Toxifix: Automating Emotionally Intelligent Comment Moderation with Reinforcement Learning from Human Feedback

Jahnavi Chintakindi
Deptartment of Data Science
University of North Texas
Denton, United States
jahnavichintakindi@my.unt.edu

Poojitha Ganta
Deptartment of Data Science
University of North Texas
Denton, United States
poojithaganta@my.unt.edu

Ramya Rangaraju
Deptartment of Data Science
University of North Texas
Denton, United States
ramyarangaraju@my.unt.edu

Stephen F. Wheeler
Deptartment of Data Science
University of North Texas
Denton, United States
stephen.wheeler@unt.edu

Ravi Varma Kumar Bevara
Deptartment of Data Science
University of North Texas
Denton, United States
ravivarmakumarbevara@my.unt.edu

Krishna Annavaram
Deptartment of Data Science
University of North Texas
Denton, United States
krishnaannavaram@my.unt.edu

***Abstract*—Digital platforms have become an important medium for communication. Although they offer clear advantages in the broader dissemination of content, the past decade has witnessed an alarming rise in toxic interactions, including cyberbully and harassment. The detection of online toxicity presents challenges because of its complex and context-dependent characteristics. Given the severe social repercussions of exposure to online toxicity, reliable models and algorithms are essential for the detection and analysis of such interactions within the expansive and continuously evolving realm of social networks.**

Using psychological and social theories, this research has developed a comprehensive definition of toxicity. From this foundation, it strives to create a sophisticated classifier capable of discerning between toxic and nontoxic remarks, ultimately boosting the accuracy of evaluating toxic behavior.Comparative analysis was conducted using BERT, DistilBERT, and RoBERTa. Based on the percentage of toxic content detected, DistilBERT performs better than the other two methods. Furthermore, according to evaluation metrics, multilingual model training outcomes, and confusion matrices, XLM-RoBERTa demonstrates better result validity compared to the other approaches.

***Index Terms*—Toxic content, deep learning, natural language processing, transformer models, prompt RLHF, BERT, RoBERTa.**

I. Introduction

Communication is one of the basic necessities of everyone's life, and responsibility depends on the management and control of social media. Due to the development of communication technology Via social media platforms, individuals representing a multitude of cultures, age brackets, communities, genders, and global regions engage with one another in a digital realm. Online social networks are incontestably some of the most culturally impaction technological breakthroughs we have seen in the 21st century.

Social media's origins date back to 1844 with the telegraph's electronic dots, even preceding the Internet. Early digital social media, like Bulletin Board Systems (BBS), revolutionized user interactions. Key milestones followed: Usenet in 1979, LinkedIn and Facebook in 2003, YouTube in 2005, Twitter in 2006, WhatsApp in 2009, Instagram in 2010, and Telegram in 2013. These platforms break geographical borders, allowing global content sharing and interaction, fundamentally altering communication with the rise of the Internet and the spread of social networks. Agushaka et al., 2023; Bonetti and Martínez-Sober, 2023

The probability of online bullying and suffering has divergent opinions, which inhibits the outflow of thoughts. Locales struggle to effectively progress discussions, which forces many networks to restrict or eliminate user comments. Digital technology revolutionizes our way of connecting, but it also spreads abuse like wildfire, from hate speech to harassment. With the flood of content on the Internet, only powerful computational tools can tackle such abuse effectively. Yet, moderating online toxicity presents technical, social, legal, and ethical hurdles.

Lately, hate speech and offensive content have surged online, targeting individuals or groups. Toxic comments, rude and provoking, drive users away from discussions. As information and communication technology skyrockets, most people freely share opinions online, fueling the spread of hate speech across platforms. A large proportion of online comments in the public domain are constructive; however, a significant proportion are toxic in nature. The current situation is very productive and will improve the quality of human life, but could also be dangerous and destructive Abbasi and Javed, 2022.

Researchers are tackling this issue with machine learning algorithms to identify toxic content. Digital information was meant to enhance decision-making and cultural exchange. However, in today's fast-paced digital world, misinformation endangers global political, social, and economic stability. AI

and machine learning are key in combating false news, forming the backbone of fact-checking systems to ensure information credibility.Ünver, 2023.

