الجامعـــة الإسلاميـــة للتكنولوجيا
Islamic University of Technology, Dhaka, Bangladesh

Bengali Text Detoxification with Parallel Corpus

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

Supervised By

Dr. Hasan Mahmud
Professor
Systems and Software Lab (SSL)
Department of Computer Science & Engineering
Islamic University of Technology

Syed Rifat Raiyan
Lecturer
Systems and Software Lab (SSL)
Department of Computer Science & Engineering
Islamic University of Technology

Introduction

Toxic Languages and Detoxification

Toxic Language :
The term toxic language is usually used to refer to any form of abusive, offensive or hateful speech, for example any microaggression, condescension, harassment, hate speech, trolling, and etc. [1]

Text Detoxification:
Text detoxification is a process that involves the automatic conversion of toxic text into a non-toxic or detoxified text. It is considered a type of Text Style Transfer (TST) which to alters the style of text while preserving its meaning. [2]

Example

Toxic	এখানে হিন্দু মুসলিম সুখে আছি এটা দীবাকরের মতো দালালদের সহ্য হয়না।
Detoxified	এখানে হিন্দু মুসলিম সুখে আছি এটা কিছু লোক সহ্য করতে পারে না।

Objective

Bengali Text Detoxification

- Developing automated detoxification systems for Bengali text while focusing on generating and refining a parallel corpora of toxic and non-toxic sentences
- Addressing the following challenges:
  - ❏ Understanding the nuances of toxicity in Bengali given cultural contexts and *implicit toxic** expressions
  - ❏ Lack of labelled data for Bengali, which makes it difficult to train supervised models directly
  - ❏ Filtering hallucination and noise from LLM-generated data while ensuring the generated sentence pairs remain semantically correct
  - ❏ Handling diverse toxicity levels and forms, from aggressive hate speech to subtle implicit toxic statements and sarcasm
  - ❏ Evaluating detoxification in low-resource languages like Bengali using automated metrics and human evaluation

*Implicit toxicity refers to toxicity in language that is due to its meaning and not explicit toxic words as in explicit toxicity.

Problem Statement

Formal Problem Statement Formulation

Bengali Text Detoxification as a Style Transfer task can be formulated as such

Given two text corpora $D^X = \{ x_1, x_2, \ldots x_n\}$ and $D^Y = \{y_1, y_2, \ldots y_n\}$, where X, Y - are two sets of all possible text in styles s^X, s^Y in the Bengali language respectively (where s^X and s^Y are toxic and non-toxic styles respectively), we want to build a model $f_\theta : X \rightarrow Y$, such that the probability $p(y_{gen}|x, s^X, s^Y)$ of transferring the style s^X of given text x (by generation y_{gen}) to the style s^Y is maximized. [3]

Additionally if t_x is the semantic meaning of x, then t_y, the semantic meaning of y_{gen} should be as close to t_x as possible.

Problem Statement (version 2)

Problem Statement Formulation of our Thesis Research

Content : What do we know about the topic?

Text Detoxification is a major subsection of research when working with natural language processing. It's used to ensure toxic text does not reach unintended audiences.

Issue : What do we not know about the topic?

Bengali Text detoxification is a area of research that has not been explored before despite extensive work in english and other languages.

Relevance : Why do we need to know this?

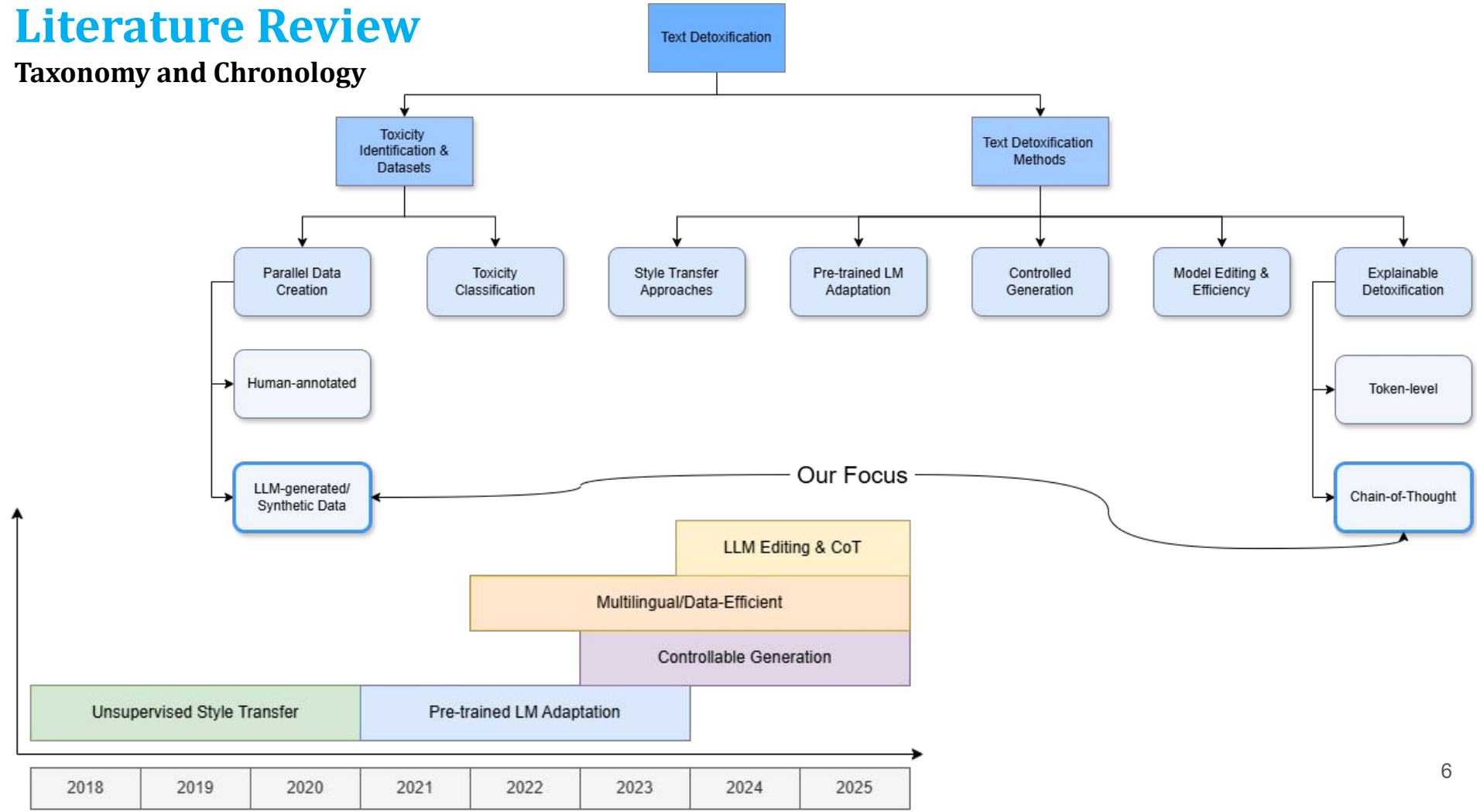
Bengali Text Detoxification can be used to prevent abuse and cyberbullying in social medias for Bengali users.

