Gender Bias Amplification in Natural Language Processing

Lindsey Ehrlich

Abstract

Recent research in machine learning bias raises concerns for bias amplification: the tendency of the model to predict an outcome at a higher rate than statistically expected. Bias amplification in image-classification was studied extensively by Hall et al. (2022). In this work, we extend the previous study to the natural language processing (NLP) space, performing a controlled experiment designed to answer how bias in the training data affects the model's amplified predictions. We use the WinoBias dataset, introduced by Zhao et al. (2018), with BERT and the scikit-learn Python library to train a pronoun prediction model on different iterations of synthetically biased data. We find similar results to the previous study—bias amplification steadily increases then sharply decreases as bias increases in the training dataset—cementing previous analysis as the norm across both computer vision and NLP models. Our results indicate a common trend in bias amplification that we hope will inform future bias mitigation efforts.

1 Introduction

As machine learning capabilities continues to expand in both breadth and scale, fairness is becoming an increasing concern. Several recent works have revealed that not only do machines learn biases, but they can *amplify* biases as well[5, 4]. By definition, bias amplification refers to the tendency of the model to predict outcomes at a higher rate than statistically expected. Bias amplification is problematic since it can exaggerate existing stereotypes[1] or yield inconsistent results between users[2].

This work focuses specifically on associative gender bias, an extensively-studied form of bias in the machine learning community[3]. Broadly, gender bias refers to the tendency of a model to favor one gender over the other. *Associative* gender bias refers to the tendency of a model to associate a gender with a certain act, word, or setting. For example, given an image of someone cooking, predicting women at a higher rate than men is a stereotypical association[4].

We are interested in determining how varying bias in the training dataset affects the bias amplification of the model's predictions. Several works examine bias amplification in image-recognition tasks[2]. So, in this work, we focus on analyzing bias amplification in natural language processing specifically. We design a simple pronoun-prediction task to hone in on the independent variable in this case: training dataset bias. We synthetically manipulate the dataset to inject biases in an iterative fashion in order to observe the effects of increasingly polarized training data.

2 Experiments

We explore how varying bias in the training dataset affects the bias amplification in the model. To this end, we propose a simple experiment to study the sole effects of bias in the training dataset. After appropriate pre-processing, we use a generic logistic regression (LR) model from scikit-learn to conduct the experiment. For evaluation, we measure both the accuracy and bias amplification of the model's output to better interpret the results from multiple angles.

2.1 Experimental Setup

We use the WinoBias dataset, a collection of 3,160 sentences each containing an occupation associated with either a male or female ground truth label.

