Gendered Versus Genderless Pronouns: An Empirical Study on How NLP Models Exhibit Bias with Gender Embedded Contexts

Ethan Cao
B.S. Data Science & Computer Science
UVA
School of Data Science

Jacquelyn Xu
B.S. Commerce & Computer Science
UVA
McIntire School of Commerce

Alex Schwartz
B.S. Computer Science & Statistics
UVA
School of Engineering

December 3rd 2025

1 Abstract

The advent of NLP models has led to widespread adoption of automated toxicity detection and sentiment analysis on social media platforms. However, prior work has shown that these models are prone to bias, often mislabeling toxicity due to the presence of specific keywords or named entities whose social associations shift the perceived meaning of a sentence. In response to this, we apply the Perturbation Sensitivity Analysis framework introduced by Prabhakaran et al. to a dataset of Reddit comments, in order to examine whether languages with explicit gendered pronouns exhibit different toxicity and sentiment labeling behavior than languages with strictly gender neutral pronouns. Specifically, we compare two gendered languages (English and Dutch) against two gender-neutral languages (Turkish and Finnish), using parallel translations of the same base sentences. Our results show that sentences realized in gendered pronoun languages are more likely to receive high toxicity scores and display greater instability under score perturbations than their gender-neutral counterparts, suggesting that both grammatical gender and language choice can amplify bias in NLP-based content moderation.

2 Research Question

Does an NLP model for a language that has only gender neutral pronouns exert bias when it comes to classifying sentiment and toxicity as an English model with sentences that either contain male or female pronouns? And when we extend this setup to multiple languages (Finnish, Turkish, and Dutch), do we see systematic differences in how these models behave across gendered and genderless pronoun systems?

3 Data

Our dataset comes from the Reddit comment subset of the RTGender corpus introduced by Voigt et al. This collection contains naturally occurring Reddit comments that are at most 50 tokens in length and include at least one English third-person singular pronoun (“he”, “she”, “him”, “her”, “his”). We separate these comments into male–pronoun and female–pronoun groups, and subsequently translate each sentence into Turkish, Finnish, and Dutch to create multilingual sentence datasets for our perturbation sensitivity analysis.

4 Motivation

Comparing how an English NLP categorizes the toxicity and sentiment of sentences with male and female gendered pronouns to a NLP for a language without gendered pronouns examines how bias can still emerge from gendered word associations beyond explicit gendered markers. Gender bias in NLP models is not limited to explicit mentions of gender (ie woman, man, gendered pronouns, etc), and can also arise from associations between certain words and societal stereotypes. For example, occupations such as “doctor” are stereotypically associated with men in many Western contexts. In turn, even if a sentence about a doctor uses gender neutral language, an NLP model may still interpret the sentence in a way

that reflects a male bias. This pattern of behavior is important because it shows that simply eliminating explicit gender markers may not be enough to prevent an NLP from reinforcing gender bias in the decisions it makes. Further understanding how gender bias arises from word associations in NLPs allows these models to be developed in ways that do not perpetuate gendered stereotypes.

5 Literature Review

Prior research has shown that NLP systems trained on large-scale text corpora often inherit and amplify social and linguistic biases present in the underlying data. Such biases can manifest differently across languages, across demographic groups, and across model architectures, therefore motivating a closer examination of multilingual model behavior in toxicity and sentiment classification.

Toxicity in Multilingual Machine Translation at Scale (Costa-Jussà et al., 2023). This study demonstrates that toxicity can be unintentionally introduced or amplified during machine translation, particularly for low-resource languages. The authors show that mistranslations, model hallucinations, and asymmetric training data coverage produce systematically higher toxicity scores when text is translated across languages. Their findings highlight the need to examine how toxicity models behave in multilingual contexts, especially when the same semantic content is expressed in grammatically different languages.

Cross-Lingual Transfer Can Worsen Bias in Sentiment Analysis (Goldfarb-Tarrant et al., 2023). This study investigates how multilingual sentiment models transfer bias across languages. In languages with scarce training data, models often rely on representations inherited from higher-resource languages, which can introduce gender and racial biases. The authors show that cross-lingual transfer sometimes exacerbates disparities rather than mitigating them, underscoring the structural sensitivity of sentiment models to language-specific features.