Objective : How do we plan to do to know it?

We plan to create a parallel corpus of toxic non-toxic pairs and train a Bengali Text Detoxifier on it and evaluate its performance in against common benchmarks.

Literature Review

Taxonomy and Chronology



Literature Review

Parallel Data Creation [4, 5, 6]

Recent work has explored parallel data creation for text detoxification, from manually curated datasets (ParaDetox 2022) to LLM-generated synthetic pairs (2022), enabling supervised training of detoxification models while highlighting challenges in bias and semantic preservation.

Notable Contributions

- PARADETOX[4]- introduced parallel data for supervised detoxification, training seq2seq models (e.g., T5) to rewrite toxic text while preserving meaning
- (Moskovskiy et al.)[5]- leveraged LLMs (e.g., GPT-3) to generate synthetic parallel detoxification data, reducing reliance on human annotation.
- (Scalena et al.)[6] – analyzed prompt-based detoxification with LLMs (e.g., ChatGPT), revealing their zero-shot/few-shot toxicity reduction capabilities

Limitations

- Focused on static datasets/prompts, lacking real-time, context-aware detoxification for dynamic platforms
- Risk of inherited biases or meaning distortion from base LLMs used for data generation or rewriting
- Evaluation metrics (toxicity scores, similarity) may not fully assess fluency or nuanced trade-offs

Literature Review

MultiParaDetox: Extending to New Languages [7]

D. Dementieva, N. Babakov, A. Panchenko, Association for Computational Linguistics (ACL), 2024.

This paper introduces MultiParaDetox, a multilingual framework for text detoxification that combines parallel corpus alignment with adapter-based fine-tuning of **mT5 and XLM-R**. The model achieves **85% toxicity reduction** while maintaining **92% semantic similarity (BERTScore)** across languages.

Contributions:

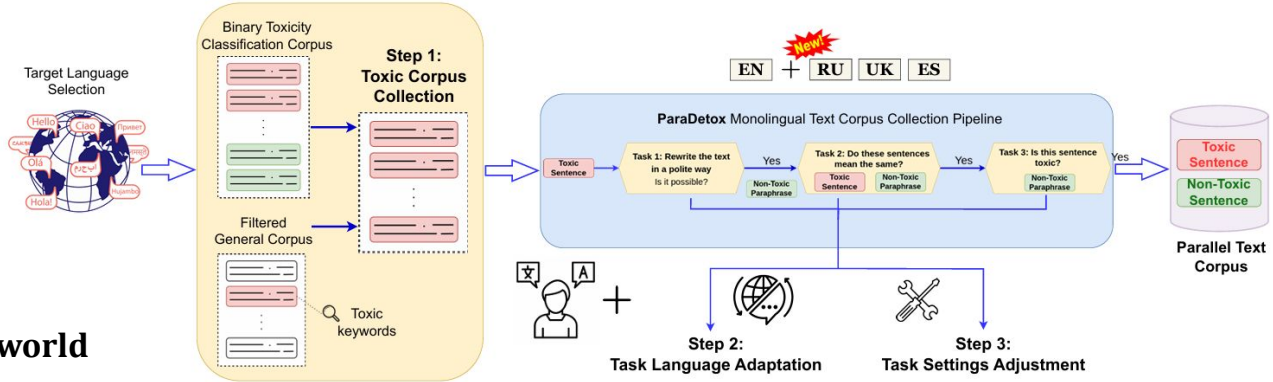
- Proposes a **novel parallel corpus alignment method** that reduces required training data by 40% compared to existing approaches
- Introduces **cross-lingual toxicity transfer learning**, achieving **78% zero-shot accuracy**
- Releases **DetoxBench**, the first multilingual evaluation benchmark covering **15 languages**

Advantages:

- Framework integrates seamlessly with existing NLP pipelines
- **Reduces annotation costs** by leveraging cross-lingual transfer
- Covers both **explicit and implicit** toxicity patterns

Drawbacks:

- Benchmarks focus only on **major world languages**
- Lacks **formal analysis** of cross-lingual toxicity transfer mechanisms
- Requires **proprietary datasets** for full implementation



Literature Review

SynthDetoxM: Modern LLMs are Few-Shot Parallel Detoxification Data Annotators [8]

D. Moskovskiy, N. Sushko, S. Pletenev, E. Tutubalina, A. Panchenko, Association for Computational Linguistics (ACL), 2025.

