Team Pratham at TextDetox CLEF 2025/Multilingual Text Detoxification 2025: Exploring different Methods for Multilingual Text Detoxification via Prompted MT0-XL,Lexical Filtering and Backtranslation.

Notebook for the PAN Lab at CLEF 2025

Pratham Shah[1,*,†], Vatsal Shah[1,†] and Sunil Kale[2]

[1]VJTI, INDIA
[2]CS Department, VJTI, INDIA

Abstract

This paper presents our solution to the Multilingual Text Detoxification task at PAN 2025, developed by Team Pratham. We implemented a hybrid system that combines prompted inference using the MT0-XL-detox-orpo[1] model with targeted lexical filtering to improve detoxification quality across 15 languages. To address challenges in code-mixed and morphologically rich languages, we introduced a backtranslation-based pipeline for six languages and developed handcrafted toxic word dictionaries (for Hinglish) and used multilingual-toxic-lexicon[2] for fine-grained filtering. Our system relies on zero-shot prompting without additional fine-tuning, leveraging multilingual transformer generalization along with rule-based post-processing. Evaluated on both automatic and human metrics, our approach achieves strong performance, including a J-score of 0.676 on the final test set, demonstrating near state-of-the-art performance across several languages.

Keywords

PAN 2025, Detoxification, MT0-XL, Lexical Filtering, XLM-R, LaBSE, Multilingual NLP, Backtranslation, NLLB-600M

1. Introduction

Toxic language on the internet has become a growing concern, affecting how people communicate and engage online. From hate speech to subtle offensive remarks, such content can make digital spaces feel hostile and unsafe. As online platforms become more global, tackling toxicity in multiple languages becomes even more important—but also more complex due to differences in grammar, slang, and cultural nuance. One solution that goes beyond simply blocking or deleting content is text detoxification—rewriting harmful language into a more neutral, respectful form while preserving the original meaning. This not only helps maintain healthy conversations but also reduces the need for harsh censorship. In this paper, we present our solution submitted to the PAN 2025 Multilingual Text Detoxification task. Our system is built upon a two-stage hybrid pipeline that integrates both neural and rule-based approaches. First, we leverage the multilingual instruction-tuned language model MT0-XL to generate detoxified outputs using task-specific prompting strategies. MT0-XL operates in zero-shot inference mode, eliminating the need for additional fine-tuning. Second, we apply language-specific lexical filtering techniques to identify and sanitize residual toxic expressions that may bypass model filtering, especially in morphologically complex or code-mixed inputs.

To further enhance performance for challenging languages such as Hinglish, Hebrew, Japanese, Italian,

[1] `s-nlp/mt0-xl-detox-orpo` on HuggingFace: https://huggingface.co/s-nlp/mt0-xl-detox-orpo
[2] `textdetox/multilingual_toxic_lexicon` on HuggingFace: https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.
†These authors contributed equally.
npshah_b22@ce.vjti.ac.in (P. Shah); vpshah_b22@ce.vjti.ac.in (V. Shah); sdkale@ce.vjti.ac.in (S. Kale)
https://www.researchgate.net/profile/Sunil-Kale-2 (S. Kale)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

French, and Tatar, we introduce a backtranslation-based pipeline involving intermediate detoxification in a high-resource pivot language(English) followed by retranslation into the target language using NLLB-600M[1] model. This approach is supported by custom toxic word dictionaries(for Hinglish) and multilingual lexicons for finer-grained control.

Our system supports detoxification in 15 languages, covering both high- and low-resource settings. The evaluation results based on standard automatic metrics - XCOMET[2]/Chrf for fluency, LaBSE[3] for semantic similarity, and a fine-tuned XLM-R[4] classifier for toxicity - demonstrate that our hybrid method achieves strong performance, including a final J score of 0.676 on the PAN 2025 test set, securing 4th position in the final Leaderboard. These results highlight the effectiveness of combining large multilingual transformers with lightweight, language-aware filtering mechanisms.

2. Previous Work

Text detoxification has been approached through a variety of modeling strategies, broadly categorized into *token-level filtering*, *sequence-to-sequence rewriting*, and *controlled generation*. Early efforts relied on encoder-only models such as **BERT**, which were used to identify and filter toxic content at the token or sentence level. These approaches typically lacked generative capabilities and were limited to binary classification or masking.

With the advent of powerful encoder-decoder models, **sequence-to-sequence (seq2seq)** approaches became more popular. Models like **T5**, **mT5**, and **mBART**[5] enabled the rewriting of toxic sentences into non-toxic paraphrases by training on parallel corpora or through fine-tuned task-specific pipelines. These methods allowed for greater flexibility in preserving meaning while transforming sentence tone or style.