Biases Toward African American English (AAE) in Toxicity and Sentiment Models (2024). This more recent research that examines NLP performance on African American English finds that toxicity and sentiment classifiers frequently misinterpret AAE expressions as hostile, despite neutral or positive intent. These errors arise from limited linguistic diversity in training corpora and from distributional biases that penalize non-standard dialects. This work demonstrates how variability in linguistic form, which is independent of meaning, can heavily influence toxicity scores, thereby motivating our investigation into grammatical gender effects.

Multilingual LLMs and Global Coverage (2024). These recent surveys on fairness in multilingual LLMs emphasize that model performance varies substantially across languages due to uneven data availability, morphological complexity, and inconsistent tokenization behavior. These studies highlight that languages with rich morphology or limited training data often receive poorer performance or inconsistent classification outputs, further motivating the need to examine gendered vs. gender-neutral languages.

Are All Languages Equally Hard to Language-Model? (Cotterell et al., 2018) Cotterell and colleagues show that languages differ fundamentally in modeling difficulty due to variations in orthography, morphology, and information density. Their results demonstrate that certain linguistic structures inherently challenge NLP models more than others. This motivates our comparison between English and Dutch (languages with gendered pronoun systems) and Turkish and Finnish, which are morphologically distinct and grammatically genderless.

Overall, these studies indicate that NLP systems do not treat languages or linguistic structures uniformly. Toxicity and sentiment models are sensitive to morphological complexity, training resource availability, and sociolinguistic features embedded in text that yields suboptimal or simply inaccurate classification results.

6 Gaps in current literature:

Nevertheless, from the existing researches demonstrated so far, while it is often implied by the authors, we discovered little to no grounded research into the biases that may be exerted in subtle contexts of "gendered pronouns". This prompted us to investigate how gender biases could arise in third person sentence contexts where pronouns are actively adopted. To this end, we decided to investigate how gender biases may be propagated through the active use of gender-based third person pronouns, by comparing it to other languages that does not implement the

use of third-person gendered pronouns. With our extended experiments, we also bring in Dutch as another language with gendered third person pronouns, which gives us a way to check whether English-specific patterns actually generalize to another gendered language or if they are language-dependent.

7 Methodology

In this study, we extend the Perturbation Sensitivity Analysis (PSA) framework into a **multilingual setting** to evaluate the robustness and fairness of natural language models across languages **with neutral gender pronouns**. We introduce perturbations through semantic-preserving translations of the same input sentence in languages that exclusively contain gender neutral pronouns (Turkish and Finnish), and also compare these against a second gendered language (Dutch) alongside with English to see how far these effects travel.

The reason we chose Dutch as a second gendered language alongside English is to provide more comparability between gendered and non-gendered pronoun languages. In which Dutch, alongside other Norwegian languages such as Danish and Swedish, has an exclusively designed gender pronoun system for third person pronouns, without the presence of a 'third' gender or gendered nouns, which is very common germanic languages. Moreover, Dutch has its own fine-tuned toxicity and sentiment analysis NLPs that makes it easy to integrate into the analytics pipeline.

Step 1: Preparing the sentence dataset

To effectively compare the outcome of the model classifications and evaluate whether there is a discrepancy due to the existence of gendered pronouns, we separate the original social media commentary dataset by equally dividing it into male and female sections. We intend to follow the same procedure as mentioned in the Perturbation Sensitivity Analysis paper (Prabhakaran et al 2021) with respect to only preserving sentences that are 50 words long and contain third person pronouns.

Step 2: Multilingual Sentence Generation

We begin from an English corpus of social media comments drawn from the RTGender dataset, specifically the `reddit_responses.csv` file. Following the original Text Bias setup, we extract sentences that contain at least one third-person singular gendered pronoun ("he", "she", "him", "her", "his"). We discard sentences longer than 50 tokens and then split the resulting set into two groups: sentences containing at least one male pronoun ("he", "his", "him") and sentences containing only female pronouns ("she", "her"). From each group we subsample an equal number of sentences (e.g., 125 each) to form the *male-pronoun* and *female-pronoun* English sentence sets.

Each English sentence x in these two sets is then translated into Turkish and Finnish using Hugging Face's `transformers pipeline` for machine translation. For English to Turkish we use the translation model `Helsinki-NLP/opus-mt-tc-big-en-tr`, for English to Finnish we use `Helsinki-NLP/opus-mt-en-fi`. And finally for English to Dutch, we use `Helsinki-NLP/opus-mt-en-nl` for translation.

