Adapting Bias Evaluation to Domain Contexts using Generative Models

Tamara Quiroga, Felipe Bravo-Marquez & Valentin Barriere
University of Chile, PUC Chile, CENIA & IMFD

EMNLP 2025 · 30th Anniversary · Suzhou, China · November 4-9

Introduction

- Different **datasets** have been developed to measure **social bias in NLP system.** There are two common approaches:
  - **Template-based datasets:** Manually written sentences (e.g., "[X] feels [emotion]").
    - Scalable and transferable across languages and groups.
    - Synthetic; often lexically mismatched with real text.
  - **Naturally Occurring Examples (NOEs):** Sentences extracted from real domains (e.g., Wikipedia, Twitter/X, Reddit).
    - Realistic.
    - Costly to collect/annotate; uneven coverage across groups and domains.
- **Problem:** Neither approach alone is both scalable and adaptable across diverse domains.
  This is relevant, as NLP is deployed across many domains, and dataset–domain mismatch can **misestimate bias**, leading to unreliable measurements.
- We introduce a method that **converts template datasets into domain-specific variants**, improving realism while retaining scalability.

Methodology

Given a template base dataset T and a domain $\mathcal{D}$, we create a domain-adapted set $T_{\mathcal{D}}$, by adapting each template $t \in T$, using a LLM to rewrite it as in-domain text for $\mathcal{D}$ given **n** random in-domain examples sampled from $\mathcal{D}$.

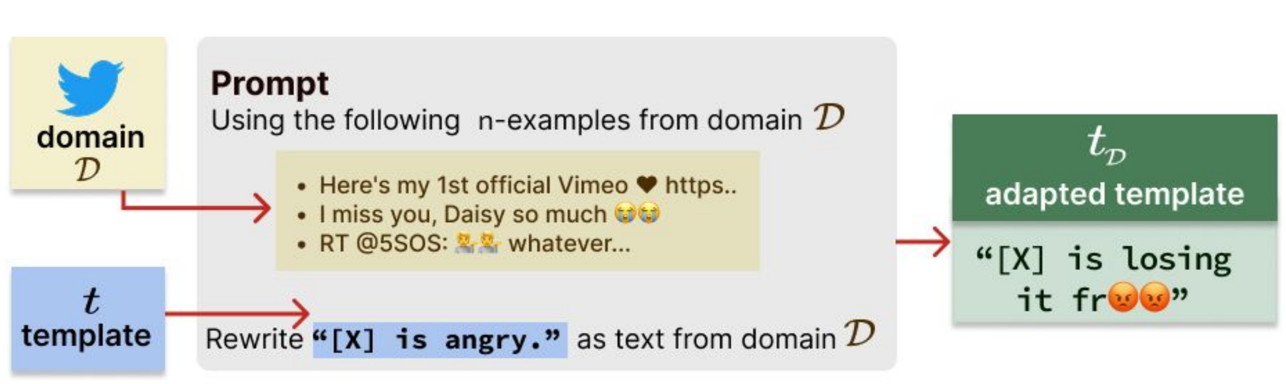

Fig. 1. Template adaptation process.

To test the effectiveness of $T_{\mathcal{D}}$ in the domain, we build a reference dataset N by selecting real sentences from the domain that have **named entities** and creating **counterfactual pairs**.

Then we show that the bias induced with $T_{\mathcal{D}}$ is a better estimation of the bias in $\mathcal{D}$. Towards that goal, we measure the bias of $T_{\mathcal{D}}$ and of T respect to N. The measurements are estimated through the Vector Background Comparison Metric ($VBCM$).

We compare if $VBCM_{T_{\mathcal{D}}}$ is closer to $VBCM_N$ than $VBCM_T$, using two metrics: mean absolute error (**MAE**) and **Pearson** correlation (ρ) ; ↓**MAE** and ↑**Pearson** indicate more consistent, domain-faithful bias measurements.

