

TDDE09

Analyzing and Mitigating Fairness Issues in NLP Models

Varun Gurupurandar, Arman Mohammadi, Vivienne Schwabe and Gemma
Sempere

What is bias in NLP models?

The systematic tendency in a model to favor one demographic group/individual over another, which can be mitigated but may well lead to unfairness - R. Ribón Fletcher et al.

The screenshot shows a machine translation interface with tabs for PERSIAN - DETECTED, PERSIAN, ENGLISH, PERSIAN, and SPANISH. The ENGLISH tab is selected. The interface displays a list of Persian sentences and their English translations. The Persian sentences all use the gender-neutral pronoun 'او' (او). The English translations use gender-specific pronouns 'He' or 'She' to reflect societal stereotypes. A text box on the left states: 'All source sentences have the same gender-neutral pronoun'. A text box on the right states: 'Translated sentences receive gender-specific pronouns, reflecting societal stereotypes'.

PERSIAN - DETECTED	PERSIAN	ENGLISH	PERSIAN	SPANISH
او مدیر است	He is the manager			
او پرستار است	She is a nurse			
او دکتر است	He is a doctor			
او زیبا است	She is beautiful			
او ناز است	She is cute			
او بامزه است	He is funny			
او نابغه است	He is a genius			

Example of gender bias in language translation (NLP)

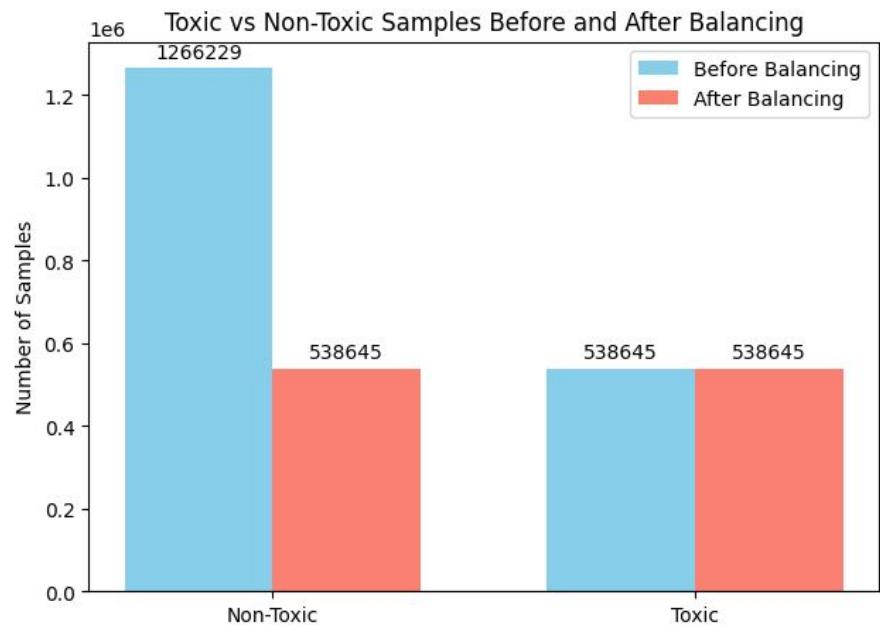
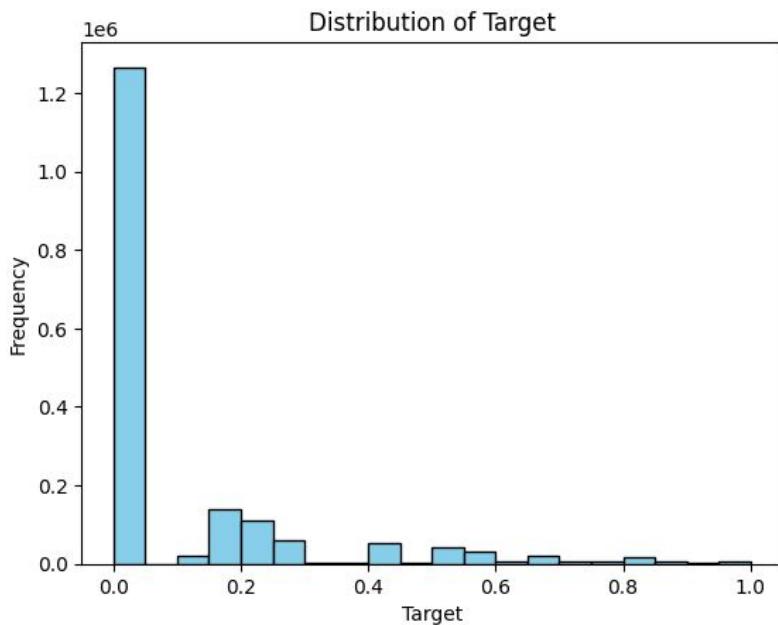
Selection of dataset: Jigsaw Comment toxicity classification

```
Data columns (total 45 columns):  
 #   Column  
 ---  -----  
 0   id  
 1   target  
 2   comment_text  
 3   severe_toxicity  
 4   obscene  
 5   identity_attack  
 6   insult  
 7   threat  
 8   asian  
 9   atheist
```

```
text:  haha you guys are a bunch of losers.  
target:  0.8936170212765957  
Is it an insult?  0.8723404255319149  
Is it a thread?  0.0  
Is it attacking to any identity?  0.0212765957446808
```

```
df['toxic_label'] = (df['target'] >= 0.1).astype(int)
```