The translation pipelines are applied separately to the male-pronoun and female-pronoun English sentences, yielding translated sets in Turkish, Finnish, and Dutch for both male-pronoun and female-pronoun sentences. Turkish and Finnish are chosen because their third-person pronouns are grammatically genderless, so the translated sentences serve as gender-neutral variants of the original English sentences while preserving core semantics as much as possible. Dutch, by contrast, retains gendered pronouns and therefore acts as a second test case for gendered languages.

Step 3: Sentiment and Toxicity Evaluation

To evaluate potential NLP model biases across gendered (English, Dutch) and genderless (Turkish, Finnish) languages, we employ separate transformer-based classifiers for toxicity and sentiment in each language. All models are accessed via the Hugging Face `pipeline` API that convert model outputs into scalar scores for toxicity and sentiment .

- **Toxicity Classification:**
  - *English:* We use `unitary/toxic-bert`, a BERT-based model fine-tuned for toxic comment classification. The pipeline is instantiated with `task="text-classification"`, and we use the returned `"score"` as a continuous toxicity score $f_{\text{tox}}(x) \in [0, 1]$.
  - *Turkish:* We use `fc63/turkish-toxic-language-detection` with `return_all_scores=True`. Our helper function `p_toxic_turkish` extracts the probability associated with the toxic class ("LABEL_1") and treats it as the Turkish toxicity score $f^{\text{tr}}_{\text{tox}}(x) \in [0, 1]$.

  - *Finnish:* We use `TurkuNLP/bert-large-finnish-cased-toxicity` via `task="sentiment-analysis"`, which returns a label and a confidence `"score"`. We interpret this confidence score as a Finnish toxicity score $f^{\text{fi}}_{\text{tox}}(x) \in [0, 1]$.
  - *Dutch:* For Dutch, we use `ml6team/robbert-dutch-base-toxic-comments` and again take the model's confidence for the toxic class as the Dutch toxicity score $f^{\text{nl}}_{\text{tox}}(x)$.

- **Sentiment Analysis:**
  - *English:* We use `distilbert/distilbert-base-uncased-finetuned-sst-2-english`, a DistilBERT model fine-tuned on SST-2. The pipeline returns a sentiment label ("POSITIVE" or "NEGATIVE") and a confidence `"score"`; our helper function `p_sentiment_english` uses this confidence value as a scalar English sentiment score.
  - *Turkish:* We use `savasy/bert-base-turkish-sentiment-cased` with `task="sentiment-analysis"`. The pipeline returns a label and confidence score, and `p_sentiment_turkish` records the confidence as the Turkish sentiment score.
  - *Finnish:* We use `nisancoskun/bert-finnish-sentiment-analysis-v2`. As with the other models, `p_sentiment_finnish` extracts the confidence associated with the predicted label as the Finnish sentiment score.
  - *Dutch:* For Dutch, we employ a Dutch sentiment classifier `DTAI-KULeuven/robbert-v2-dutch-sentiment` and interpret the confidence on the predicted sentiment label as $f^{\text{nl}}_{\text{sent}}(x)$.

You would notice here that we are using a separate NLP classification model for each language, this is due to the fact that there does not exist a perfect NLP model with perfect f1 score for all languages, meaning that the models only perform well with certain languages (i.e. English as a more popular language), rather than other less commonly used language (i.e. Turkish and Finnish). Hence, to reduce bias and potential confounding factors of the NLP models themselves being poorly trained on certain languages, thereby yielding bad or inaccurate classifications, we decided to pick a tailored NLP model for each classification purposes for each language. Dutch is simply treated as another language with its own dedicated model.

Step 4: Model Evaluation with Perturbation Sensitivity Analysis

We adopt the Perturbation Sensitivity Analysis (PSA) framework from the original paper to quantify how much model predictions change when we replace gendered English pronouns with genderless translations. For each English sentence $\mathbf{x}$ (with a male or female pronoun) and its corresponding translated variants $x^{(\text{tr})}$ (Turkish), $x^{(\text{fi})}$ (Finnish), and $x^{(\text{nl})}$ (Dutch), we compute model outputs $f(x)$ and $f(x^{(\ell)})$ for both the toxicity and sentiment models, where ℓ indexes the target language.