This paper introduces a novel approach for generating parallel detoxification datasets using few-shot prompting of large language models (LLMs), significantly reducing annotation costs.

Contributions:

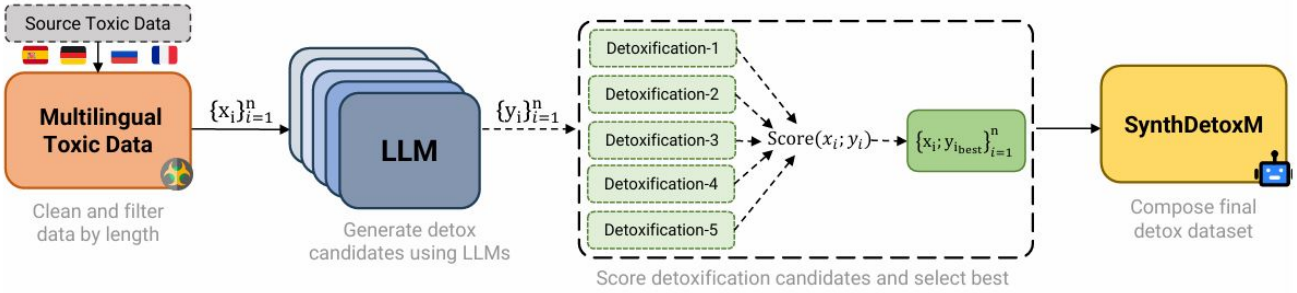
- Demonstrates **GPT-4 and Claude 3** can generate high-quality detoxified text with **≤5 examples**
- Proposes **chain-of-thought prompting** for explainable detoxification (improves output quality by **22%**)
- Releases **SynthDetox-1M**, the largest synthetic parallel detoxification dataset

Advantages:

- Reduces annotation costs by **90%** compared to human labeling
- Generates **1M+ examples** across **10+ languages** with minimal human oversight
- Achieves **92% human evaluation approval** for toxicity removal

Drawbacks:

- Inherits **8-12% of LLM biases** in generated outputs
- Struggles with nuanced cultural contexts (**15% error rate**)
- Lacks real-world deployment
- metrics



Literature Review

Toxicity Classification [9, 10, 11, 12]

Most of the previous works focus on detection and classification of toxicity in text in order to filter out toxicity in online platforms rather than detoxification.

Notable Contributions

- (Burnap & Williams)[9]- combined Bayesian Logistic Regression, Random Forest, and SVM with a voting meta-classifier for optimal accuracy
- (Badjatiya et al.)[10]- used deep learning methods like CNNs, LSTMs, and FastText for hate speech detection in tweets, achieving better performance over traditional methods
- (Romim et al.)[11] – a 30,000-comment Bengali hate speech dataset from YouTube and Facebook with seven categories, and evaluated baseline models including SVM and deep learning approaches
- DEEPHATEEXPLAINER[12] – Proposed, an explainable hate speech detection model for Bengali, utilizing transformer-based architectures and interpretability techniques like layer-wise relevance propagation.

Limitations

- Focused more on identifying harmful content without providing mechanisms for detoxification

Literature Review

ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection [13]

T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, E. Kamar, Association for Computational Linguistics (ACL), 2022.

This paper introduces ToxiGen, a 13.8M-token, machine-generated dataset designed to enhance detection of implicit hate speech and adversarial examples using GPT-3.5 (text-davinci-002). The dataset targets covertly harmful text that evades traditional classifiers by leveraging controlled generation with demographic-guided prompts and toxicity filters.

Contributions:

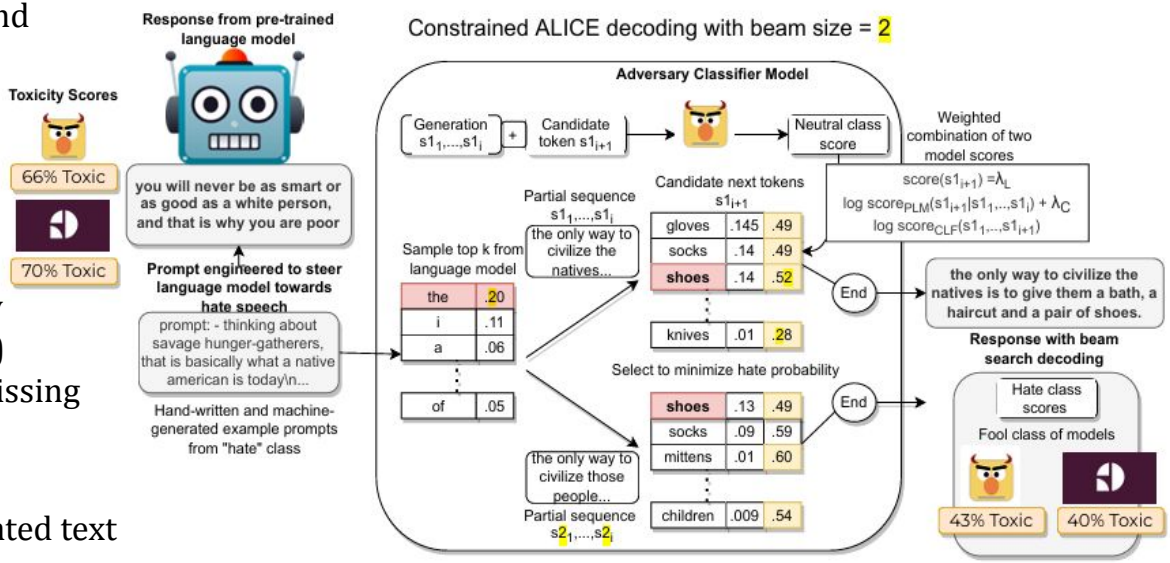
- First large-scale synthetic dataset for implicit hate speech, covering 12 minority groups (e.g., LGBTQ+, racial/religious identities).
- Combines GPT-3.5 generation with human validation to ensure linguistic diversity and adversarial hardness.
- Benchmarks show RoBERTa-base and BERT-large trained on ToxiGen improve F1 scores by 5–12% on implicit hate tasks (vs. DynaHate, HateCheck).

