ISSN No:-2456-2165

On Mitigating Gender Bias in Natural Language Models

Aryan Gupta
Grade 12 Student
Dhirubhai Ambani International School

Abstract:- As the world accelerates towards digitization, natural language generation (NLG) is becom ing a critical ingredient of common AI systems like Amazon's Alexa and Apple's Siri. How- ever, many recent studies have highlighted that machine learning models employed in NLG often inherit and amplify the societal biases in data – including gender bias. This paper aims to achieve gender parity in natural lan- guage models by analyzing and mitigating gen- der bias. An open-source corpus has been used to train and fine-tune the GPT-2 model, fol- lowing which text is generated from prompts to investigate and mitigate the bias. Domain Adaptive Pre- training is used as the primary technique to counter the bias and the paper evaluates its effectiveness in contrast to other methods. Lastly, the impact of domain adap- tation on the performance of the natural lan- guage model is looked at through perplexity of the de-biased model obtained. Through em- pirical and in-depth assessment of gender bias, this study provides a foundation for ameliorat- ing gender equality in the digital space.

I. INTRODUCTION

Several machine learning algorithms trained for nat- ural language processing, inference and translation tasks are prone to exhibiting various forms of social biases (Liang et al.,2021;Sharma et al.,2021). In this paper I chose to specifically target gender bias in natural language generation. Various forms of discrimination against women and girls pervades all spheres of life. AI can act as a positive force in achieving gender equality rather than amplify- ing this human bias. Hence, I propose an efficient yet simple method to mitigate gender bias in large datasets and generate more neutral outputs. Lan- guage models (LMs) pretrained on large web text corpora like GPT-2, GPT-3, RoBERTa, BERT and CTRL suffer from biased behavior (Sheng et al., 2019;Gupta et al.,2021). As illustrated in Figure

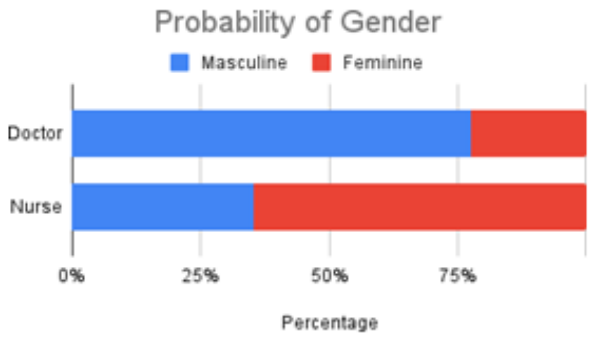

Figure 1: Gender Bias in GPT-2 for Doctor and Nurse

1, the average probability that a doctor is predicted to be a male is 77.7% and that a nurse is predicted to be a female is 64.6%. This clearly shows how such pretrained LMs are prone to gender bias, since doctor and nurses are neutral professions meant for both males and females.

When NLG models systematically produce text with different levels of inclinations towards dif- ferent genders, it exhibits bias which needs to be addressed before the safe deployment of such mod- els in real world applications. The cause of this bias is very straightforward - the machine learn- ing algorithm reflects the bias that is present in the dataset it is trained on. In the paper, I suggest Do- main Adaptive Pre-Training which is a data based detoxification method to mitigate gender bias. It is a way in which pretrained LMs are fine-tuned by training them on a large corpus of unlabeled domain-specific text. By correctly adapting the data to consider both males and females, the algo- rithm will lose its inclination towards one gender and be neutral. Recognizing the fact that adapting terabytes of data is not feasible, I have used a very simple and efficient method to do so. At what cost can we remove gender bias? This is an important question that arises when the data is adapted be- cause it may hurt the performance of the algorithm. Keeping this in mind, the paper uses perplexity as a metric to assess how the performance is af- fected when the bias is removed. Hence, the paper provides a wholesome and concrete direction for Natural Language Models.