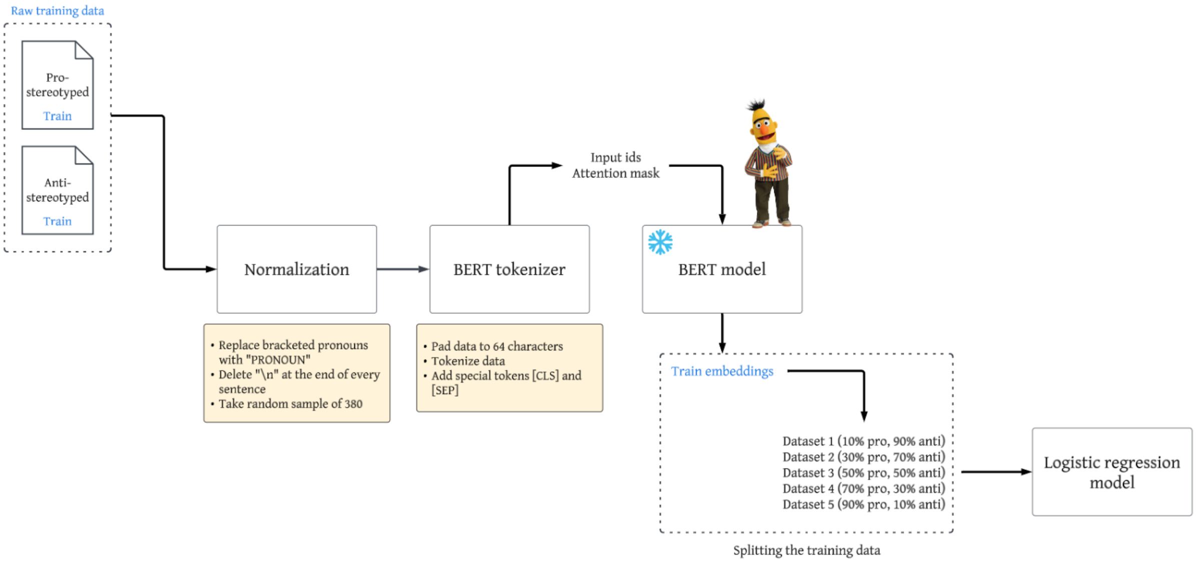

Raw training data
Pro-stereotyped
Train
Anti-stereotyped
Train
Normalization
• Replace bracketed pronouns with "PRONOUN"
• Delete "\n" at the end of every sentence
• Take random sample of 380
BERT tokenizer
• Pad data to 64 characters
• Tokenize data
• Add special tokens [CLS] and [SEP]
Input ids
Attention mask
BERT model
Train embeddings
Dataset 1 (10% pro, 90% anti)
Dataset 2 (30% pro, 70% anti)
Dataset 3 (50% pro, 50% anti)
Dataset 4 (70% pro, 30% anti)
Dataset 5 (90% pro, 10% anti)
Splitting the training data
Logistic regression model

Figure 1: **Training Data Procedure**. Flowchart shows how the training data is processed, including pre-processing steps necessary for text inputs (e.g. normalization, embedding generation) as well as the data split and eventual logistic regression (LR) model training. A simple LR model from scikit-learn was used.

Though the dataset contains two types of distinct sentence structures (labeled type 1 and type 2), we focus on only type 1 sentences for simplicity. The dataset is pre-split into four subsets for convenience: anti-stereotyped training data, anti-stereotyped testing data, pro-stereotyped training data, and pro-stereotyped testing data. As the names suggest, pro-stereotyped sentences play into societal biases while anti-stereotyped sentences brush against them. For example, a pro-stereotyped sentence is "[the developer] argued with the designer because [he] did not like the design" while the corresponding anti-stereotyped sentence is "[the developer] argued with the designer because [she] did not like the design." In the former example, the entity of interest (contained in brackets for clarity) is labeled as a male to agree with the stereotype. The latter example switches the pronoun to achieve the opposite.

To prepare the data, we normalize each sentence by replacing the bracketed pronouns with "PRONOUN," deleting the new line at the end of every sentence, and taking a random sample of 380 sentences. We then feed the data through the BERT tokenizer to generate input ids and an attention mask. We feed the input ids and attention mask through the BERT model to generate embeddings. For the training data, we synthetically split the embeddings into five skews of pro- and anti-stereotyped sentences: 10 percent pro, 90 percent anti; 30 percent pro, 70 percent anti; 50 percent pro, 50 percent anti; 70 percent pro, 30 percent anti; 90 percent pro, 10 percent anti.

2.2 Experimental Procedure

We train five different LR models, each on a different pro- versus anti-stereotyped skew. The task is to predict the gender (either "he" or "she") for the entity of interest in the sentence. We use the same test embeddings to generate pronoun predictions for each model. We then measure the individual accuracy of predictions on pro-stereotyped and anti-stereotyped sentences. We also measure the bias amplification for each model by simply subtracting the percent (in decimals) of pro- (or anti-) stereotyped sentences from the percentage in the model predictions. We choose this metric as a simplified version of previous