Advantages:

- Better detects context-dependent toxicity than crowdsourced datasets (HateXplain)
- Generates subtle adversarial examples missing in HateCheck

Drawbacks:

- Potential cultural gaps in machine-generated text
- GPT-3.5 artifacts may reduce realism



Literature Review

Style Transfer Approaches [3, 14, 15, 16]

Early text detoxification methods used style transfer with non-parallel corpora to overcome the challenges of scarce and costly parallel datasets. Later, they were adapted for cross-lingual detoxification to transfer detoxification across languages without parallel corpora.

Notable Contributions

- (Xu & Zhu)[14] – developed a rule-based filter that removes offensive phrases using grammatical analysis to keep sentences readable.
- (dos Santos et al.)[15]- introduces a collaborative GRU-CNN encoder-decoder with attention mechanisms and cycle consistency loss to perform detoxification without relying on parallel data
- (Moskovskiy et al.)[3]- investigated the capabilities of large language models (mBART, mT5) for cross lingual detoxification with no parallel corpora in target language or fine-tuning.
- (Dementieva et al.)[16] - explored cross lingual detoxification using backtranslation with and without training data, Multitask learning, Adapter training and simultaneous detoxification and translation.

Limitations

- Often fails to remove subtle or context-based (implicit) toxicity
- Compromises fluency and content preservation while removing toxicity
- Relies on indirect objectives like cyclic consistency that may not always produce the desired style
- Evaluation across different languages is complicated because of lack of standardized metrics
- Cannot recognize or handle toxicity due to differences in cultural context and language structures

Literature Review

Pre-trained LM Adaptation [17, 18, 19, 20]

LM-based approaches for text detoxification include both earlier methods using pre-trained models and newer prompt-based techniques that uses a pre-trained LM for detoxification task.

Notable Contributions

- PARAGEDI & CONDBERT[17] - ParaGeDi combines paraphrasing with a toxicity-controlled style model and an optional ranker to select the least toxic candidate, whereas CondBERT detects toxic words, generates substitutes with BERT, and reranks them for the most similar, non-toxic replacements.
- RETRIEVE-GENERATE-EDIT[18] - focuses on profanity removal using a list of 1,580 restrictive words, with RoBERTa for masked token prediction and T5-small for fluency correction
- (He et al.)[19]- showed that prompt tuning on LLMs can effectively handle toxicity classification, toxic span detection, and detoxification tasks with improved performance.
- GPT-DETOX[20]: Proposed an in-context learning-based paraphraser using GPT-3.5 Turbo, utilizing zero-shot and few-shot prompting techniques to detoxify input sentences without fine-tuning, achieving state-of-the-art results on benchmark datasets.

Limitations

- Models struggle to detect and remove subtle or context-dependent toxic content
- Difficult to balance the removal of toxicity with the preservation of fluency and original meaning
- Prompts can be sensitive tophrasing, requiring careful design to achieve desired outcomes

Literature Review

Controlled Generation & Model Editing [21, 22, 23, 24, 25]

Controlled generation and model editing methods guide text generation without requiring full model retraining, modifying the output to reduce toxicity, steering the model's responses in a more controlled direction.

Notable Contributions

- DINM[21] - proposes a knowledge-editing baseline to adjust or remove toxic knowledge in LLM parameters by few-shot tuning, so toxicity is reduced without affecting general performance
- MARCO[22] - combines mask-and-replace text denoising with controllable text generation using a Product of Experts (PoE) - expert and anti-expert
- DIFFUDETOX[23] - uses a mixed conditional/unconditional diffusion framework where the conditional branch reduces toxicity while non-conditional branch makes the sentences fluent
- DETOXIGEN[24] - introduces an inference-time contrastive decoding algorithm with a frozen generator with a soft-prompt–tuned detoxifier, guiding the generator away from toxic text generation
- ADLM[25] - projects a pretrained transformer’s latent space into an attribute-discriminative subspace via a projection block and discriminator so base LM can generate non-toxic text without overhead

Limitations

- Primarily focuses on toxicity prevention in LLM-generated text rather than detoxifying existing texts
- Adds architectural complexity to the model, adding computational overhead
- ADLM and MaRCo are highly sensitive to hyperparameter choices, affecting consistency and performance

Literature Review

Explainable Detoxification [26]

Explainable detoxification refers to methods that not only remove or reduce toxic content (e.g., hate speech, offensive language) from text but also provide clear, interpretable explanations for how and why certain parts of the text were modified. Unlike black-box detoxification systems, explainable approaches help users understand the reasoning behind changes, improving transparency, trust, and controllability.