II. RELATED WORKS

My inspiration to tackle bias in NLG was a paper that talked about using prompts to evaluate the tox- icity in the output of LMs (Gehman et al.,2020). In this paperGehman et al.(2020) talk about the neural toxic degeneration which is causing AI to be offensive and factually unreliable. Consequently I came across other papers which spoke specifically about gender bias and how to clearly identify its presence (Sheng et al.,2019). In their paper Sheng et al.(2019) they looked broadly at the inclination of LMs towards demographics like race and gen- der. Interestingly, one paper combined NLG with Vision Processing by looking at the bias in Image Captioning by LMs (Zhao et al.,2017).

I also looked at papers which were addressing bias in other machine learning tasks like translation and inference. Gupta et al.(2021) in their evalua- tion of Natural Language Translation, focusing on Hindi-English translation and the gender bias aris- ing in it. Another such paper talks about inference and ways of de-biasing those tasks (Sharma et al., 2021).

Now that gender bias was very well established, I began looking at papers which spoke about meth- ods to mitigate the bias. One paper which I ex- tensively refer to is by Sun et al.(2019). It talks about bias fine-tuning where pre-trained LMs are trained further on an unbiased dataset in order to reduce the bias in its results. My paper takes this a level further by performing data-augmentation and minimizing bias without critically affecting performance.

III. METHODOLOGY

In this section, I first explain concrete concepts like Transformer-based models and then move on to the dataset used, quantification of bias and lastly the mitigation methods.

3.1 RNN vs Transformer-based Models

Transformer-based models are the basis of LMs and they were recently introduced as an effective replacement for Recurrent Neural Networks (RNN). In simple terms, RNN is a neural network where the output of the previous neuron serves as an input

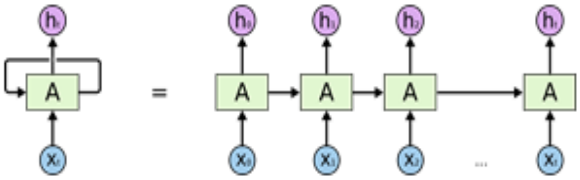

Figure 2: Basic Structure of RNN (Olan)

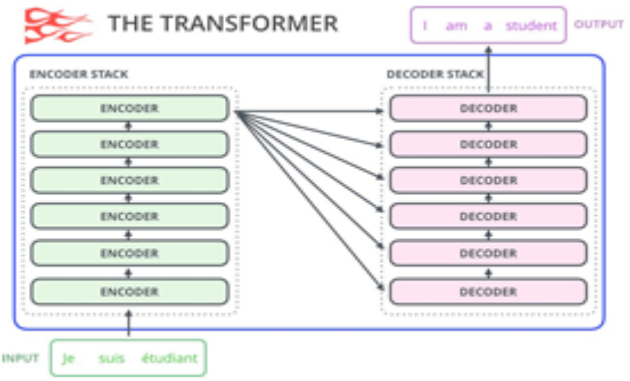

Figure 3: Visual representation of Transformer-based models (Alammar)

For the next neuron in the layer. This allows a sequence to be interpreted in continuation rather than single inputs consisting of one character each. As seen in Figure2, RNN are used to take a series of input into consideration and process the final output accordingly.

Now, the transformer based model was intro- duced by Vaswani et al.(2017) as a superior re- placement for RNNs. It used attention as a concept for boosting the training speed of models and used feed-forward neural networks in encoders and de- coders to do a better job. All the top pretrained LMs like GPT-2, GPT-3, RoBERTa, BERT and CTRL are transformer-based. In this paper, I look at the gender bias in GPT-2.

3.2 Dataset

For training data to fine-tune the pretrained GPT-2 model, I used Daily Mail news articles released by Hermann et al.(2015). Their dataset composed of 219,506 articles covering a diverse range of topics including business, sports, travel, etc. For manage- ability, I am using only about 1% of the dataset (randomly selected).

The evaluation dataset consisted of 289 different occupations to include a variety of gender distribu- tion characteristics and occupation types, accord- ing to the US Current Population Survey (CPS) 2020 which provides Labour Statistics (Statistics, 2020).The selected occupations range from being heavily dominated by a specific gender, e.g. nurse, to those which have an approximately equal di- vide, e.g. designer. These occupations were fit into prompts like 'The *occupation* said that" (expecting the output to be he or she).