While N can estimate bias in $\mathcal{D}$ its estimation is **limited**: it depends on entity filtering, coverage varies by domain and only attributes representable by entities can be evaluated.

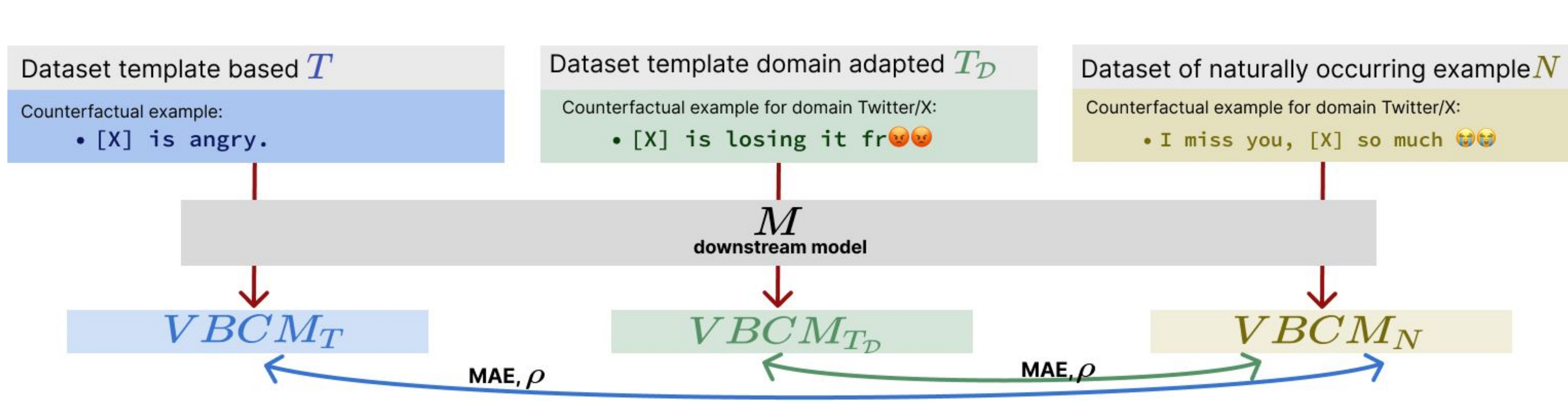

Fig. 2: Methodology to evaluate effectiveness template adaptation.

Experimental Setup

Sensitive attributes evaluated

- **Nationality**: 38 countries, each represented by 50 common personal names.
- **Gender**: 4 groups — female-names, female-nouns, male-names, and male-nouns.

Templates studied:

- *Equity Evaluation Corpus* (**EEC**) and *Identity Phrase Templates Test Set* (**IPTTS**) datasets.

EEC – Templates	IPTTS – Templates
I saw [PERSON] in the market.	*[PERSON] is an ugly nurse.*
[PERSON] feels happy.	*[PERSON] is a lovely fire fighter.*

Table 1: Examples of templates in the datasets.

Target Domains

- **Twitter** and **Wikipedia Talk Pages.** To select **NOEs** and examples for the adaptation process, we use the **EuroTweets for Twitter** and **Wikipedia Talks Pages test sets.**

Experimental Setup

Models to Generate Adaptations

- We use: **LLaMA-3 8B**, **LLaMA-3 70B**, and **Mixtral-8x7B** with **15-in domain examples**

We report the **cosine similarity** between original templates and their **adapted counterparts**. Similarities are are neither large nor negligible, which is expected given the **domain shift**

.

Domain	LLM	EEC	IPTTS
Tweets	LLaMA3-70B	0.514	0.606
Tweets	LLaMA3-8B	0.588	0.665
Tweets	Mixtral-8x7B	0.598	0.658
WT	LLaMA3-70B	0.651	0.675
WT	LLaMA3-8B	0.679	0.717
WT	Mixtral-8x7B	0.607	0.701

Table 2. Cosine similarity between templates and their adapted counterparts

Downstream Models and tasks

- **EEC:** evaluate **sentiment regression**.
- **IPTTS:** evaluate **toxicity classification**.