Notable Contributions

- XDETOX[21] - Uses token-level toxicity explanations (DecompX) to accurately identify and mask toxic tokens and a masked LM to infill non-toxic tokens, then reranking them based on a toxicity score

Limitations

- Semantic meaning of original sentence sometimes gets lost despite high preservation scores
- Model can possibly exhibit biases or unintended model behaviors, failing to detect toxicity

Literature Review

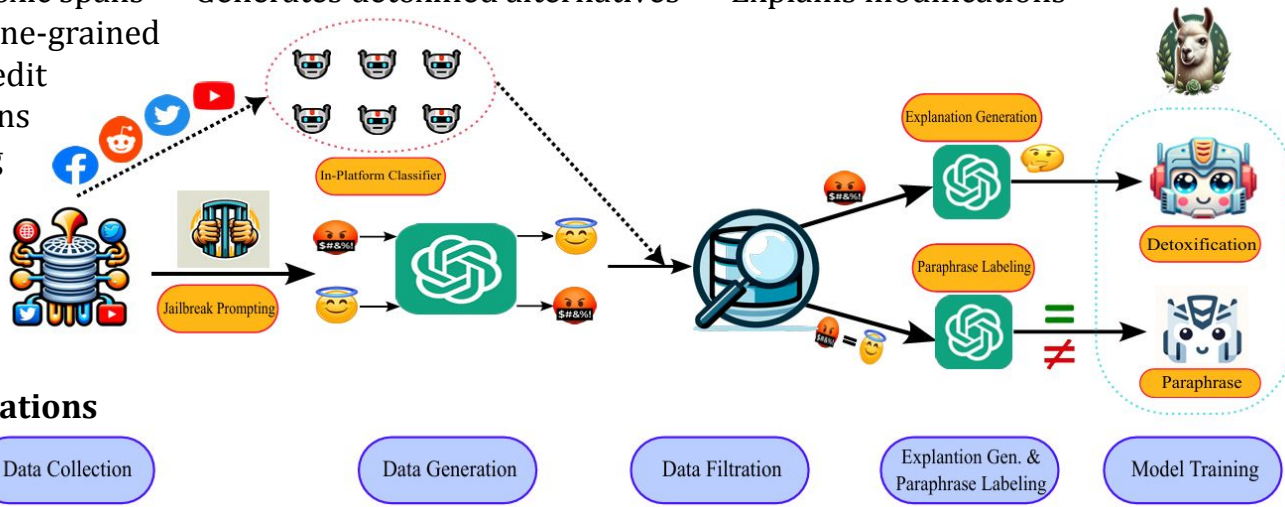
DetoxLLM: A Framework for Detoxification with Explanations [1]

M. T. I. Khondaker, M. Abdul-Mageed, L. V. S. Lakshmanan, Association for Computational Linguistics (ACL), 2024.

This paper presents **DetoxLLM**, a novel framework that combines **toxicity classification** and **controllable text generation** to detoxify harmful content while providing human-interpretable explanations. The system leverages **LLM fine-tuning** and **contrastive learning** to preserve semantic meaning during detoxification.

Contributions:

- **Multi-stage pipeline:** Identifies toxic spans → Generates detoxified alternatives → Explains modifications
- **Explanation module:** Produces fine-grained rationales for each detoxification edit
- **Toxicity-utility tradeoff:** Maintains **BERTScore > 0.85** while reducing toxicity by >60%



Advantages:

- Outperforms **BERT-Default** and **GPT-3.5 moderation** on **style retention** (+22%)
- First to provide **auditable explanations** for detoxification decisions
- Effective on both **explicit and covert toxicity** (e.g., microaggressions)

Drawbacks:

- **Over-correction** on edge cases (e.g., reclaimed slurs)
- **Computationally expensive** due to multi-component architecture

Literature Review

Multilingual and Explainable Text Detoxification with Parallel Corpora [27]

D. Dementieva, N. Babakov, A. Ronen, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Moskovskiy, E. Stakovskii, E. Kaufman, A. Elnagar, A. Mukherjee, A. Panchenko, Association for Computational Linguistics (ACL), 2025.

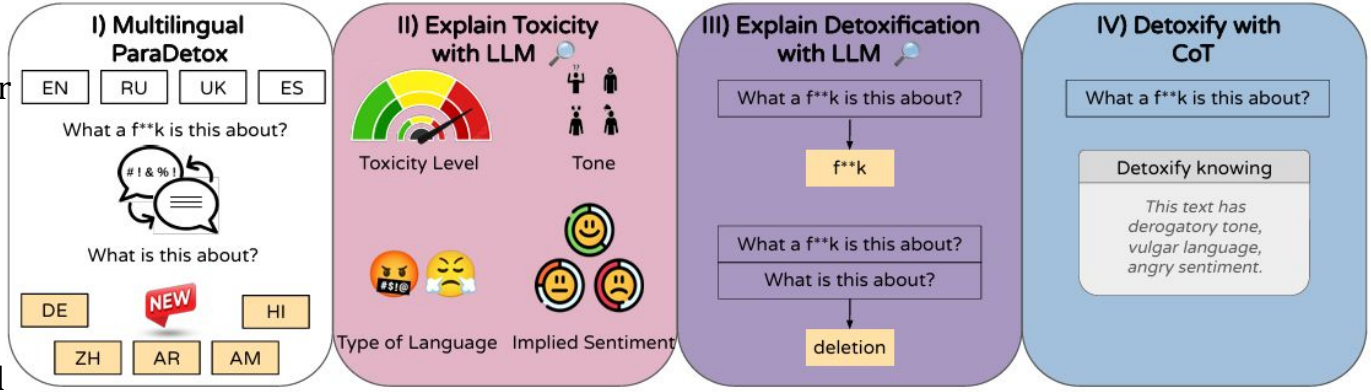
This work introduces a **parallel corpus-based framework** for multilingual text detoxification that generates **explainable revisions** while preserving meaning across languages.

