Personalisation or Prejudice?

Addressing Geographic Bias in Hate Speech Detection using Debias Tuning in Large Language Models

Funded by the European Union

UK Research and Innovation

Usage/License

Available via Hugging Face Transformers. Released under **Apache 2.0** and **Llama 3.1** licenses. Code is provided under Apache 2.0 on GitHub. Users should consider potential biases and domain-dependent performance variations.

Download

The models can be accessed and downloaded from their HuggingFace collection page: https://huggingface.co/collections/HYBRIDS/hate-speech

Researchers

- Paloma Piot
- Patricia Martín Rodilla
- Javier Parapar.

These researchers are affiliated with the Information Retrieval Lab at the University of A Coruña.

Summary

This model is designed for hate speech detection and incorporates geographic and language contextualisation analysis as described in the paper. Experiments show that adding country-based personae or using non-English languages can introduce prediction inconsistencies and bias. To mitigate this, the model includes a debias-tuning strategy that penalises inconsistent outputs between contextualised and non-contextualised inputs, resulting in more stable and fair classifications across scenarios.

Key Features

- Geographic persona probing: Systematically tests LLMs' responses when prompted with country-specific identity cues.
- Cross-lingual evaluation: Examines performance differences when evaluating hate speech in English vs. translated texts.
- Debias tuning method: A custom fine-tuning loss that penalises classification inconsistency between contextual and non-contextual inputs.
- Empirical benchmark: Uses diverse LLMs in controlled experimental settings to quantify bias and mitigation effects.

What We Offer

- Open-source implementations of experiments and debias tuning techniques.
- Pretrained model variants supporting geographic debiasing (with and without language context) on HuggingFace.
- Reproducible evaluation scripts and data processing pipelines.

Collaboration Objectives

- Researchers in bias, fairness, and multilingual NLP to extend the debias-tuning framework;
- Moderation practitioners to apply debiased models in real workflows;
- Experts in dataset augmentation and bias measurement to expand geographic bias research;
- Cross-disciplinary collaborations on fairness in automated content moderation.

hybrids

HYBRIDS Website
www.hybridsproject.eu

Get in Touch
paloma.piot@udc.es