To curtail info misuse, collaboration between governments, organizations, and tech firms is vital, pooling resources and insights. It's crucial to have regulations ensuring ethical tech use, guarding against misuse while upholding free speech. Recently, there's been a surge in using both classical ML and cutting-edge DL techniques to spot toxic online chatter. Popular ML classifiers in NLP include logistic regression, random forest, support vector machines, and the top-tier transformer architecture for detecting toxic social media contentBonetti and Martínez-Sober, 2023; Giridhar and Singh, 2023; Maity and More, 2024; Shukla and Arora, 2023.

ML models are turbocharged with LSI for dimensional shrinkage and LDA for pinpointing topics, auto-classifying tweets by theme. AI and machine learning supercharge digital forensics, pushing tool capabilities skyward, yet we must prepare for methods that can juggle even wilder data sizes and attribute swingsAgushaka et al., 2023; Khan et al., 2024. The forensic acquisition of data from digital devices, particularly social media content, has received increasing interest in academic literature due to the vast nature of the platforms and content available.

AI's reinforcement learning with human feedback merges human insight with machine learning prowess. It involves training an AI agent to make proper decisions after receiving feedback. Several authors have explored the application of reinforcement learning from human feedback techniques for this purpose Kim et al., 2023; Lin and Chien, 2024.

II. Statement of the Problem

Deeming language offensive is all about perspective, shifting with the user, place, culture, and history. Crafting a completely automated moderation system? That means diving deep into these social media waters. Toxic comments run rampant, threatening free speech and user mental health. Driven by the surge in online toxicity and antisocial behavior, our research tackles this urgent issue in a realm almost everyone frequents now. Such negative behavior erodes democracy globally and fuels societal divides.

Research on toxic comment classification has become particularly active since 2018. Detrimental content can adversely affect an individual's mental health, leading to harassment, psychological disorders, depression, and abuse, outcomes that can have irreversible consequences. Although social networks offer a wide range of opportunities for developing real-world applications that benefit society and humanity, they also have a darker side. This includes objectionable and harmful content such as hate speech, rumors, fake news, toxic comments, aggression, and cyberbully.

RLHF taps into human insight to directly steer and refine AI through reinforcement learning. This approach boosts AI's adaptability and personalization by infusing it with human expertise. From revolutionizing sectors like education and healthcare to transforming game AI and NLP, RLHF holds game-changing potential. It steers AI towards ethical focus and societal relevance but comes with hurdles like scaling issues, feedback biases, and ethical dilemmas.

III. Review of Literature

Social networks allow everyone to voice their thoughts and critiques for the world to see. Influence can spark campaigns to promote new ideas. Nowadays, almost anyone can post comments, love notes, or vile messages, fueling division and tension. Spotting toxic remarks is vital to curtail their negative effects. Therefore, identifying harmful comments online is crucial to foster healthier and more inclusive dialogue.

Various approaches for toxic comment detection have been proposed, often focusing on classification, feature dimension reduction, and feature importance evaluation. Poojitha et al. Poojitha and Charish, 2023 deployed machine learning models to filter offensive language and safeguard users against harassment. Their classifier distinguished toxic and non-toxic comments, providing clarity in evaluating online discourse. Zaheri et al. Zaheri et al., 2023 leveraged NLP methods to sort text data into toxic and safe labels.

Abbasi et al. Abbasi and Javed, 2022 explored modern deep learning algorithms for multi-label toxic language detection, analyzing classification accuracy across various architectures. Ahmad Khan et al. Khan et al., 2024 investigated ensemble and data augmentation methods to address imbalanced datasets using SMOTE and random oversampling. Onan et al. Onan, 2023 unveiled an innovative GTR-GA system, fusing graph neural networks with genetic algorithms to produce top-notch augmented data for sentiment analysis and text classification.

Glazkova et al. Glazkova and Morozov, 2023 Leveraged transformer-based models for extracting keyphrases and creating concise summaries from scholarly texts. Benchmark analysis proved supervised models outperformed unsupervised ones.. Kim et al. Kim et al., 2023 developed ToxiGen-ConPrompt, a pre-trained model enhanced leveraging AI-crafted data enhances the ability to spot implicit hate speech.