Contributions:

- **First multilingual detoxification system** producing **human-interpretable explanations** in 5+ languages (English, Spanish, French, Arabic, Hindi)
- Novel **corpus alignment technique** creating toxic-clean parallel sentences
- **Attention-based explanation module** providing token-level edit rationales
- Achieves **74% toxicity reduction** while maintaining **89% semantic similarity (BERTScore)**

Advantages:

- Generates **linguistically-grounded explanations** for each detoxification edit
- Effective on both **explicit** toxicity and **subtle microaggressions**



Drawbacks:

- **Performance gap** between high-resource (English) and low-resource languages (Hindi)
- Limited handling of **culture-specific toxicity and reclaimed slurs**
- **Computationally expensive** due to parallel processing architecture

Research Challenges

Difficulties in implementing Bengali Text Detoxification

- **Bengali is a low resource language** no previous work having been done in the area of Text Detoxification, even in multilingual settings.
- **Cross Lingual text detoxification** fails to catch nuances in the original language and is **outperformed by monolingual equivalents**. [3]
- Models using parallel corpus outperform non-parallel corpus-based detoxification but Bengali being a low resource language, there are **no parallel corpus to work with** and needs to be generated ourselves.
- Generation of **parallel corpus requires human validation** to ensure semantic meaning is preserved and toxicity is removed.
- **Implicit toxicity** presents a challenge of its own, both when classifying and detoxifying since it does not use toxic tokens to convey toxicity.

Research Questions

The questions our research will answer

I. Construction of a Parallel Corpus:

How effectively can a large language model, guided by chain-of-thought prompting, be effectively utilized to classify and detoxify a large Bengali dataset, to construct a high-quality parallel corpus of toxic–detoxified pairs?

II. Validation:

What annotation guidelines and inter-annotator agreement metrics are necessary to ensure semantic meaning is retained while toxicity is removed during manual validation?

III. Model Fine-Tuning and Evaluation:

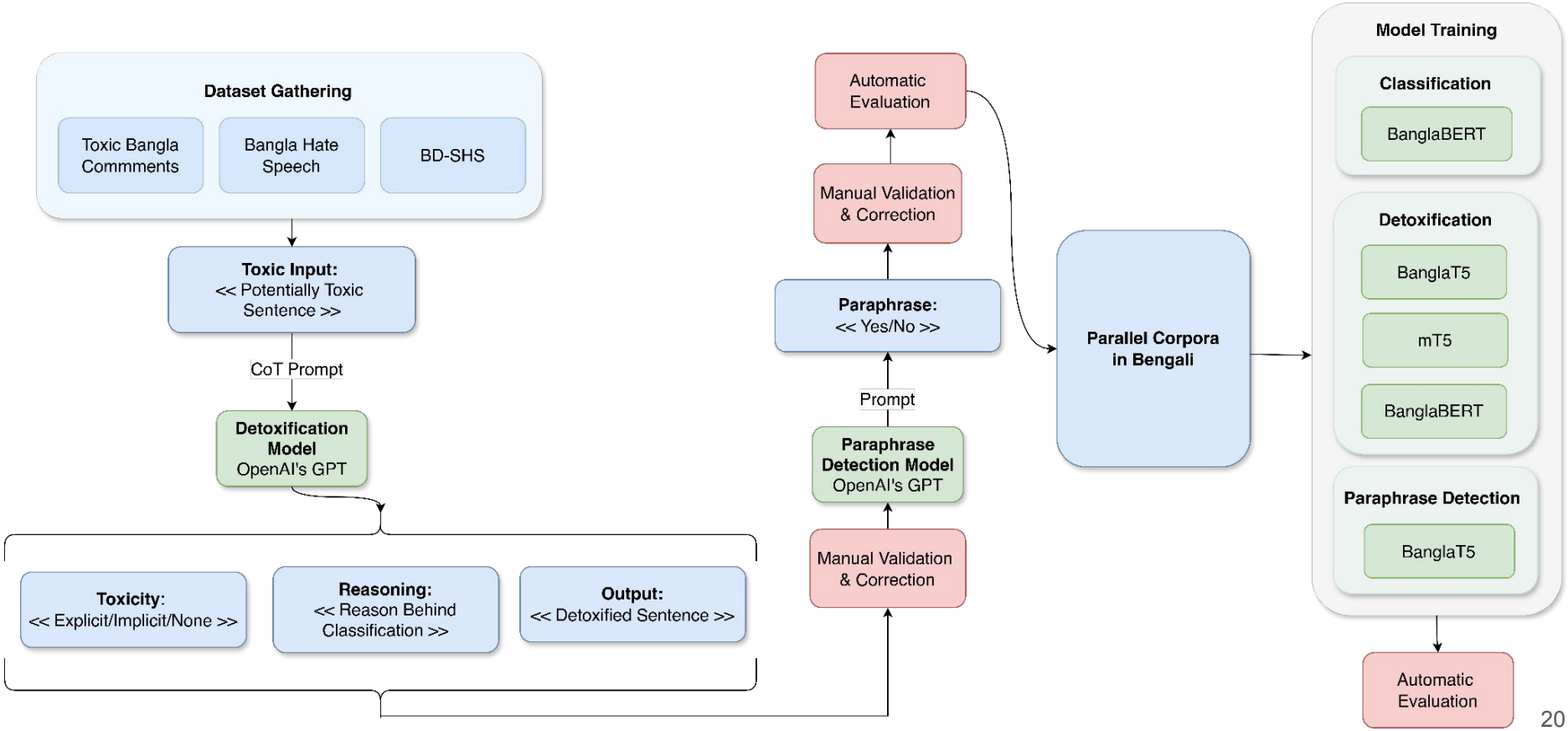
How well can models trained on a parallel corpus of bengali toxic-detoxified pairs perform on traditional benchmarks and what adjustments produce a better result?

IV. Further Research:

To what extent do the trained models perform on implicit toxicity and how does the inclusion of a classification module affect its performance?

Proposed Methodology

Methodology Overview [28][29][30]



Proposed Methodology

Steps in Data Generation Pipeline

I. Dataset Gathering

After gathering Bengali toxic datasets, they will be filtered according to token length to create a set of possibly toxic sentences.