Toxic - Non-toxic → balancing our dataset



How do we measure bias?

Word Embedding Association Test (WEAT)

Dog *Cat*

$$\text{Cosine similarity: } s(v_1, v_2) = \frac{v_1^T v_2}{\|v_1\| \|v_2\|}$$

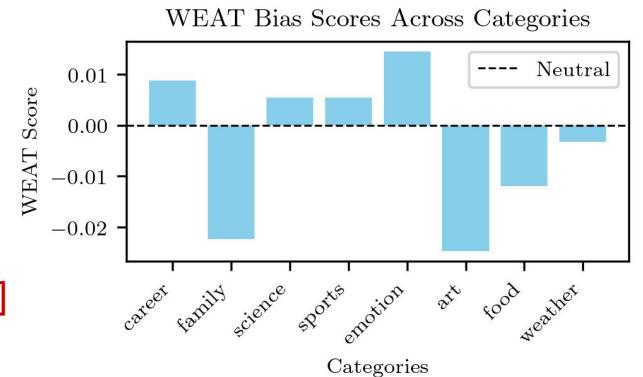
["man", "he", "him", "father", ...] ["woman", "she", "her", "mother", ...]

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

Football

WEAT in context: $WEAT(C) = \frac{1}{|C|} \sum_{v_c \in C} b(v_c)$

Sports context: ["soccer", "basketball", "tennis", ...]



Data: <https://huggingface.co/fse/word2vec-google-news-300>

Sentence Embedding Association Test (SEAT)

["He is a leader", ...] ["She is a doctor, ..."]

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

↓ ↑
"This job requires intelligence."

Model selection

Model selection

- Pre-trained BERT base model [1]
 - Transformer model
 - Pre-trained on large corpus of english data
- Fine-tune base model on Jigsaw dataset
- Classification task: toxicity detection

[1]: <https://huggingface.co/google-bert/bert-base-uncased>

Mitigation strategies

Mitigation strategies for bias

- **Pre-processing algorithms**
 - Data augmentation
- In-processing algorithms
- Post-processing algorithms

Data augmentation

- **Idea:** remove bias from training data
- Create augmented dataset biased towards underrepresented group
- Train on union of original and augmented data
- Examples with gender swapping:

“He went to the university.” → “She went to the university.”

“The woman likes cooking.” → “The man likes cooking.”

Results

Results after gender swapping

Categories: - negative: negatively associated adjectives
- positive: positively associated adjectives

