Analyzing and Mitigating Fairness Issues in NLP Models

Varun Gurupurandar and Arman Mohammadi and Vivienne Schwabe and Gemma Sempere
Linköping University
vargu125@student.liu.se and arman.mohammadi@liu.se and vivsc641@student.liu.se
and gemse337@student.liu.se

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 and illustrate 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 Introduction

Natural Language Processing (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, undermine of user trust, and reinforce harmful stereotypes in automated decision-making processes (Sheng et al., 2019).

Focusing on the evaluation and mitigation of bias in NLP, this study uses a toxicity classification model as a practical context to investigate how biases can be measured and reduced. Specifically, we analyze bias within a 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, SEAT. Furthermore, we explore data augmentation as a pre-processing strategy, particularly in gender and political dimensions.

2 Data and processing

This study employs the "Jigsaw Unintended Bias in Toxicity Classification" dataset, released as part of a Kaggle competition organized by Jigsaw and Google's Conversation AI team (Xiao et al.). The dataset comprises over 2 million online comments annotated for toxicity, alongside floating-point identity attributes that represent the probability of a comment referencing specific demographic groups (e.g., gender, race, political affiliation).

For this analysis, only the toxicity label was used, binarized at a 0.1 threshold to classify comments as toxic or non-toxic. Due to a pronounced class imbalance — approximately 90% of the comments being non-toxic—the dataset was balanced to contain an equal number of examples in each class. This step was taken to prevent skewed model training and to ensure more reliable bias evaluation.

3 Evaluation metrics

In order to assess biases in word embeddings, **?** 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:

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

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.

4 Model

The model selected for this project is a pre-trained Bidirectional Encoder Representations from Transformers (BERT) base model (Devlin et al., 2019). This model was pre-trained on a large collection of English data. The specific BERT model employed in this study is the "bert-base-uncased" model from HuggingFace (Devlin et al., 2019).

In this project, the pre-trained BERT model was fine-tuned on the Jigsaw dataset, as presented in section 2. This fine-tuned model is subsequently utilized for a classification task, specifically the detection of toxicity.

5 Mitigation strategy

In the present project, a pre-processing algorithm called data augmentation has been implemented. This approach is motivated by the observation that datasets often exhibit an imbalance in the number of references per group. Consequently, data aug mentation 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, which 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).

In this project, data augmentation is employed to mitigate gender bias and political bias. It is important to note that the scope of this project is limited to the consideration of binary groups. Consequently, the augmentation is constrained to the utilization of male and female genders, along with left and right political orientations.

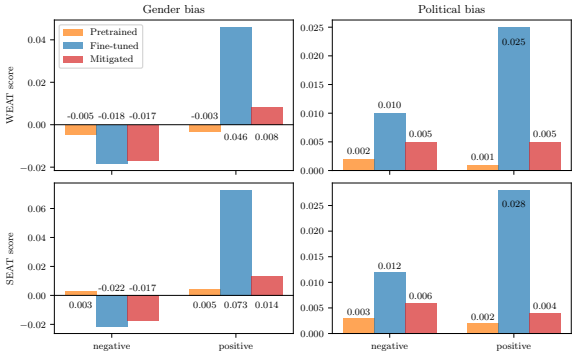

Figure 1: Bias evaluation of the model using WEAT and SEAT scores.

6 Results

This section presents bias evaluation results using WEAT and SEAT metrics, derived from BERT's final hidden state embeddings. Pre-trained, fine-tuned, and bias-mitigated models, all trained with identical hyperparameters, were compared. An augmented dataset matching the original's size may slightly lower performance due to less informative or synthetic samples. Nevertheless, classification accuracy on the test set remained stable at approximately 0.75 across all configurations.

The left column of Figure 1 presents results of a gender bias evaluation in the context of positive and negative adjectives. The fine-tuned model shows a clear shift, with positive adjectives more strongly linked to male terms and negative adjectives to female terms. The right column displays results from a political bias evaluation targeting U.S. political affiliations. Consistent with gender bias observations, fine-tuning resulted in a noticeable shift in term representation. The model showed heightened bias in both positive and negative contexts, with a stronger association of these terms toward the conservative party.

7 Discussion and Conclusion

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.

References

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

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.

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.

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.

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