Translate, then Detect: Leveraging Machine Translation for Cross-Lingual Toxicity Classification

Samuel J. Bell[⋆†] **Eduardo Sánchez**[⋆†‡] **David Dale**[†]
Pontus Stenetorp[‡] **Mikel Artetxe**[§] **Marta R. Costa-jussà**[†]
[†]Meta [‡]University College London
[§]University of the Basque Country (UPV/EHU)
{eduardosanchez, alastruey, chrisropers, costajussa}@meta.com
p.stenetorp@cs.ucl.ac.uk mikel.artetxe@ehu.eus

Abstract

Multilingual toxicity detection remains a significant challenge due to the scarcity of training data and resources for many languages. While prior work has leveraged the *translate-test* paradigm to support cross-lingual transfer across a range of classification tasks, the utility of translation in supporting toxicity detection at scale remains unclear. In this work, we conduct a comprehensive comparison of translation-based and language-specific/multilingual classification pipelines. We find that translation-based pipelines consistently outperform out-of-distribution classifiers in 81.3% of cases (13 of 16 languages), with translation benefits strongly correlated with both the resource level of the target language and the quality of the machine translation (MT) system. Our analysis reveals that traditional classifiers outperform large language model (LLM) judges, with this advantage being particularly pronounced for low-resource languages, where `translate-classify` methods dominate `translate-judge` approaches in 6 out of 7 cases. We additionally show that MT-specific fine-tuning on LLMs yields lower refusal rates compared to standard instruction-tuned models, but it can negatively impact toxicity detection accuracy for low-resource languages. These findings offer actionable guidance for practitioners developing scalable multilingual content moderation systems.

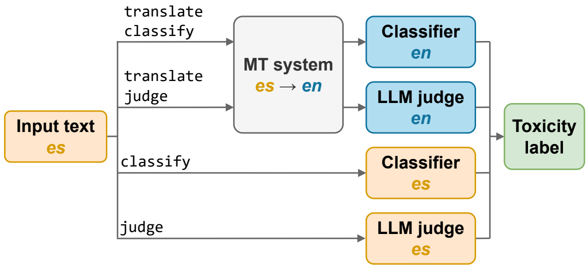


Figure 1: Across 17 languages, we evaluate toxicity detection using translation-based pipelines (`translate-classify`, `translate-judge`) against classifying in the original language (`classify`, `judge`). In this example, text in Spanish (es) is optionally translated to English (en) before classification.

1 Introduction

Detecting instances of toxic, abusive, or hateful content at scale is a challenging problem with important, real-world implications for content moderation. In a multilingual setting, however, toxicity detection is often rendered particularly difficult due to a paucity of labeled data for lower-resourced languages. In parallel, recent years have seen the scaling up of machine translation (MT) systems to cover a vast array of world languages (e.g., NLLB-Team et al., 2022), offering a potential pathway toward leveraging cross-lingual transfer for improved multilingual toxicity detection.

In monolingual non-English settings, cross-lingual transfer has already proven useful for toxicity detection (Eskelinen et al., 2023; Kobellarz and Silva, 2022), aligning with broader analyses of translation's utility for cross-lingual transfer across a range of classification tasks (Artetxe et al., 2023; Etxaniz et al., 2023b; Ponti et al., 2021). Specifically, Artetxe et al. (2023) compare *translate-test* (translating a sample before zero-shot classification) against *translate-train* (translating a sample before classification with a classifier *finetuned on translation data*) and find that *translate-test* is competitive as long as translation quality is sufficient.

While cross-lingual classification has been widely studied in other NLP tasks, toxicity detection presents distinctive challenges that warrant separate investigation. Toxic language is culturally and contextually grounded, with expressions, slurs, and taboos that often lack direct equivalents across languages, making transfer more brittle than for semantically simpler labels. Online toxicity also frequently involves code-switching, orthographic

⋆Joint first author

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1249–1264
November 8–9, 2025. ©2025 Association for Computational Linguistics

variation, and deliberate obfuscation, which may be less common in other tasks. Moreover, toxicity labels are inherently subjective and shaped by cultural norms, leading to potential label drift when transferring across languages. These factors, combined with the high stakes of moderation errors, make cross-lingual transfer in toxicity detection both consequential and scientifically challenging.

In this work, we present an empirical exploration of translation for multilingual toxicity detection, through the lens of the practitioner for whom labeled data may be unavailable—a particularly common scenario when working with lower-resourced languages—by comparing translation-based pipelines against a variety of off-the-shelf multi- and monolingual classifiers. Across 27 pipelines spanning five MT systems and nine toxicity classifiers—including both traditional classifiers and large language model (LLM) judges—we evaluate the benefit of cross-lingual classification in 17 languages with varying levels of resources.

Our results suggest that leveraging translation is an effective method for multilingual toxicity detection (§4.1), with benefits scaling in line with increasing language resources and MT system quality (§4.2). Motivated by these results, we study the issue of refusal rates and its mitigation via MT supervised finetuning (MT-SFT), as well as the downstream effect of MT-SFT on toxicity detection performance (§4.3). Finally, we explore classifying using LLM judges and compare them to traditional toxicity classifiers (§4.4). We conclude with practical recommendations for deploying multilingual toxicity detection systems at scale.

2 Related work

2.1 Multilingual toxicity detection

Multilingual toxicity detection is widely used in cases like content moderation or faithful translation (e.g. Costa-jussà et al., 2023). Prior work has either trained models using multilingual corpora of labeled training data (e.g. Hanu, 2020), or sought to exploit cross-lingual transfer via monolingual finetuning of multilingual foundation models (e.g. XLM-ROBERTa; Conneau et al., 2020a). Multilingual evaluation datasets exist for toxicity detection (e.g. Kivlichan et al., 2020; Gupta, 2021) alongside those used for text detoxification (Dementieva et al., 2024b, 2025). In this work, we evaluate a representative sample of off-the-shelf traditional classifiers, including cross-lingual, and both mono- and multilingual classifiers, across a wide variety of languages.