We then compute the following Perturbation Sensitivity Analysis metrics:

- **Score Sensitivity (ScoreSens):** For each sentence, we compute the difference between the English score and the target-language score, e.g. $f(x) - f(x^{(\text{tr})})$, $f(x) - f(x^{(\text{fi})})$, or $f(x) - f(x^{(\text{nl})})$. We treat these differences as per-example sensitivities and summarize them by their mean (and distribution) across all sentences, separately for male-pronoun and female-pronoun sentences.

- **Score Deviation (ScoreDev):** For each sentence, we take the set of scores across languages, $\{f(x), f(x^{(\text{tr})}), f(x^{(\text{fi})}), f(x^{(\text{nl})})\}$ and compute the standard deviation of these values. We then average this per-sentence standard deviation over the dataset, yielding an aggregate measure of how much model scores vary across language variants.

- **Score Range (ScoreRange):** For each sentence, we compute the range of scores across language variants, e.g. $\max_\ell f(x^{(\ell)}) - \min_\ell f(x^{(\ell)})$. In pairwise comparisons (English vs. a single target language) this reduces to the absolute difference $|f(x) - f(x^{(\ell)})|$. We average this per-sentence range over all sentences to obtain a global measure of cross-lingual score spread.

- **Label Distance (LabelDist):** We convert scores into binary labels by applying a threshold c (for toxicity: toxic vs. non-toxic; for sentiment: positive vs. non-positive). For each pair of languages (e.g., English vs. Turkish, English vs. Dutch), we define the set of indices of sentences labeled as positive by each model, and

compute the Jaccard distance between these sets. This *LabelDist* measures the proportion of sentences whose predicted class label flips when moving from English to a genderless translation (or to Dutch). We compute LabelDist across a range of thresholds $c \in [0.3, 0.9]$ to analyze how sensitive classification decisions are to both the choice of threshold and the choice of language.

Step 5: Evaluation and visualization of results

After gathering model classification results, we could compare and visualize how sensitive the model is towards sentences with gendered pronouns and sentences with gender neutral pronouns, and determine the underlying bias that may occur due the involvement of gender as a factor. The conclusion will ideally be backed up by results generated from multiple languages, and the average of all of those classification scores will be taken as a final value determining the scores assigned to gender neutral sentences. This is to ensure that the model isn't just malfunctioning due to the absence of one particular kind of language in the model's original training dataset. With Dutch added as a second gendered pronoun language alongside with English, we have an extra gendered language to show that some of the strongest toxicity effects are not just English-specific results, but are those that can also appear (and even get worse) in other languages.

8 Results and Findings

From our experimentation, the key discovery is that sentences containing female third-person gendered pronouns tend to exhibit greater context shifts and instability when evaluated across languages, particularly when compared against their gender-neutral counterparts. Across all four languages in our study, we observe systematic differences in how models assign sentiment and toxicity scores, with language structure and pronoun systems jointly influencing model behavior.

This suggests that the presence of gendered pronouns can meaningfully affect how NLP models interpret the overall tone of a sentence. At the same time, the results indicate that differences between languages themselves can amplify or modulate these effects: some language–model combinations react more strongly to the same underlying content than others. As a result, both genderedness and linguistic context contribute to variation in toxicity and sentiment classification, and these effects can compound each other to a certain extent when compared together.

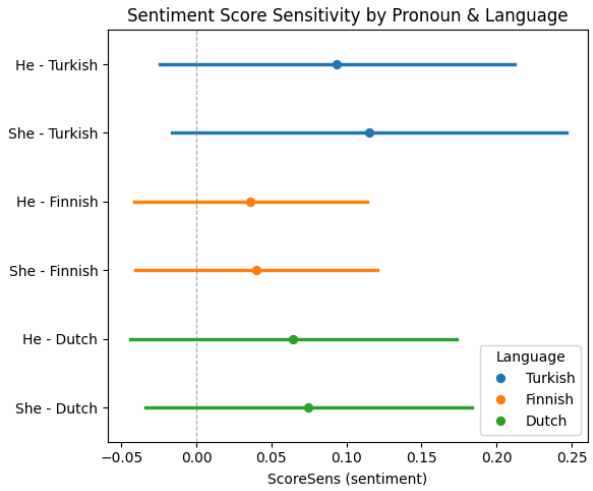

Figure 1: Sentiment ScoreSens by pronoun and language, demonstrating systematic score shifts between English, Dutch, Turkish, and Finnish realizations of the same sentences.