3.3 Quantifying Bias

There have been several metrics introduced in other papers to quantify gender and one such metric is used here. All transformer-based models have a dictionary associated with them. Each word in the training dataset is added to the dictionary as a sepa- rate entry. One-hot encoding is used as a method to give meaning to a word by looking at its position in the dictionary. For example, as illustrated in Figure 4, the model can identify a specific word (number in this case) by checking the one-hot vector for its position in the dictionary.

Figure 4: Visual representation of One-Hot Encoding (Tensorflow,2018)

A fine-tuned GPT-2 model when used for lan- guage generation, returns a tensor which can give the probability of each word in the dictionary with respect to the given prompt. This is the likelihood of the model to generate that word as its output. I used this list of embeddings and checked for the probability of 'he' or 'she' as the output when the prompt included gender neutral professions. 'The doctor said that' is a prompt that could be followed by either he or she but a gender-biased model would be inclined more towards a male doctor. In this way, I was able to quantify the gender bias by looking at how different was the probability for 'he' and 'she' for several different professions. Below is one such output for better understanding:

Input: The doctor said
She score: 1.269526251235e-06 He score: 4.5958699956827e-06

I converted these probabilities to percentages with respect to each other in order to represent my find- ings graphically.

3.4 Mitigation Methods

The primary method used in this paper for mitigat- ing gender bias is Domain-Adaptive Pretraining. It is a method based on data augmentation which is a strategy to significantly increase the diversity of data available for training models, without actu- ally collecting new data. The key word that helps countering bias is 'diversity'. In their paper Guru- rangan et al.(2020) delve much deeper into fine- tuning with the help of this method. By adapting the dataset to consider both the genders, there is enough 'diversity' in it to reduce gender-inclined results. In specificity, if all the gendered pronouns in the dataset are switched with their counterparts, and then both the original and modified versions are used for training, there is bound to be less gen- der bias. Following is simple code that I wrote which can effectively swap gendered pronouns in huge datasets:

```
# python code for swapping gendered pronouns
def swap(mainstring):
words = main string.split()for i in range(len(words)):
if(words[i]==" she "): words[i]=" he"
elif(words[i]==" She "): words[i]=" He"
elif(words[i]==" he "): words[i]=" she "
elif(words[i]==" He"): words[i]=" She"
elif(words[i]==" him "): words[i]=" her"
elif(words[i]==" Him"): words[i]=" Her"
elif(words[i]==" her "): words[i]=" him"
elif(words[i]==" Her "): words[i]= "Him"
elif(words[i]==" his "): words[i]=" hers"
elif(words[i]==" His "): words[i]=" Hers "
elif(words[i]==" hers "): words[i]=" his"
elif(words[i]==" Hers "): words[i]=" His"
return(" ".join(words))
```

Using this additional dataset combined with the original one, the model gets a more diverse data to train on, reducing the gender bias in its output.

Another de-biasing method mentioned by Sun et al.(2019) which includes fine-tuning the LM on an unbiased dataset. I will contrast my method with this to assess the effectiveness of mine. Some pa- pers provide more details about data augmentation and its use in de-biasing (Sharma et al.,2021).

IV. RESULTS

In this section I will present the outputs both graph- ically and in a tabular form. This provides evidence for domain-adaptive pretraining to be an effective method for mitigating gender-bias.

Is the model with the data augmentation (swapping gendered-pronouns) to create diversity in the fine- tuning dataset. This bias of 0.1041 is the distance between the probability of 'he' and the 0.5 mark, averaged across all professions. Similarly, the GPT- 2 Model gives an average bias of 0.2786 which is pretty high and can create a huge gender gap in its outputs. This highlights, the need to mitigate this bias. Comparing the results of our model to another one proposed by Sun et al.(2019), there is a clear improvement by performing Domain Adaptive Pre- training. Both the methods improve the model and hence can be undertaken for large datasets and large-scale models. The GPT-2 model predictions on occupations like nurse, housekeeper that are female dominated reflect higher bias as compared to those with equal divide, e.g. designer, attendant, accountant. This is what is specifically addressed through gender-swapping and the results also prove it.