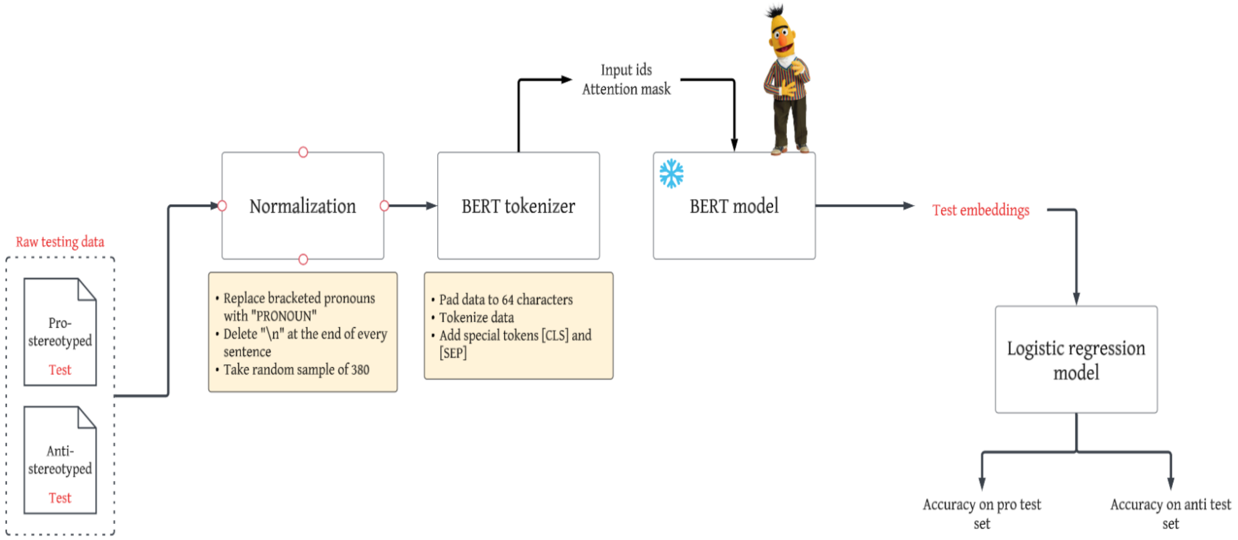

Raw testing data
Pro-stereotyped
Test
Anti-stereotyped
Test
Normalization
• Replace bracketed pronouns with "PRONOUN"
• Delete "\n" at the end of every sentence
• Take random sample of 380
BERT tokenizer
• Pad data to 64 characters
• Tokenize data
• Add special tokens [CLS] and [SEP]
Input ids
Attention mask
BERT model
Test embeddings
Logistic regression model
Accuracy on pro test set
Accuracy on anti test set

Figure 2: **Testing Data Procedure**. Flowchart shows how the testing data is processed, including preprocessing steps necessary for text inputs (e.g. normalization, embedding generation) as well as the accuracy tests on the logistic regression (LR) model's predictions. A simple LR model from scikit-learn was used.

metrics[4]. Correctly, the metric detects 0 bias amplification if the model predictions and training dataset contain the same amount of bias. The metric adjusts proportionally to discrepancies in bias between the model predictions and training dataset. Note that this metric specifically measures bias *amplification*; it does not imply anything about the actual bias present.

2.3 Results

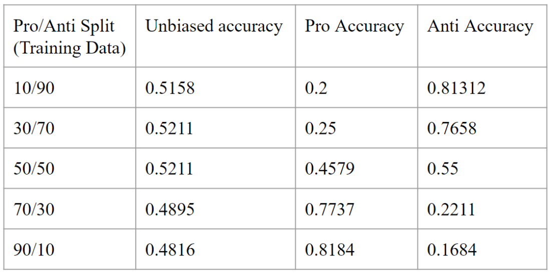
Initially puzzling, the model's accuracy seemed to hover around 50%. However, when we looked at the prediction accuracies for pro- and anti-stereotyped sentences individually, we find the expected result: as the training dataset becomes more biased, the model's predictions proportionally increase in bias.