According to Figure 1, translating English sentences into Turkish and Finnish induces consistent shifts in average sentiment scores, while Dutch shows comparable patterns of sensitivity. Across all four languages, sentences containing female pronouns ("she," "her") exhibit slightly larger average sensitivities than those containing male pronouns. This indicates that sentiment models tend to be marginally more unstable with female-referenced contexts, regardless of language.

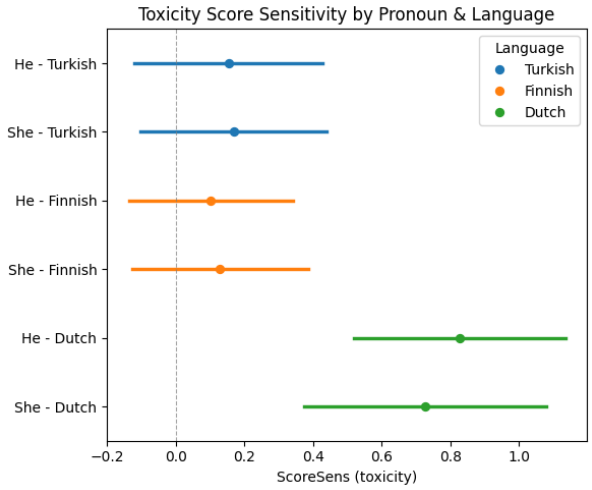

Figure 2: Toxicity ScoreSens by pronoun and language, showing how toxicity scores vary across English, Dutch, Turkish, and Finnish realizations of the same sentences.

Figure 2 shows that toxicity scores are substantially more sensitive to language variation than sentiment scores. Both English and Dutch, which are the two gendered pronoun languages, exhibited much larger Score Sensitivity values than Turkish and Finnish. This means that the toxicity models for gendered languages respond more strongly to the same underlying content, and small differences in phrasing or pronoun context produce larger changes in toxicity predictions. As before, female pronoun sentences yield slightly higher sensitivities, though the dominant pattern is the broader separation between gendered and gender-neutral languages. It could also be said that the language of Dutch or their respective NLP models exhibited bias that confounded by other hidden factors, hence yielding such an extraordinary outlier outcome.

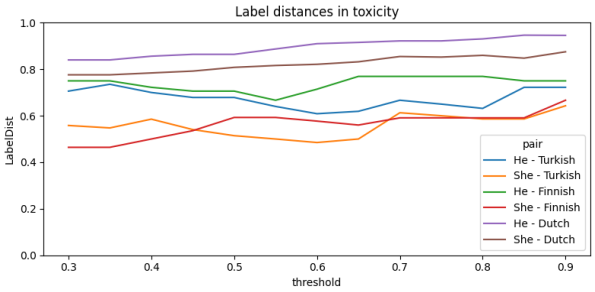

Figure 3: LabelDist for toxicity across thresholds, summarizing how often labels flip between English, Dutch, Turkish, and Finnish variants of the same sentences.

Looking at label distance, according to Figure 3, toxicity classification is highly unstable across languages, particularly at higher decision thresholds. LabelDist values increase steadily as thresholds become stricter, demonstrating that a substantial portion of sentences flip from toxic to non-toxic (or vice versa) depending on the language in which they are realized. What's notable is that both English and Dutch consistently produce higher label distances than Turkish and Finnish, showing that toxicity models for gendered languages tend to be more reactive and volatile compared to non-gendered languages.

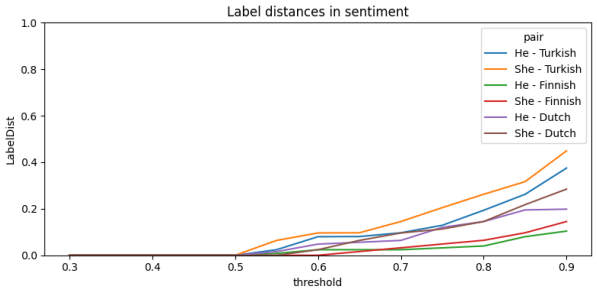

Figure 4: LabelDist for sentiment across thresholds, comparing how often sentiment labels flip across English, Dutch, Turkish, and Finnish variants.

