
INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

# Fairness in Multilingual Toxicity Detection

Final Submission

Miss. Panuja Paskkaran

Supervised by

Mr. John Sriskandarajah

Submitted in partial fulfilment of the requirements for the BEng in Software Engineering degree at the University of Westminster.

**April 2025**

**Abstract**

As toxic and harmful content on online platforms is an ever-growing concern, the cross-lingual aspect of it makes it even graver in nature. Existing toxicity detection models are mostly English centric and tend to have limited performance across different languages, leading to biased moderation. This discrepancy results in inappropriate flagging or ignoring of the content in less-represented languages. Moreover, many toxicity detection datasets suffer from class imbalance or lack cultural context, making it challenging to produce a system that is accurate and fair across languages.

To address these difficulties, we employ the XLM-RoBERTa Model an advanced transformer architecture for multi-lingual tasks. The dataset was cleaned, then augmented in two ways, back-translation in five languages (French, Italian, Turkish, Russian, Portugues) and synonym replacement using a BERT based contextual augmenter. These techniques provided additional data diversity and compensated the class imbalance. This new dataset of 100k samples was tokenized, stratified to train/ validation/ test sets, trained with Hugging Face's Trainer API and adamw optimizer trained in mixed precision with gradient accumulation to make better use of a single GPU.

After training, the fine-tuned model performed exceptionally well on tests, achieving 91.4% accuracy, 95.6% precision, 95.6% recall, and an F1-score of 95.6%. These metrics demonstrate high, consistent performance across all six toxicity categories. The combination with a multilingual transformer model yields significant improvements in fairness as well as classification performance in toxicity detection systems.