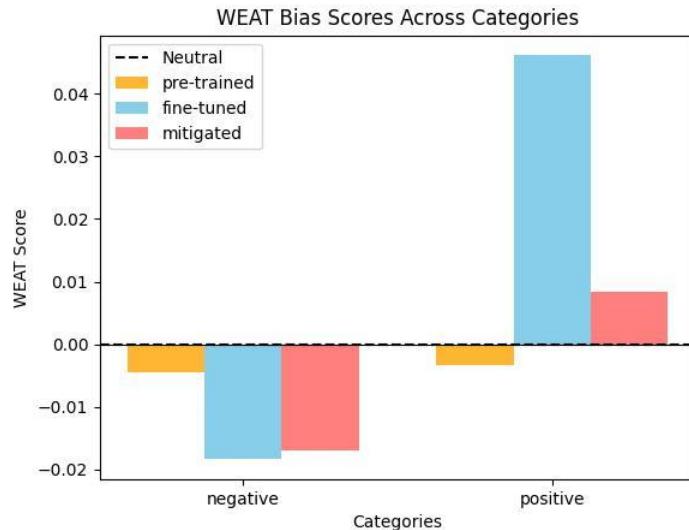


Fig. 1: WEAT scores for pre-trained, fine-tuned and mitigated model

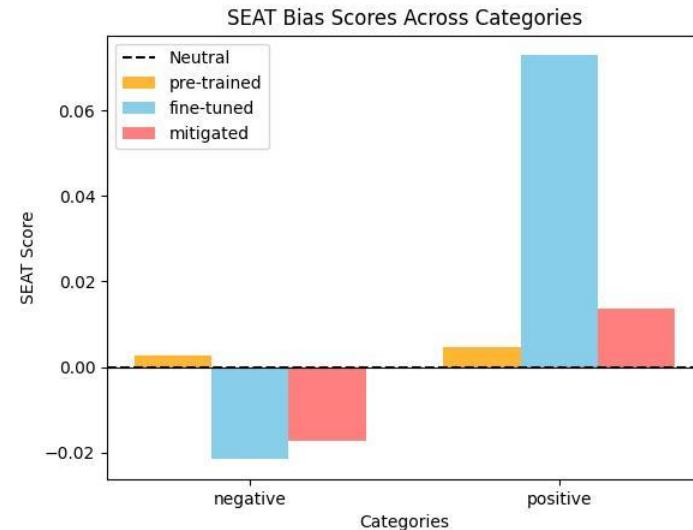
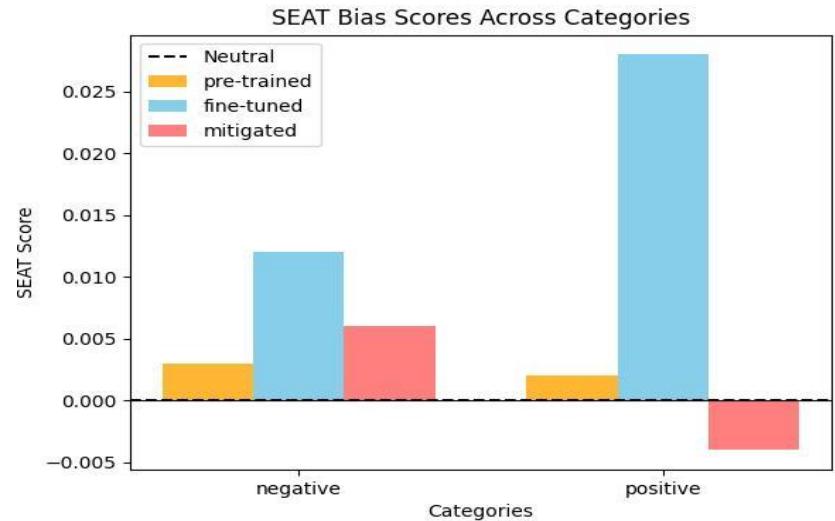
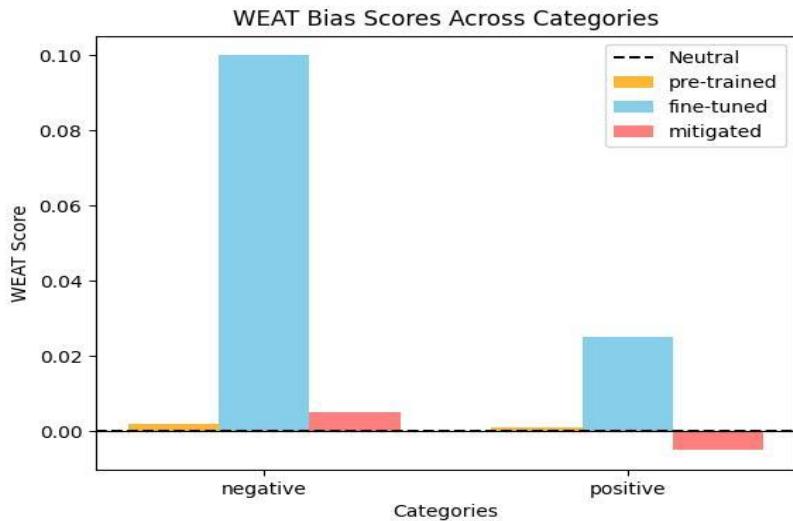


Fig. 2: SEAT scores for pre-trained, fine-tuned and mitigated model

Results after Political swapping



Conclusion/Discussion

- Training machine learning models can bring unwanted and unnoticed bias
- Bias mitigation improved fairness metrics in our work
- Can we claim now that the model is unbiased?

Thank you!

Any questions?