2.2 Cross-lingual classification

Early approaches to cross-lingual classification relied on bilingual lexicons and statistical methods to project documents into a shared feature space (Rapp, 1995; Dumais et al., 1997; Gliozzo and Strapparava, 2006). The introduction of cross-lingual word embeddings (Mikolov et al., 2013; Faruqui and Dyer, 2014; Ammar et al., 2016) enabled models trained in one language to be applied to others through a shared vector space. Prior to multilingual encoders, transfer was typically achieved via MT, either by translating the training data into the target language (*translate-train*) or by translating inputs into the source language at inference (*translate-test*) (Wan, 2009; Prettenhofer and Stein, 2010).

Multilingual sentence encoders such as LASER (Artetxe and Schwenk, 2019) and mBERT (Devlin et al., 2019) demonstrated the feasibility of direct zero-shot transfer without translation. XLM (Lample and Conneau, 2019) introduced translation language modeling to improve alignment, and XLM-R (Conneau et al., 2020b) showed consistent gains from scaling model and data size. Artetxe et al. (2020) provided a systematic comparison of translate-train and translate-test, while Etxaniz et al. (2023a) revisited translate-test with modern neural MT, finding it competitive for low-resource and distant languages.

Recent work explores large multilingual LLMs (Muennighoff et al., 2022) and parameter-efficient adaptation methods (Pfeiffer et al., 2020), aiming to combine the flexibility of fine-tuning with the scalability of zero-shot prompting.

3 Methods

We evaluate the performance of toxicity detection *pipelines*, where a pipeline comprises a binary toxicity classifier and an optional MT system. In many languages—and particularly for lower-resourced languages—labeled data for toxicity detection is unavailable, precluding the training and deployment of specialized classifiers and motivating the consideration of translation-based pipelines. As such, we are principally interested in comparing pipelines in the following three regimes:

`classify (ID)` An untranslated, in-distribution (ID) sample is classified in the source language

Language Code	Language	FineWeb-2 Docs	Dataset	No. Samples
am	Amharic	280,355	Amharic Hate Speech (Ayele et al., 2023)	1,501
ar	Arabic	57,752,149	L-HSAB (Mulki et al., 2019)	5,846
de	German	427,700,394	GermEval 2018 (Wiegand et al., 2018)	3,398
es	Spanish	405,634,303	Jigsaw Multilingual (Kivlichan et al., 2020)	8,438
fr	French	332,646,715	Jigsaw Multilingual (Kivlichan et al., 2020)	10,920
he	Hebrew	13,639,095	OffensiveHebrew (Hamad et al., 2023)	500
hi	Hindi	20,587,135	MACD (Gupta et al., 2022)	6,728
it	Italian	219,117,921	Jigsaw Multilingual (Kivlichan et al., 2020)	8,494
kn	Kannada	2,309,261	MACD (Gupta et al., 2022)	6,587
ml	Malayalam	3,406,035	MACD (Gupta et al., 2022)	5,170
pt	Portuguese	189,851,449	ToLD-Br (Leite et al., 2020)	21,000
ru	Russian	605,468,615	Russian Language Toxic Comments (Belchikov, 2019)	14,412
ta	Tamil	5,450,192	MACD (Gupta et al., 2022)	6,000
te	Telugu	2,811,760	MACD (Gupta et al., 2022)	6,000
th	Thai	35,949,449	Thai Toxicity Tweet Corpus (Sirihattasak et al., 2018)	2,794
tr	Turkish	88,769,907	Jigsaw Multilingual (Kivlichan et al., 2020)	14,000
uk	Ukrainian	47,552,562	TextDetox 2024 (Dementieva et al., 2024a)	5,000

Table 1: Toxicity datasets used per language, including number of samples, and number of documents in FineWeb-2 as a measure of language resourcedness.

using a classifier trained on data from the same distribution (e.g., evaluating a classifier on French social media posts that has been trained on French social media posts).

`classify (OOD)` An untranslated, out-of-distribution (OOD) sample is classified in the source language using a classifier trained on data from a different distribution (e.g., evaluating a classifier on French video comments that has been trained on French social media posts).

`translate-classify` The sample is translated into English using an MT model before being classified in English, using a toxicity classifier that supports English. No evaluated classifiers have been trained on translated data.

While we expect finetuned classifiers to exhibit the strongest performance while operating ID, it is relative to the far more common OOD scenario (i.e., where no suitably finetuned classifier is available to process the source language) that we expect `translate` pipelines to offer significant utility.

3.1 Evaluation

We evaluate various pipeline implementations across several languages and datasets, each of which comprise text samples x_i and gold toxicity labels y_i. Each pipeline, given a sample, produces a continuous score corresponding to toxicity.

Pipeline performance To avoid the need for thresholding, we evaluate pipeline performance via the Area Under the Receiver Operating Characteristic curve (AUC), which provides a continuous measure of how well the pipeline can separate toxic from non-toxic samples. The AUC is defined as:

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t)$$

where $\text{TPR}(t)$ and $\text{FPR}(t)$ are the true positive and false positive rates at threshold t.

When comparing pipelines, we typically evaluate the benefit of using one pipeline over another by way of change in AUC. For two pipelines, P_A and P_B,

$$\Delta\text{AUC}(P_A, P_B) = \text{AUC}(P_A) - \text{AUC}(P_B)$$

We evaluate all possible combinations of pipeline and dataset where the supported pipeline language matches the dataset's language.

Language resources We evaluate the role of language resourcefulness on pipeline performance, where we roughly approximate the number of available resources using the amount of documents available in FineWeb2 (Penedo et al., 2025), a large-scale dataset of web text sourced from various CommonCrawl snapshots.

Translation system quality Following standard practice (e.g., Kocmi et al. 2024), we additionally evaluate the quality of translations into English using the CometKiwi-DA-XL (Rei et al., 2023) quality estimation model, evaluated on the BOUQuET (Omnilingual MT Team et al., 2025) dataset.

Classifier	Supported Languages	Base Model	Training Dataset
xlm-r-finetuned-toxic-political-tweets-es	es	XLM-RoBERTa	Tweets by Spanish politicians
distilbert-base-multilingual-cased-toxicity	102 languages	DistilBERT multilingual	Jigsaw
distilbert-base-german-cased-toxic-comments	de	German DistilBERT	Various incl. GermEval 2018
russian_toxicity_classifier (Dementieva et al., 2022)	ru	RuBERT	Russian Language Toxic Comments
xlmr-large-toxicity-classifier	am, ar, de, en, es, hi, ru, uk, zh	XLM-RoBERTa	TextDetox 2024 (Dementieva et al., 2024b)
amharic-hate-speech	am	Amharic RoBERTa	Amharic Hate Speech
multilingual-toxic-xlm-roberta (Hanu, 2020)	en, es, fr, it, pt, ru, tr	XLM-RoBERTa	Jigsaw Multilingual
toxic-bert (Hanu, 2020)	en	BERT	Jigsaw

Table 2: Open-source toxicity classifiers evaluated in this work.