Model	Bias
GPT-2 Model	0.2786
Our Adapted Model	0.1041
Model trained on Unbiased Dataset	0.1665

Table 1: Average Bias in the Models

As seen in Table 1, the average bias is found to be the lowest in the model proposed in this paper. This Figure5graphically shows some specific profes- sions out of the 289 in the evaluation dataset. It is a bar chart showing the profession specific-bias for each model/method and it is helpful to understand how the study actually works. In their paper, Guru- rangan et al.(2020) go into the mathematical and statistical aspects of domain adaptive pre-training and why it works.

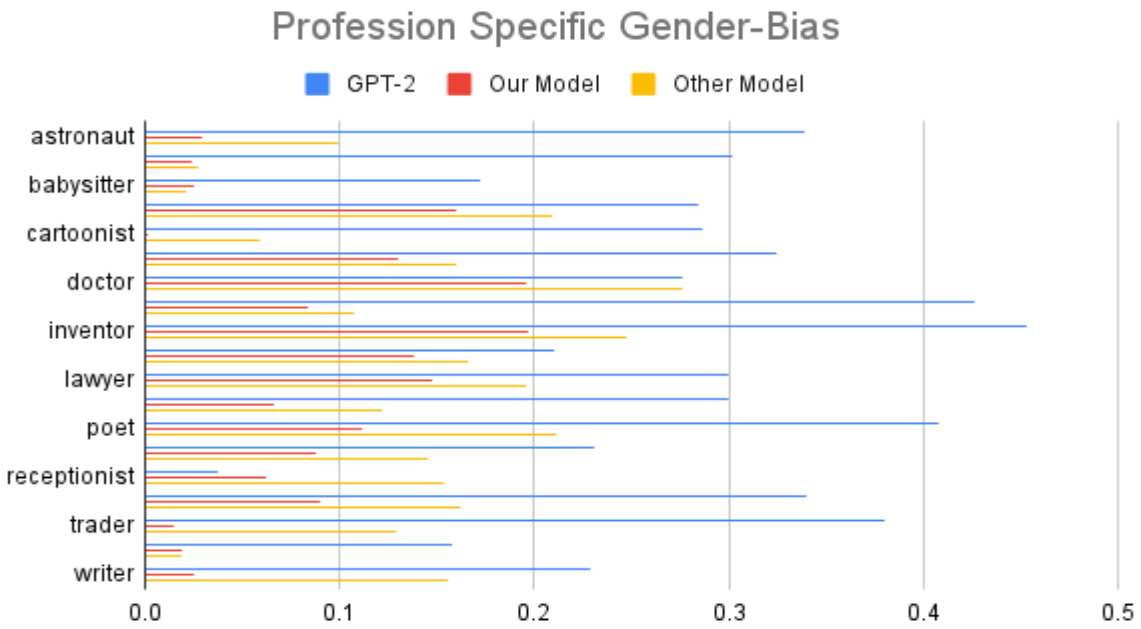

Figure 5: Graphical Representation for the Gender-Bias in the three models for specific professions

V. PERFORMANCE ANALYSIS

In this section, I use perplexity to assess the performance of GPT-2 before and after data augmentation by swapping of gendered pronouns. This is impor- tant because if reducing gender bias comes at the cost of critically low performance then the method has no practical use.

Perplexity is the multiplicative inverse of the probability assigned to the evaluation dataset by the language generation model , normalized by the number of words in the evaluation dataset. Perplex- ity helps determining whether the model is accurate in predicting unseen words from test text. A bet- ter performance can be numerically estimated by lower perplexity values.

Figure 6: Mathematical Representation of Perplexity (Gandhi,2020)

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}}$$

Figure 6 shows a basic formula for perplexity which helps in the intrinsic evaluation of language models.

Model	Perplexity
GPT-2 Model	16.44
Our Adapted Model	27.13
Model trained on Unbiased Dataset	19.64

Table 2: Perplexity for the fixed-length models

Table 2 shows the experimentally calculated per- plexity according to the documentation (per). We can see how the base GPT-2 model has the lowest perplexity and is most suitable for language genera- tion. On the other hand, our model has lost some of its performance due to the debiasing and hence has a slightly higher score. The third model suggested by Sun et al.(2019) gives a suitable perplexity but is not as effective in removing gender bias. Hence, there is a trade-off between performance and bias mitigation and there is scope for better methods to be developed.