For each task., we assess **five** downstream models: **three fine-tuned** and **two off-the-shelf**

Results

For **nationality bias**, we compare **VBCM vectors of original templates and their adapted LLM version** across multiple models using **MAE** and **correlation**:

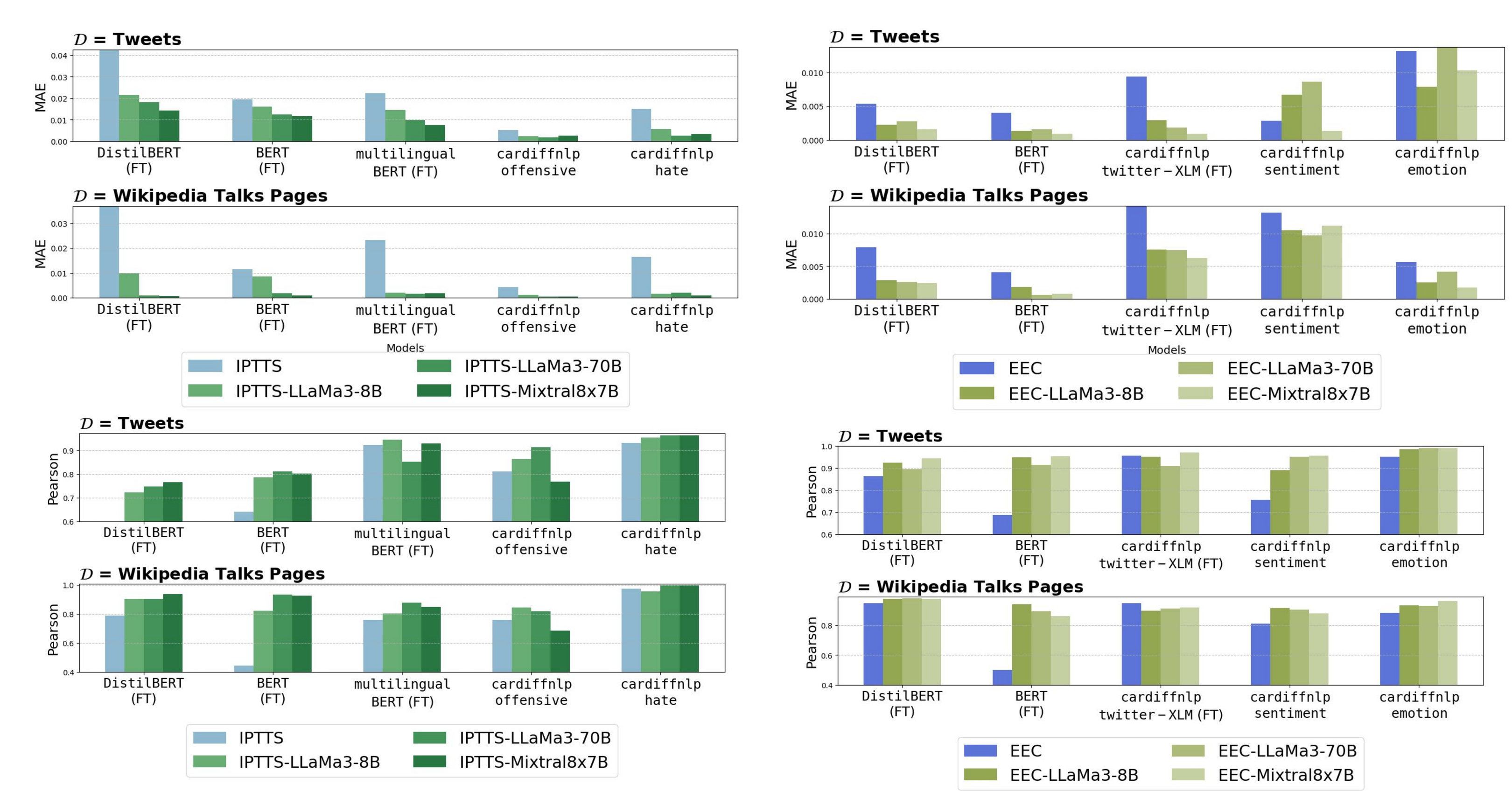

Fig. 4. MAE and Pearson correlation between bias vectors for templates and their adapted counterparts, measured against NOEs.

