Analyzing and Mitigating Fairness Issues in NLP Models

Abstract

The presence of bias in Natural Language Processing (NLP) has the potential to result in the unfair treatment of various demographic groups, thereby reinforcing societal inequalities and reducing trust in AI applications. This study identifies two key metrics commonly used to evaluate bias in NLP models and examines the effectiveness of data augmentation as a mitigation strategy. The "Jigsaw Unintended Bias in Toxicity Classification" dataset is selected as a benchmark for assessing political and gender biases. The evaluation indicates that, while the dataset introduces bias, data augmentation can partially mitigate its effects.

1 Describe

In this section, the project question is introduced, the utilized methods are described, and both the results and their interpretation are presented.

1.1 Introduction

NLP systems have become increasingly integrated into a wide range of applications. However, a growing body of research has demonstrated that these systems can perpetuate or amplify societal biases present in the data they are trained on (Caliskan et al., 2017). Such biases can lead to unfair treatment of certain demographic groups, undermining the trust of users, and reinforce harmful stereotypes in automated decision-making processes (Sheng et al., 2019).

This study uses a toxicity classification model as a practical context to investigate how biases can be evaluated and reduced. Specifically, the analysis of bias is done within a Bidirectional Encoder Representations from Transformers (BERT)-based model fine-tuned on the "Jigsaw Unintended Bias in Toxicity Classification" dataset. To assess the degree of bias, we employ metrics, such as the Word Embedding Association Test (WEAT) and its sentence-level extension, Sentence Embedding Association Test (SEAT). Furthermore, we explore data augmentation as a pre-processing strategy, particularly in gender and political dimensions.

The research question is of particular interest to me, because bias in automated content moderation systems can lead to unfair treatment of certain demographic groups, reduce user trust, and reinforce harmful stereotypes. Given the growing reliance on AI-driven moderation in online platforms, ensuring fairness in these models is crucial for maintaining inclusive and equitable digital spaces. From a research perspective, this study contributes to the broader field of bias evaluation and mitigation in NLP by applying a data augmentation strategy to a new dataset and analyzing its effectiveness in reducing bias. While existing studies have proposed various bias mitigation techniques, their effectiveness can vary depending on the dataset and application context (Sun et al., 2019; Liang et al., 2021).

1.2 Method

This study utilizes the "Jigsaw Unintended Bias in Toxicity Classification" dataset, which contains over 2 million online comments labeled for toxicity (Xiao et al.). For this analysis, toxicity labels were binarized at a 0.1 threshold, and the dataset was balanced to ensure equal representation of toxic and non-toxic examples, preventing model bias due to class imbalance.

A BERT-base-uncased model from HuggingFace (Devlin et al., 2019) was fine-tuned on this dataset for a classification task, specifically toxicity detection.

To assess bias, the WEAT was used, which quantifies bias by computing the cosine similarity between word vectors, determining how closely words from a given target category (e.g., male or female) align with attribute categories (e.g., positive or negative associated adjectives) (Schmahl et al., 2020). A negative bias score indicates a stronger association with female attributes, while a positive score suggests a male bias. The corresponding formulas are provided in Appendix A.

Building on WEAT, the SEAT extends this methodology to sentence embeddings by applying the same association test to sentence-level representations (May et al., 2019).

To mitigate bias, this study applies data augmentation, a pre-processing technique that balances dataset representation by creating an augmented dataset, which has a bias towards the underrepresented group (Sun et al., 2019). The model is then trained on the union of the original and augmented dataset (Sun et al., 2019). In this project this approach is implemented to address gender (male/female) and political (left/right) bias. It is important to note that the scope of this project is limited to the consideration of binary groups.

1.3 Results

This section presents bias evaluation results using WEAT metrics from BERT's final hidden state embeddings. Pre-trained, fine-tuned, and bias-mitigated models were compared with identical hyperparameters. Classification accuracy remained stable at 0.75 across all configurations.

Gender Bias Figure 1a shows that the WEAT score increased after fine-tuning, with bias mitigation reducing the score but not to pre-trained levels. The same holds true for the SEAT scores, compare Figure 2.

Political Bias Figure 1b shows a similar pattern, with the WEAT score increasing after fine-tuning and decreasing after bias mitigation, though remaining above the pre-trained baseline.

1.4 Analysis

This study demonstrates that fine-tuning a classifier on domain-specific datasets introduces measurable bias into a relatively neutral pre-trained BERT model. WEAT and SEAT metrics revealed increased bias in gender and political dimensions post-fine-tuning. A mitigation strategy effectively reduced this bias, with post-mitigation evaluations showing lower bias scores. The pre-trained BERT model exhibited low WEAT and SEAT scores, likely due to bias-aware pre-training corpora and filtering advancements.

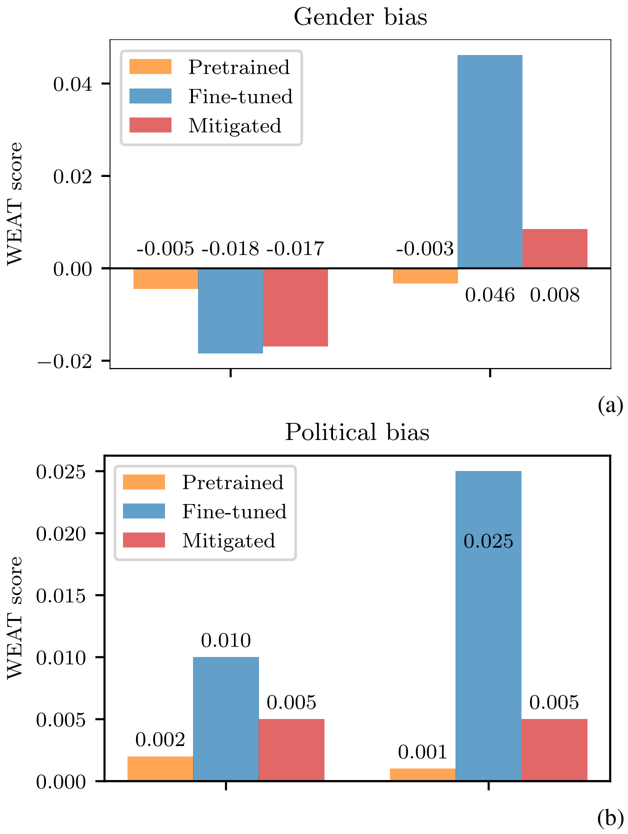


Figure 1: Bias evaluation of the model using WEAT scores.

2 Examine

This section provides a critical analysis of the personal project experience, linking it to both the course material and relevant researched literature.

2.1 Literature base

Throughout this project, we built upon concepts introduced in the course while expanding our understanding through independent research. While the course provided an introduction to bias in large language models, the knowledge was not sufficient to directly apply to our project. To bridge this gap, we conducted a literature review on bias in NLP, drawing insights from studies such as (Liang et al., 2021; Warchol, 2020; Chang et al., 2019). This foundational research enabled us to define a clear project pipeline consisting of dataset selection, bias mitigation strategies, and evaluation metrics. The gained knowledge about these topics can be found in Sections 2.2, 2.5, and 2.4.

One area where the course provided a strong foundation was model architecture. We had studied Transformer-based models in depth, which directly influenced our choice to use BERT. Given the nature of our dataset and the need for contextualized embeddings, BERT was a well-suited model. The knowledge gained in class about fine-tuning pretrained models helped us efficiently adapt BERT to our task.

2.2 Dataset

A crucial step in our project was identifying a dataset containing measurable bias. We explored several datasets and analyzed their suitability for bias evaluation. I specifically investigated the dataset and paper *REDDITBIAS: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models* (Barikeri et al., 2021), reproducing their results using their codebase. This experience significantly enhanced my ability to work with external code, troubleshoot errors, and adapt existing implementations. Despite these efforts, we ultimately chose a different dataset.

2.3 BERT

Our `bert-base-uncased` model from HuggingFace was selected based on prior coursework and its compatibility with our dataset. The course had introduced BERT's architecture, explaining its ability to generate context-dependent token representations. Additionally, our experience implementing a BERT model from a GPT architecture in the first advanced lab further strengthened our confidence in working with transformer-based models.

A key learning moment for me was fine-tuning. Before the project, I perceived fine-tuning as a highly complex process. However, applying it in practice clarified that it primarily involves additional training epochs on domain-specific data. This practical exposure clarified the process and reinforced my understanding of transfer learning.

2.4 Evaluation metrics

Unlike model architecture, bias evaluation was not covered in depth in the course, requiring additional research. We explored WEAT, a widely used bias evaluation method. Caliskan et al. adopted a psychological test, the Implicit Association Test (IAT), which measures subconscious gender bias in humans, to measure bias in word embeddings. Implementing WEAT and SEAT required us to understand the formulas and the inputs to those tests.

2.5 Mitigation strategy

Mitigating bias was another aspect not extensively discussed in the course. To address this, I conducted an in-depth literature review on bias mitigation techniques, classifying methods into preprocessing, in-processing, and post-processing approaches (Bellamy et al., 2019). I found out in (Sun et al., 2019) that the pre-processing approach data augmentation is motivated by the observation that datasets often exhibit an imbalance in the number of references per group. Consequently, data augmentation attempts to remove the bias from the training data so that the model is trained on data with an equal representation of both groups. In order to achieve this objective, an augmented dataset is created. This augmented dataset has a bias towards the underrepresented group. The model is then trained on the union of the original and augmented dataset (Sun et al., 2019). This understanding helped me to implement the data augmentation for our project.

We specifically implemented gender swapping, where gendered terms were replaced to balance representation. For example, we swapped *he* → *she, man* → *woman*, but also *gentleman* → *lady, king* → *king* and vice versa. However, I later realized that some substitutions were ineffective or even counterproductive. For instance, terms like *king* and *queen* were unlikely to contribute meaningfully to bias mitigation due to their low frequency in the dataset.

Moreover during my research I found out that one has to pay attention to some phrases, as they might not make sense after gender swapping. For example swapping the gender in the phrase "*she* gives birth" to "*he* gives birth" is nonsense and is not enriching the dataset (Sun et al., 2019). This highlights the limitations of simple word swaps and the need for more sophisticated augmentation techniques, such as more context-aware augmentation or sentence rewriting using language models.

In addition to data augmentation, I attempted adversarial debiasing, an in-processing mitigation strategy. Adversarial debiasing introduces an adversary model that learns to predict protected attributes (e.g., gender or political affiliation) from the main model's representations. The objective is to reduce the adversary's accuracy, forcing the main model to learn representations independent of bias-related attributes (Bellamy et al., 2019).

I implemented this approach and ran several training iterations. However, the training process was highly unstable, requiring careful hyperparameter tuning. Additionally, adversarial training is computationally expensive, and due to limited computational resources, I was unable to achieve meaningful improvements before the project deadline. This experience underscored the trade-off between complexity and feasibility in real-world bias mitigation. In future work, I would explore adversarial debiasing again or try a post-processing mitigation strategy, which might not need additional training like Equalized odds postprocessing (Hardt et al., 2016) or Individual and Group Debias (Lohia et al., 2019). I would also try a hybrid approach for more robust mitigation with the goal to decrease the score in WEAT and SEAT even more approaching the pre-trained score.

2.6 Future Improvements

This project significantly deepened my understanding of bias in NLP, fine-tuning of transformer models, and bias evaluation techniques. While the course provided a solid theoretical foundation, hands-on implementation revealed complexities not immediately apparent in lectures. In retrospect, I would refine the data augmentation strategy by analysing the dataset first and then including domain-specific terms. Additionally, expanding evaluation metrics beyond WEAT and SEAT, such as analyzing model outputs for biased predictions, could provide a more comprehensive bias assessment. Moreover I would have liked to have more time for adversarial debiasing or investigate other in-processing and post-processing techniques.

Overall, the project reinforced the importance of balancing theoretical knowledge with practical application and highlighted the challenges of bias mitigation in real-world NLP systems.

3 Articulate

A key learning from this project was developing a deeper understanding of bias in NLP models — both in how it manifests and how it can be mitigated. I initially viewed bias in large language models as a well-defined problem with clear solutions. However, through hands-on experimentation, I realized that bias is a complex, multifaceted issue that requires a combination of mitigation and evaluation techniques to be properly addressed. For example our results showed that the mitigation strategy reduced the score of the WEAT, but the pre-trained BERT model still outperformed the mitigated one. Additionally, I gained a much clearer understanding of fine-tuning transformer models and the practical challenges of applying bias mitigation strategies in real-world scenarios.

This learning came through a combination of literature research, implementation, and experimentation. The initial literature review helped establish a theoretical foundation, but practical application revealed the limitations of existing methods. For example implementing fine-tuning on BERT helped me understand that adapting pre-trained models is less complex than I initially thought, as it mainly involves additional training on a specific dataset. The biggest challenge was bias mitigation, where I experimented with data augmentation (gender swapping) and adversarial debiasing. While gender swapping produced measurable improvements, adversarial debiasing did not yield meaningful results due to long training times and the difficulty of properly balancing the adversarial loss. By that I learned that research is not always as easy and straight forward as it looks in research papers, but often is more complex and requires much more work and time.