Model	Type
Llama 3.1 8B Instruct (Grattafiori et al., 2024)	LLM
Gemma 3 4B Instruct (Gemma Team et al., 2025)	LLM
GPT-4o (OpenAI, 2024)	LLM
NLLB 200 3.3B (NLLB-Team et al., 2022)	NMT

Table 3: Translation systems evaluated in this work.

3.2 Datasets

We curate a set of ten toxicity benchmarks for evaluating pipeline performance, spanning 17 languages, where each dataset comprises samples of text with gold labels indicating toxicity. Benchmarks were identified via searching related work on toxicity detection and by searching the Hugging Face datasets catalog. We limited our search to only datasets comprising natural human data, and to those where the gold labels are produced by human annotators, such that datasets comprising model-generated or otherwise synthetic text or labels were discarded. Datasets were restricted to those with a permissive license, where data provenance was clearly indicated, and where the data is readily-accessible online. This resulted in the following benchmarks: Amharic Hate Speech (Ayele et al., 2023); GermEval 2018 (German; Wiegand et al. 2018); Jigsaw Multilingual (Spanish, French, Italian, and Turkish partitions only; Kivlichan et al. 2020); L-HSAB (Levantine Arabic; Mulki et al. 2019); MACD (Hindi, Kannada, Malayalam, Tamil, and Telugu; Gupta et al. 2022); Offensive-Hebrew (Hamad et al., 2023); ToLD-Br (Brazilian Portuguese; Leite et al. 2020); Russian Language Toxic Comments (Belchikov, 2019); Thai Toxicity Tweet Corpus (Sirihattasak et al., 2018); and TextDetox 2024 (Ukrainian partition only; Dementieva et al. 2024a). See Table 1 for full details.

Across all datasets, only the test partition is used for evaluation. Where a toxicity classifier is trained on data that includes one of our benchmark's training partitions, we consider that classifier to be operating ID. Otherwise, as the classifier has been trained on data unlike the benchmark, we consider it to be operating OOD. See Table 2 for the training data used to produce each classifier.

For the purposes of our evaluation, we intentionally avoid drawing a distinction between toxicity detection and and hate speech detection. While hate speech and toxic or offensive are distinct concepts (Davidson et al., 2017; Waseem et al., 2017)—with hate speech typically being interpreted as directed toward a specific group (Davidson et al., 2017; Röttger et al., 2021)—in practice, most evaluation datasets use the terms toxicity, abusive or offensive language, and hate speech almost interchangeably (Fortuna et al., 2020; Banko et al., 2020). As a result, we consider datasets spanning toxicity and hate speech detection, and expect minimal difference in findings between tasks.

3.3 Toxicity classifiers

We consider eight open-source toxicity classifiers, including English-language, non-English monolingual, and multilingual, all of which are available on Hugging Face. See Appendix A.1 for selection

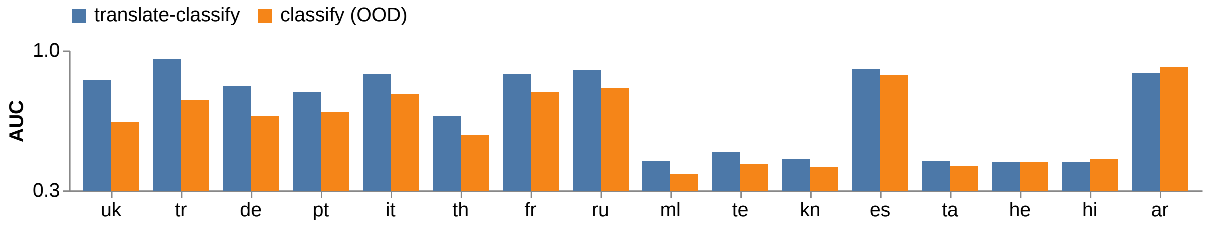


Figure 2: AUC of best possible `translate-classify` pipeline (over all combinations of translation systems and English toxicity classifiers) and best possible `classify (OOD)` pipeline (over all OOD toxicity classifiers). **The `translate-classify` approach wins across 13 out of 16 evaluated languages.**

criteria and Table 2 for full details of all classifiers considered.

All classifiers evaluated make use of pretrained Transformer-based encoder models as a backbone, such as BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), ROBERTa (Liu et al., 2019), or the multilingual XLM-ROBERTa (Conneau et al., 2020a), some of which have undergone additional fine-tuning on language specific corpora, such as Russian RuBERT (Kuratov and Arkhipov, 2019). All classifiers are then finetuned on a portion of labeled toxicity data, such as detailed in §3.2.

3.4 Translation systems

For `translate-classify` pipelines, we translate samples into English with a translation system before classifying the translations with an English-supporting classifier. We evaluate four different translation systems (see Table 3), including both encoder-decoder MT systems (NMT) and decoder-only (i.e., LLM) translation systems. In the NMT category, we use NLLB 200 3.3B (NLLB-Team et al., 2022). We evaluate three LLM systems (two open-weights and one behind-API): Llama 3.1 8B Instruct (Grattafiori et al., 2024), Gemma 3 4B Instruct (Gemma Team et al., 2025) and GPT-4o (OpenAI, 2024). The following prompt is used to produce the translations:

```
Translate the following sentence from
    ↪ {{lang}} into English. Respond
    ↪ only with the translation into
    ↪ English, without any additional
    ↪ comments.
{{sentence}}
```

4 Experiments

4.1 Translated pipelines often win

We compare the AUC of the best `translate-classify` pipeline (the best possible combination of translation system and toxicity classifier) against the best possible `classify` pipeline (the best toxicity classifier that supports each language).