Measuring bias amplification, our results are similar to those previously found[2]. The bias amplification steadily increases until a threshold point (around 70% bias) when it sharply decreases. The results show that one the dataset becomes *too* biased, there is a point when the bias amplification is 0 and even negative. However, as before, it is important to note that bias amplification is not a measure of bias itself. It is still imperative that we find ways to mitigate bias regardless of the bias amplification present.

3 Conclusion and Future Work

Our experiment undoubtedly has some limitations, namely the small scale. The dataset is comparably small, so replicating our experiment with a larger dataset would serve as valuable confirmation of our findings. Furthermore, interested in only the translation of the training data to bias amplification, we did not investigate the inner mechanics of the model. To clarify, we examined the amplification behavior of the model, and not the reasons for the observed behavior. We can speculate that the reasons for bias amplification lie in the model's architecture, but without further research, we cannot pinpoint the exact culprit(s).

We hope our work serves as a building block toward further investigations in bias amplification specific to NLP tasks.



Pro/Anti Split (Training Data)	Unbiased accuracy	Pro Accuracy	Anti Accuracy
10/90	0.5158	0.2	0.81312
30/70	0.5211	0.25	0.7658
50/50	0.5211	0.4579	0.55
70/30	0.4895	0.7737	0.2211
90/10	0.4816	0.8184	0.1684

Figure 3: **Accuracy Results**. Table displays accuracy results for each of the five models. While the overall (unbiased) accuracy remained the same, measuring the prediction accuracy on pro- and anti-stereotyped sentences individually gives more insight. As the percentage of pro-stereotyped sentences in the training dataset increases, the percentage of accurate predictions on pro-stereotyped sentences also increases. The same trend is apparent for anti-stereotyped sentences.

4 Acknowledgements

We thank Dr. Russakovsky and the Princeton Laboratory Learning Program for the opportunity to conduct research over the summer. We thank Angelina Wang and Austin Hanjie for their technical expertise and guidance. We especially thank Allison Chen for her excellent mentoring and support.

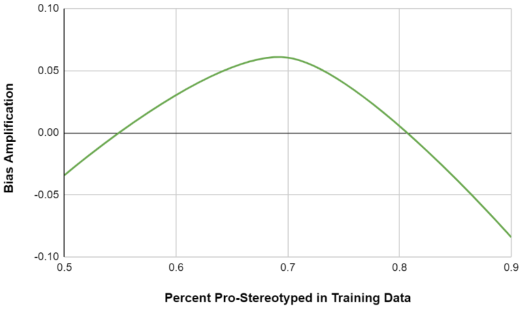


Figure 4: **Bias Amplification Results**. Graph plots the bias amplification in the model's predictions with respect to the amount of pro-stereotyped sentence bias in the training data. As expected, we see a steady increase in bias amplification, but then the amplification unexpectedly decreases as we near the extrema.

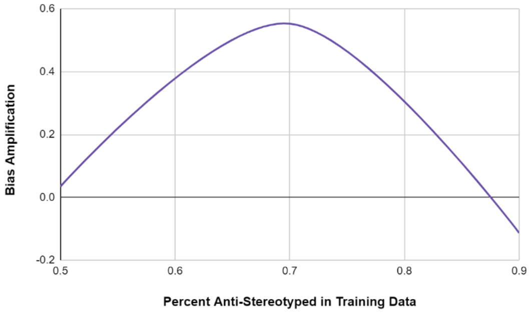


Figure 5: **Bias Amplification Results**. Graph plots the bias amplification in the model's predictions with respect to the amount of anti-stereotyped sentence bias in the training data. We see a similar trend to the pro-stereotyped bias amplification plot.

References

[1] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv:1911.03842*, 2019. 1

[2] Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. A systematic study of bias amplification. *arXiv:2201.11706*, 2022. 1, 3

[3] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv:1906.08976*, 2019. 1

[4] Angelina Wang and Olga Russakovsky. Directional bias amplification. *arXiv:2102.12594*, 2021. 1, 3

[5] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017. 1