Several shared tasks have accelerated progress in this domain. Notably, the **RUSSE-2022** and **PAN 2024** detoxification tracks highlighted the effectiveness of **instruction-tuned models** such as **MT0**, which showed strong performance across multiple languages without the need for extensive fine-tuning. These tasks also emphasized the importance of *semantic preservation*, *toxicity reduction*, and *fluency* in generated outputs.

Our work builds on these foundations but introduces two key distinctions. First, we apply **MT0 in a zero-shot inference mode**, leveraging its instruction-following capability without additional fine-tuning. This reduces dependence on task-specific datasets and allows for broader generalization across languages. Second, we enhance this with a **handcrafted, language-specific lexical filtering layer**, which targets residual toxic expressions not effectively handled by generic models—particularly in low-resource and code-mixed scenarios.

While prior works have explored backtranslation or multilingual adaptation, few have combined large instruction-tuned models with custom, linguistically informed detoxification strategies. By doing so, our approach addresses persistent challenges in *multilingual detoxification*, particularly for languages with limited training data, complex morphology, or informal syntactic patterns.

3. Methodology

3.1. Prompted Detoxification with MT0-XL

We used the s-nlp/mt0-xl-detox-orpo model from Hugging Face Transformers as the backbone of our multilingual detoxification pipeline. This model is an instruction-tuned variant of MT0-XL, optimized using Direct Preference Optimization (ORPO) specifically for detoxification tasks. We chose this model

[1] `facebook/nllb-200-distilled-600M` on HuggingFace: https://huggingface.co/facebook/nllb-200-distilled-600M
[2] `myyycroft/XCOMET-lite` on HuggingFace: https://huggingface.co/myyycroft/XCOMET-lite
[3] `sentence-transformers/LaBSE` on HuggingFace: https://huggingface.co/sentence-transformers/LaBSE
[4] `textdetox/xlmr-large-toxicity-classifier-v2` on HuggingFace: https://huggingface.co/textdetox/xlmr-large-toxicity-classifier-v2
[5] `textdetox/mbart-detox-baseline` on HuggingFace: https://huggingface.co/textdetox/mbart-detox-baseline

due to its robust multilingual support, strong zero-shot performance, and proven ability to generate fluent and semantically consistent detoxified text using natural language prompts.

The model was deployed in inference-only mode, with no additional fine-tuning. Instead, we leveraged its instruction-following capabilities by designing a set of language-specific prompts such as "Detoxify:", "Entgiften:", and "Desintoxicar:", covering all 15 languages in the PAN 2025 task. These prompts were organized in a manually curated dictionary and dynamically injected into each input at runtime.

We implemented a modular inference pipeline in PyTorch using Hugging Face Transformers. Our custom DetoxificationDataset class handled multilingual tokenization, batch collation, and prompt templating. This enabled efficient large-scale inference across all languages in a unified framework. Our contributions also included:

- **C**rafting a multilingual prompt strategy to optimize zero-shot detoxification
- **D**esigning and integrating a batch-wise inference system for scalable deployment
- **D**eveloping post-processing modules including lexical filtering and toxicity masking
- **C**onducting detailed evaluation using XLM-R for toxicity, LaBSE for semantic similarity, and XCOMET and Chrf for fluency.

Although MT0-XL was able to generate outputs for all 15 target languages, it did not perform equally well across them. Specifically, its effectiveness dropped for few languages—Hinglish, Hebrew, Japanese, Chinese, and Tatar. In these cases, the model often struggled to completely eliminate toxic expressions or to maintain fluency and semantic coherence. This can be attributed to factors such as linguistic complexity, code-mixing, morphological richness, and the model's limited exposure to these languages during pretraining.

To overcome these limitations, we extended our system with two complementary strategies: language-specific lexical filtering and a backtranslation-based detoxification pipeline. Initially, these approaches were developed and evaluated independently. However, we ultimately combined them to form a hybrid model that addressed both structural and linguistic challenges more effectively.

This hybrid approach significantly improved output quality—particularly for the more complex or underrepresented languages—by first translating toxic inputs into a high-resource pivot language for intermediate detoxification, and then retranslating them back into the original language. To further refine the results, we applied custom toxic word dictionaries (e.g., for Hinglish) and multilingual lexicon-based filters to detect and eliminate residual toxicity in the final output.