Results In Fig. 2, we evaluate `translate-classify` in the common scenario where a language-specific finetuned toxicity classifier is unavailable, i.e., where classifiers are operating OOD with respect to either their source language or training domain, `classify (OOD)`. We observe that in such a scenario, the best `translate-classify` pipeline outperforms the best `classify (OOD)` pipeline across 13 of 16 languages considered (81.3%). Reducing a degree of freedom by using a fixed classifier, `distilbert-base-multilingual-cased-toxicity`, `translate-classify` still outperforms `classify` in 12 of 16 languages (75%; see Fig. S1).

In Fig. 3 we evaluate translated pipelines in scenarios where a language-specific finetuned classifier *is* available (`classify (ID)`), though we note that this is far from the case for the majority of languages. Here, `translate-classify` still offers a robust baseline, outperforming finetuned `classify (ID)` pipelines across three out of seven languages. See Table S1 for full results over all languages.

4.2 Translation benefit scales with resources

Next, we explore which factors determine the success of `translate-classify` pipelines. To allow for consistent comparison across languages and control for variability in classifier performance, we now limit ourselves to two fixed classifiers: for `translate-classify` we use the English classifier, `toxic-bert`, while for `classify` we use our most multilingual classifier, `distilbert-base-multilingual-cased-toxicity`. We evaluate the role of language resourcefulness and translation quality on change in AUC between pipelines, as specified in §3.1.

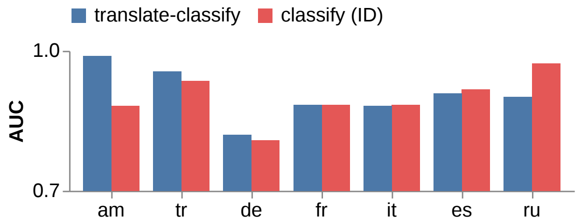


Figure 3: AUC of best possible `translate-classify` pipeline (over all combinations of translation systems and English toxicity classifiers) and best possible `classify (ID)` pipeline (over all ID toxicity classifiers). **The `translate-classify` approach still wins across three of seven languages where in-distribution finetuned classifiers are available.**

Results In Fig. 4, we observe that the relative benefit of `translate-classify` over `classify`, as measured by the change in AUC, is higher for better-resourced languages. This is consistent across four different translation systems, including both LLM and NMT systems. After fixing the best performing classifiers, we notice that the relative benefit of translation for some languages is affected, suggesting the framework is susceptible to model selection to maximize gains.

Similarly, in Fig. 5 we see that the relative benefit of `translate-classify` increases with the quality of translations in each language, across both LLM and NMT systems. We note a higher sensitivity to both language resourcefulness and translation quality for the NMT system, NLLB, compared with LLM systems.

4.3 MT-SFT reduces refusal and improves performance

When using safety-tuned LLMs for translation, a we noticed that key risk is *refusal*: the model declines to translate inputs containing harmful or toxic content, which can severely limit coverage in toxicity detection. We examine whether finetuning for MT can mitigate this problem by comparing two `translate-classify` pipelines: (1) `translate-classify (Llama 3)`, which uses translations from a standard instruction-tuned LLM (Llama 3.1 8B Instruct), and (2) `translate-classify (+TowerBlocks/MT)`, which uses translations from the same base model after supervised finetuning (MT-SFT) on the TowerBlocks/MT dataset (Alves et al., 2024) (see Appendix A.2 for details). Both pipelines feed translations to a fixed English-only classifier, `toxic-bert`, to isolate translation effects, and are compared against a direct multilingual `classify` pipeline using `distilbert-base-multilingual-cased-toxicity`.

Refusal detection We use `Minos` (Suphavadeeprasit et al., 2025) to assign each translation output $y_i = T(x_i)$ a refusal probability $P_r(y_i)$. The refusal rate is defined as:

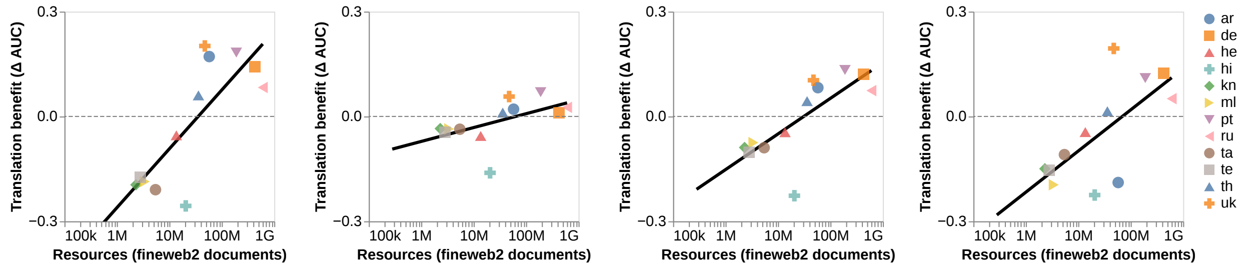
$$\mathrm{R}(T) = \frac{1}{N}\sum_{i=1}^{N}[P_r(T(x_i)) > 0.95],$$

where a 0.95 threshold minimizes false positives. For two systems T_A and T_B, the difference in refusal rates is:

$$\Delta\mathrm{R}(P_{T_A}, P_{T_B}) = \mathrm{R}(P_{T_A}) - \mathrm{R}(P_{T_B}).$$

Refusal results As shown in Fig. 8, `translate-classify (+TowerBlocks/MT)` reduces refusal rates in *every* language compared with `translate-classify (Llama 3)`. The reduction scales approximately log-linearly with language resources (Fig. 9a), indicating that MT-SFT particularly benefits high-resource languages where refusals are rarer but still impactful. Lower refusal means more toxic content is actually processed by the classifier, directly improving pipeline coverage.

Human verification of refusal mitigation To validate both the accuracy of our automated refusal detection and the effectiveness of MT-SFT in addressing refusals, we conducted a targeted human annotation study. For each dataset, we randomly sampled up to 5% of the content flagged as refusals by the base Llama 3.1 8B Instruct model, with a minimum of 10 examples per dataset. Annotators manually verified whether each flagged case was indeed a refusal, then examined translations of the same inputs generated by the MT-finetuned model. As shown in Table 4, the refusal detector achieved perfect true positive rates for Thai, German, and Ukrainian, and high —though not perfect— accuracy for Malayalam and Levantine Arabic, where some false positives were observed. Importantly, the MT-finetuned model produced valid translations for *all* annotated examples, yielding a true negative rate of 100% across every language in the sample. This confirms that, at least for the languages tested, MT-SFT can completely eliminate refusals observed in the base instruction-tuned model, turning previously blocked content into usable inputs for the downstream classifier.



