Probing of Language Model Representations for Bias Certification

Vansh Gupta, Tamás Visy, Susanna Di Vita, Noah Pfenniger

1 Introduction

Despite the efficiency of transformer-based models like *BERT* and LLMs like *GPT*, these models often learn and perpetuate gender biases, affecting downstream tasks. This work investigates gender-related biases in LM representations through probing. Our main contributions include identifying gender biases through non-binary association tests for detailed layer-wise analysis, enhancing probe robustness by training on a diverse dataset and introducing a novel bias ranking system using new group fairness metrics, establishing a fairness certification framework inspired by the *OT-IND-CPA* security model to assess and certify fairness concerning gender biases.

2 Data

Probe Dataset: The probe training datasets consist of the *Stanford Sentiment Treebank (SST)* and the *TweetEval* dataset. To achieve the goal of an unbiased dataset, a series of *filters* have been implemented including personal names, pronouns, titles, relationship terms, occupations, and transsexual terms. Resulting in a dataset of 99,769 training sentences, with 67,348 from *SST* and 32,421 from *TweetEval*. Of these, 27,803 (28%) had one or more masks.

Minimal Groups: To assess potential biases in language models, a dataset of minimal groups was constructed using *templates* that generate sentences with various subjects and adjectives. The methodology, based on the *Equity Evaluation Corpus* and *LABDet*, utilized 11 template sentences. Multiple gender-specific reference phrases and titles were employed, with 20 random names (10 female and 10 male) sampled from the *UCI* name dataset.

Template	Example Sentence
The situation makes {} feel {}	The situation makes him feel sad

Table 1: Example template sentence used for minimal groups

To ensure comprehensive sentiment coverage, 20 words each for positive, negative, and neutral sentiments were selected, which resulted in 12,280 evaluation sentences, split equally between male and female subjects. Ultimately, there were 4,040 test samples each for positive and negative sentiments and 4,200 for neutral sentiment.

3 Method Overview

Our **probes** are based on a variety of classifiers, such as *logistic regression*, *Gaussian Naive-Bayes* or a *MLP*. We train them on the encodings from a specific layer of the LM under investigation, then predict sentiments for each encoded test sentence.

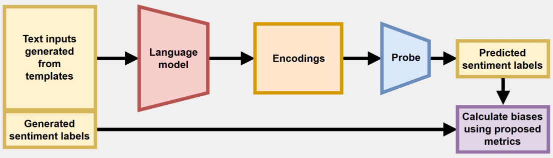
Text inputs generated from templates
Generated sentiment labels
Language model
Encodings
Probe
Predicted sentiment labels
Calculate biases using proposed metrics

Figure 1: Overview of probing using a trained probe to evaluate biases of an LM.

We then evaluate the results using

- **demographic parity**, which measures fairness by maintaining consistent probabilities of bias detection across genders.
- **equalized odds**, which ensures that the model's true and false positive rates are the same for all genders.
- and **average relative sentiment change**, which measures how consistently the model predicts sentiment for neutral sentences with and without the masking of group identifiers.

4 Fairness Certification

Training Phase: The attacker trains a probe $\mathcal{P}$ on an unbiased dataset $\mathcal{D}_{train}$, achieving high accuracy p.
Query Phase: The attacker selects a group of sentences S, where each s_i differs only by elements of G. The attacker sends S to the oracle $\mathcal{O}$.
Oracle Response: The oracle randomly selects and encodes a sentence, returning the encoding e.
Attacker's Task: The attacker must identify the group element g_i used. The advantage is:

$$\mathrm{Adv}_{\mathcal{M}}^{\mathcal{P}} = \left| P(\mathcal{P}(e) = g_i) - \frac{1}{|G|} \right|$$

An LM $\mathcal{M}$ is (t, p, ϵ) fair if any probe trained in t FLOPs with validation accuracy p has at most advantage ϵ.

Advantage Derivation: For $y = Y$ being the sentiment chosen by the attacker, using Bayes theorem and Total Probability Rule:

$$P(\mathcal{P}(e) = g_i \mid y = Y) = \frac{P(y = Y \mid \mathcal{P}(e) = g_i)}{\Sigma_{g_i \in G} P(y = Y \mid \mathcal{P}(e) = g_i)}$$

With known sentence encoding:

$$P(y = Y \mid \mathcal{P}(e) = g_i) = \begin{cases} \hat{p} & Y = 1 \\ 1 - \hat{p} & Y = -1 \end{cases}$$

Range: $\mathrm{Adv}_{\mathcal{M}}^{\mathcal{P}} \in [0, 1 - \frac{1}{|G|}]$

5 Results and Discussion

We find that MLP-based probes generally perform best at predicting sentiments based on the LMs encodings.

Model	Rank	Label				Probability			
		ARSC	DP	EON	EOP	ARSC	DP	EON	EOP
GloVe	*1 (Baseline)*	*0.13*	*-0.1*	*-0.1*	*-0.11*	*0.05*	*-0.04*	*-0.04*	*-0.05*
BERT	5	-0.02	0.01	0.02	0.0	-0.01	0.01	**0.01**	0.01
ELECTRA	2	**-0.07**	**0.04**	0.0	**0.08**	**-0.03**	**0.02**	0.01	**0.03**
GPT2	3	-0.04	0.01	**0.04**	-0.02	-0.01	0.0	-0.0	0.0
RoBERTa	6	-0.01	0.0	-0.01	0.01	-0.01	0.0	-0.0	0.01
XLNet	4	-0.03	0.01	0.0	0.01	-0.01	0.01	0.0	0.01

Table 1: Results of language models' embeddings from their final layer probed using an MLP.

The differences in bias trends of LMs can be linked to each model's training objectives and architectures. *BERT* and *RoBERTa*, trained with robust pre-training objectives and extensive datasets, tend to generalize better and mitigate biases through balanced representations. Conversely, other models like *GloVe* and *ELECTRA*, inadvertently amplify certain biases due to the respective training methodologies.

6 Conclusion

The probing analysis effectively identified the varying degrees of bias present in different language models. The results demonstrated that *BERT* and *RoBERTa* exhibited minimal bias, indicating their suitability for gender-neutral language processing applications. In contrast, a simpler model, *GloVe*, showed the highest bias among the LMs, suggesting that more advanced LMs learn to disregard biases in the data.

Additionally, inspired by the *OT-IND-CPA* security model, we established an extensible framework to certify the fairness of language models. Higher probe accuracies in models like *BERT* and *RoBERTa*, along with their minimal epsilon values, indicate strong fairness in their representations and a reduced risk of information leakage.