3.2. MT0-XL + Language-Specific Lexical Filtering

3.2.1. Language-Specific Lexical Filtering

For languages like Hinglish, we curated custom toxic word dictionaries tailored to the unique challenges of code-mixed language. The decision to build a handcrafted list arose from the limited coverage of existing lexicons, which often missed informal, transliterated, or culturally specific slurs commonly used in toxic Hinglish expressions. Since we are familiar with the language, we manually analyzed toxic sentences from the dataset and extracted additional toxic terms that were frequently missed by the model. These words were then included in our dictionary to ensure the model would delete or mask them effectively during post-processing.

For other languages such as Hebrew, Japanese, and Chinese, we leveraged publicly available toxic lexicons from Hugging Face and similar resources. During post-processing, we applied a combination of strategies—including masking, deletion, and soft lexical replacements—to eliminate residual toxic expressions that were not fully addressed by MT0-XL, as illustrated in Figure 1 below. This post-processing step proved critical for enhancing detoxification quality, especially in morphologically rich or low-resource languages where generative models alone were insufficient.

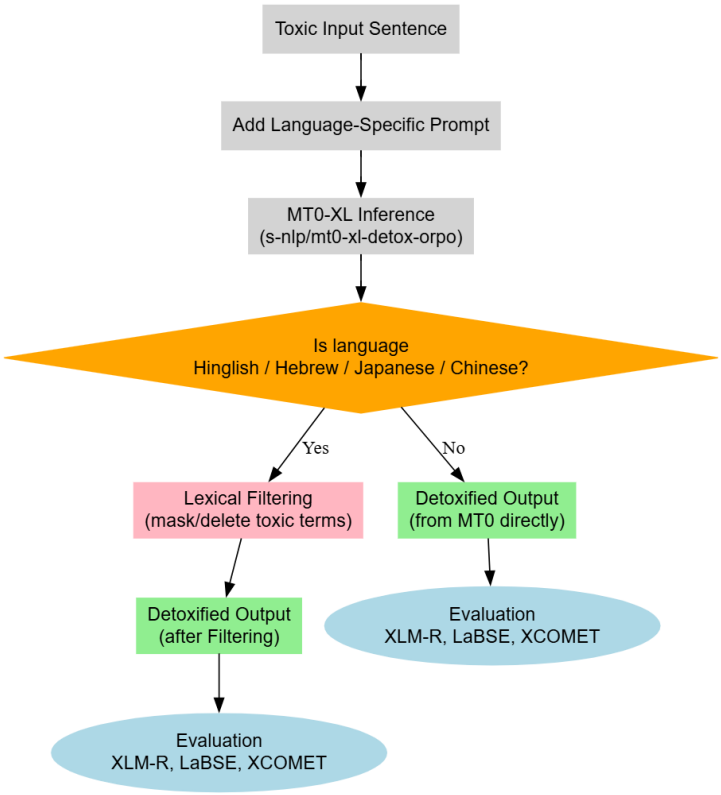

Figure 1: Overview of the detoxification pipeline using MT0 and lexical filtering.

Table 1
Results across 15 languages using XCOMET and ChrF for fluency, LaBSE for similarity, STA for toxicity neutrality, and J-score as the official metric.

Language	XCOMET ↑	ChrF ↑	Similarity (LaBSE) ↑	Toxicity (STA) ↓	J-Score ↑
Amharic	0.743	0.476	0.785	0.810	0.486
Arabic	0.872	0.784	0.902	0.913	0.724
German	0.951	0.815	0.933	0.843	0.750
English	0.896	0.730	0.875	0.918	0.729
Spanish	0.880	0.674	0.865	0.901	0.696
French	0.908	0.816	0.914	0.902	0.752
Hebrew	0.705	0.300	0.695	0.812	0.413
Hindi	0.849	0.712	0.863	0.844	0.631
Hinglish	0.697	0.553	0.776	0.615	0.339
Italian	0.867	0.756	0.927	0.929	0.749
Japanese	0.860	0.718	0.923	0.741	0.591
Russian	0.896	0.755	0.884	0.944	0.755
Tatar	0.830	0.764	0.883	0.780	0.584
Ukrainian	0.898	0.795	0.917	0.937	0.776
Chinese	0.845	0.479	0.844	0.732	0.530

3.2.2. Evaluation Results

Our system demonstrated robust performance across the 15 evaluated languages, with an overall J-score average of 0.634, and an XCOMET fluency score average of 0.846.(Table 1)

High-resource languages such as Ukrainian (J = 0.776), Russian (0.755), German (0.750), French (0.752), and Italian (0.749) achieved the highest J-scores, supported by strong performance across all sub-metrics,

including fluency (XCOMET greater than 0.89), semantic similarity (LaBSE greater than 0.91), and high toxicity neutrality (STA greater than 0.92).