The main course objective was to enable us to work with scientific information within the context of NLP, which this project directly supported. The project required extensive literature review, helping me develop skills in finding relevant scientific papers, assessing their contributions, and synthesizing knowledge from different sources. While implementing multiple bias mitigation and evaluation strategies showed me that not all theoretical methods are implemented as easy as I think. The project and the labs have showed me, that implementation requires a in-depth understanding of the concepts.

3.1 References

References

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy,

J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. 2019. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. 2019. Bias mitigation post-processing for individual and group fairness. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2847–2851.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Katja Geertruida Schmahl, Tom Julian Viering, Stavros Makrodimitris, Arman Naseri Jahfari, David Tax, and Marco Loog. 2020. Is Wikipedia succeeding in reducing gender bias? assessing changes in gender bias in Wikipedia using word embeddings. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 94–103, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Simon Warchol. 2020. Bias in nlp embeddings. Accessed: 2025-03-26.

Yao Xiao, Yaoyao Chang, Cheng Peng, Siyu Li, and Zhiyu Yuan. Jigsaw unintended bias in toxicity classification.

List of Acronyms

BERT Bidirectional Encoder Representations from Transformers

IAT Implicit Association Test

NLP Natural Language Processing

SEAT Sentence Embedding Association Test

WEAT Word Embedding Association Test

A Appendix

In order to assess biases in word embeddings, Caliskan et al. (2017) proposed a method called Word Embedding Association Test (WEAT). The relationship between a pair of words in the embedding dimension, represented with the vectors v_1 and v_2 is measured by the cosine similarity:

$$s(v_1, v_2) = \frac{v_1^T v_2}{\|v_1\| \|v_2\|} \quad (1)$$

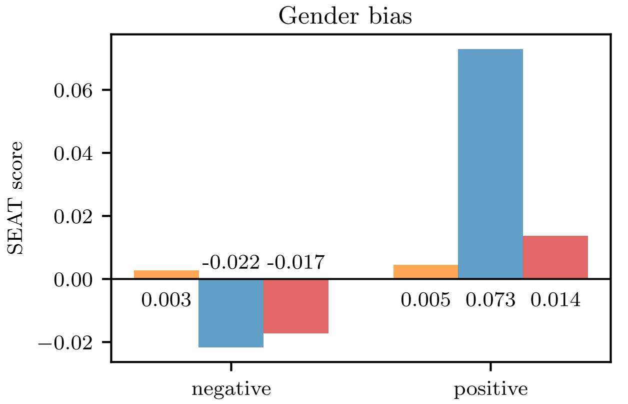
Let v_c represent a word from context C (e.g., "football" from "sports" context), while v_m denote a list of words from one end of the biased target spectrum, such as a male-specific word (e.g., "he" or "his") in gender bias evaluation, and v_f a female-specific word (e.g., "she" or "her"). The gender bias for the mentioned context is then computed using the following equation:

$$b(v_c) = \frac{1}{|M|} \sum_{v_m \in M} s(v_c, v_m) - \frac{1}{|F|} \sum_{v_f \in F} s(v_c, v_f) \tag{2}$$

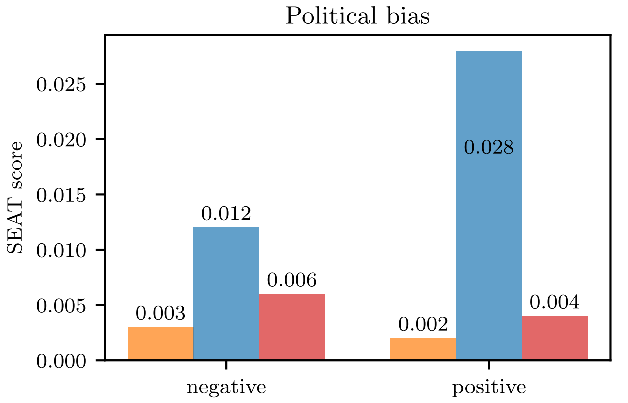
Here, a negative value indicates that the category word is female-biased, while a positive value indicates a male bias. This score is averaged over all words in the context C to obtain the bias score $b(C)$:

$$b(C) = \frac{1}{|C|} \sum_{v_c \in C} b(v_c) \tag{3}$$

The Sentence Encoder Association Test (SEAT) extends WEAT to sentence embeddings by applying the same methodology to sentence-level representations (May et al., 2019). This score is computed similarly to WEAT but accounts for the contextual nature of sentence embeddings.



(a)



(b)

Figure 2: Bias evaluation of the model using SEAT scores.