(a) GPT-4o (b) Llama 3.1 8B Instruct (c) Gemma 3 4B Instruct (d) NLLB

Figure 4: Change in AUC (i.e., translation benefit) between `translate-classify` pipelines with a fixed English classifier, `toxic-bert`, and `classify` pipelines with a fixed multilingual classifier, `distilbert-base-multilingual-cased-toxicity`, as a function of language resources, over four translation systems **(a)** GPT-4o, **(b)** Llama 3.1 8B Instruct, **(c)** Gemma 3 4B Instruct, and **(d)** NLLB. **Translation benefit is increased for higher resourced languages.**

(a) Llama 3.1 8B Instruct (b) NLLB

Figure 5: Change in AUC (i.e., translation benefit) between `translate-classify` pipelines and `classify` pipelines, as a function of English translation quality measured by `CometKiwi-DA-XL`, over two translation systems **(a)** Llama 3.1 8B Instruct, and **(b)** NLLB. **Translation benefit increases with translation quality for both LLM-based and NMT systems.**

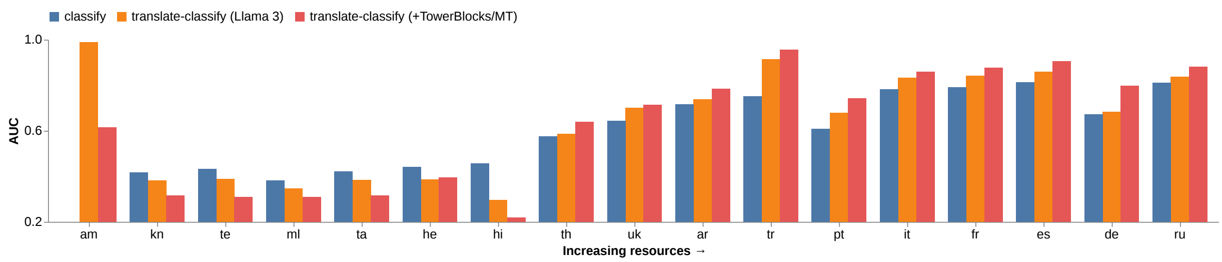


Figure 6: AUC of `translate-classify (Llama 3)` and `translate-classify (+TowerBlocks/MT)` using a fixed English classifier, `toxic-bert`, and a `classify` pipeline using a fixed multilingual classifier, `distilbert-base-multilingual-cased-toxicity`. **Using a finetuned LLM for translation improves pipeline performance for higher-resourced languages.**

MT-SFT improves performance for high resource languages In addition to lowering refusals, MT-SFT also improves classification accuracy. In Fig. 6, `translate-classify (+TowerBlocks/MT)` achieves higher AUC than `translate-classify (Llama 3)` for 11 of 17 languages, with gains concentrated in high-resource settings. When measured against the multilingual `classify` baseline, `translate-classify (+TowerBlocks/MT)` shows even stronger sensitivity to language resource availability (Fig. 7).

4.4 LLM judges underperform on lower-resourced languages

Given the strong performance of LLMs across a range of tasks, we additionally compare pipelines based on traditional classifiers vs. zero-shot LLM judges.

Specifically, we analyze the performance of

Name	TPR (`Llama 3.1`)	TNR (`+TowerBlocks/MT`)
TH	100%	100%
DE	100%	100%
UK	70%	100%
ML	75%	100%
AR	40%	100%

Table 4: Analysis of human annotations of refusal predictions, showing True Positive Rate (TPR) of Llama 3.1 8B Instruct (`Llama 3.1`) and the True Negative Rate (TNR) of the same model finetuned on TowerBlocks/MT (`+TowerBlocks/MT`). **Refusal detection is highly accurate for Thai, German, and Ukrainian. Even for Malayalam and Levantine Arabic where the refusal detector produces some false positives, the finetuned model never refuses.**

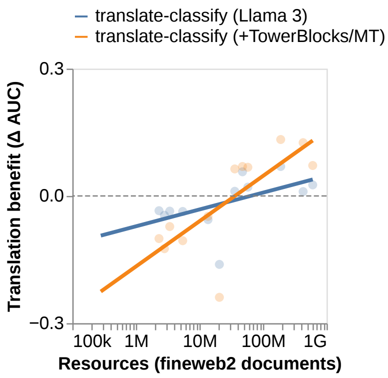


Figure 7: Change in AUC (i.e., translation benefit) of `translate-classify (Llama 3)` and `translate-classify (+TowerBlocks/MT)` pipelines with a fixed English classifier against a fixed `classify` pipeline with a fixed multilingual classifier, as a function of language resources. **The `translate-classify (+TowerBlocks/MT)` pipeline performance is more sensitive to available language resources.**

two judge-based pipelines: (1) `judge`, where an LLM is prompted to classify untranslated samples, and (2) `translate-judge`, where the LLM is prompted to classify samples that have been translated into English. The judge in both cases is Llama 3.1 8B Instruct with the following custom prompt, where `{{sample}}` is either the translated or original sample according to pipeline:

```
You are a toxicity classifier. Classify
    ↪ the following text as toxic or
    ↪ not toxic. Respond with 'toxic'
    ↪ or 'not toxic' with no other
    ↪ comments or explanation.
Text: {{sample}}
Classification:
```

Translations are also produced using Llama 3.1 8B Instruct. We compare both judge pipelines against a `classify` pipeline based on the overall best English classifier, `toxic-bert`, coupled again with Llama 3.1 8B Instruct for fair comparison. We evaluate all pipelines using accuracy as opposed to AUC, as the LLM judge emits only a label, rather than a continuous score.

Results Figure 10 illustrates notable patterns in the comparative performance of `translate-judge` and `judge` pipelines. Across all languages, translation-based approaches narrowly outperform their untranslated counterparts; however, this advantage becomes pronounced in low-resource settings, where `translate-judge` completely dominates, outperforming `judge` in 6 out of 7 low-resource languages. Similarly, `translate-classify` pipelines provide a slight overall edge compared to both `judge` and `translate-judge`, but the margin is especially significant for low-resource languages, where `translate-classify` overwhelmingly wins (again in 6 out of 7 cases). These results further indicate that multilingual capabilities in LLMs are not homogeneously distributed: while MT models demonstrate broader multilingual reach, toxicity classification performance by LLMs is markedly less consistent across lower-resource languages.