In contrast, Hinglish (0.339), Hebrew (0.413), Chinese (0.530), and Amharic (0.486) still scored significantly lower in J-score, reflecting challenges in handling code-mixed, low-resource, or morphologically complex languages. These results reaffirmed the necessity of our backtranslation pipeline, which helped mitigate but not entirely resolve residual toxicity and fluency degradation in such languages.

The average J-score for non-pivot languages was 0.656, while non-pivot languages scored 0.571 when using only MT0-XL, demonstrating the effectiveness of our adaptive strategy in challenging linguistic settings.

3.3. MT0-XL + Backtranslation based Detoxification

3.3.1. Backtranslation-Based Detoxification

While MT0-XL exhibited strong performance in multilingual detoxification tasks, its direct application to certain languages revealed notable shortcomings. Specifically, for **Italian, French, Japanese, Tatar, Hinglish, and Hebrew**, we found that direct inference often failed to sufficiently reduce toxicity. These failures were especially pronounced in cases involving *code-mixed expressions* (as seen in Hinglish), *informal or non-standard usage*, and *morphological or syntactic complexity* that MT0-XL may not have been adequately trained to handle.

To address these challenges, we designed a **backtranslation-based detoxification pipeline** that incorporates a **high-resource pivot language**—either **English or Arabic**, depending on the language family and characteristics of the input. English was used for Indo-European and code-mixed languages (e.g., French, Italian, Hinglish, Tatar, Japanese), while Arabic served as a more linguistically aligned pivot for Semitic (e.g., Hebrew). This choice aimed to maximize semantic fidelity during translation and to improve detoxification effectiveness by operating in a language space where models were better optimized.

Pipeline Design and Rationale:

1. **Translation to Pivot Language (English or Arabic):**
   The first step involved translating the original toxic sentence into a high-resource pivot language. The rationale here was to project the sentence into a more semantically normalized and resource-rich space where detoxification models could operate more reliably. This process helped reduce linguistic noise, addressed irregularities such as code-mixing or morphological inflection, and enabled a more consistent application of detoxification techniques.
   The final step of backtranslation was performed using NLLB-600M. Although larger models like NLLB-3.3B were available, we opted for NLLB-600M due to hardware limitations. Despite its smaller size, NLLB-600M offered reliable performance across a wide range of languages, making it a practical choice for multilingual backtranslation in resource-constrained settings.
2. **Detoxification in Pivot Language:**
   The translated sentence was then passed through a dedicated detoxification model. We evaluated:
   - `s-nlp/mt0-xl-detox-orpo`
   - `s-nlp/bart-base-detox`

   The choice of detoxification model was informed by performance metrics such as semantic preservation, fluency, and toxicity reduction. `MT0-XL-detox-orpo` consistently produced better outputs, particularly across pivot languages. Its instruction-tuned nature allowed it to effectively follow detoxification commands without needing task-specific fine-tuning.
3. **Backtranslation to Original Language:**
   Once detoxified, the sentence in the pivot language was backtranslated into the original language using **NLLB-600M**. This model was chosen for its strong multilingual capabilities and support for low-resource languages. Backtranslation served two purposes:

- It restored the original language while maintaining the non-toxic transformation.
- It minimized information loss or distortion by leveraging a robust multilingual model trained on diverse parallel corpora.

Experiments and Analysis The detoxification pipeline was designed with a pivot-language strategy to address the performance disparity between high-resource and low-resource languages. English and Arabic were chosen as pivots due to their strong alignment with specific language families, which aids in preserving both sentence structure and meaning during translation. Detoxifying content in the pivot language allows for the use of well-trained detoxification models without the need to fine-tune for each individual target language. Furthermore, translation back to the original language is handled by NLLB-600M, which ensures that the detoxified sentence is rendered fluently and faithfully, thereby improving the overall quality of the final output.