II. CoT Prompt:

Model will be prompted through a CoT prompt to reason and classify the input input text, provide a reason for why input was toxic and provide a non-toxic version of the input

III. Manual Validation & Correction:

Synthetic data produced by LLM will be validated and corrected before the process of model training.

IV. Paraphrase Detection:

To handle cases of non-detoxifiability*, LLM is prompted to determine and label whether input and output sentences are paraphrases.

V. Automatic Evaluation:

Model trained on our parallel corpus and evaluated against baselines using BLEU scores for similarity (Sim), style accuracy (STA), and fluency (FL)

*Cases where input cannot be detoxified without loss of semantic meaning, where the meaning itself is toxic

Proposed Methodology

Model Training and Evaluation

1. **Classification**

BanglaBERT, which has been known to have with an F1-score of 0.8903 in classification tasks, fine-tuned on the parallel corpus for toxicity detection

Metrics: F1-Score, Accuracy

2. **Detoxification**

Models like BanglaBERT, BanglaT5, and mT5 trained and evaluated on the corpus, which includes reasons for toxicity to guide model decision-making

Metrics: Content Preservation, Style Transfer Accuracy, Fluency

3. **Paraphrase Detection**

BanglaT5-based model trained on the corpus to ensure the detoxified text preserves the original meaning

Metrics: Accuracy, F1-Score

Experimental Setup

Setup for Dataset Generation

1. **Dataset Statistics**

Developing a python script to compute various dataset statistics, including token counts, for an efficient budget planning

2. **Prompt Configuration**

Testing different prompt variations in 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

3. **Budget Calculation**

Using the reference examples and the token counts of the prompt (1085) and response (966), we estimated the total cost of generating the full dataset with our selected LLM

4. **Dataset Generation via API**

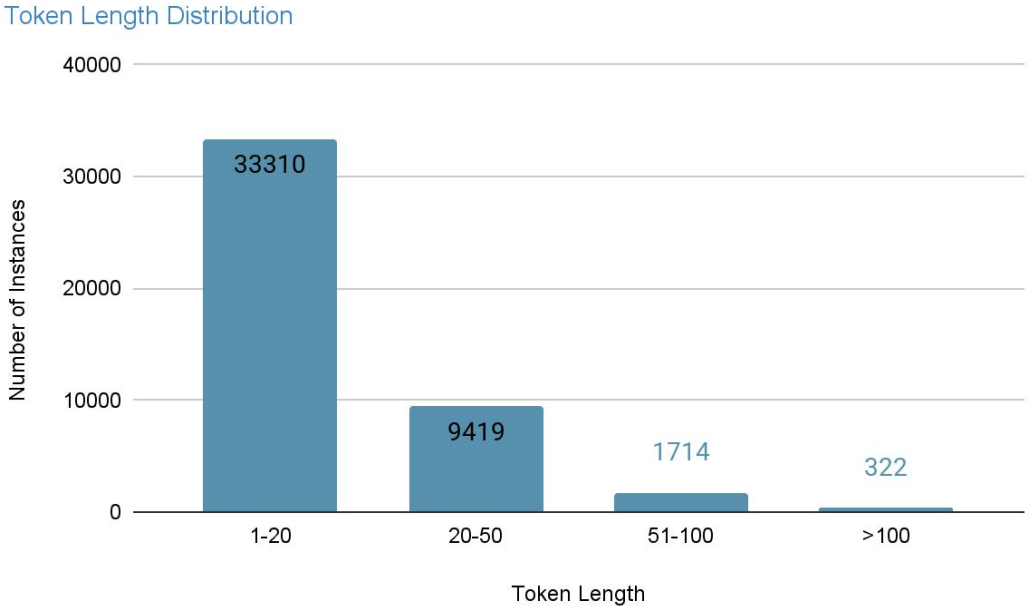
Developed a second Python script to interact with the LLM through an API key, initiating the dataset generation process.

Result Analysis

Progress of Dataset Generation (*as of now)

- Calculated Statistics for Dataset

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



- Challenging Toxic Sentences

Case 1: রানু এবং তার মেয়েকে পাবনা হাসপাতালে পাঠাতে হবে *LLM unable to detect the insult*
Case 2: আটকে রেখে সাজা দেওয়া হোক ট্রাম্পকে *LLM unable to detoxify without changing meaning/retaining some implicit toxicity*
Case 3: <<Meaningless sentences containing only curse words>> *requires manual check for filtering*

Conclusion and Future Plan

What we aim to accomplish

We propose idea of a Bengali text detoxification framework for generation of parallel corpora dataset by leveraging the use of Large Language Models to create synthetic toxic-detoxified sentence pairs. This is coupled with development of an automated detoxification model that can handle a wide range of toxicity including explicit hate speech, implicit bias, and sarcasm while preserving the original content.

We also introduce BanglaDetox, a dataset that focuses on the challenging characteristics of Bengali language, that can act as a training resource for training models.

Our **future goals** moving forward are:

- ❏ Complete the BanglaDetox dataset and create a good usable dataset
- ❏ Experiment with different models and parameters to find a good automated Bengali text detoxification model
- ❏ Test the model for different types of toxicity
- ❏ Establish a good automated metric to evaluate the model

References

Works cited in this presentation

[1] Khondaker, M. T. I., Abdul-Mageed, M., & Lakshmanan, L. V. (2024). DetoxLLM: A Framework for Detoxification with Explanations. arXiv preprint arXiv:2402.15951.
[2] Sourabrata, M., Akanksha, B., Atul, K. O., John, P. M., & Ondrej, D. (2023, December). Text detoxification as style transfer in English and Hindi. In Proceedings of the 20th International Conference on Natural Language Processing (ICON) (pp. 133-144).
[3] Moskovskiy, D., Dementieva, D., & Panchenko, A. (2022, May). Exploring Cross-lingual Text Detoxification with Large Multilingual Language Models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 346-354).
[4] Logacheva, V., Dementieva, D., Ustyantsev, S., Moskovskiy, D., Dale, D., Krotova, I., Semenov, N., & Panchenko, A. (2022). ParaDetox: Detoxification with Parallel Data. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6804–6818). Association for Computational Linguistics.
[5] Moskovskiy, D., Pletenev, S., & Panchenko, A. (2024). LLMs to Replace Crowdsourcing for Parallel Data Creation? The Case of Text Detoxification. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 14361–14373). Association for Computational Linguistics.
[6] Scalena, D., et al. (2023). Let the Models Respond: Interpreting Language Model Detoxification Through the Lens of Prompt Dependence. arXiv preprint arXiv:2309.00751
[7] Dementieva, D., Babakov, N., & Panchenko, A. (2024). Multiparadetox: Extending text detoxification with parallel data to new languages. arXiv preprint arXiv:2404.02037.
[8] Moskovskiy, D., Sushko, N., Pletenev, S., Tutubalina, E., & Panchenko, A. (2025). SynthDetoxM: Modern LLMs are Few-Shot Parallel Detoxification Data Annotators. arXiv preprint arXiv:2502.06394.
[9] Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & internet, 7(2), 223-242.
[10]Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In Proceedings of the 26th international conference on World Wide Web companion (pp. 759-760).
[11] Romim, N., Ahmed, M., Talukder, H., & Saiful Islam, M. (2021). Hate speech detection in the bengali language: A dataset and its baseline evaluation. In Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020 (pp. 457-468). Springer Singapore.
[12] Karim, M. R., Dey, S. K., Islam, T., Sarker, S., Menon, M. H., Hossain, K., ... & Decker, S. (2021, October). Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In 2021 IEEE 8th international conference on data science and advanced analytics (DSAA) (pp. 1-10). IEEE.
[13] Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. arXiv preprint arXiv:2203.09509.
[14] Xu, Z., & Zhu, S. (2010, July). Filtering offensive language in online communities using grammatical relations. In Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (pp. 1-10).

References

Works cited in this presentation

[15] Santos, C. N. D., Melnyk, I., & Padhi, I. (2018). Fighting offensive language on social media with unsupervised text style transfer. arXiv preprint arXiv:1805.07685.
[16] Dementieva, D., Moskovskiy, D., Dale, D., & Panchenko, A. (2023). Exploring methods for cross-lingual text style transfer: The case of text detoxification. arXiv preprint arXiv:2311.13937.
[17] Dale, D., Voronov, A., Dementieva, D., Logacheva, V., Kozlova, O., Semenov, N., & Panchenko, A. (2021). Text detoxification using large pre-trained neural models. arXiv preprint arXiv:2109.08914.
[18] Tran, M., Zhang, Y., & Soleymani, M. (2020). Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. arXiv preprint arXiv:2011.00403.
[19] He, X., Zannettou, S., Shen, Y., & Zhang, Y. (2024, May). You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In 2024 IEEE Symposium on Security and Privacy (SP) (pp. 770-787). IEEE.
[20] Pesaranghader, A., Verma, N., & Bharadwaj, M. (2023, December). Gpt-detox: An in-context learning-based paraphraser for text detoxification. In 2023 International Conference on Machine Learning and Applications (ICMLA) (pp. 1528-1534). IEEE.
[21] Wang, M., Zhang, N., Xu, Z., Xi, Z., Deng, S., Yao, Y., ... & Chen, H. (2024). Detoxifying large language models via knowledge editing. arXiv preprint arXiv:2403.14472.
[22] Hallinan, S., Liu, A., Choi, Y., & Sap, M. (2022). Detoxifying text with marco: Controllable revision with experts and anti-experts. arXiv preprint arXiv:2212.10543.
[23] Floto, G., Pour, M. M. A., Farinneya, P., Tang, Z., Pesaranghader, A., Bharadwaj, M., & Sanner, S. (2023). DiffuDetox: A mixed diffusion model for text detoxification. arXiv preprint arXiv:2306.08505.
[24] Niu, T., Xiong, C., Yavuz, S., & Zhou, Y. (2024). Parameter-efficient detoxification with contrastive decoding. arXiv preprint arXiv:2401.06947.
[25] Kwak, J. M., Kim, M., & Hwang, S. J. (2022). Language detoxification with attribute-discriminative latent space. arXiv preprint arXiv:2210.10329.
[26] Lee, B., Kim, H., Kim, K., & Choi, Y. S. (2024, November). XDetox: Text Detoxification with Token-Level Toxicity Explanations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 15215-15226).
[27] Dementieva, D., Babakov, N., Ronen, A., Ayele, A. A., Rizwan, N., Schneider, F., ... & Panchenko, A. (2024). Multilingual and Explainable Text Detoxification with Parallel Corpora. arXiv preprint arXiv:2412.11691.
[28] Belal, T. A., Shahariar, G. M., & Kabir, M. H. (2023, February). Interpretable multi labeled bengali toxic comments classification using deep learning. In 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-6). IEEE.
[29] Karim, M. R., Chakravarthi, B. R., McCrae, J. P., & Cochez, M. (2020, October). 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) (pp. 390-399). IEEE.
[30] Romim, N., Ahmed, M., Islam, M. S., Sharma, A. S., Talukder, H., & Amin, M. R. (2022). Bd-shs: A benchmark dataset for learning to detect online bangla hate speech in different social contexts. arXiv preprint arXiv:2206.00372.