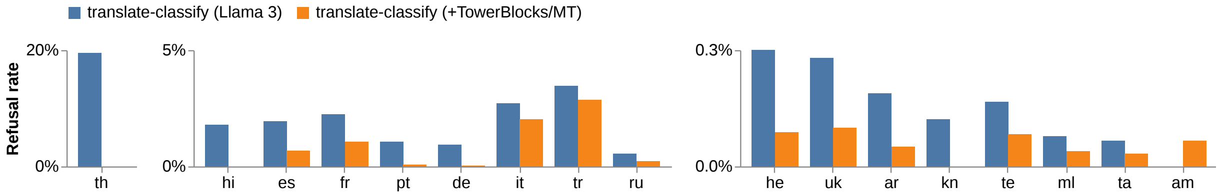
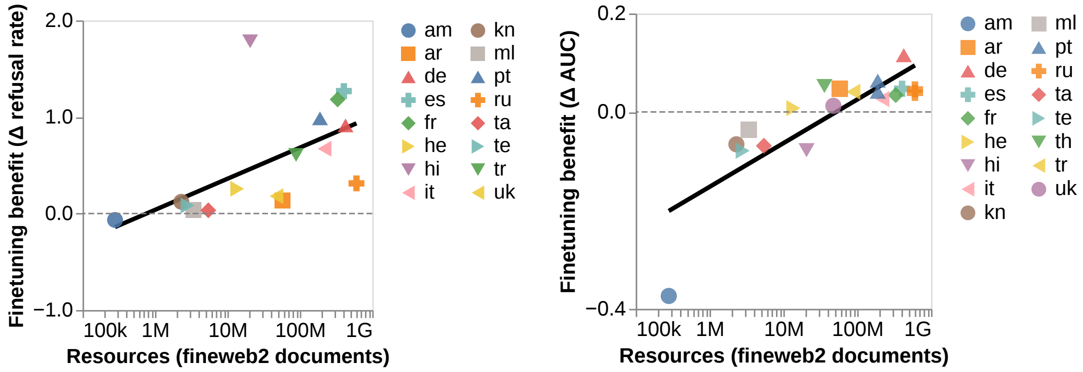


Figure 8: Translation refusal rate of `translate-classify (Llama 3)` and `translate-classify (+TowerBlocks/MT)` pipelines. Note three separate scales for legibility. **Using a finetuned LLM for translation reduces refusal rates.**



(a) Δ refusal

(b) Δ AUC

Figure 9: Change in **(a)** translation refusal rate and **(b)** AUC of a `translate-classify (+TowerBlocks/MT)` pipeline against a `translate-classify (Llama 3)` pipeline, both with a fixed English classifier, `toxic-bert`, as a function of language resources. **The benefit of using a finetuned LLM for translation, in terms of both refusal rates and improved performance, increases for with language resources.**

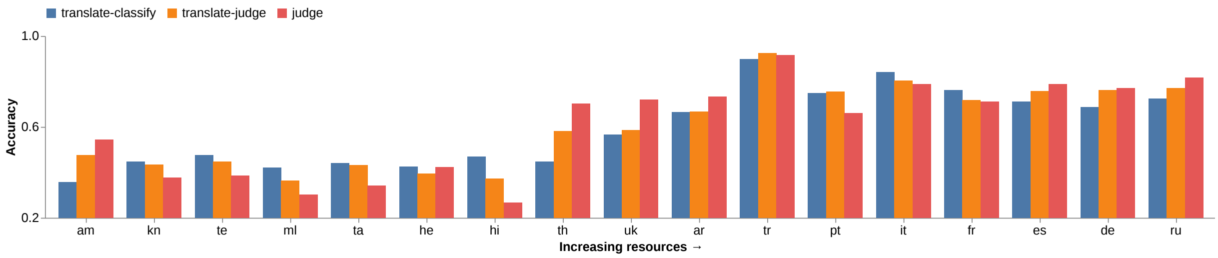


Figure 10: Accuracy of `translate-judge`, `judge`, and `translate-classify` with a fixed English classifier, `toxic-bert`, all using a Llama 3 for translation. **Translation with traditional classifiers outperforms LLM judges for most lower resourced languages.**

5 Discussion

Across ten benchmarks spanning 17 languages, our analysis suggests that translation-based approaches can be successfully leveraged to support multilingual toxicity detection at scale. Specifically, we observe that `translate-classify` pipelines outperform `classify (OOD)`, a non-finetuned classifier operating OOD (i.e., an off-the-shelf model) in the majority of cases, and can even occasionally outperform `classify (ID)`, dedicated finetuned classifiers evaluated ID. The relative benefit of using `translate-classify` over `classify` pipelines increases with both a language's available resources and the quality of the translation system. This may suggest that while translation may be an effective strategy *in general*, it does have the potential to increase performance disparities between better- and worse-resourced languages. We additionally note that using an MT-finetuned LLM for translations can further drive up pipeline performance, in part, by reducing refusal rates, but that this benefit appears to be reserved for higher-resourced languages. Finally, we evaluate the utility of an LLM judge approach over traditional (e.g., BERT-based) classification, finding that in lower-resourced languages, translate-classify consistently outperforms.

Practical recommendations We make four practical recommendations for practitioners looking to deploy multilingual toxicity detection at scale.

1. At the very least, translate-classify pipelines using traditional classifiers and LLM-based translation should be considered a robust baseline.

2. If fine-tuning on dedicated data is unavailable, a translate-classify pipeline is likely to provide a strong first choice of model, particularly in languages where translation quality is high.

3. If operating on a higher-resourced language, making use of an MT-finetuned LLM may offer some performance improvements over a standard instruction-tuned LLM, particularly in the scenario where refusal rates can be reduced.

4. Unlike many other NLP tasks, an LLM judge demonstrates only a limited performance advantage on select higher-resourced languages when compared to traditional (e.g., BERT-based) classifiers.