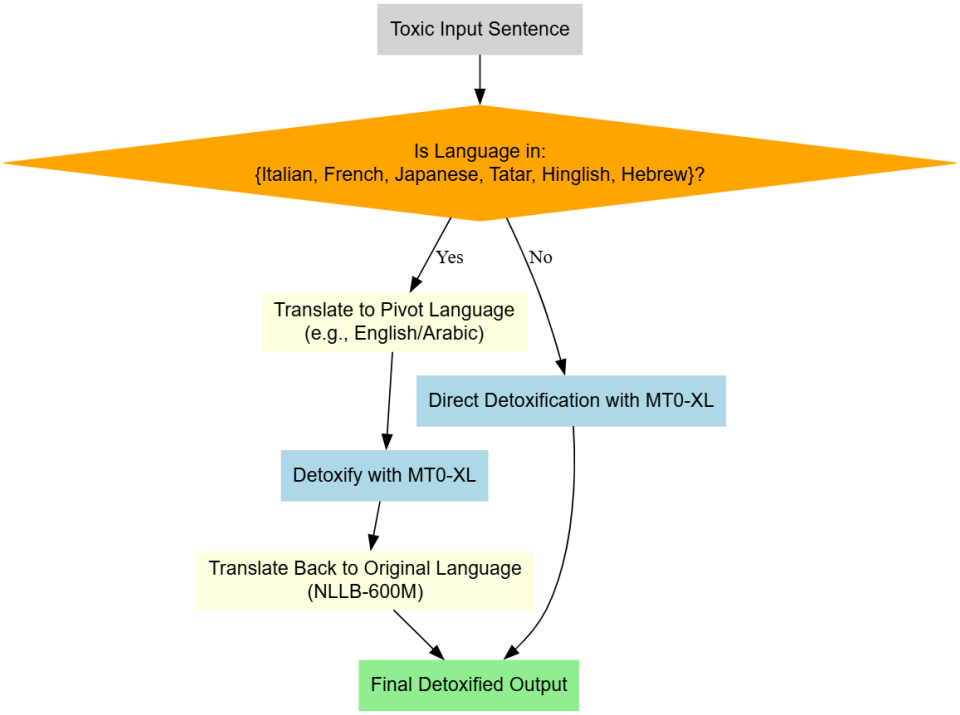

Figure 2: Overview of the detoxification pipeline using MT0/Bart-Base and Backtranslation methodology.

Observed Improvements: This approach led to noticeable improvements in detoxification, especially for linguistically complex or low-resource languages that exhibited residual toxicity when directly prompted. The pipeline demonstrated particular effectiveness in handling **code-mixed inputs**, such as Hinglish, due to the normalization effects introduced during the pivot translation process. Semantic fidelity was well-maintained, with detoxified outputs retaining the original intent while reducing or eliminating offensive content. Overall, the use of English and Arabic as pivot languages not only simplified the implementation but also enhanced the generalizability of the solution across diverse language families.

In summary, the proposed **MT0-XL + Backtranslation strategy with language-aware pivot selection** provides a scalable and effective solution for multilingual detoxification, especially in settings where direct inference with instruction-tuned models proves insufficient.

Languages Using Backtranslation:

- Italian
- French

- Japanese
- Tatar
- Hinglish
- Hebrew

3.3.2. Direct Prompt-Based Detoxification

For the remaining nine languages—English, Hindi, Russian, Spanish, German, Arabic, Ukrainian, Chinese, and Amharic—we used direct prompting with MT0-XL without any intermediate translation. This was effective due to the availability of training data for these languages and better generalization of MT0's multilingual capabilities.

The prompt-only method provided fluent and contextually relevant detoxified outputs in most cases, with minimal need for further correction.

3.3.3. Evaluation Results

Our system demonstrated strong overall performance across the 15 target languages, with high J-scores in high-resource languages such as Ukrainian (0.786), Russian (0.758), German (0.746), French (0.777), and English (0.725). These results reflect high fluency (XCOMET > 0.89), semantic similarity (LaBSE > 0.87), and effective toxicity reduction (STA > 0.90) in those languages.(Table 2)

Languages such as Amharic (0.488), Hebrew (0.486), Chinese (0.530), and particularly Hinglish (0.323) showed lower J-scores, indicating that these morphologically rich or code-mixed languages required additional post-processing. The impact of our backtranslation pipelines was especially visible in improving scores for challenging languages like Tatar (0.650) and Japanese (0.584).

The XCOMET across all languages was consistently higher than (0.84), confirming the fluency of the outputs, while ChrF was mainly used for early-stage validation. The composite J-score validated our hybrid model's robustness, with a clear margin between pivot and non-pivot languages, justifying the tailored detoxification strategies.

Table 2
Results across 15 languages. XCOMET and ChrF evaluate fluency; LaBSE for similarity; STA for toxicity neutrality. J-Score is the official composite metric.