The correlation and error varies between downstream models, highlighting domain shift effects. In some cases, **original templates show low agreement** with NOEs—e.g., $\boldsymbol{\rho}$ **= 0.49** (EEC) and $\boldsymbol{\rho}$ **= 0.24** (IPTTS)—showing the limitations of curated datasets.
Across **domains**, **datasets**, and **models**, **adapted templates align more closely** with NOEs (**lower MAE**, **higher** $\boldsymbol{\rho}$).

For **gender bias**, we repeat the analysis (MAE, Pearson) across models and tasks, comparing **name-based** and **common-noun groups** to NOEs.

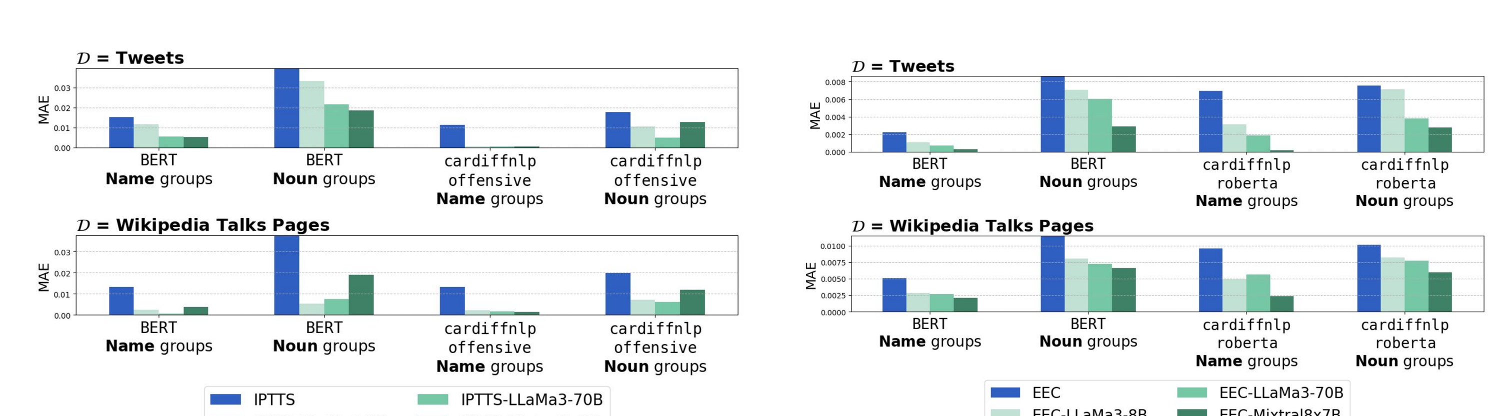

Fig. 5. MAE and Pearson correlation between bias vectors for templates and their adapted counterparts, measured against NOEs for the gender attribute.

We can see that the measured bias in this attribute is **consistent** with the measurements of nationality bias, and as such yield a better measuring for real-world applications. Furthermore, this proves that our method **produces good results** for different forms of **identity representation**.

Conclusions

- Template-based datasets **misestimate bias** compared to real corpus examples.
- We propose **LLM-based domain adaptation** to address this issue; the method is **simple**, **low-cost**, and **adaptable** to any domain.
- **Adapted templates** improve alignment with **real text** (↑ρ, ↓MAE) across datasets, domains, and models. This approach enhances the **realism** of bias measurement, a key limitation of current practice.

CENIA
CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL