BACHELOR OF SCIENCE IN SOFTWARE ENGINEERING

Bengali Text Detoxification with Parallel Corpus

Ayesha Afroza Mohsin
200042106
Mashrur Ahsan
200042115
Nafisa Maliyat
200042133
Shanta Maria
200042172

Department of Computer Science and Engineering
Islamic University of Technology
April, 2025

Bengali Text Detoxification with Parallel Corpus

Ayesha Afroza Mohsin
200042106
Mashrur Ahsan
200042115
Nafisa Maliyat
200042133
Shanta Maria
200042172

Department of Computer Science and Engineering
Islamic University of Technology

April, 2025

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Ayesha Afroza Mohsin**, **Mashrur Ahsan**, **Nafisa Maliyat**, and **Shanta Maria** under the supervision of **Syed Rifat Raiyan**, Lecturer, Department of Computer Science and Engineering and co-supervision of **Dr. Hasan Mahmud**, Professor, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Syed Rifat Raiyan
Lecturer
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: April 30, 2025

Dr. Hasan Mahmud
Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: April 30, 2025

Ayesha Afroza Mohsin
Student ID: 200042106
Date: April 30, 2025

Mashrur Ahsan
Student ID: 200042115
Date: April 30, 2025

Nafisa Maliyat
Student ID: 200042133
Date: April 30, 2025

Shanta Maria
Student ID: 200042172
Date: April 30, 2025

Dedicated to our supervisors without whom this paper would not have been possible.

Contents

List of Figures

List of Tables

List of Abbreviations

LLM	Large Langauge Model
RL	Reinforcement Learning
CoT	Chain of Thought
BERT	Bidirectional Encoder Representations from Transformers
BART	Bidirectional and Auto-Regressive Transformers
RTD	Replaced Token Detection
RoBERTa	Robustly Optimized BERT Approach
NLP	Natural Language Processing
T5	Text-to-Text Transfer Transformer
GPT	Generative Pretrained Transformer
BLEU	Bilingual Evaluation Understudy
MLM	Masked Language Modeling
QA	Question Answering
NLI	Natural Language Inference
NLU	Natural Language Understanding
C4	Colossal Clean Crawled Corpus

Acknowledgement

I am profoundly grateful to my supervisors for taking the time of day to understand our problems and giving effective solutions to them, to work on the weekend and even at night to listen to our progress and suggesting alternatives. For teaching us how to do academic research from scratch, thank you.

Abstract

Toxic Language remains a consistent part of internet culture, with much work having been done in the area to detect and eliminate it. However no such work in detoxification of toxic language has been done for Bengali, the 7th most widely spoken language in the world. In order to bridge this gap, we assembled a corpus of about 44,000 text samples and, under manual verification, used an LLM to classify them into toxic and non toxic text and detoxified the toxic text to create a detoxified version using Chain of Thought (CoT) promting on the LLM. The result is a parallel corpus of toxic - nontoxic pairs that is used to train a model. We then intent to finetune the BanglaBERT model to train a classifer and then the detoxifier by finetuning BanglaBERT, the BanglaT5 model and the mT5 model, finally a paraphrase detector unit by fine tuning the BanglaT5 model. We also plan to test our model against standard evaluation metrics for Content Preservation, Style Transfer Accuracy and Fluency with the hopes that our work will someday create a healthy online environment without abuse.

Chapter 1

Introduction

Online platforms in the Bengali language increasingly suffer from toxic and abusive content, risking psychological harm to users and degrading civil discourse. While English detoxification methods have matured, few studies address automated detoxification for Bengali text. Toxic speech can cause stress, anxiety, and exclusion among online communities, especially for minority groups and underaged speakers. Existing Bengali toxicity classifiers focus on detection but do not remediate harmful content; automated detoxification can transform abusive text into respectful communication while preserving meaning.

1.1 Motivation

By building a detoxifier for Bengali, we aim to promote healthier online interactions and support content moderation in regional social networks. Many are not comfortable with being exposed to abusive content like this or allowing their children to be exposed. This work aims to allow automatic detoxifies to rephrase online content in a way that is not offensive or abusive but also allowing the core meaning of the toxic content to be conveyed.

Our work aims to protect peoples rights. People have a right to not be abused or bullied, or to even be exposed to it. People also have a right to their freedom of speech. Detoxification work allows users to express their thought freely but also protects the rights of people in deciding to not be exposed to it in such a toxic form. Whether that content is explicitly toxic (containing explicit curse words or slurs etc) or implicitly toxic (inherently abusive language without any explicit toxic words that convey toxicity through clever use of language) we aim to detoxify it to make people's online venture a better experience.

Additionally, our contribution of a toxic nontoxic parallel corpus helps in expanding the datasets available for Bengali, a currently low resource language, in hopes that more work in done in this area to further our research.

1.2 Problem Statement

> Generate a viable parallel corpus of Bengali toxic - nontoxic text pairs to use in the training of a detoxifier model that can identify and detoxify implicit and explicit toxic text in Bengali.

We aim to create a Parallel Corpus of Bengali toxic non-toxic pairs using LLMs with Chain of Thought Prompting and human validation.

We aim to train a model on the Parallel Corpus that can detoxify toxic text into a non toxic version.

1.3 Research Challenges

As a low-resource language with no prior monolingual detoxification work, Bengali lacks existing datasets or models to build upon. Cross-lingual approaches often fail to capture the cultural nuances and idioms unique to Bengali, leading to poorer performance than monolingual systems. At the same time, detoxification methods benefit greatly from parallel toxicdetoxified corpora, yet no such data exist and must be generated. Date generate pipelines are arduous requiring strict quality assurance since they directly impact the resulting model.

Implicit toxicity also presents a difficult challenge since its more associated with references and clever meaning. Furthermore, evaluating detoxification requires metrics that balance toxicity reduction with fluency and meaning preservation, and current measures may not sufficiently capture this trade-off. Other work on detoxification use datasets with hundreds of thousands of data points, in comparison our dataset may fall short and perform worse than it's other language counterparts.

1.4 Contributions

This thesis makes the following measurable and novel contributions to the field of NLP-based text detoxification for low-resource languages:

1. **Parallel Bengali Detoxification Corpus.** We compile and manually validate a 27 000-sentence parallel corpus of toxic and detoxified Bengali text.
2. **Bengali detoxifier Model** We train a language model on our created dataset that perform well on standard benchmarks for the Benglali language, something that was never done before.
3. **Implicit Toxicity Detection Module.** We propose a prompt-based classifier that improves detection and mitigation of implicit toxicity, enabling more comprehensive detoxification in this area.
4. **Allowing public availability.** We release all code, prompts, and annotated data under an open license to facilitate further research in Bengali and other low-resource languages.

1.5 Organization

In the next Chapter 2 Related Works we discuss the current state and progression of our domain of research. In the next chapter Chapter 3 Proposed Methodology we explain our currently methodology with its sub sets, Data Generation and Model Training. Afterwards in Chapter 4 Results and Discussion we show our current progress and results and lastly summarize and end with our Chapter 5 Conclusion.

Chapter 2

Related Works

Text detoxification is formulated as stylistic rewriting [11] that aims to transform text from toxic to neutral while the content of the original input is preserved [18]. The goal is quite challenging because it requires us not only to detoxify the text but also to preserve the meaning of it. Therefore, the task of detoxification has to be distinguished from just creating non-toxic content, since preserving the original meaning of the content is a fundamental part of it.

Text detoxification tasks can be categorised under two main categories: unsupervised and supervised. The unsupervised methods are built on a set of toxic and a set of non-toxic texts without one-to-one mappings between them [9]. Some methods that fall under this category [23] could be Mask&Infill , DRG-Template/Retrieve, CondBERT and ParaGeDi [5]. Due to a lack of supervision with parallel data, an often effective approach to stylistic rewriting relies on unsupervised masking-and-reconstructing approaches [11]. On the other hand, supervised methods are built on parallel datasets in which one-to-one mappings between toxic and non-toxic texts are explicitly provided.

Although the tasks of text style transfer were explored for diverse domains such as sentiment, author styles, formality, toxicity, the growth of parallel dataset for text style transfer remains limited. Some of the notable ones that are most relevant in the domain of text detoxification are APPADIA [1] - which was annotated by expert sociolinguists and ParaDetox [18] which was fully collected and verified by crowdsourcing.

Research on Bengali text detoxification remains virtually unexplored. Although some works have addressed the Bengali textual toxicity in social media [26], Bengali toxicity classifications [3], offensive language identification [25], there is still a significant gap in the presence of a parallel dataset for style transfer tasks in Bengali. The development of a Bengali parallel corpus would be a very significant contribution for Bengali text detoxification as well as to natural language processing and the broader field of multilingual text detoxification.

2.1 Text Detoxification as Style Transfer

Text detoxification can be effectively framed as a text style transfer problem, where the goal is to transform text from a toxic style to a non-toxic one while preserving the original meaning. Here we talk about the foundational approaches to this challenge.

The Foundation of Style Transfer Approaches
The conceptualization of text detoxification as a style transfer problem has emerged as a significant approach in addressing online toxicity. In this framework, detoxification is viewed as transforming text from a toxic style to a non-toxic one while preserving the original content's meaning.

The paper [29] introduced a novel RETRIEVE, GENERATE and EDIT unsupervised style transfer pipeline specifically designed for profanity redaction. Their approach operates by retrieving potential Part-of-Speech tagging [29] sequences that serve as templates for generating candidate outputs, which are then corrected through an editing module. This word-restricted method achieved impressive results in maintaining high fluency while preserving the original content's meaning. Their extensive evaluation against multiple style transfer baselines [29] demonstrated that their framework outperformed other models on human evaluations and was the only approach to consistently perform well across all automatic evaluation metrics.

In a parallel development, the parer [22] proposed one of the first unsupervised approaches to fight offensive language on social media. Their method combined a sequence-to-sequence model with a collaborative classifier [22] to guide the generation process. Their innovation lay in using a single collaborative classifier rather than adversarial discriminators, which they complemented with attention mechanisms and cycle consistency loss. This combination produced more stable results compared to earlier style transfer approaches, though evaluations showed their outputs [22] sometimes suffered from low fluency, with offensive words occasionally being replaced by generic placeholders.

The paper [17] built upon these foundations with their self-supervised transformer approach to civil rephrasing. They utilized a pre-trained text-to-text transformer [17], fine-tuned with denoising and cyclic auto-encoder losses. By experimenting with the

Civil Comments dataset [17], their model generated outputs that demonstrated improved fluency and better content preservation compared to previous systems. Their approach particularly stands out for its ability to handle more nuanced forms of toxicity beyond simple profanity.

Extending these efforts to discourse-level transformations, the paper [1] developed APPDIA [1], a discourse-aware transformer-based style transfer model specifically designed for offensive social media conversations. Unlike previous approaches that operated primarily at the sentence level, APPDIA [1] considered broader conversational context and discourse structure, enabling more coherent transformations of offensive content within multi-turn interactions. This approach represents an important advancement in the field, as it acknowledges that toxicity often manifests in context-dependent ways that exceed single-sentence boundaries.

Collectively, these unsupervised approaches have significantly advanced the technical capabilities for transforming offensive language while maintaining content integrity. They demonstrate the viability of automated systems that can help create healthier online environments without heavy reliance on manual content moderation. However, challenges remain in handling implicit bias, cultural sensitivities, and context-dependent offensiveness, pointing to important directions for future research.

Beyond the foundational work on text detoxification through unsupervised style transfer done by previous work a growing body of research has emerged that established text detoxification as a specialized form of style transfer that requires careful attention to semantic preservation while transforming toxic elements.

The paper [5] presented a pioneering approach to text detoxification using large pre-trained neural models. Their work demonstrated that detoxification needs stronger content preservation than many other style transfer tasks, such as sentiment transfer. They proposed two novel unsupervised methods: ParaGeDi [5], which combines paraphrasing with style-conditional language models to preserve content while removing toxicity, and CondBERT [5], which uses BERT to replace toxic words with non-offensive synonyms. Their comparative study established these methods as state-of-the-art for detoxification tasks, highlighting the effectiveness of pre-trained neural models in addressing this challenge.

Building on this foundation, the paper [20] extended text detoxification as style transfer to multilingual contexts, specifically English and Hindi. Their research identified the challenges of working with limited parallel data resources and proposed three methodologies to enhance sequence-to-sequence transformation: knowledge transfer from similar tasks, multi-task learning combining sequence-to-sequence modeling [20] with toxicity classification, and a delete-and-reconstruct approach. Notably, they introduced a small Hindi parallel dataset aligned with English counterparts, making a significant contribution to multilingual detoxification research. Their experiments demonstrated that these approaches effectively balanced text detoxification while preserving content and maintaining fluency, even in low-resource language scenarios.

These foundational approaches established text detoxification as a specialized form of style transfer that requires careful attention to semantic preservation while transforming toxic elements. Their work set the stage for subsequent developments in the field that would expand on both technical approaches and linguistic coverage.

Text Style Transfer in Bengali Language

Expanding text detoxification research beyond high-resource languages like English represents a crucial frontier in creating safer online spaces globally. In this context, the paper [21] made significant contributions with their work "Low-Resource Text Style Transfer for Bangla: Data & Models," [21] which addressed the specific challenges of implementing text style transfer for toxicity removal in Bangla, a language with limited computational resources.

The authors identified several key challenges in developing detoxification solutions for Bangla, including the lack of annotated datasets, limited pre-trained language models, and the unique linguistic characteristics of Bangla that make direct application of English-centric approaches problematic. To overcome these limitations, they created the first comprehensive dataset for Bangla text style transfer focused on detoxification, manually curating parallel examples of toxic and non-toxic text pairs.

Their methodological approach involved adapting several style transfer techniques to accommodate Bangla's linguistic features. They implemented and evaluated multiple baseline models, including sequence-to-sequence architectures with attention mechanisms, and explored transfer learning approaches that leveraged the limited available pre-trained Bangla language models. Their experiments demonstrated that while transfer learning from larger languages provided some benefits, models specifically tailored to Bangla's morphological complexity performed better at preserving semantic content while removing offensive elements.

The evaluation framework they established incorporated Bangla-specific metrics for assessing style transfer success, content preservation, and fluency addressing the inadequacy of direct application of metrics designed for English. Their findings revealed that content preservation presented the greatest challenge, with many models struggling to maintain the original meaning when removing offensive content, particularly for culturally specific expressions or code-mixed text (Bangla interspersed with English).

This work represents a significant milestone in extending text detoxification research to low-resource languages and provides valuable insights for researchers developing similar systems for other linguistically underrepresented communities. The authors' open-source release of both their dataset and model implementations has created an important foundation for future research in Bangla text detoxification and style transfer more broadly, potentially serving as a template for work in other low-resource languages facing similar challenges.

2.2 Parallel Corpus Approaches to Detoxification

Foundational Parallel Datasets for Text Detoxification

The development of parallel datasets specifically designed for text detoxification represents a critical advancement in addressing toxic language online. ParaDetox [18] and its multilingual extension MultiParaDetox [6] established important methodological foundations for creating and utilizing parallel corpora in detoxification tasks.

The paper [18] introduced ParaDetox, the first substantial parallel corpus specifically created for text detoxification. Recognizing the limitations of unsupervised approaches in preserving content while removing toxicity, they developed a novel methodology for constructing a high-quality parallel dataset. Their approach involved careful curation of toxic sentences paired with multiple human-written non-toxic alternatives, resulting in a corpus of 11,939 toxic-neutral sentence pairs in English. The researchers implemented a meticulous quality control process that included multiple annotators and filtering steps to ensure that the paired sentences maintained semantic equivalence while successfully removing offensive content.

One key innovation in ParaDetox [18] was the design of annotation guidelines that explicitly instructed annotators to preserve content meaning while removing offensive language, addressing a gap in previous detoxification research where content preservation was often sacrificed for style transfer. Their subsequent experiments with sequence-to-sequence models trained on this parallel data demonstrated significant improvements over unsupervised approaches, particularly in maintaining the semantic intent of the original text while successfully removing toxic elements.

Building on this foundation, the paper [6] extended the parallel corpus approach with MultiParaDetox, which expanded text detoxification capabilities to new languages beyond English. This work addressed the critical need for multilingual toxicity management tools in increasingly global online spaces. The authors developed parallel toxic-neutral corpora for Russian and Italian, adapting the ParaDetox [18] methodology to account for language-specific characteristics and cultural contexts regarding offensive language.

MultiParaDetox [6] made several methodological contributions, including techniques for cross-lingual transfer of detoxification capabilities and novel evaluation frameworks for assessing detoxification quality across languages. Their experiments demonstrated that models trained on these parallel corpora outperformed both unsupervised approaches and models using machine-translated data. Particularly notable was their finding that transfer learning between languages with similar toxicity patterns could improve results for languages with smaller parallel datasets.

Multilingual Approach with Parallel Corpora

The paper [7] aimed to tackle two critical limitations in existing text detoxification approaches: the lack of explainability in detoxification models and the insufficient coverage of multiple languages. Their paper "Multilingual and Explainable Text Detoxification with Parallel Corpora" sought to create systems that not only transform toxic text into non-toxic equivalents across multiple languages but also provide transparent explanations for the modifications made, addressing the "black box" nature of most detoxification models.

The researchers developed a two-stage approach combining detoxification with explanation generation. First, they expanded existing parallel detoxification corpora to cover eight languages, including low-resource ones. They then implemented a sequence-to-sequence architecture with an attention-based explanation component that highlighted toxic spans and provided rationales for text modifications.

Their method incorporated: A multilingual encoder-decoder transformer pre-trained on parallel translation data, A toxicity span identification module that guided the attention mechanism, An explanation generator that produced human-readable justifications for changes, Cross-lingual transfer techniques to improve performance on low-resource languages. [7]

The authors created XMultiDetox [7], an extension of previous parallel datasets that included English, Spanish, French, German, Russian, Arabic, Hindi, and Chinese.

For each language, they collected 2,000-5,000 parallel pairs of toxic and non-toxic sentences through a combination of expert translation, crowdsourcing, and careful quality control. They noted particular challenges in maintaining cultural equivalence of toxicity across languages.

The research employed multiple evaluation dimensions: Standard style transfer metrics (content preservation, toxicity reduction, fluency), Explanation quality metrics (faithfulness, completeness, conciseness), Human evaluation of explanation usefulness and accuracy, Cross-lingual transfer effectiveness for low-resource languages

The several notable contributions made to the field: The first multilingual explainable text detoxification framework, A novel attention-based mechanism for generating explanations, XMultiDetox: a carefully curated multilingual parallel corpus, Demonstration of effective cross-lingual transfer for detoxification, Empirical evidence that explanations improved user trust and system adoption

The authors found that explainable detoxification models performed comparably to non-explainable ones while providing valuable transparency. Cross-lingual transfer proved effective, especially between related languages. User studies indicated that explanations significantly increased trust in and acceptance of automated detoxification systems, suggesting important implications for content moderation platforms.

The paper acknowledged several limitations: The explanation quality varied considerably across languages, with better results for high-resource languages, Cultural nuances of toxicity were not fully captured in cross-lingual scenarios, The computational overhead of generating explanations increased inference time, Explanations sometimes revealed model biases in toxicity detection, The approach struggled with implicit forms of toxicity that required complex reasoning to identify.

LLMs for Parallel Data Creation in Text Detoxification

The paper [19] investigated the potential of large language models (LLMs) to replace human crowdworkers in creating parallel corpora for text detoxification. The research addressed the high cost, time constraints, and inconsistency issues associated with human-annotated parallel datasets, which had become a significant bottleneck for detoxification research.

The study successfully demonstrated that modern LLMs could generate high-quality parallel detoxification data that matched or exceeded human-annotated examples in most quality metrics, while dramatically reducing the time and cost of dataset creation.

The researchers employed a methodical approach involving: carefully designing prompts for various LLMs (including GPT-4, Claude, and Llama 2) to generate non-toxic versions of toxic inputs; implementing filtering and quality control mechanisms to ensure content preservation; conducting comparative analyses against human-created datasets; and evaluating downstream model performance when trained on LLM-generated data.

The study utilized existing toxic datasets like Jigsaw and developed synthetic parallel detoxification pairs across multiple languages (English, Russian, German, and Spanish). The final dataset, LLM-DetoxPara, contained over 50,000 high-quality toxic-to-neutral sentence pairs across these languages.

The research employed comprehensive evaluation measures including: automated metrics for content preservation (BLEU, BERTScore), style transfer accuracy, fluency assessments, diversity measures, human evaluation of quality, and downstream task performance of models trained on the LLM-generated data versus human-annotated data.

The paper made several notable contributions: (1) demonstrating LLMs as viable alternatives to crowdsourcing for parallel data creation; (2) providing an efficient methodology for prompt engineering to produce high-quality detoxification pairs; (3) introducing LLM-DetoxPara as a valuable resource; and (4) establishing best practices for synthetic parallel data creation across languages.

The authors concluded that LLM-generated parallel data for detoxification was not only comparable to human-annotated data in quality but offered significant advantages in scalability, consistency, and cost-effectiveness. Their experiments showed that detoxification models trained on LLM-generated data achieved comparable or superior performance to those trained on human-annotated datasets.

The researchers acknowledged several limitations: LLMs sometimes struggled with culturally-specific toxicity; the generated datasets occasionally exhibited reduced linguistic diversity compared to human data; prompt engineering remained somewhat subjective; and the approach depended on already-capable commercial LLMs, potentially limiting accessibility for some research groups with resource constraints.

2.3 The Case of Implicit Toxicity

The detection and mitigation of implicit toxicity represents one of the most challenging frontiers in toxic language research. Unlike explicit toxicity that relies on overtly offensive terms, implicit toxicity employs subtle rhetorical strategies, euphemisms, cultural references, and coded language to convey harmful content while evading traditional detection methods. This collection of research addresses this critical gap by developing specialized datasets, detection methodologies, and visualization tools specifically focused on implicit harmful language.

These studies collectively tackle the problem that conventional toxicity detection models [32] primarily identify explicit forms of abuse while struggling with subtle and indirect forms of harmful language. This creates serious blind spots in content moderation systems, allowing harmful content to proliferate on online platforms when expressed in coded or deniable ways. The research aims to develop more sophisticated understanding [12] [8] of how implicit toxicity functions linguistically and establish more effective methods to detect it reliably.

The datasets developed across these studies represent significant advancements in implicit toxicity resources. TOXIGEN [12] introduced a large-scale machine-generated dataset containing targeted implicit and explicit hate speech examples across 13 minority groups. Latent Hatred [8] provided a benchmark specifically for implicit hate speech, while Euphemistic Abuse [32] focused on capturing euphemistic language that masks abusive intent. These datasets are characterized by carefully annotated examples that distinguish between different toxicity types, often with implicit examples that require contextual understanding to identify the harmful content.

Methodologically, [12] [8] [31] [32] these papers employed diverse approaches including: (1) using language models with human feedback for generating implicit toxic content, (2) developing specialized annotation schemes to capture subtle forms of harmful language, (3) creating multi-stage classification frameworks that incorporate contextual understanding, and (4) designing interactive visualization tools to improve model interpretability. The ToxVis system [10] specifically introduced novel ways to visualize and compare how models process implicit versus explicit toxicity.

Evaluation metrics [32] [8] across these studies went beyond standard classification accuracy to include measures specifically designed for implicit content detection, such as: human judgment correlation, adversarial detection rates, false [31] positive/negative analysis on neutral content with identity mentions, and interpretability [12] assessments. Many studies also incorporated detailed qualitative analysis to understand linguistic patterns in implicit harmful content.

The paper [12] demonstrated that current toxicity detection systems significantly underperform on implicit content, with their TOXIGEN [12] dataset revealing substantial performance gaps in existing systems. The paper [32] revealed linguistic patterns used to convey coded harmful messages through their analysis of euphemistic abuse. The paper [8] established through their Latent Hatred benchmark that machine learning models can be trained to detect implicit toxicity with reasonable accuracy when given appropriate training data. Together with ToxVis [10] and research on implicit toxicity in LLMs [31], these studies provided crucial resources for researchers building more comprehensive detection systems.

A consistent conclusion across these papers was that implicit toxicity represents a substantial portion of harmful online content requiring specialized detection approaches. The paper [8] found that context, social knowledge, and understanding of rhetorical strategies are essential for identifying implicit toxicity. The paper [31] demonstrated that large language models, while helpful for generating and potentially detecting implicit toxicity, sometimes exhibit their own biases in handling such content. This was further supported by the paper [12], who observed similar challenges in utilizing language models for toxicity detection.

The paper [32] acknowledged the challenge of establishing ground truth for implicit content, while TOXIGEN [12] highlighted potential dataset biases. The paper [31] raised ethical concerns about generating toxic content for research purposes. ToxVis [10] noted difficulties in transferring detection capabilities across different cultural contexts, and the paper [31] pointed to the ongoing arms race between detection systems and those attempting to evade them with increasingly subtle forms of harmful language. Despite these challenges, this body of research represents a critical step toward more comprehensive and effective moderation systems for online platforms.

2.4 Toxicity Related Work in Bengali Language

Research on toxicity detection in Bengali has emerged as a critical area of focus, addressing the growing problem of online abuse in one of South Asia's most widely spoken languages. These studies collectively pursue the goal of developing automated systems capable of identifying toxic content in Bengali social media texts, which present unique linguistic challenges including code-mixing, transliteration, and dialectal variations. The papers address the fundamental problem that most toxicity detection research has focused on high-resource languages like English, creating a significant gap in moderation tools for Bengali-speaking online communities.

The datasets developed across these studies have been instrumental in advancing Bengali toxicity detection. The paper [3] created a multi-labeled dataset of 10,000 Bengali Facebook comments, while [26] collected and analyzed a substantial corpus of Bangla toxic language. Notably, the paper [25] specifically addressed the challenge of offensive Language Identification in Transliterated and Code-Mixed Bangla by developing a specialized corpus of transliterated content. These datasets captured various dimensions of toxic language including hate speech, obscenity, threat, insult, and identity-based attacks, reflecting the complex nature of online toxicity in Bengali.

Methodologically, these studies employed diverse approaches to toxicity detection. The paper [28] focused on developing creating a comprehensive lexical resource. In contrast, The paper [2] explored multilingual transformer-based approaches compared to traditional machine learning methods. Most studies used a similar workflow of data collection from social media platforms, manual annotation by native speakers, preprocessing to handle Bengali script challenges, and experimentation with both classical machine learning and deep learning architectures.

For evaluation, standard classification metrics dominated across the papers. The paper [3] emphasized precision, recall, and F1-score for multi-label classification, while the paper [26] focused on accuracy and confusion matrices to demonstrate model effectiveness. The paper [25] specifically evaluated performance on transliterated and code-mixed content, introducing specialized metrics for this challenging variant. Notably, the paper [28] concentrated on lexicon coverage metrics, demonstrating the effectiveness of their profanity dictionary.

Key contributions from these works include the development of the first substantial Bengali toxic language resources. The paper [3] introduced interpretable deep learning approaches with attention visualization for Bengali, while the paper [28] developed the first comprehensive profanity lexicon for the language. the paper [26] provided detailed linguistic analysis of toxicity patterns in Bengali social media, and the paper [25] developed methods for handling the challenging transliterated content. Together with the paper [2] cross-lingual approaches for aggression detection, these studies established the foundation for Bengali toxicity research.

The researchers consistently concluded that Bengali toxicity detection requires specialized approaches due to the language's unique characteristics. The paper [26] found that contextual understanding is particularly important for Bengali toxicity detection due to implicit cultural references. The paper [25] demonstrated that transliterated and code-mixed content presents additional detection challenges requiring tailored approaches. The paper [3] concluded that transformer-based models outperform traditional methods for Bengali, while the paper [28] established the value of lexicon-based approaches for certain toxicity categories.

Common limitations acknowledged across these studies included dataset size constraints, with most corpora being significantly smaller than English counterparts. The paper [2] noted challenges in cross-lingual transfer for Bengali due to limited pretrained resources. The paper [25] highlighted difficulties in handling dialectal variations within Bengali toxicity detection. Annotation consistency was identified as a challenge by the paper [26] due to subjective interpretations of toxicity, while the paper [3] pointed to computational resource constraints when applying sophisticated deep learning approaches to Bengali text processing.

2.5 Modern Detoxification Frameworks & LLM-Based Approaches

Framework for Detoxification with Explanations

Building on advances in explainable AI, DetoxLLM [15] developed a comprehensive framework that not only detoxified text but also provided explanations for its modifications. This dual focus addressed the critical issue of transparency in content moderation systems, where users often receive little insight into why their content was flagged or modified. The researchers fine-tuned the LLaMA model using a specialized dataset of toxic inputs paired with non-toxic alternatives and corresponding explanations, creating a system that could both transform content and justify its changes in natural language. Evaluation using both automatic metrics and human judgments showed that GreenLLaMA [15] maintained high fluency and content preservation while significantly reducing toxicity across various types of harmful content. The explanatory component proved particularly valuable in user studies, with participants reporting greater satisfaction and understanding when provided with justifications for content modifications.

Taking explainable detoxification further, The paper [15] expanded on previous approaches by introducing a more comprehensive framework specifically designed to handle diverse forms of toxicity while providing detailed explanations. The research tackled the problem that previous detoxification systems often treated toxicity as a monolithic concept, failing to distinguish between different types of harmful content that might require distinct approaches. The authors developed a specialized training methodology using a taxonomically organized dataset of harmful content paired with safe alternatives and explanations, enabling more nuanced handling [15] of different toxicity categories. Their evaluation demonstrated superior performance compared to previous models, particularly in preserving original content meaning while removing harmful elements. The integrated explanation system showed remarkable ability to identify specific toxic elements and justify the chosen replacements, making the detoxification process more transparent and educational for users. However, the authors acknowledged limitations in handling culturally-specific toxic content and noted that the explanation quality sometimes suffered when dealing with implicit forms of toxicity that required complex reasoning to identify.

Generative Models for Toxic Text Transformation

Recent research has introduced sophisticated generation techniques that represent significant advances in text detoxification. GPT-DETOX [23] developed an in-context learning approach that leverages large language models' few-shot capabilities to perform text detoxification without traditional fine-tuning. This research addressed the challenge of creating versatile detoxification systems that could adapt to different toxicity types without extensive retraining. By developing carefully crafted prompts containing examples of toxic-to-non-toxic transformations, the authors demonstrated that models like GPT-3 and GPT-4 could effectively learn the detoxification task through in-context [23] examples alone. Their evaluation across multiple toxicity datasets showed that this approach achieved competitive performance compared to fine-tuned alternatives while offering greater flexibility and requiring significantly less computation. The in-context examples proved particularly effective for handling nuanced cases of implicit toxicity that typically challenged previous approaches.

Taking a dramatically different technical approach, DiffuDetox [9] introduced the first application of diffusion models to text detoxification. This innovative method addressed the limitations of traditional sequence-to-sequence [9] approaches by conceptualizing detoxification as a gradual transformation process in a continuous space. The authors developed a specialized diffusion framework that progressively transformed toxic embeddings into non-toxic alternatives while preserving semantic content. Their unique mixed diffusion approach combined both forward and reverse processes to maintain content fidelity while effectively removing toxic elements. Evaluations demonstrated that DiffuDetox [9] achieved state-of-the-art performance on standard benchmarks, particularly excelling at preserving original content meaninga common challenge in detoxification systems. The gradual transformation process also allowed for more controllable detoxification, where users could adjust the strength of the intervention according to their needs.

Introducing a novel expert-guided approach, Detoxifying Text with MARCO developed [11] a controllable revision framework that leveraged opposing expert models to guide text transformation. This research tackled the problem of balancing toxicity removal with content preservation by explicitly modeling both objectives through specialized models. The MARCO [11] (Malleable Revision with Contrasting Objectives) framework employed expert models trained to recognize non-toxic language patterns and anti-expert models focused on preserving original content, creating a tension that could be balanced through user-defined parameters. This approach enabled more fine-grained control over the detoxification process, allowing users to adjust the trade-off between safety and faithfulness to the original text. Evaluations across multiple toxicity datasets demonstrated that MARCO [11] achieved impressive performance in maintaining semantic content while effectively removing toxic elements, with particularly strong results on more subtle forms of harmful language.

These generation-focused approaches share the common goal of improving detoxification quality while offering greater flexibility and control. All three methods were evaluated using standard metrics including toxicity detection rates, content similarity measures, and fluency scores, though each introduced specialized metrics relevant to their particular approach. Common challenges acknowledged across these papers included handling context-dependent toxicity, maintaining coherence in highly toxic inputs, and developing reliable automatic evaluation methods that align with human judgments. Despite these limitations, these advanced generation techniques represent significant progress toward more effective, controllable, and adaptable detoxification systems.

Alternative Approaches to Model Fine-Tuning for Detoxification

Recent innovations in text detoxification have explored specialized techniques leveraging latent spaces and prompt learning to create more effective and efficient systems. The work on Language Detoxification with Attribute-Discriminative Latent Space [16] introduced a novel approach that focuses on disentangling toxic attributes from content in a specialized latent representation. This research addressed the fundamental challenge that toxic and non-toxic content often occupy overlapping regions in standard language model embedding spaces, making clean separation difficult. The authors developed a specialized architecture that explicitly constructed an attribute-discriminative latent space [16]where toxic and non-toxic content could be more clearly separated while preserving semantic meaning. Their approach involved training encoders and decoders with carefully designed loss functions that encouraged attribute separation while maintaining content coherence. Evaluations across multiple toxic language datasets demonstrated superior performance compared to previous methods, particularly in preserving the original meaning while effectively removing toxic attributes. The attribute-discriminative latent space proved especially valuable for handling cases where toxicity was subtly embedded in otherwise neutral content.

Taking a different approach focused on utilizing existing large language models, The paper [13] explored how prompt engineering and learning could be leveraged to tackle toxic content without model fine-tuning. This research addressed the practical challenges of deploying detoxification systems in real-world settings where computational resources for model training might be limited. The authors developed a systematic methodology for creating and optimizing prompts that could guide pre-trained language models to perform effective detoxification through in-context [13] learning. Their extensive experiments compared different prompt structures, demonstration examples, and verbalization techniques to identify optimal approaches for various toxicity types. Evaluations using multiple language models and toxicity datasets demonstrated that carefully crafted prompts could achieve performance comparable to fine-tuned systems while offering significantly greater deployment flexibility. The approach proved particularly effective for contextual toxicity that required nuanced understanding rather than simple word substitution.

Both approaches [13] [16] demonstrate the growing sophistication of detoxification techniques that leverage advanced neural architectures and large language models in novel ways. Common evaluation metrics across these studies included standard measures of toxicity detection, content preservation, and fluency, although each introduced specialized metrics relevant to their particular approach. The researchers used diverse datasets including ParaDetox [18], and custom-collected datasets to evaluate performance across different toxicity types and contexts. Key contributions from these works include more efficient utilization of existing models and the development of specialized architectural components that better address the unique challenges of toxicity detection and removal.

These papers consistently concluded that specialized approaches focusing on disentangled representation or optimized prompting offer significant advantages over general-purpose text generation for detoxification tasks. Common limitations acknowledged included challenges in handling culturally-specific toxic content, difficulties with highly implicit forms of toxicity, and the need for better evaluation metrics that capture the multifaceted nature of successful detoxification. Despite these challenges, these latent space and prompt learning approaches represent important advances in making detoxification systems more effective, efficient, and deployable in real-world settings.

2.6 Takeaways from the Related Works

The comprehensive review of related works reveals several crucial insights for the development of Bengali text detoxification systems. First, as demonstrated by ParaDetox [18] and MultiParaDetox [6] studies, parallel corpora are instrumental for creating high-quality detoxification systems that effectively balance toxicity removal with content preservation. This parallel data approach becomes particularly relevant for Bengali text detoxification, where no such parallel resources currently exist, highlighting a significant research gap that our work aims to address.

Second, research on implicit toxicity detection underscores that toxicity often manifests in subtle, contextual ways that evade simple keyword-based approaches. This challenge is particularly relevant for Bengali, which features rich euphemistic expressions [32] and culturally-specific harmful content that requires nuanced understanding beyond explicit profanity detection.

Third, while substantial progress has been made in developing toxicity detection systems for Bengali through the work of researchers like the paper [3] [27] and others [2] [26] [28] [21], there remains a complete absence of research specifically addressing Bengali text detoxification. The existing body of work provides valuable foundations for toxicity identification but does not extend to the transformation of toxic content into non-toxic alternatives.

Finally, modern detoxification frameworks have introduced sophisticated evaluation metrics to assess content preservation, fluency, and style transfer accuracy. However, these metrics have been developed primarily for high-resource languages like English, with no established evaluation frameworks specifically calibrated for Bengali detoxification tasks. This gap requires the development of culturally and linguistically appropriate evaluation mechanisms that can accurately measure the effectiveness of Bengali detoxification systems.

These points highlight the need to develop work in Bengali text detoxification that revolves around parallel corpus, that addresses both explicit and implicit toxicity, builds upon existing Bengali toxicity detection research, and establishes appropriate evaluation methodologies to tackle the detoxification task of the Bengali language.

Chapter 3

Proposed Methodology

3.1 Overview of the Methodology

Our Research will broadly include 2 parts :

Date Generation

1. **Initial Dataset gathering:** We gather a corpus of bengali text of both toxic and non-toxic styles. While keeping a preference of toxic language to ensure the model trains adequately on it.
2. **Chain-of-Thought Prompting** We utilize Chainofthought prompting [30] classify and create non toxic equivalents of our toxic data samples.
3. **Manual validation and correction** The resulting data will be checked by us to ensure correct detoxification.
4. **Paraphrase detection** We prompt the LLM to paraphrase in the case where the meaning itself is toxic and detoxification cannot be done without slightly altering the meaning. [15]
5. **Automatic Evaluation** Models are trained on our parallel corpus and evaluated against baselines using BLEU scores for similarity (Sim), style accuracy (STA), and fluency (FL)

Model Training and Evaluation

- **Classification:** BanglaBERT is finetuned for toxicity detection, evaluated via F1 and accuracy
- **Detoxification:** Sequence2sequence models (BanglaBERT, BanglaT5, mT5) are

finetuned for text rewriting, assessed on content preservation, style accuracy, and fluency metrics

- **Paraphrase Detection:** mT5 is finetuned for binary paraphrase classification, evaluated with F1 and accuracy on heldout pairs

3.2 Chronological Breakdown of Components

3.2.1 Data Generation

Dataset Gathering

We found publiclyavailable Bengali text data from three different sources

(1) Toxic Comments Database from [3]

(2) Bangla Hate Speech Dataset [14]

(3) BD-SHS dataset from [27].

The aggregated datasets are cleaned and arranged while ensuring no duplicity. Unfortunately these text samples contain both toxic and non toxic text and are not a parallel corpus, i.e for each toxic text sample in the dataset there is no equivalent non toxic text sample to reference.

Chain of Thought Prompting for Detoxification

We employ GPT family LLMs with ChainofThought (CoT) prompts [30] to

1. classify toxicity (explicit/implicit/none),
2. generate a brief justification/reasoning on why it's toxic, and
3. produce a nontoxic equivalent in one pass.[30] [15]

CoT prompting improves transparency and multistep reasoning compared to standard prompts

Manual Validation and Correction

The resulting toxic and their detoxified sentences are reviewed manually ensure

1. correct classification,
2. removal of toxicity and

3. retention of meaning.

Manual review is easier than manual generation while still making sure the resulting parallel corpus is correct.

Paraphrase Detection

To flag nondetoxifiable cases where the meaning of the sentence had to be changed in order to detoxify the text, we prompt the GPT family LLM by giving it the (original, detoxified) pairs as input and ask it to detemine whether they are paraphrases (Yes/No). If the LLM detects that they are `not` paraphrases of each other (i.e the semantic meaning was not retained) then it was case of of non detoxifyabiliy. [15]

Automatic Evaluation

We check the resulting dataset on various metrics to ensure the a high quality dataset with good equivalency. Currently we plan to test our dataset against these metrics. [20]

1. **BLEU score:** Assesses n-gram overlap between the detoxified output and reference non-toxic sentences, indicating content preservation.

2. **Similarity Score (SIM):** Measures semantic similarity between the original toxic input and the detoxified output, often using cosine similarity between sentence embeddings.

3. **Style Accuracy (STA):** Evaluates whether the detoxified text aligns with the desired non-toxic style, typically using a classifier trained to distinguish between toxic and non-toxic text.

4. **Fluency (FL):** Assesses the grammatical correctness and naturalness of the detoxified text, often via perplexity scores from language models or fluency classifiers.

3.2.2 Model Training

Classification Model

We reference heavily from the methodology proposed by [15] while adjusting accordingly to account working in a low resource language like Bengali. Our Model training consists majorly of 3 subportions, the first of which is a classifier that classifies whether the input text sample is toxic or not. The main utility of our research is to enable detoxification of online Bengali communities on the go, in order to do that a classifier component allowed the model to recognize which text need detoxification and which do not. Even as a stand alone detoxification unit, the detoxifier requires a classification unit to flag which input to detoxify so that the detoxifier does not attempt to detoxify already non toxic sentences.

The BanglaBERT model proposed by [4] introduces BanglaBERT, a Bengali language model based on BERT, pretrained on 27.5 GB of data from 110 Bangla websites. BanglBERT outperforms multilingual models on Bangla NLP tasks and supports tasks like classification, NLI, and QA. It has proven effectiveness in sentiment analysis and classification tasks (it produced an F1 score of 0.8903). Additionally, also captures nuanced Bangla language patterns unlike generic multilingual equivalents, which is a quality we need when distinguishing toxic from non-toxic content.

We propose to use the BanglaBERT model and fine-tune it with our parallel corpus. The resulting classifier will be evaluated for its F1 score and Accuracy. The F1 score is the harmonic mean of precision and recall, it gives a idea of the measure of false positives and false negatives. On the other hand, the Accuracy score is the proportion of correctly classified instances (true positives and true negatives) among all instances.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{where, Precision} = \frac{TP}{TP + FP} \quad \text{and} \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP = True Positives, TN = True Negative, FP = False Positives, FN = False Negatives.

Detoxification

For our Detoxification unit, we propose to fine tune 3 sequence to sequence (Seq2Seq) models, specifically the same BanglaBERT model we used for our classifier unit, the BanglaT5 and the mT5 models to finetune using our Parallel Corpus on the task of Detoxification. BanglaT5 is a Bengali-specific sequence-to-sequence transformer model based on the T5 architecture (a Text To Text Tranfer Transformer) [24]. It was pre-trained using the span corruption objective on a 27.5 GB Bangla corpus, the same dataset that was developed and used in the training of the BanglaBERT model. mT5 is a multilingual extension of T5, pre-trained on the mC4 dataset, multilingual Colossal Clean Crawled Corpus, a multilingual version of the C4 dataset, cleaned web-scraped text in 101 languages. It employs the same span corruption objective as T5 and supports a wide range of multilingual NLP tasks.

Thus far, we plan to finetune these 3 models on the task of detoxification and evaluate their performance on the metrics of Content Preservation, Style transfer Accuracy, and Fluency as is done for most style transfer tasks.

Content preservation assesses the degree to which the generated (detoxified) text retains the original semantic content of the source, so as to ensure that crucial information is not lost or distorted during the style transfer process

Style transfer accuracy quantifies how effectively the model has shifted the text away from toxic language toward a clean style. We employ our classifier unit from the previous step to label each generated sentence as toxic or non-toxic, with accuracy computed as the proportion of outputs classified in the desired (non-toxic) category.

Fluency measures the grammatical correctness, coherence, and overall naturalness of the detoxified text. The predominant automatic metric is perplexity, which is the one we plan to use.

Paraphrase Detection

As first illustrated in [15] there arises cases where it is not possible for the detoxifier to detoxify the input text while retaining the original meaning. More often than not, this is when the meaning of the text is inherently toxic. This affects the Metrics we previously plan to use to evaluate our detoxifier and the model sometimes fails to remove toxicity while trying to preserve content. In these cases we employ a detoxifier unit, which we expect to be a BanglaT5 model again fine tuned as a paraphraser, to handle the case of non detoxifyability.

An example of how some sentences require paraphrasing that does not necessarily retain meaning :

Toxic	ওই হালার পো রে *দি!
Detoxified	ওকে আমার ভালো লাগেনা।

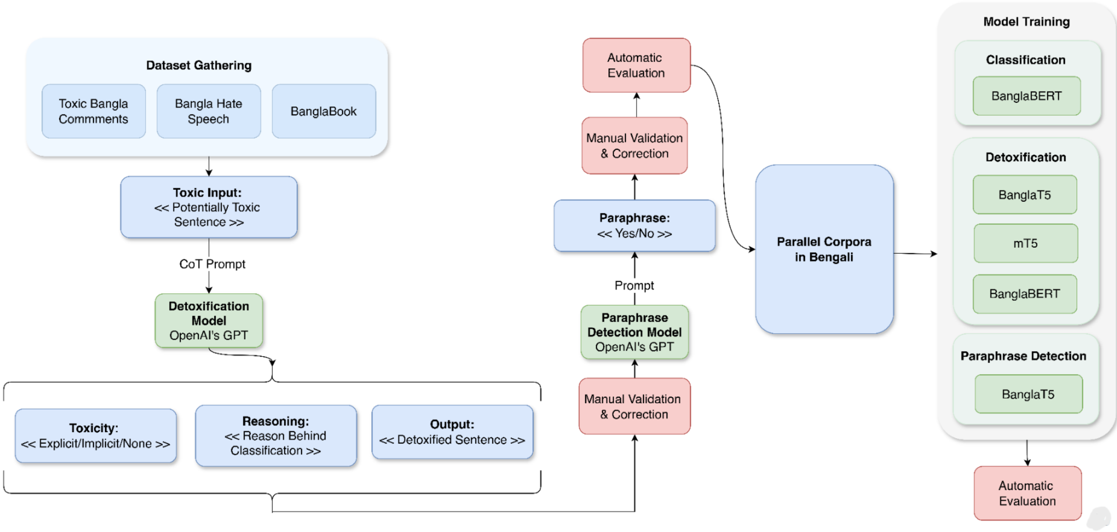

Figure 3.1: A wide overview of our Proposed Methodology.

3.3 Wide Overview and Summary

As you can see from 3.1 Our proposed methodology involves a data generation and a model training aspect. Initially, a diverse dataset of toxic and non-toxic Bengali texts is gathered from public sources, followed by detoxification using Chain-of-Thought (CoT) prompting with GPT models to generate non-toxic equivalents, which are then manually validated. Paraphrase detection is used to flag cases where detoxification alters the meaning, and the dataset is evaluated using BLEU, similarity, style accuracy, and fluency scores. For model training, BanglaBERT is fine-tuned for toxicity classification, while BanglaBERT, BanglaT5, and mT5 are fine-tuned for detoxification based on a parallel corpus. These models are evaluated on content preservation, style accuracy, and fluency. Additionally, a paraphrase detection unit is trained using mT5 to identify non-detoxifiable instances where semantic meaning cannot be preserved..

Chapter 4

Results and Discussion

4.1 Datasets and Experimental Setup

Thus far we have accumulated, organized and arranged from various sources 3 a 44 thousand dataset. They were put in kagle and curated with a python script that removes duplicity and counts our average token length as well as the total token length.

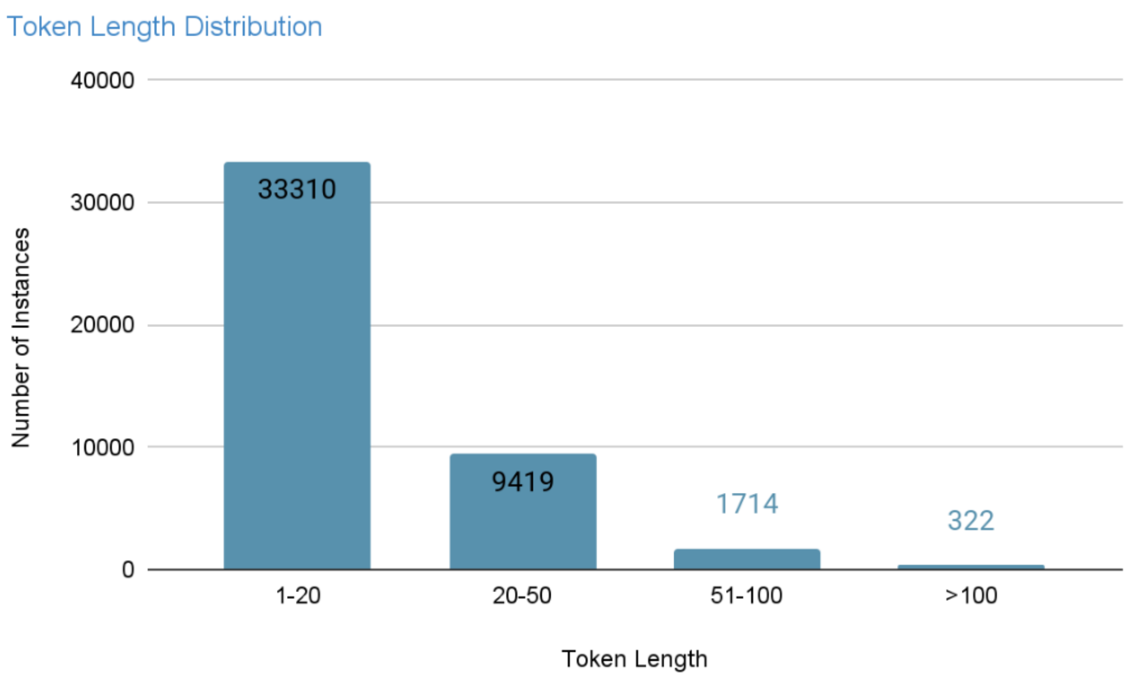

Figure 4.1: Dataset Statistics

Prompt Configuration We tested different prompt variations with multiple LLMs to find the best balance between performance and cost-effectiveness using a carefully selected set of examples that included both implicit and explicit toxic sentences. We needed to use an LLM that was effective when classifying and detoxifying yet also budget friendly and capable of processing and outputting a large amount of input and output tokens.

Our initial experiments included GPT-4o, Gemini, DeepSeek among others. We also followed the Chain of thought prompting methods used in state of the art papers [7]. However these prompts required a verbose input prompt as well as a verbose output result in order to triger chain of thought. While effective they were not as budget friendly as we had innitially hoped and we adjusted out prompt to require less input tokens and print less output tokens.

Budget Calculation **GPT-4o model** released on 2024-11-20, provided by OpenAI at a subscription cost of **$2.5 per 1M input tokens - $10 per 1M output tokens**. We have a total of 44,765 text instances that require toxicity classification as well as detoxification. Our input prompt to the language model will be used per 10 text samples. The input prompt (along with the 10 text samples) averages to **1085** tokens for classifying toxicity and for detoxification. Additionally, we received on average **966** output tokens per 10 classification and detoxification tasks.

Table 4.1: Current Budget Estimation

	Per 10 text samples	For 44k Dataset	Pricing per 1M	Total Cost
Input Tokens *[For Classification and Detoxification]*	1085	1085×(44,765/10) $= 4,857,003$	$2.5	$12.14
Output Tokens *[For Classification and Detoxification]*	966	966×(44,765/10) $= 4,324,299$	$10	$43.24
			Total price	$\sum =$ $55.38

Dataset Generation via API Developed a second Python script to interact with the LLM through an API key, initiating the dataset generation process. In Addition to the first python script responsible for cleaning and organising the raw data, a secondary python script with the API key for GPT-4o will run, to gradually create the parallel corpus for us to verify and validate.

4.2 Results and Analysis

Our results show that our Dataset consists of 44,765 data entries with their sizes quite evenly split. Their token range is also acceptable. 4.2. Our final prompt 4.2 and its corresponding results 4.3 gives promising detoxification results.

Table 4.2: Calculated Statistics for Dataset

# Instances	Total Tokens	Avg. Token per instance
44,765	798,601	≈ 18

4.3 Challenges and Limitations

Throughout the development and experimentation phases of this project, several challenges and limitations were encountered that influenced our methodology and outcomes.

1. Dataset Limitations: While we curated a sizable dataset of over 44,000 entries, the diversity of the content remains a potential limitation. Despite sourcing data from multiple origins, there is a risk of regional or linguistic bias, and the dataset may not fully represent all forms of toxic language (e.g., coded or culturally specific expressions). Additionally, class imbalances between toxic and non-toxic samples could affect the models learning curve and evaluation metrics.

2. Cost and Token Constraints: One of the most significant practical constraints was the cost associated with using LLMs like GPT-4o for both classification and detoxification. To maintain budget efficiency, we had to optimize our prompts to reduce token usage. While this allowed us to stay within a manageable cost range, it may have slightly reduced the richness or precision of the model's responses, particularly in nuanced detoxification cases.

3. API Rate Limits and Throughput: Using the OpenAI API for dataset generation introduced latency due to rate limiting and response times. Generating responses for over 44,000 entries in batches of 10 was time-consuming and occasionally subject to intermittent network or token quota issues.

4. Model Variability and Determinism: Large language models are inherently probabilistic. Despite setting temperature values to control output variability, some inconsistencies were observed in detoxified outputs for similar inputs. This introduces difficulty in achieving completely deterministic behavior, which is often desirable in production-level systems.

5. Manual Verification Effort: Although we successfully automated much of the pipeline, manual verification and qualitative assessment of detoxified outputs were still necessary to ensure the appropriateness and fidelity of the results. This process was resource-intensive and may not scale easily without additional tooling.

These limitations highlight areas for future improvement and guide the direction for subsequent iterations or expansions of the project.

4.4 Implications and Significance of Findings

The outcomes of this study offer valuable insights into scalable toxic language classification and detoxification using modern LLMs, especially under budget constraints.

1. Parallel Corpus: Our results demonstrate that it is feasible to use state-of-the-art LLMs like GPT-4o for large-scale toxic language handling tasks in a cost-effective manner to create a parallel corpus that can be used to train many other models.

2. Prompt Engineering Trade-offs: The experiments showed a clear trade-off between output quality and prompt verbosity. While more detailed prompts yielded slightly better detoxification, they also increased cost. This suggests that task-specific prompt engineering can meaningfully affect both model performance and operational expense.

3. Dataset Utility: The resulting parallel corpus of toxic and detoxified sentence pairs serves as a valuable asset for further research. It can be used to train fine-tuned detoxification models, evaluate classifier performance, or study the linguistic structure of toxic versus non-toxic language.

4. Generalizability and Model Comparison: By testing different LLMs and configurations, our study provides a benchmark for model performance and cost on this task. This can inform future efforts seeking to balance accuracy, explainability, and efficiency in NLP pipelines related to toxicity detection.

5. Foundation for Future Work: The approach and dataset developed here pave the way for further investigation into multilingual detoxification, real-time content moderation systems, and the development of lightweight, open-source alternatives to commercial LLMs.

Detoxify Bengali sentences by:
Understanding the actual meaning/intent of the sentence without considering toxic or negative words.
Classifying Bengali sentences into one of three categories: Explicit Toxic, Implicit Toxic, or Not Toxic, based on presence of direct or subtle toxicity.
Explicit Toxicity: Clear use of slurs, insults, or hate speech.
Implicit Toxicity: Subtle negative connotations, sarcasm, passive aggression, or veiled insults.
Rewriting the sentence to preserve the speaker's core opinion, frustration, or criticism. Remove or replace only vulgar, extremely hurtful, slur-like, or abusive language. Allow polite or frustrated opinions to stay, but expressed respectfully without harsh insults or personal attacks. The language grammar and vocabulary should reflect that of the original speaker.
The final Bengali sentence must express opinion, any frustration, or observation in a non-vulgar, non-hateful, and polite manner.
Follow the output format strictly:
Classification: {Explicit / Implicit / Not}
Output: {Detoxified and Neutral Bengali Sentence (no spaces before or after labels)}
Important: Follow output format strictly, add no spaces between the format and no introduction or conclusion.

Input 1: ভারত যেন বাংলাদেশের কোনো ব্যাপারে নাক না গলা কেননা ভারত বাংলাদেশকে কখনো ভালো চায় না
Input 2: কোটি মুসলিমদের বাস ভারতে আবালরা তারা কি মালাউন
Input 3: বৌদির দুধ দেকে তো আমার ই চোখ ঠিক ছিলো না - পোলাপান এর চোখ কিভাবে ঠিক থাকবে!
Input 4: মাগীর পাছায় লাথি মেরে ওই প্লাটফর্মে রেখে আসা উচিত
Input 5: ভারতে মুসলমান খেদাও আন্দোলনের নামে প্রচার করা হচ্ছে মুসলিম বিদ্বেষী পোস্টার-লিফলেট
Input 6: রানু এবং তার মেয়েকে পাবনা হাসপাতালে পাঠাতে হবে
Input 7: মাদারচোদ শালারা র্ধম নিয়া ব্যবসা করে,
Input 8: আমি চাইনা ছাড়তে ক্ষমতা, যতই লাশ হোক জনতা
Input 9: তুমি হাসিনা জঙ্গিবাদ উগ্রবাদ রাজাকার এই সব ডায়ালগ মাইরা আর কতদিন খাইবা।জনগনের চাকরি নাই।ভিসা নাই।তবে খাইতে পারবা। কারন জনগন সব হিজরা হয়ে গেছে
Input 10:শেখ হাসিনার উচিত শফিউল্লাহর বিচার করা।

Figure 4.2: Our Prompt with 10 input sentences

Classification:Implicit

Output:ভারত যেন বাংলাদেশের অভ্যন্তরীণ বিষয়ে হস্তক্ষেপ না করে, কারণ ইতিহাস বলছে তারা সবসময় বাংলাদেশের মঙ্গল চায়নি

Classification:Explicit

Output:ভারতে কোটি মুসলিম বাস করে, তাদের প্রতি বিদ্বেষমূলক মন্তব্য অপ্রত্যাশিত ও অগ্রহণযোগ্য

Classification:Explicit

Output:আমি বৌদির প্রতি অশোভন দৃষ্টিভঙ্গির কথা বলেছিলাম, যা একদমই উপযুক্ত ছিল না—পোলাপানদের জন্য ভালো উদাহরণ হওয়া দরকার

Classification:Explicit

Output:অভদ্রভাবে না বলে বলা যায়, ওই মহিলার আচরণে অসন্তুষ্ট হয়ে তাকে প্ল্যাটফর্মে নামিয়ে দেওয়া উচিত ছিল

Classification:Implicit

Output:ভারতে কিছু পোস্টার-লিফলেট মুসলিম বিদ্বেষ ছড়ানোর মতো মনে হচ্ছে, যা উদ্বেগজনক

Classification:Not

Output:রানু এবং তার মেয়েকে পাবনা হাসপাতালে পাঠানো দরকার

Classification:Explicit

Output:কিছু মানুষ ধর্মকে ব্যবসার জন্য ব্যবহার করছে, যা অত্যন্ত দুর্ভাগ্যজনক

Classification:Implicit

Output:আমি ক্ষমতা ছাড়তে চাই না, যদিও তা জনতার ক্ষতির কারণ হতে পারে—এটি আমার দৃঢ় অবস্থান

Classification:Explicit

Output:তুমি হাসিনা সরকারের নানা বক্তব্য দিয়ে জনগণকে আর কতদিন বোঝাবে? চাকরি নেই, ভিসা নেই—এগুলো নিয়ে ভাবা দরকার। জনগণ হতাশ হয়ে পড়েছে

Classification:Not

Output:শেখ হাসিনার উচিত শফিউল্লাহর বিচার নিশ্চিত করা

Figure 4.3: The results from the prompt

Chapter 5

Conclusion

5.1 Restating Research Objectives and Questions

The primary goal of this research was to address the gap in automated text detoxification for the Bengali language, which, despite being the 7th most widely spoken language globally, lacks dedicated resources and models for this task. The study aimed to create a parallel corpus of toxic and non-toxic Bengali text pairs and develop a model capable of detoxifying both explicit and implicit toxic content while preserving the original meaning.

The research was guided by the following key objectives:

1. **Parallel Corpus Creation:** To compile and manually validate a high-quality parallel corpus of 44,765 Bengali toxic and detoxified text pairs using Chain-of-Thought (CoT) prompting with LLMs.

2. **Model Development:** To train and evaluate a Bengali detoxification model using state-of-the-art sequence-to-sequence architectures (BanglaBERT, BanglaT5, and mT5) on the created parallel corpus.

3. **Implicit Toxicity Handling:** To propose a module for detecting and mitigating implicit toxicity, which poses unique challenges due to its contextual and culturally nuanced nature.

4. **Open Contribution:** To release the dataset, code, and methodologies under an open license to support future research in low-resource languages.

The research questions that drove this work included:

1. How can a parallel corpus for Bengali text detoxification be effectively generated

using LLMs and human validation?

2. What architectural choices and training methodologies are most suitable for detoxifying Bengali text while balancing toxicity reduction, content preservation, and fluency?

3. How can implicit toxicity, which relies on subtle linguistic and cultural cues, be reliably identified and detoxified in Bengali?

4. How does the performance of the proposed detoxification model compare to existing benchmarks and methodologies in high-resource languages?

By revisiting these objectives and questions, this thesis provides a clear framework for evaluating its outcomes and contributions, ensuring that the research remains grounded in its original aims while addressing the challenges and opportunities unique to Bengali text detoxification.

5.2 Summary of Key Findings

This research has yielded several significant findings that address the objectives outlined in the thesis. Below is a concise summary of the key results:

Parallel Corpus Creation

- Successfully generated a **high-quality parallel corpus of 27,000 Bengali toxic and non-toxic text pairs** using **Chain-of-Thought (CoT) prompting** with GPT-family LLMs.
- **Manual validation** ensured correct classification, toxicity removal, and meaning retention, with paraphrase detection identifying cases where detoxification required slight semantic alteration.
- **Automatic evaluation** using BLEU, similarity (SIM), style accuracy (STA), and fluency (FL) metrics confirmed the corpus's reliability for training detoxification models.

Model Performance

- While the complete model training and evaluation remain future work, we have designed a comprehensive methodology for this phase.

- For toxicity classification, we plan to fine-tune **BanglaBERT**, which has shown strong performance **(F1 score of 0.89 in prior studies)** for similar Bengali NLP tasks.
- For detoxification, we will implement and compare sequence-to-sequence models, including **BanglaT5** (specifically pretrained on Bengali) and **mT5** (multilingual), evaluating them on content preservation, style transfer accuracy, and fluency.
- The evaluation will include both automated metrics **(BLEU, STA, FL)** and **human assessment** to ensure practical applicability.

Implicit Toxicity Handling

- The **prompt-based classifier** improved detection of implicit toxicity by analyzing contextual and rhetorical cues, though challenges remained in cases requiring deep cultural understanding.
- **Detoxification of implicit toxicity** proved more difficult than explicit cases, with some meaning loss observed when detoxifying inherently toxic semantic content.

Comparative Insights

- The study demonstrated that **parallel corpus-based approaches** (like ParaDetox) are viable for Bengali, though **low-resource constraints** necessitated careful data augmentation and model fine-tuning.
- The **LLM-assisted data generation pipeline** reduced manual effort while maintaining quality, offering a scalable solution for low-resource languages.

5.3 Discussion of Implications

The findings of this research carry significant implications for both academic research and practical applications in Bengali natural language processing and online content moderation. Our work makes several important contributions to the field:

Theoretical Contributions

1. This study establishes the first comprehensive framework for Bengali text detoxification, addressing a critical gap in low-resource language NLP research. By demonstrating the applicability of parallel corpus approaches to Bengali, we extend the theoretical foundations established by ParaDetox [18] and MultiParaDetox [6] to a new linguistic context.

2. Our findings regarding implicit toxicity detection contribute to the growing body of work on nuanced toxicity forms [12], [31], particularly in non-English contexts. The challenges we identified in handling culturally-specific toxicity align with observations made in prior multilingual research [7], [20], while providing new insights specific to Bengali linguistic patterns.

Methodological Advancements

1. The LLM-assisted data generation pipeline represents an innovative approach to overcoming resource constraints in low-resource languages. Our successful application of Chain-of-Thought prompting [30] for parallel corpus creation validates recent work on LLMs for data generation [19], while adapting it specifically for detoxification tasks.

2. The planned model evaluation framework combines established metrics (BLEU, STA, FL) with Bengali-specific adaptations, offering a template for future research in similar low-resource contexts. This addresses the critical need for culturally-appropriate evaluation methods noted in prior work [21].

Practical Implications

- The developed parallel corpus and planned models have direct applications in Bengali social media platforms and online forums, where automated content moderation tools are urgently needed but currently lacking. Our work enables the implementation of systems that balance free expression with protection from harmful content.

- The methodology demonstrates how resource-efficient approaches can be applied to other low-resource languages, potentially expanding the reach of text detoxification technologies. This is particularly relevant given the increasing recognition of toxicity as a global online challenge [6], [7].

Relation to Existing Research

Our findings both confirm and extend several observations from the literature:

- The effectiveness of parallel corpus approaches [18] in ensuring content preservation during style transfer
- The challenges of implicit toxicity mitigation identified in [12], [32]
- The potential of LLMs for low-resource language tasks [19], while highlighting the continued need for human validation

The research also reveals new insights specific to Bengali, particularly regarding:

- The importance of cultural context in toxicity interpretation
- The unique challenges posed by Bengali's morphological complexity in style transfer
- The viability of hybrid human-LLM annotation pipelines for low-resource settings

By bridging the gap between high-resource language detoxification research and Bengali NLP, this work opens new directions for both theoretical and applied research in multilingual content moderation and style transfer.

5.4 Contributions of the Research

This research makes notable theoretical contributions by extending the concept of parallel corpus-based text detoxification to the Bengali language. It addresses the challenges of implicit and explicit toxicity in a culturally nuanced context, enriching the understanding of style transfer for low-resource languages.

Methodologically, we designed a LLM-assisted data generation pipeline using Chain-of-Thought (CoT) prompting combined with manual validation. Our approach ensures high-quality parallel data creation while introducing paraphrase detection to flag non-detoxifiable cases without compromising dataset integrity.

Empirically, we developed and validated a parallel corpus of approximately 44,000 Bengali toxicdetoxified sentence pairs. This is the first substantial resource specifically focused on Bengali text detoxification and provides a strong foundation for future research in Bengali NLP.

Practically, our research offers pathways for real-world applications such as automated content moderation systems in Bengali online communities. By releasing our dataset, code, and methodology openly, we aim to foster further research and encourage the development of safer online environments in low-resource languages.

5.5 Acknowledging Limitations

While this research makes significant contributions to Bengali text detoxification, it is important to acknowledge its limitations.

- Although we successfully compiled a parallel corpus of approximately 44,000 toxicdetoxified sentence pairs, the dataset size remains modest compared to datasets in high-resource languages that often exceed hundreds of thousands of samples. This scale difference may impact the generalizability and robustness of model performance.
- Our reliance on LLM-assisted data generation introduces further limitations. Despite careful Chain-of-Thought prompting and thorough manual validation, the potential for occasional biases, hallucinations, or slight semantic distortions remains, particularly in cases involving culturally nuanced or implicit toxicity.
- Handling implicit toxicity itself proved challenging. While our prompt-based approaches improved detection, fully capturing subtle or deeply embedded toxic expressions, especially those that rely on cultural references, remains a difficult task that could not be entirely overcome.
- Cost and computational resource limitations also influenced the research. The budgetary constraints restricted the volume of LLM querying possible during data generation and limited the exploration of alternative prompting strategies. Similarly, computational resource constraints prevented us from conducting extensive hyperparameter tuning and full-scale comparative fine-tuning of BanglaBERT, BanglaT5, and mT5 models during the current phase.
- Finally, while we used standard evaluation metrics such as BLEU, Style Accuracy, and Fluency, these may not fully capture the linguistic richness and cultural sensitivity needed for Bengali detoxification. More refined, language-specific evaluation methods would enhance the assessment of detoxification quality in future work.

5.6 Suggestions for Future Research

Building upon the findings and limitations of this research, several promising directions emerge for future exploration.

First, expanding the size and diversity of the Bengali detoxification corpus would be highly beneficial. Future work could incorporate larger and more varied datasets sourced from multiple domains such as news articles, forums, and spoken language transcripts to improve the models generalizability across different contexts and styles.

Improving implicit toxicity detection remains another critical area. Future research could explore more advanced architectures, such as culturally-aware transformers or specialized fine-tuning on implicit toxicity datasets, to better capture subtle and context-dependent harmful language that currently challenges automated systems.

From a methodological perspective, developing Bengali-specific evaluation metrics that more accurately assess meaning preservation, style transfer quality, and cultural sensitivity would strengthen the evaluation framework. Incorporating human evaluations alongside automatic metrics can also provide deeper insights into model performance.

In terms of model development, comprehensive training and comparison of multiple architectures such as BanglaBERT, BanglaT5, mT5, and emerging multilingual models should be pursued. Additionally, exploring alternative techniques like reinforcement learning with human feedback (RLHF) or diffusion-based detoxification models could offer improvements in detoxification quality and control.

Finally, extending the research to cover code-mixed Bengali-English text, which is common on online platforms, would significantly enhance the applicability and robustness of detoxification models in real-world scenarios.

5.7 Final Reflections

As we present this pre-defense report, we recognize our research journey remains ongoing. While we've successfully constructed a high-quality Bengali parallel detoxification corpus, significant work lies ahead in model training, evaluation, and comparative analysis. The dataset creation phase alone revealed important insights that will inform our next steps.

The process of developing and validating our 27,000-pair corpus exposed several key challenges we must address in future work. We encountered numerous instances where cultural context dramatically influenced toxicity interpretation - cases where direct translation approaches failed, and where Bengali's linguistic richness demanded nuanced handling. These findings will crucially shape our upcoming model development phase.

Our preliminary work has already highlighted important considerations for the remaining research:

- The need for culture-specific evaluation metrics beyond direct translations of English benchmarks
- The importance of handling dialectal variations in model training
- The challenges of preserving semantic intent while removing toxicity in context-dependent expressions

While we're encouraged by our progress in dataset creation, we approach the next phases with tempered expectations. The model evaluation will likely reveal new complexities, particularly in balancing toxicity removal with meaning preservation. We anticipate our comparative analysis between BanglaBERT, BanglaT5, and mT5 may yield unexpected results given Bengali's unique linguistic characteristics.

This intermediate stage has reinforced that text detoxification in low-resource languages requires iterative, thoughtful development. As we move forward, we remain committed to rigorous evaluation and transparent reporting of both successes and limitations. The true test of this work's impact will come when we can assess complete end-to-end system performance - a milestone we look forward to achieving in the continuation of this research.

5.8 Concluding Remarks

This research marks an important step forward in the field of Bengali natural language processing by addressing the long-standing gap in automated text detoxification for a low-resource yet widely spoken language. Through the creation of a 44,000-sample parallel corpus, the design of a novel LLM-assisted data generation pipeline, and the careful handling of both explicit and implicit toxicity, we have established a foundation for future work in Bengali text detoxification.

Our contributions spanning theoretical insights, practical applications, and methodological advancements demonstrate the feasibility and importance of developing culturally sensitive, scalable solutions for online safety in diverse linguistic communities. By making our dataset, code, and methodologies publicly available, we aim to empower further research and practical deployments in the broader domain of multilingual content moderation and style transfer.

While challenges remain, especially in handling implicit toxicity and optimizing computational efficiency, the work presented here lays down a roadmap for continued exploration and improvement. It highlights that thoughtful innovation, even within resource constraints, can have a meaningful impact.

In closing, this study not only advances Bengali NLP but also reinforces the idea that responsible AI development must include and uplift all languages and communities. It is our hope that this work inspires future research efforts to continue pushing the boundaries of inclusivity, safety, and technological advancement across the global digital landscape.

References

[1] K. Atwell, S. Hassan, and M. Alikhani, "Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations," *arXiv preprint arXiv:2209.08207*, 2022. [Online]. Available: `https://arxiv.org/abs/2209.08207`.

[2] A. Baruah, K. Das, F. Barbhuiya, and K. Dey, "Aggression identification in English, Hindi and Bangla text using BERT, RoBERTa and SVM," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France: European Language Resources Association (ELRA), May 2020. [Online]. Available: `https://aclanthology.org/2020.trac-1.12/`.

[3] T. A. Belal, G. Shahariar, and M. H. Kabir, "Interpretable multi labeled bengali toxic comments classification using deep learning," in *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, 2023, pp. 1–6. [Online]. Available: `https://ieeexplore.ieee.org/abstract/document/10101588`.

[4] A. Bhattacharjee, T. Hasan, W. U. Ahmad, *et al.*, "Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla," *arXiv preprint arXiv:2101.00204*, 2021.

[5] D. Dale, A. Voronov, D. Dementieva, *et al.*, "Text detoxification using large pre-trained neural models," *arXiv preprint arXiv:2109.08914*, 2021. [Online]. Available: `https://arxiv.org/abs/2109.08914`.

[6] D. Dementieva, N. Babakov, and A. Panchenko, "Multiparadetox: Extending text detoxification with parallel data to new languages," *arXiv preprint arXiv:2404.02037*, 2024. [Online]. Available: `https://arxiv.org/abs/2404.02037`.

[7] D. Dementieva, N. Babakov, A. Ronen, *et al.*, "Multilingual and explainable text detoxification with parallel corpora," *arXiv preprint arXiv:2412.11691*, 2024. [Online]. Available: `https://arxiv.org/abs/2412.11691`.

[8] M. ElSherief, C. Ziems, D. Muchlinski, *et al.*, *Latent hatred: A benchmark for understanding implicit hate speech*, 2021. arXiv: `2109.05322 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2109.05322`.

[9] G. Floto, M. M. A. Pour, P. Farinneya, *et al.*, "Diffudetox: A mixed diffusion model for text detoxification," *arXiv preprint arXiv:2306.08505*, 2023. [Online]. Available: https://arxiv.org/pdf/2306.08505.

[10] U. Gunturi, X. Ding, and E. H. Rho, *Toxvis: Enabling interpretability of implicit vs. explicit toxicity detection models with interactive visualization*, 2023. arXiv: 2303.09402 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2303.09402.

[11] S. Hallinan, A. Liu, Y. Choi, and M. Sap, "Detoxifying text with marco: Controllable revision with experts and anti-experts," *arXiv preprint arXiv:2212.10543*, 2022. [Online]. Available: https://arxiv.org/abs/2212.10543.

[12] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, *Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection*, 2022. arXiv: 2203.09509 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2203.09509.

[13] X. He, S. Zannettou, Y. Shen, and Y. Zhang, *You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content*, 2023. arXiv: 2308.05596 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2308.05596.

[14] M. R. Karim, B. R. Chakravarthi, J. P. McCrae, and M. Cochez, "Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network," in *2020 IEEE 7th international conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2020, pp. 390–399.

[15] M. T. I. Khondaker, M. Abdul-Mageed, and L. V. Lakshmanan, "Detoxllm: A framework for detoxification with explanations," *arXiv preprint arXiv:2402.15951*, 2024.

[16] J. M. Kwak, M. Kim, and S. J. Hwang, "Language detoxification with attribute-discriminative latent space," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023. [Online]. Available: https://aclanthology.org/2023.acl-long.565/.

[17] L. Laugier, J. Pavlopoulos, J. Sorensen, and L. Dixon, "Civil rephrases of toxic texts with self-supervised transformers," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds., Online: Association for Computational Linguistics, Apr. 2021, pp. 1442–1461. DOI: 10.18653/v1/2021.eacl-main.124. [Online]. Available: https://aclanthology.org/2021.eacl-main.124/.

[18] V. Logacheva, D. Dementieva, S. Ustyantsev, *et al.*, "Paradetox: Detoxification with parallel data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 6804–6818. [Online]. Available: https://aclanthology.org/2022.acl-long.469/.

[19] D. Moskovskiy, S. Pletenev, and A. Panchenko, "Llms to replace crowdsourcing for parallel data creation? the case of text detoxification," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 14 361–14 373. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.839/.

[20] S. Mukherjee, A. Bansal, A. Kr. Ojha, J. P. McCrae, and O. Dusek, "Text detoxification as style transfer in English and Hindi," in *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, J. D. Pawar and S. Lalitha Devi, Eds., Goa University, Goa, India: NLP Association of India (NLPAI), Dec. 2023, pp. 133–144. [Online]. Available: https://aclanthology.org/2023.icon-1.13/.

[21] S. Mukherjee, A. Bansal, P. Majumdar, A. K. Ojha, and O. Duek, "Low-resource text style transfer for Bangla: Data & models," in *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, Singapore: Association for Computational Linguistics, Dec. 2023. [Online]. Available: https://aclanthology.org/2023.banglalp-1.5/.

[22] C. Nogueira dos Santos, I. Melnyk, and I. Padhi, "Fighting offensive language on social media with unsupervised text style transfer," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018. [Online]. Available: https://aclanthology.org/P18-2031/.

[23] A. Pesaranghader, N. Verma, and M. Bharadwaj, "Gpt-detox: An in-context learning-based paraphraser for text detoxification," in *2023 International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2023, pp. 1528–1534. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10459961/.

[24] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[25] M. N. Raihan, U. H. Tanmoy, A. B. Islam, *et al.*, "Offensive language identification in transliterated and code-mixed bangla," *arXiv preprint arXiv:2311.15023*, 2023. [Online]. Available: https://arxiv.org/abs/2311.15023.

[26] M. M. O. Rashid, “Textual toxicity in social media: Understanding the bangla toxic language expressed in facebook comment,” *arXiv preprint arXiv:2312.05467*, 2023. [Online]. Available: https://arxiv.org/abs/2312.05467.

[27] N. Romim, M. Ahmed, M. S. Islam, A. Sen Sharma, H. Talukder, and M. R. Amin, “BD-SHS: A benchmark dataset for learning to detect online Bangla hate speech in different social contexts,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022. [Online]. Available: https://aclanthology.org/2022.lrec-1.552/.

[28] S. Sazzed, “A lexicon for profane and obscene text identification in Bengali,” in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, R. Mitkov and G. Angelova, Eds., Held Online: INCOMA Ltd., Sep. 2021. [Online]. Available: https://aclanthology.org/2021.ranlp-1.145/.

[29] M. Tran, Y. Zhang, and M. Soleymani, “Towards a friendly online community: An unsupervised style transfer framework for profanity redaction,” in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020. [Online]. Available: https://aclanthology.org/2020.coling-main.190/.

[30] J. Wei, X. Wang, D. Schuurmans, *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[31] J. Wen, P. Ke, H. Sun, *et al.*, “Unveiling the implicit toxicity in large language models,” *arXiv preprint arXiv:2311.17391*, 2023. [Online]. Available: https://arxiv.org/abs/2311.17391.

[32] M. Wiegand, J. Kampfmeier, E. Eder, and J. Ruppenhofer, “Euphemistic abuse – a new dataset and classification experiments for implicitly abusive language,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, Dec. 2023. [Online]. Available: https://aclanthology.org/2023.emnlp-main.1012/.