Language	XCOMET ↑	ChrF ↑	Similarity (LaBSE) ↑	Toxicity (STA) ↓	J-Score ↑
Amharic	0.738	0.474	0.784	0.821	0.488
Arabic	0.864	0.779	0.898	0.915	0.716
German	0.947	0.803	0.930	0.845	0.746
English	0.891	0.725	0.872	0.921	0.725
Spanish	0.873	0.666	0.861	0.899	0.686
French	0.903	0.822	0.928	0.924	0.777
Hebrew	0.707	0.425	0.747	0.874	0.486
Hindi	0.848	0.711	0.863	0.846	0.633
Hinglish	0.684	0.521	0.749	0.631	0.323
Italian	0.834	0.737	0.915	0.944	0.724
Japanese	0.861	0.740	0.931	0.724	0.584
Russian	0.894	0.754	0.884	0.950	0.758
Tatar	0.836	0.775	0.912	0.847	0.650
Ukrainian	0.895	0.795	0.915	0.953	0.786
Chinese	0.845	0.479	0.844	0.731	0.530

3.4. Hybrid Model: Prompted MT0-XL, Lexical Filtering, and Backtranslation

To further enhance detoxification quality across diverse linguistic scenarios, we developed a **hybrid pipeline** that combines three complementary components: **prompted inference using MT0-XL**,

backtranslation, and **language-specific lexical filtering**. Each component addresses specific limitations when used in isolation and, when integrated, provides a more robust and adaptive detoxification mechanism across both high-resource and low-resource languages.

Prompted MT0-XL Inference: At the core of our hybrid strategy lies the `s-nlp/mt0-xl-detox-orpo` model, used in *zero-shot* mode with carefully crafted detoxification prompts for each target language. This approach leverages MT0-XL's multilingual generalization capabilities without requiring task-specific fine-tuning, making it a scalable and language-agnostic solution. While MT0-XL alone achieved strong performance on many languages, it often struggled to eliminate deeply embedded or culturally nuanced toxic phrases, particularly in *code-mixed* or *morphologically rich* languages.

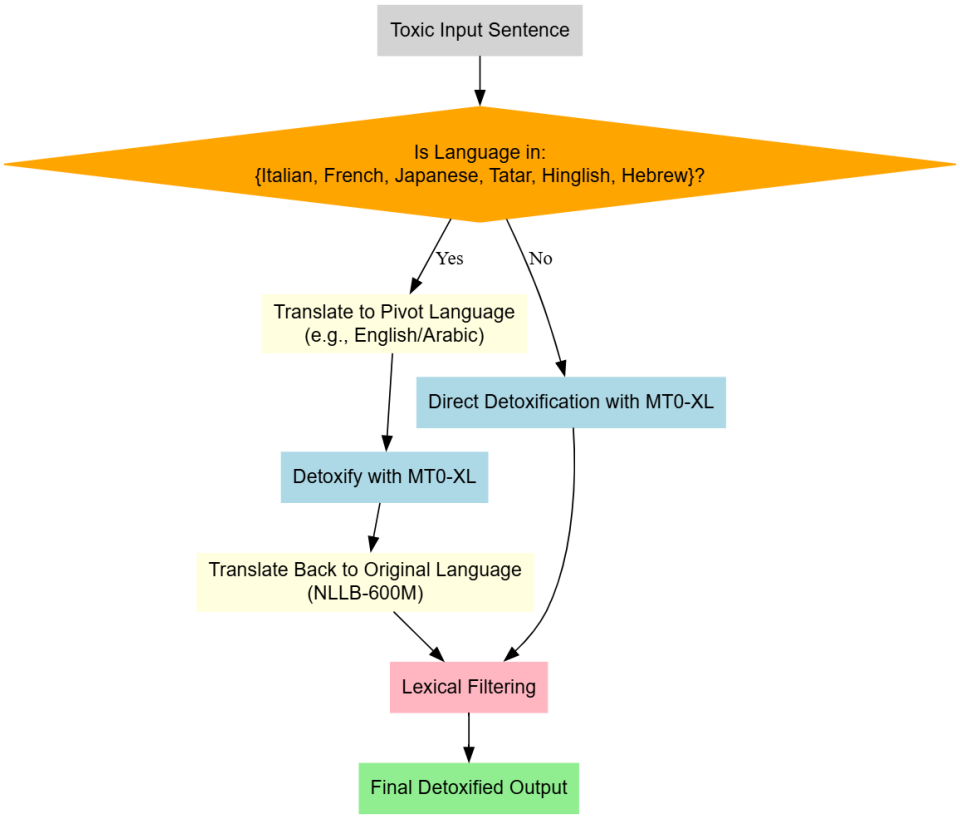

Figure 3: Overview of the detoxification pipeline of the Hybrid Approach.

Backtranslation: To further improve output quality—particularly in terms of fluency and deep detoxification—we integrated a *backtranslation stage* into the pipeline. The MT0-XL output was first translated into a high-resource *pivot language* (typically English or Arabic), then detoxified again using MT0-XL, and finally *backtranslated* into the original language using `NLLB-600M`. This step helped restore syntactic fluency and eliminate persistent toxic artifacts, especially in *low-resource* or *syntactically irregular* languages. We applied this technique to six languages where direct MT0 prompting was insufficient: **Italian, French, Japanese, Tatar, Hebrew, and Hinglish**.

Lexical Filtering: To address remaining toxicity that persisted even after backtranslation or prompting, we added a final *language-specific lexical filtering* step. This component scans the generated output for known toxic terms using curated dictionaries and either removes or replaces them with neutral alternatives. The filtering process is tailored to each language:

- For **Hinglish**, we manually constructed a *synthetic toxic-neutral dictionary* based on common code-mixed expressions. This dictionary captured informal, transliterated, and culturally grounded toxic phrases frequently missed by generative models.
- For **other languages**, we utilized publicly available *toxicity lexicons* hosted on Hugging Face. These lexicons were manually vetted and compiled from authoritative sources to ensure broad and language-specific coverage of offensive expressions.

In summary, our final system applies:

- **Direct prompted MT0-XL inference** for 9 pivot languages,
- **Backtranslation refinement** for six selected languages (Italian, French, Japanese, Tatar, Hebrew, Hinglish), and
- **Lexical filtering** as a final sweep to sanitize residual toxicity in all outputs.

This **hybrid pipeline** ensures a balanced trade-off between scalability, generalization, and linguistic sensitivity, enabling strong performance across a wide variety of linguistic contexts.

Benefits of the Hybrid Approach:

- Combining MT0-XL's generative power with explicit lexical filtering significantly improved detoxification effectiveness for **code-mixed and informal inputs**.
- Lexical resources enhanced **coverage of language-specific toxicity**, especially in edge cases where generative models underperformed.
- Backtranslation further **improved fluency and semantic stability**, serving as a final layer of refinement.
- The pipeline remained scalable and **language-agnostic**, with minimal dependence on fine-tuned models.

In summary, this hybrid framework demonstrated superior performance across a wide variety of languages and toxicity types, particularly for informal, low-resource, or culturally specific inputs. The integration of curated lexical knowledge and generation-based detoxification proved to be a powerful combination for robust multilingual text detoxification.

3.4.1. Evaluation Results

Our multilingual detoxification system achieved a strong overall performance with an average **J-score of 0.640** and a high **XCOMET fluency score of 0.846**. The model excelled particularly on high-resource languages such as **French (J = 0.773)**, **Ukrainian (J = 0.770)**, **German (J = 0.758)**, **Russian (J = 0.753)**, **Italian (J = 0.739)**, and **English (J = 0.728)**, where it delivered fluent, semantically similar, and effectively detoxified outputs.(Table 3)

In contrast, performance was noticeably lower for **code-mixed** or **morphologically rich languages** like **Hinglish (J = 0.333)**, **Hebrew (0.480)**, **Chinese (0.541)**, **Amharic (0.489)**, and **Japanese (0.565)**, which posed greater challenges due to limited training data, informal phrasing, or complex grammar.

Notably, the subset of languages processed with our backtranslation pipeline—referred to as *pivot languages*—achieved a higher average J-score of **0.675**, compared to **0.587** for *non-pivot languages* that used direct MT0-XL prompting. These results validate the effectiveness of our hybrid strategy, where backtranslation and lexical filtering significantly improve detoxification quality for linguistically complex or low-resource languages.

Table 3
Test phase results across 15 languages using Hybrid Approach. XCOMET and ChrF evaluate fluency; LaBSE for semantic similarity; STA for toxicity reduction; J-score is the official metric.

Language	XCOMET ↑	ChrF ↑	Similarity (LaBSE) ↑	Toxicity (STA) ↓	J-Score ↑
Amharic	0.736	0.463	0.771	0.835	0.489
Arabic	0.866	0.766	0.890	0.915	0.713
German	0.949	0.807	0.926	0.860	0.758
English	0.894	0.731	0.873	0.922	0.728
Spanish	0.878	0.665	0.860	0.907	0.695
French	0.908	0.827	0.929	0.915	0.773
Hebrew	0.725	0.442	0.753	0.828	0.480
Hindi	0.844	0.703	0.856	0.851	0.627
Hinglish	0.699	0.550	0.773	0.609	0.333
Italian	0.859	0.742	0.914	0.939	0.739
Japanese	0.857	0.733	0.927	0.709	0.565
Russian	0.895	0.749	0.882	0.943	0.753
Tatar	0.841	0.769	0.908	0.819	0.634
Ukrainian	0.895	0.790	0.909	0.938	0.770
Chinese	0.844	0.473	0.843	0.749	0.541