Zhang et al. Zhang et al., 2024 addressed the class imbalance issue in abusive language detection, introducing external data augmentation methods using abusive lexicons. Patel et al. Patel and Pramanik, 2025 Put several machine learning models to the test with the Jigsaw dataset. Naïve Bayes led the pack in accuracy, with XGBoost just a step behind.

Lin et al. Lin and Chien, 2024Utilized cutting-edge AI for spotting toxic remarks and harmful news, achieving 94% accuracy in toxicity and 81% in content classification. Prabha et al. Prabha and Yadav, 2025 utilized SVM and word embedding models for toxic comment classification, with SVM outperforming other techniques.

Neoga et al. Neoga and Baruah, 2024 proposed a hybrid architecture combining BiLSTM and CNN, showing notable performance gains for toxicity detection. Dutta et al. Dutta and Neoga, 2024 focused on regional language (Assamese) using SVM, demonstrating superior accuracy and F1-score. Shahid et al. Shahid and Umair, 2024 used transformer-based models

for toxicity detection in Urdu, leveraging deep learning for robust binary classification.

Jesica et al. Jessica and Sugiarto, 2024 implemented BERT-CNN and BERT-LSTM hybrids to detect harmful content. Their models effectively combined contextual embeddings with sequential modeling capabilities. Shukla et al. Shukla and Arora, 2023 developed systems to classify comments involving racial, religious, or caste-based abuse using both classical and deep learning approaches.

Maity et al. Maity and More, 2024 employed BiLSTM for multilabel classification of harmful comments, achieving 95% accuracy with optimized architectures. Giridhar et al. Giridhar and Singh, 2023 designed LSTM and LSTM-CNN hybrids to flag toxic posts on social platforms. Asif et al. **b25** introduced a deep learning framework for detecting trolls and text-embedded toxic images, contributing to more effective content moderation.

Finally, Mezghani et al. Mezghani and Elleuch, 2024 developed a multilingual toxic comment detection system for Tunisian dialect using SVM, BLSTM, and hybrid models, achieving high classification accuracy and demonstrating the importance of regional language support in toxicity detection systems.

IV. Objectives and Existing Methods of the Study

The exponential rise of toxic interactions on social media has highlighted the need for intelligent, scalable, and emotionally sensitive content moderation tools. While many existing solutions rely on rule-based filtering or simple classification, they often lack the capability to transform toxic content into constructive dialogue. This research proposes a multi-stage pipeline that incorporates Real-time toxic comment detection and rewriting using cutting-edge NLP models and RL trained on human feedback. The study ultimately aims to advance online discourse moderation by introducing human-aligned NLP methods that foster safer digital spaces.

Key Objectives of the Study

- To paraphrase identified toxic comments using instruction-based, few-shot prompting tailored to emotional intelligence.
- To investigate the effectiveness of transformer-based classifiers in identifying online toxicity.
- Evaluate datasets detecting harmful content.
- To develop a reward-based evaluation system that considers empathy, bias, hallucination, and semantic preservation.
- To apply RLHF techniques to refine paraphrased outputs using human-aligned feedback.
- To deploy a live demonstration of the system on Hugging Face for real-world testing.
- To develop a cutting-edge toxic comment detection system for social media.
- Compare RLHF with cutting-edge NLP to spot harmful social media content.

A. Existing Methods

The challenge of moderating toxic content stems from the vast volume, velocity, cultural diversity, and contextual variability of online material. Manual removal is not feasible due to the multilingual nature of content, the labor intensity, and the associated costs and training required. Even after detecting detrimental content, current manual moderation techniques are inadequate and unsustainable.

Semi-automated moderation approaches have shown limited success, emphasizing the need for fully automated detection and moderation methods. With the increasing prevalence of toxic comments, there is a growing demand for accurate and scalable solutions. These harmful interactions seriously endanger free speech and users' mental health. Harnessing comparative analysis insights can help craft better strategies to combat toxicity and promote inclusive digital spaces.

This research uses cutting-edge techniques like deep learning, RLHF, and NLP to sort toxic from harmless comments. A major focus is on pinpointing highly sensitive algorithms for spotting toxicity. RLHF offers an accelerated and targeted learning approach by integrating human expertise, which enhances decision-making, customer interaction, and operational efficiency. By aligning with human knowledge, RLHF reduces the need for trial-and-error methods, enabling safer and more cost-effective systems.

