2 Dataset Debiasing Performance and Bias

Dataset Debiasing is a pre-processing method. We applied the methods mentioned in section 5 on the Raw Dataset post initial data wrangling. Then we measured the performance of the model and the bias in the results.The results have been tabulated in Table 2.

Classification Accuracy - Baseline			0.81
Quantitative measure of Bias in each class			
Fear	Anger	Sadness	Joy
49	30	20	1

Table 2: Dataset Debiasing Results

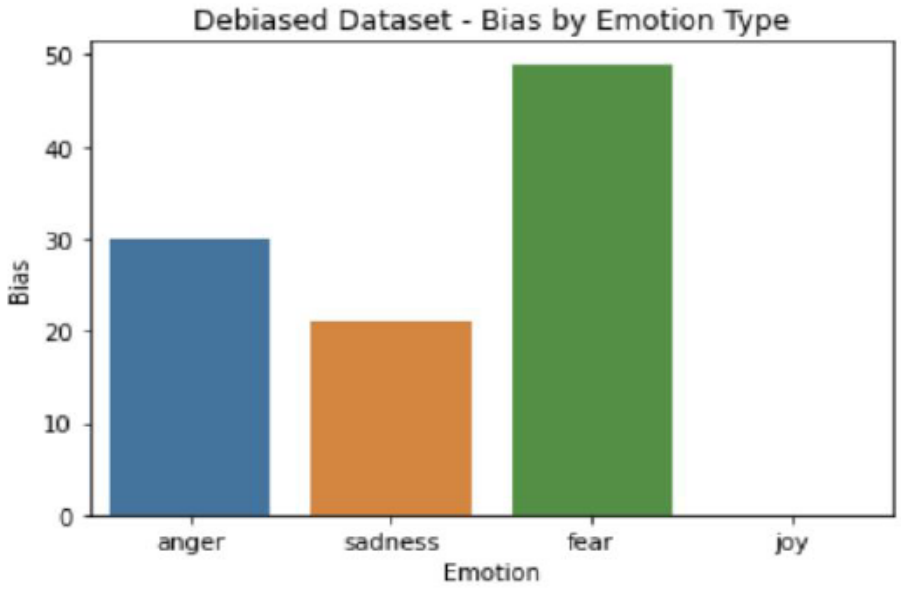


Figure 9: Bias Post Dataset Debiasing

8.3 Hard Debiasing Performance and Bias

Hard Debiasing is a post-processing method. We initially generated the word embeddings using a standard Word2Vec model. Then we applied the methods mentioned in section 6 on these embeddings, following which we loaded the embeddings and performed our classification task. Then we measured the performance of the model and the bias in the results. The results have been tabulated in Table 3.

Classification Accuracy - Baseline			0.80
Quantitative measure of Bias in each class			
Fear	Anger	Sadness	Joy
58	38	25	1

Table 2: Hard Debiasing Results

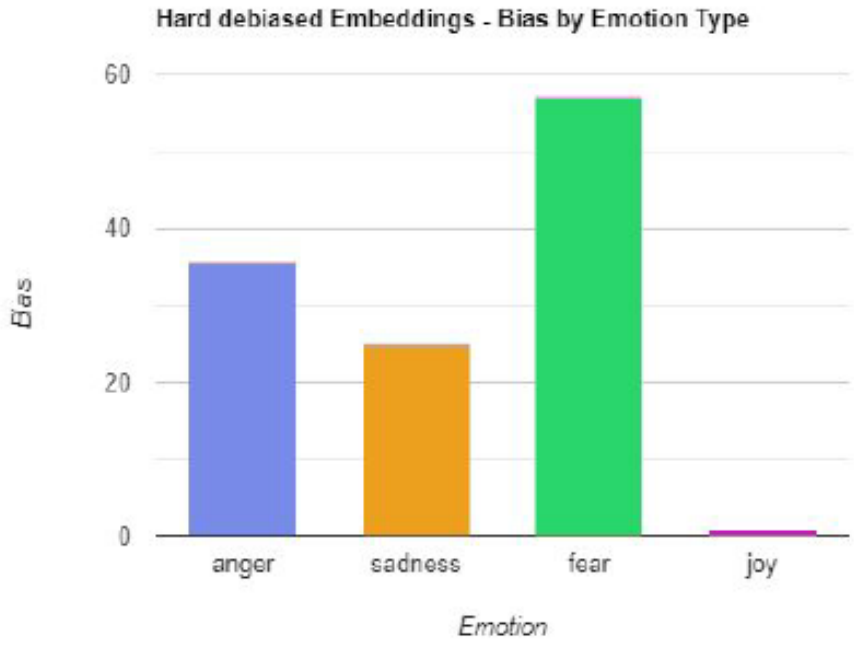


Figure 10: Bias Post Hard Debiasing

8.4 Comparative Analysis

We could see that, compared to baseline levels of total bias, there was a reduction in bias when we applied hard and dataset debiasing. A comparative change in the percentage of Bias can be seen in below figure.

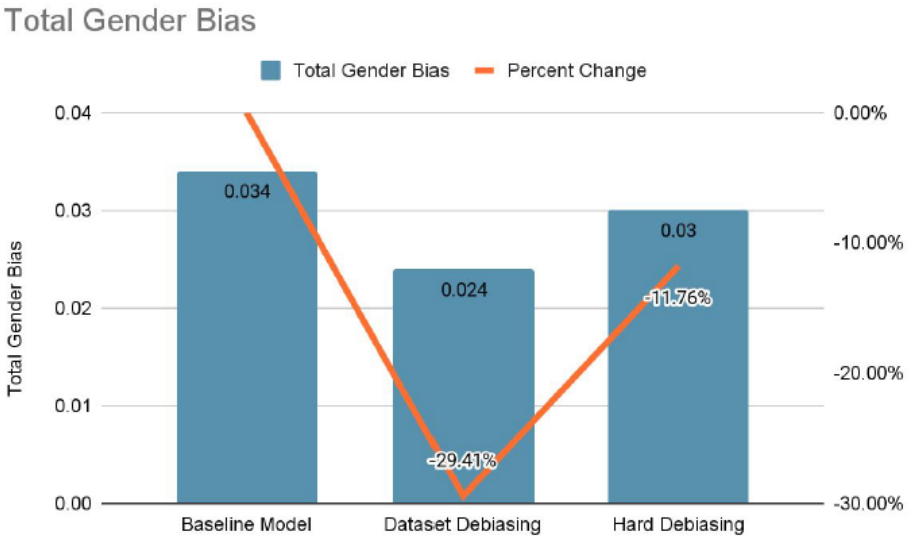


Figure 11: Percentage Change of total bias in each method

Calculating the accuracy of the resulting model post each debiasing method, we noticed that there was no significant change in levels of performance. This is along the expected lines of our hypothesis, a comparison of the results can be seen in the figure below.

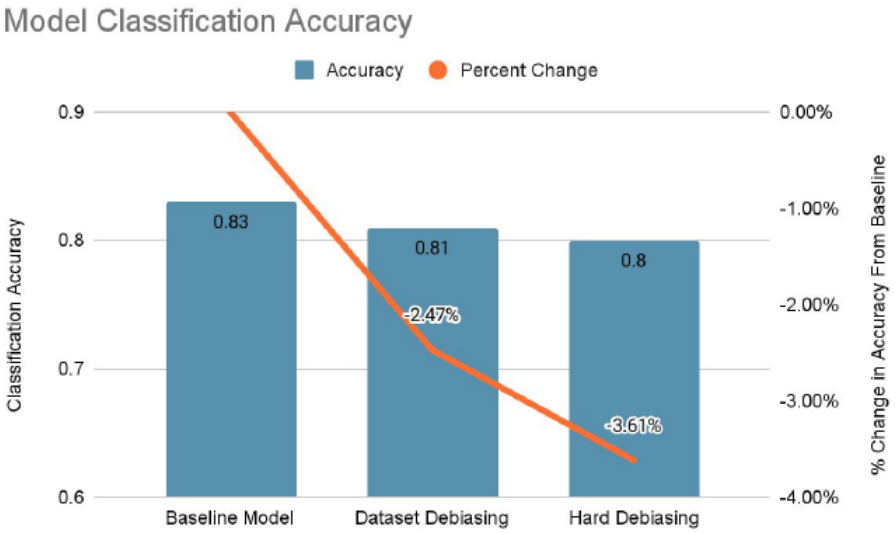


Figure 12: Comparison of Model Performance

9 Conclusions and Future Work

We derive the following two conclusions for the debiasing methods we implemented -