3.5. Evaluation Pipeline

We evaluated outputs using:

- **Toxicity** – using a fine-tuned XLM-R classifier
- **Similarity** – using LaBSE cosine similarity
- **Fluency** – using XCOMET and Chrf

Our primary evaluation was conducted using the official metrics provided by the PAN 2025 organizers: XCOMET for fluency, LaBSE for semantic similarity, and a fine-tuned XLM-R toxicity classifier for toxicity reduction. We acknowledge that XCOMET is the official fluency metric, as it is specifically designed to reflect human preferences in multilingual generation tasks. However, during the early stages of development—prior to integration of the XCOMET evaluation module—we also used ChrF as a lightweight proxy for fluency, due to its wide availability, simplicity, and relatively strong correlation with fluency in prior multilingual text generation tasks.

No external annotated dataset was used for evaluation; instead, we used the development set provided by the PAN 2025 task organizers, which includes pairs of toxic and reference-neutral sentences in 15 languages. All similarity and fluency metrics were computed on this set. Once the official evaluation codebase and XCOMET integration were available, we adopted it as our primary fluency metric, and updated our analysis accordingly. The reported results and final system evaluation are therefore fully aligned with the official PAN evaluation protocol.

4. Results

Our submitted system ranked **4th overall** on the official PAN 2025 Multilingual Text Detoxification task leaderboard during the final test phase. Despite being a lightweight and inference-only approach without any model fine-tuning, our hybrid pipeline performed competitively against other state-of-the-art systems. This placement reflects the robustness and effectiveness of combining prompted MT0-XL inference, backtranslation, and language-specific lexical filtering across both high-resource and low-resource languages.(Table 4)

However, the organizers also introduced a complementary evaluation using **LLM-as-a-Judge**, leveraging a fine-tuned LLaMA-3.1-8B-Instruct model to simulate human-like assessment. Under this subjective,

Table 4
Results of the Test Phase, 2025. Scores are computed using XCOMET (fluency), LaBSE similarity, STA toxicity reduction, and final J-score.

Language	XCOMET ↑	ChrF ↑	Similarity (LaBSE) ↑	Toxicity (STA) ↓	J-Score ↑
Amharic	0.743	0.476	0.785	0.810	0.486
Arabic	0.872	0.784	0.902	0.913	0.724
German	0.951	0.815	0.933	0.843	0.750
English	0.896	0.730	0.875	0.918	0.729
Spanish	0.880	0.674	0.865	0.901	0.696
French	0.908	0.816	0.914	0.902	0.752
Hebrew	0.697	0.299	0.693	0.828	0.416
Hindi	0.849	0.712	0.863	0.844	0.631
Hinglish	0.687	0.551	0.773	0.659	0.356
Italian	0.867	0.756	0.927	0.929	0.749
Japanese	0.860	0.718	0.923	0.741	0.591
Russian	0.896	0.755	0.884	0.944	0.755
Tatar	0.830	0.764	0.883	0.780	0.584
Ukrainian	0.898	0.795	0.917	0.937	0.776
Chinese	0.805	0.469	0.830	0.790	0.533

human-aligned evaluation, our system ranked **15th out of 32 teams** on non-pivot languages, and **7th out of 32 teams** on pivot languages.

This discrepancy highlights an important insight: while our approach excels under automatic metrics, particularly for high-resource languages, its relative effectiveness decreases in cross-lingual and code-mixed scenarios when judged by human preferences. Despite this, our system still outperformed several baselines (e.g., baseline_mt0, baseline_gpt4) and remained competitive across both evaluation settings.

These results validate the strength of our hybrid methodology while also pointing toward future improvements—especially in enhancing style transfer fluency and human alignment in low-resource or informal language settings.

5. Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to assist with language editing and minor rephrasing. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Generative AI Detection, Multilingual Text Detoxification, Multi-author Writing Style Analysis, and Generative Plagiarism Detection, in: Advances in Information Retrieval, Lecture Notes in Computer Science, Springer Nature Switzerland, 2025, pp. 434–441.

[2] D. Dementieva, V. Protasov, N. Babakov, N. Rizwan, I. Alimova, C. Brune, V. Konovalov, A. Muti, C. Liebeskind, M. Litvak, D. Nozza, S. Khan, S. Takeshita, N. Vanetik, A. Ayele, F. Schneider, X. Wang, S. Yimam, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the Multilingual Text Detoxification Task at PAN 2025, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.