However, incorporating human feedback introduces challenges, including scalability, bias, and ethical considerations. RLHF enables systems to be trained from human demonstrations, where the agent interacts with its environment and receives evaluative feedback based on its actions. This interaction makes RLHF a promising paradigm for building AI systems that align with human values and societal norms.

Unlike traditional reinforcement learning, which relies solely on environmental rewards and penalties, RLHF incorporates human feedback to resolve ambiguities in reward functions. RL, the foundation of RLHF, involves training agents to optimize decision-making policies through repeated interactions. Policy optimization techniques allow iterative refinement of agent behavior to maximize long-term rewards.

Empowering RLHF models to generalize across domains and transfer knowledge between systems holds potential for transforming human-guided AI. Despite its advantages, RLHF must be deployed responsibly, considering the technical, ethical, and societal complexities it introduces.

B. Conceptual Frameworks

AI tools are crushing it, quickly and accurately spotting harmful content on social media. But we've got to dive deeper to craft models that grasp language subtleties in varied settings.

Building trust and fairness in AI decision-making processes is critical to real-time content moderation. Designing ethical, robust, and context-aware systems that reflect societal values will be essential for future applications.

V. Research Design and Methodology

This study creates a smart system that analyzes user posts to spot harmful patterns.. By setting a threshold on objectionable posts, the system aims to ensure a safer and healthier environment on social media platforms. The broader objective is to empower social media systems with real-time, automated toxicity detection that not only enhances individual well-being but also promotes a constructive digital space.

Toxicity detection can be approached as a dual-problem task: (1) identifying the source(s) of toxic content, and (2) recognizing the potentially vulnerable victim(s). Addressing these challenges requires advanced Natural Language Processing (NLP) techniques and deep learning models capable of capturing nuanced semantic and contextual indicators of toxicity.

Importantly, toxicity detection transcends traditional computer science or AI boundaries. Effective moderation must consider the social, cultural, and ethical dimensions underlying online communication. This necessitates a multidisciplinary approach grounded in domain-specific content analysis, human values, social norms, and group dynamics. Consequently, toxicity detection is an interdisciplinary problem informed by theoretical frameworks, empirical models, and value-aligned classifications.

A. Data Collection, Pre-Processing, and Analysis

This project employed a hybrid data acquisition strategy, integrating both benchmark datasets and real-world user-generated content. Public datasets provided pre-labeled toxic and non-toxic comment pairs, while manually scraped YouTube data introduced noisy, user-contextual inputs to simulate real-world deployment conditions.

Datasets Used:

- **Jigsaw Multilingual Dataset** (Kaggle): https://www.kaggle.com/datasets/julian3833/jigsaw-toxic-comment-classification-challenge
- **YouTube Toxic Comments Dataset** (Kaggle): https://www.kaggle.com/datasets/reihanenamdari/youtube-toxicity-data
- **Most Liked YouTube Comments Dataset** (Kaggle): https://www.kaggle.com/datasets/nipunarora8/most-liked-comments-on-youtube
- **BTS YouTube Comments Dataset** (Kaggle): https://www.kaggle.com/datasets/seungguini/bts-youtube-comments
- **Random Multi YouTube Comments (Scraped)**: https://www.kaggle.com/datasets/janvi59/random-multi-youtube-comments-scrapped
- **Human Evaluation Reward Scores Dataset (RLHF)**: Manually curated dataset for reward-based tuning

Preprocessing Steps: The following preprocessing techniques were applied across the combined datasets:

- Removal of duplicate entries and non-English comments
- Normalization of text and cleanup of punctuation
- Manual annotation of toxicity levels for scraped, unlabeled data

A total of approximately **19,300** user comments were used throughout the pipeline for training, validation, and testing phases of model development and evaluation.

B. Exploratory Data Analysis (EDA) and Hypotheses for the Study

We performed Exploratory Data Analysis (EDA) to spot toxicity trends, inspect class imbalances, and pinpoint linguistic cues that set toxic comments apart from non-toxic ones.