Limitations

While we approach multilingual toxicity detection through the lens of a practitioner making a choice between available, off-the-shelf pipeline components, this does limit our ability to analyze the role of specific finetuning details. For example, in contrast with previous work (Artetxe et al., 2023) that has contrasted cross-lingual transfer pipelines where the classifier was finetuned on either the original domain or the outputs of the translation system, we only make use of publicly-available classifiers which may be finetuned on different numbers of samples or different domains, and none of which are finetuned on translations. However, given the performance improvements offered by the translate-classify pipeline *without finetuning on translations*, we might expect a translation-finetuned classifier to further benefit the translate-classify approach.

As we note in §3.2, our work is also potentially limited by shifts in data distribution between languages. In order to identify broad trends across many languages with different levels of resources, we draw samples from different constituent datasets. These datasets, however, are drawn from different domains (e.g., social media vs. WikiMedia talk pages) with labels produced using different annotation schemas (e.g., identifying hate speech vs. toxicity). As a result, our conclusions should be interpreted as indicative of general trends about the relative utility of translation, rather than individual claims about how well translation may function on any given language. This limitation could be overcome with access to additional highly-multilingual datasets of labeled toxicity data.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Waleed Ammar, Phoebe Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 152–157, Berlin, Germany. Association for Computational Linguistics.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of EMNLP*, pages 7674–7684.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. In *Transactions of the Association for Computational Linguistics*, volume 7, pages 597–610.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. Exploring Amharic Hate Speech Data Collection and Classification Approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 49–59, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A Unified Taxonomy of Harmful Content. In

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 125–137, Online. Association for Computational Linguistics.

Anatoly Belchikov. 2019. Russian Language Toxic Comments.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Marta Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9570–9586, Singapore. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025, Abu Dhabi, UAE. Association for Computational Linguistics.

Daryna Dementieva, Valeriia Khylenko, Nikolay Babakov, and Georg Groh. 2024a. Toxicity Classification in Ukrainian. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 244–255, Mexico City, Mexico. Association for Computational Linguistics.

Daryna Dementieva, Varvara Logacheva, Irina Nikishina, Alena Fenogenova, David Dale, Irina Krotova, Nikita Semenov, Tatiana Shavrina, and Alexander Panchenko. 2022. RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2022"*, pages 114–131. RSUH.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024b. Overview of the Multilingual Text Detoxification Task at PAN 2024. *CLEF 2024: Conference and Labs of the Evaluation Forum*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, pages 18–24.

Anni Eskelinen, Laura Silvala, Filip Ginter, Sampo Pyysalo, and Veronika Laippala. 2023. Toxicity detection in Finnish using machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 685–697, Tórshavn, Faroe Islands. University of Tartu Library.

Julen Etxaniz, Mikel Artetxe, and Rodrigo Agerri. 2023a. On the effectiveness of translate-test for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9933–9947, Singapore. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023b. Do multilingual language models think better in english? *Preprint*, arXiv:2308.01223.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Gemma Team, Aishwarya Kamath, Johan Ferret, . . . , Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting lexical alignment for cross-language textual entailment. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–40, Trento, Italy. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, . . . , Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Vikram Gupta. 2021. Multilingual and multilabel emotion recognition using virtual adversarial training. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 74–85, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Hastagiri Prakash Vanchinathan, and Animesh Mukherjee. 2022. Multilingual Abusive Comment Detection at Scale for Indic Languages. *Advances in Neural Information Processing Systems*, 35:26176–26191.

Nagham Hamad, Mustafa Jarrar, Mohammad Khalilia, and Nadim Nashif. 2023. Offensive Hebrew Corpus and Detection using BERT. In *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8.

Laura Hanu. 2020. Detoxify.

Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw Multilingual Toxic Comment Classification.

Jordan K. Kobellarz and Thiago H. Silva. 2022. Should we translate? evaluating toxicity in online comments when translating from portuguese to english. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '22, page 89–98, New York, NY, USA. Association for Computing Machinery.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *Preprint*, arXiv:1905.07213.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. In *arXiv preprint arXiv:1309.4168*.

Niklas Muennighoff, Nouamane Tazi, and Sebastian Ruder. 2022. Crosslingual generalization through multitask finetuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1596–1610.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Omnilingual MT Team, Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Cynthia Gao, Jean Maillard, Alex Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. BOUQuET: Dataset, Benchmark and Open initiative for Universal Quality Evaluation in Translation. *Preprint*, arXiv:2502.04314.

OpenAI. 2024. GPT-4o System Card. *Preprint*, arXiv:2410.21276.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein

Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language. *Preprint*, arXiv:2506.20920.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of EMNLP*, pages 7654–7673.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer. *Preprint*, arXiv:2107.11353.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, Uppsala, Sweden. Association for Computational Linguistics.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up COMETKIWI: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. *Preprint*, arXiv:2309.11925.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Sugan Sirihattasak, Mamoru Komachi, and H. Ishikawa. 2018. Annotation and Classification of Toxicity for Thai Twitter. In *Proceedings of TA-COS 2018 – 2nd Workshop on Text Analytics for Cybersecurity and Online Safety*.

Jai Suphavadeeprasit, Teknium, Chen Guang, Shannon Sands, and rparikh007. 2025. Minos classifier.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 235–243, Suntec, Singapore. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. Technical report, Verlag der Österreichischen Akademie der Wissenschaften.

A Additional methods

A.1 Toxicity classifier selection

We evaluate on a sample of toxicity classifiers that are publicly-available on Hugging Face. We reviewed classifiers that matched the search terms "toxic" and "toxicity", selecting those that supported either English or one or more of the 17 languages analyzed. Classifiers were limited to those that were permissively-licensed, with clear data provenance (to allow for distinguishing between ID and OOD performance), and substantial community engagement (as measured by downloads and likes). See Table 2 for all classifiers evaluated.

A.2 MT finetuning an LLM

We used Llama 3.1 8b Instruct as our baseline model and finetuned it for 5 epochs with the MT split from Towerblocks 0.2, a multi-task, multilingual SFT dataset. We employed the AdamW optimizer with a learning rate initialized to 1×10^{-6}, β_1 and β_2 coefficients set to 0.9 and 0.95 respectively, and a weight decay of 0.1. We used a cosine annealing learning rate scheduler configured with a final learning rate scaled to 0.2 times the initial rate and a total of 1,000 warmup steps.