- **Method 1 -** The inherent bias in the dataset was reduced by **29.4%** by creating gender-specific counterparts, masking names and training BERT-NER to predict those names
- **Method 2 -** We reduce bias by **12%** in the embeddings by employing neutralise and equalise (Hard) debiasing.

We further observe that despite employing sophisticated methods like hard debiasing, the overall performance of the method was more or less the same as Dataset debiasing for this particular dataset. Neither method could completely eliminate bias. One potential reason for this outcome could be that since we are debiasing based on the narrow filter of gender, there also exist other biases like race, class, etc. that could have shared biasing influences on emotion classification.

Proposed below are a few methods that could be implemented as future work to further enhance the performance and help eliminate bias as much as possible.

- **Adversarial Debiasing of LLM:** Adversarial debiasing adds an adversarial loss function to the current loss function in an effort to make our sensitive attributes that utilize hidden correlations less predictable.
- **Ensemble Models for Bias Mitigation:** We will attempt to train an ensemble of several LLMs in which, even if each base model is biased, their averaged model may be fair as those biases can cancel each other out.

10 References

1. Pandey, P. (2019, August 6). Is your machine learning model biased? Medium. Retrieved December 6, 2022, from https://towardsdatascience.com/is-your-machine-learning-model-biased-94f9ee176b67

2. Najibi, A. (2020, October 26). Racial discrimination in face recognition technology. Science in the News. Retrieved December 6, 2022, from https://sitn.hms.harvard.edu/flash/2020/raci discrimination-in-face-recognition-technology/

3. Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. KDD 2015.

4. Datta A., Sen S., Zick Y. (2017) Algorithmic Transparency via Quantitative Input Influence. In: Cerquitelli T., Quercia D., Pasquale F. (eds) Transparent Data Mining for Big and Small Data. Studies in Big Data, vol 32. Springer, Cham

5. CS 294: Fairness in Machine Learning. CS 294 Fairness in Machine Learning. (n.d.). Retrieved September 30, 2022, from https://fairmlclass.github.io/

6. Gaut, A., Sun, T., Tang, S., Huang, Y., Qian, J., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., amp; Wang, W. Y. (2020, August 8). Towards understanding gender bias in relation extraction. arXiv.org. Retrieved September 30, 2022, from https://doi.org/10.48550/arxiv.1911.03642

7. Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., amp; Ordonez, V. (2019, October 11). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. arXiv.org. Retrieved September 30, 2022, from https://arxiv.org/abs/1811.08489

8. CS224N: Natural Language Processing with deep learning. Stanford CS 224N — Natural Language Processing with Deep Learning. (n.d.). Retrieved September 30, 2022, from https://web.stanford.edu/class/cs224n/

9. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

10. Bias EEC. Saif. (n.d.). Retrieved September 30, 2022, from https://saifmohammad.com/WebPages/Biases-SA.html

11. Wang, T. (n.d.). Double-hard debias: Tailoring word embeddings for gender bias ... - arxiv. Retrieved December 7, 2022, from https://arxiv.org/pdf/2005.00965.pdf

12. Tolga. (n.d.). Tolga-B/Debiaswe: Remove problematic gender bias from word embeddings. GitHub. Retrieved December 6, 2022, from https://github.com/tolga-b/debiaswe

13. Semeval-2018 Task 1: Affect in Tweets. Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. In Proceedings of International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, LA, USA, June 2018 from http://saifmohammad.com/WebDocs/semeval2018-task1.pdf.

14. Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

15. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2019. Proceedings of the 2019 Conference of the North. (2019).

16. Kiritchenko, S. and Mohammad, S. 2018. Examining gender and race bias in two hundred sentiment analysis systems. Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. (2018).