At the same time, Figure 4 shows that sentiment labels are more stable at lower thresholds, though label flips begin to increase around the 0.5 mark, especially for the Turkish and Dutch variants. Finnish remains the most stable overall. Both male and female pronoun sentences show similar cross-language behavior, with slightly higher flip rates for female-referenced sentences in Turkish.

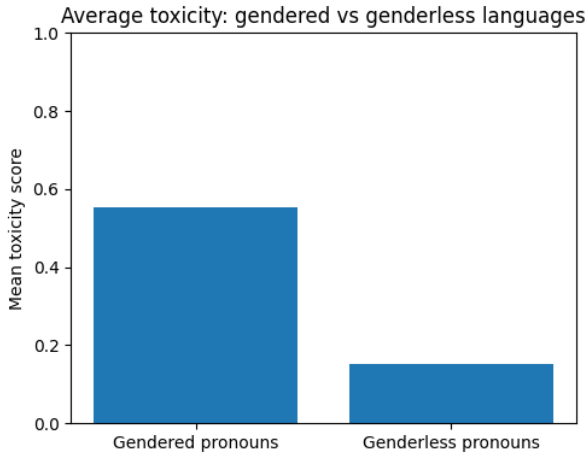

Figure 5: Average toxicity scores for gendered pronoun languages (English, Dutch) versus gender-neutral pronoun languages (Finnish, Turkish).

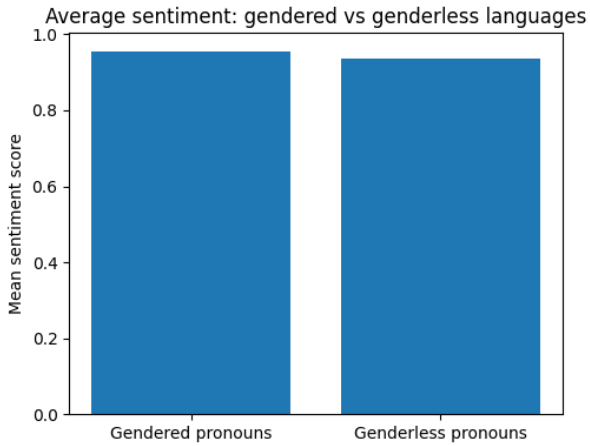

Figure 6: Average toxicity scores by language (Dutch, English, Finnish, Turkish), illustrating that gendered languages exhibit higher baseline toxicity.

Figures 5 and 6 summarize the broader trend. Languages with gendered pronouns (English and Dutch) assign notably higher average toxicity scores compared to Finnish and Turkish. Dutch, in particular, exhibits the highest baseline toxicity, while Finnish and Turkish consistently produce the lowest scores. This pattern suggests that grammatical gender correlates with higher sensitivity in toxicity classification, even when the underlying semantic content remains constant across translations.

Overall, the results demonstrate substantial differences in toxicity and sentiment classification across languages. Gendered pronoun languages tend to yield higher toxicity scores, larger perturbation sensitivities, and more frequent label flips. At the same time, sentences containing female pronouns show slightly greater instability across all languages. Together, these findings highlight how gendered language structure and pronoun use jointly influence how NLP models interpret text, reinforcing the need for multilingual fairness evaluations in real-world moderation systems.

9 Conclusion

Through our limited scope of experimentation, we did discover empirical evidence of languages with gendered pronouns exhibiting greater score sensitivity and label distance range in both toxicity and sentiment analysis. Suggesting that there does exist signs of gender bias being injected into the model as part of its original training data, as demonstrated by its skewed classification results.

Nevertheless, there can be ways that this experimentation can be elaborated on, such as introducing a wider range of languages to experiment with. Introducing more languages, and separating them into equal number batches of gender versus genderless categories, would greatly add justifications into this research as to whether gender is truly a variable in causing NLP model bias in toxicity and sentiment classification.

Another aspect that this experiment could be improved on is to utilize a singular multilingual NLP model for sentiment and toxicity classification. While such models are currently not available in the open source space, due to the variability in the accuracy of its predictions across different languages (by exhibiting optimal accuracy in mainstream languages and lower accuracy in less commonly used languages), it cannot be deployed for the purposes of this experimentation due to inherent model inaccuracies that may lead to unreliable predictions of toxicity or sentiment labels. Hence, for the purposes of this experiment as of this time, we decided to resort to using optimized and fine-tuned NLP models for each language only.