Multilingual Toxic Comment Detection: Translated vs. Original-Language Training with *Detoxify*

CIS 5300 Milestone 1

Team: *Zixuan Bian, Aria Shi, Siyuan Shen, Alex Yang*

November 4, 2025

1 Literature Review

This project studies cross-lingual toxicity detection under two competing data strategies: (i) *original-language* training and evaluation, and (ii) *translate-train/translate-test* pipelines that rely on machine translation (MT). Below we summarize three representative works that directly inform our problem setting—covering (a) toxicity datasets and bias-aware evaluation, (b) multilingual encoders for cross-lingual transfer, and (c) the effect of MT on abusive/toxic language detection.

Borkan et al. (2019): Nuanced metrics for unintended bias in toxicity classification

Problem & data. Borkan et al. introduce Civil Comments, a large-scale, human-labeled dataset for toxicity detection that includes identity-sensitive annotations and is closely related to the Jigsaw toxicity challenges [1]. The paper's central contribution is a set of bias-aware evaluation metrics (e.g., subgroup AUC, BNSP/BPSN AUC) designed to measure performance not only on the overall test set but also across identity subpopulations (e.g., terms related to race, gender, religion). **Approach & findings.** The authors benchmark linear and neural toxicity classifiers and show that models with strong overall ROC-AUC can still underperform on specific identity terms, thereby exhibiting *unintended bias*. Their metrics reveal disparities that would be invisible to aggregate scores alone and motivate training/evaluation protocols that explicitly monitor subgroup performance. For our project, this paper underpins the evaluation design: when comparing original-language versus MT-based pipelines, we should report both *overall* and *subgroup-aware* metrics to guard against performance regressions on particular communities.

Conneau et al. (2020): XLM-RoBERTa for multilingual transfer

Problem & model. Conneau et al. present XLM-RoBERTa (XLM-R), a transformer pre-trained on 100+ languages with CommonCrawl using a masked language modeling objective [**?**]. XLM-R is a drop-in multilingual encoder that supports zero-shot and few-shot transfer across languages. **Approach & findings.** The paper demonstrates that scaling both data and model capacity yields across-the-board gains on cross-lingual understanding benchmarks (e.g., XNLI, MLQA), often surpassing prior multilingual BERT variants. For toxicity detection, the practical takeaway is that a single XLM-R backbone, fine-tuned on one or several source languages, can generalize surprisingly well to new target languages without target-language labels. This directly motivates our "original-language" baseline built on XLM-R (multilingual fine-tuning across available languages) and provides a principled alternative to the translate-train paradigm.

Ranasinghe & Zampieri (2020): Cross-lingual offensive language identification

Problem & setting. Ranasinghe and Zampieri study offensive language identification in low-resource languages using cross-lingual transfer from high-resource English [2]. They evaluate two families of solutions: (i) multilingual encoders (e.g., XLM-R) fine-tuned on English (and sometimes a small amount of target-language data), and (ii) translation-based pipelines that either translate training data into the target language or translate target-language inputs into English at inference time. **Approach & findings.** The authors find that multilingual encoders fine-tuned on English often transfer well to typologically diverse, low-resource targets, and that even small target-language adaptation can yield further gains. Translation-based systems can be competitive but are sensitive to MT noise: literal or domain-mismatched translations degrade the lexical/pragmatic cues that trigger offensive/toxic classifications. For our study, this paper offers two actionable insights: (1) an XLM-R baseline trained on original-language data is a strong anchor for cross-lingual performance; (2) when using MT, adding translated text into training (*translate-train*) typically improves robustness on translated inputs compared to using MT only at test time (*translate-test*), but careful quality control is needed to mitigate MT-induced distribution shift.

2 Data Description

We rely primarily on two open datasets widely used for toxicity and offensive language classification:

- **Jigsaw Civil Comments (English)**: This dataset contains millions of English comments labeled for *toxicity*, *severe toxicity*, *obscenity*, *threat*, *insult*, and *identity-based hate* [1]. It serves as the core English source for both the baseline and cross-lingual experiments. The Civil Comments corpus also provides fine-grained annotations that allow bias and subgroup-level evaluation.

- **Multilingual Jigsaw / Toxicity 2020 extensions**: We incorporate the multilingual extension introduced by UnitaryAI's `Detoxify` repository, which includes comment-level annotations for seven languages (English, Spanish, Portuguese, Italian, French, Turkish, and Russian). These datasets were translated and quality-checked by the Jigsaw team to facilitate multilingual benchmarking.[1]

Other details please check data.md attached.

References

[1] Daniel Borkan, Lucas Dixon, Jeffrey Sorenson, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Proceedings of the Companion of The World Wide Web Conference (WWW Companion)*, 2019.

[2] Tharindu Ranasinghe and Marcos Zampieri. Multilingual offensive language identification with cross-lingual word embeddings and transformers. In *Proceedings of the 2020 EMNLP Workshop on W-NUT*, 2020.

[1] `https://github.com/unitaryai/detoxify`