B Additional results

In Table S1 we present the detailed results behind Figs. 2 and 3, showing the performance of the best-possible `translate-classify`, `classify (OOD)`, and `classify (ID)` pipelines over all languages. In Tables S2 to S4 we present the corresponding best-performing translation system and classifier combinations for `translate-classify`, `classify (OOD)`, and `classify (ID)` respectively.

In Fig. S1, we present a version of Fig. 2 but reducing one degree of freedom: rather than choosing the best-possible combination of translation system and classifier, here we choose the best possible translation system though use a fixed classifier, `distilbert-base-multilingual-cased-toxicity`. In this setting, `translate-classify` still outperforms across 12 of 16 languages.

		AUC	
Language	ID	OOD	Translated
ar	-	**0.92**	0.89
he	-	**0.44**	0.44
hi	-	**0.46**	0.44
kn	-	0.42	**0.45**
ml	-	0.38	**0.45**
pt	-	0.69	**0.79**
ta	-	0.42	**0.44**
te	-	0.43	**0.49**
th	-	0.57	**0.67**
uk	-	0.64	**0.85**
am	0.88	-	**0.99**
de	0.81	0.67	**0.82**
es	**0.92**	0.88	0.91
fr	**0.88**	0.79	0.88
it	**0.88**	0.78	0.88
ru	**0.97**	0.81	0.90
tr	0.94	0.75	**0.96**

Table S1: Best possible performance over all languages. Where a finetuned classifier isn't available, translation-based pipelines often outperform.

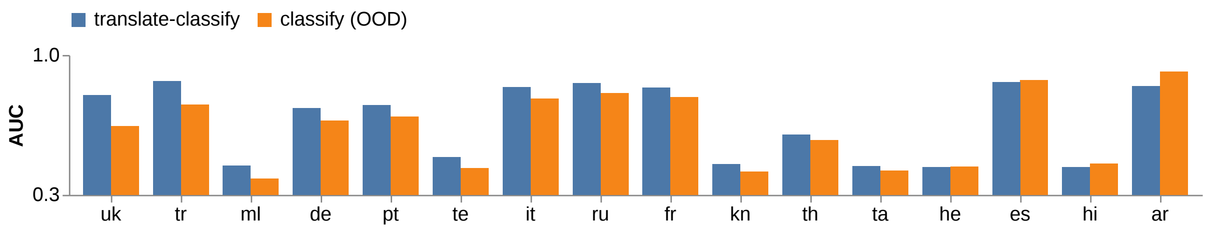


Fig. S1: Translation-based toxicity detection pipelines with a fixed English-supporting classifier, `distilbert-base-multilingual-cased-toxicity`, outperform off-the-shelf pipelines across 12 out of 16 evaluated languages.

Language	Classifier	AUC
am	textdetox/xlmr-large-toxicity-classifier	0.88
de	ml6team/distilbert-base-german-cased-toxic-comments	0.81
es	unitary/multilingual-toxic-xlm-roberta	0.92
fr	unitary/multilingual-toxic-xlm-roberta	0.88
it	unitary/multilingual-toxic-xlm-roberta	0.88
ru	s-nlp/russian_toxicity_classifier	0.97
tr	unitary/multilingual-toxic-xlm-roberta	0.94

Table S2: Best-performing ID pipeline per language.

Language	Classifier	AUC
ar	textdetox/xlmr-large-toxicity-classifier	0.92
de	citizenlab/distilbert-base-multilingual-cased-toxicity	0.67
es	textdetox/xlmr-large-toxicity-classifier	0.88
fr	citizenlab/distilbert-base-multilingual-cased-toxicity	0.79
he	citizenlab/distilbert-base-multilingual-cased-toxicity	0.44
hi	citizenlab/distilbert-base-multilingual-cased-toxicity	0.46
it	citizenlab/distilbert-base-multilingual-cased-toxicity	0.78
kn	citizenlab/distilbert-base-multilingual-cased-toxicity	0.42
ml	citizenlab/distilbert-base-multilingual-cased-toxicity	0.38
pt	unitary/multilingual-toxic-xlm-roberta	0.69
ru	citizenlab/distilbert-base-multilingual-cased-toxicity	0.81
ta	citizenlab/distilbert-base-multilingual-cased-toxicity	0.42
te	citizenlab/distilbert-base-multilingual-cased-toxicity	0.43
th	citizenlab/distilbert-base-multilingual-cased-toxicity	0.57
tr	citizenlab/distilbert-base-multilingual-cased-toxicity	0.75
uk	citizenlab/distilbert-base-multilingual-cased-toxicity	0.64

Table S3: Best-performing OOD pipeline per language.

Language	Translation system	Classifier	AUC
am	Llama 3.1 8B Instruct	unitary/toxic-bert	0.99
ar	GPT-4o	unitary/toxic-bert	0.89
de	GPT-4o	unitary/multilingual-toxic-xlm-roberta	0.82
es	GPT-4o	unitary/toxic-bert	0.91
fr	GPT-4o	unitary/toxic-bert	0.88
he	Llama 3.1 8B TowerBlocks	citizenlab/distilbert-base-multilingual-cased-toxicity	0.44
hi	Llama 3.1 8B Instruct	citizenlab/distilbert-base-multilingual-cased-toxicity	0.44
it	GPT-4o	unitary/toxic-bert	0.88
kn	Llama 3.1 8B Instruct	citizenlab/distilbert-base-multilingual-cased-toxicity	0.45
ml	Llama 3.1 8B Instruct	citizenlab/distilbert-base-multilingual-cased-toxicity	0.45
pt	GPT-4o	unitary/toxic-bert	0.79
ru	GPT-4o	unitary/multilingual-toxic-xlm-roberta	0.90
ta	Llama 3.1 8B Instruct	citizenlab/distilbert-base-multilingual-cased-toxicity	0.44
te	Llama 3.1 8B Instruct	citizenlab/distilbert-base-multilingual-cased-toxicity	0.49
th	GPT-4o	textdetox/xlmr-large-toxicity-classifier	0.67
tr	Llama 3.1 8B TowerBlocks	unitary/toxic-bert	0.96
uk	GPT-4o	unitary/multilingual-toxic-xlm-roberta	0.85

Table S4: Best-performing translated pipeline per language.