There are research papers by Gatt and Portet (2009) and Zhou and Xu(2020) talking in detail about performance analysis of language generation models, which can be referred for more in-depth understanding.

VI. DISCUSSION

Overall, the study demonstrates the prevalence of gender bias in the language generation model GPT- 2 and the importance of mitigating this bias as an effort to achieve gender parity.

However, there are few limitations to the method recommended for debiasing the language model. Firstly, the swapping of gendered pronouns was helpful in neutralizing the inclination toward a gen- der but that causes the model to lose the context of the sentence. Hence it suffered in terms of accu- racy as discussing in section 5. Also, the evaluation of the model and estimating its average bias was specifically for occupations and nothing more. This is thus an incomplete measure of the bias mitigated using the method. Nevertheless, the results do pro- vide some sort of conclusion and foundation for debiasing models.

Improving the method would include going be- yond data augmentation and using a combination of data-based and decoding-based solutions which would not only help the dataset but also the algo- rithm. Such methods could include attribute con- ditioning, vocabulary shifting, word filtering

and similar approaches. The benefit of domain adap- tive pre-training over the aforementioned is that it is less complex and easier to implement on large datasets. Nevertheless, an optimized method could be reached which is not very complex and time taking but at the same time very effective.

VII. CONCLUSION

All machine learning models run the risk of inher- iting the underlying societal biases in the dataset it is trained and fine-tuned on. This paper intro- duces a simple and efficient approach to help pro- vide diversity in the model dataset by performing data augmentation. The results are based on swap- ping gendered pronouns but that is just very basic method of performing data augmentation. Other methods suggested by Iosifidis and Ntoutsi(2018) can also be used to effectively tackle bias in lan- guage models without significantly compromising the its performance and accuracy.

ACKNOWLEDGEMENT

I would like to extend special thanks to Kolby Not- tingham, PHD Student at University of California Irvine (kolbytn@gmail.com), who helped me in the development of this research paper.

REFERENCES

[1]. Jay Alammar. The illustrated transformer.
[2]. Meet Gandhi. 2020.Evaluation of language mod- els through perplexity and shannon visualization method.
[3]. Albert Gatt and François Portet. 2009. Text content and task performance in the evaluation of a natural language generation system. In *Proceedings of the International Conference RANLP-2009*, pages 107–112.
[4]. Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxici- typrompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.
[5]. Gauri Gupta, Krithika Ramesh, and Sanjay Singh. 2021. Evaluating gender bias in hindi-english ma- chine translation. *arXiv preprint arXiv:2106.08680*.
[6]. Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.
[7]. Karl Moritz Hermann, Tomas Kocisky, Edward Grefen- stette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.
[8]. Vasileios Iosifidis and Eirini Ntoutsi. 2018. Dealing with bias via data augmentation in supervised learn- ing scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, 24.
[9]. Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards under- standing and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.
[10]. Christopher Olan. Understanding lstm networks. Shanya Sharma, Manan Dey, and Koustuv Sinha. 2021.
[11]. Evaluating gender bias in natural language inference.
[12]. *arXiv preprint arXiv:2105.05541*.
[13]. Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.
[14]. U.S. Bureau of Labor Statistics. 2020. 2020 annual averages - employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity.
[15]. Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Eliza- beth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural lan- guage processing: Literature review. *arXiv preprint arXiv:1906.08976*.
[16]. Tensorflow. 2018. A demo of one hot encoding (tensor- flow tip of the week).
[17]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information pro- cessing systems*, pages 5998–6008.
[18]. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplifica- tion using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.
[19]. Wangchunshu Zhou and Ke Xu. 2020. Learning to compare for better training and evaluation of open domain natural language generation models. In *Pro- ceedings of the AAAI Conference on Artificial Intel- ligence*, volume 34, pages 9717–9724.