Findings:

- Approximately 30.33% of the dataset was classified as toxic using the BERT model.
- Toxic comments tended to be shorter in length and exhibited higher sentiment intensity.
- Word cloud analysis highlighted a frequent presence of aggressive and hate-driven vocabulary in toxic comments.

Research Hypotheses:

- Transformer-based models (e.g., BERT, XLM-RoBERTa) will outperform traditional models in detecting online toxicity.
- Instruction-tuned paraphrasing with few-shot prompting will enable effective tone transformation.
- Reinforcement Learning from Human Feedback (RLHF) will significantly improve emotional intelligence and fairness in paraphrased outputs.

C. Core Research Concepts

Fine-Tuning Transformer Models: Fine-tuning involves adapting pre-trained transformer models to specific downstream tasks using labeled data. In this study, BERT, RoBERTa, and XLM-RoBERTa were fine-tuned for binary toxicity classification. Their contextual embeddings enabled the models to distinguish nuanced differences between offensive and non-offensive language.

Prompt Engineering for Paraphrasing: Prompt engineering refers to the design of input queries that guide large language models toward desired behavior. Instruction-based prompts, combined with few-shot examples, were used to paraphrase toxic comments into more neutral forms while preserving semantic meaning. Iterative refinement of prompts was employed to minimize hallucinations and maintain semantic fidelity.

Reinforcement Learning from Human Feedback (RLHF): RLHF introduces human judgment into the training loop by using reward signals based on human evaluations. In this work, the paraphrasing model (Granite 3.2–2B) was optimized using a composite reward function that evaluated empathy, toxicity reduction, bias mitigation, and semantic similarity.

Bias in Language Models: Bias in NLP systems refers to consistent patterns that reflect or reinforce harmful stereotypes or societal imbalances. This study incorporated fairness classifiers in Stage 3 to assess bias levels in both original and paraphrased outputs.

Hallucination in Generated Outputs: Hallucination occurs when a language model generates text that is fluent but factually incorrect or irrelevant. Natural Language Inference (NLI) models such as RoBERTa were used to detect hallucinations and ensure the paraphrased outputs did not misrepresent the original content.

Pre-Trained Transformer Models: Pre-trained transformers, rooted in the transformer architecture and honed using self-supervision on vast datasets, excel in NLP tasks like classification, sentiment analysis, and paraphrasing post fine-tuning.

- **BERT (English)**: A transformer-based model pre-trained on English corpora; fine-tuned for classifying toxic vs. non-toxic content.
- **RoBERTa (English)**: A variant of BERT with optimized pre-training over larger data and longer sequences; captures nuanced toxic expressions more effectively.
- **DistilBERT (English)**: A lightweight version of BERT retaining 97% of its language understanding, optimized for faster real-time detection.
- **XLM-RoBERTa (Multilingual)**: A cross-lingual model supporting over 100 languages, well-suited for multilingual toxicity classification.
- **mBERT (Multilingual)**: A multilingual version of BERT trained on 104 languages; capable of detecting toxicity across varied linguistic inputs.
- **mDistilBERT (Multilingual)**: A distilled variant of mBERT offering reduced latency with reasonable classification accuracy.

Toxicity Reduction: A metric quantifying how effectively a model transforms offensive or harmful content into more socially acceptable language, typically measured by a reduction in toxicity scores.

Semantic Similarity: The degree to which the paraphrased sentence retains the original meaning, commonly assessed using cosine similarity of embedding vectors (e.g., similarity > 0.85).

Empathy Improvement: The enhancement of emotional tone in the paraphrased text, measured using empathy classifiers to assess the model's emotional intelligence.

LLM Evaluation: A comparative evaluation of large language models (LLMs) such as GPT-NeoX, Claude 2, and Granite to assess their performance in moderating tone and preserving original content meaning.

D. Data Analytics and Implementation Pipeline

The system was implemented as a four-stage NLP pipeline, as illustrated in Fig. **??**. The architecture includes components for classification, paraphrasing, evaluation, and RLHF refinement, hosted in a scalable cloud environment.

1) Stage 1: Toxic Comment Classification: **Goal:** Identify whether online comments are toxic or non-toxic.

Datasets: All datasets mentioned in Section 4.1 were used for model training and evaluation.

Models Trained:

- **English Models:** BERT, RoBERTa, DistilBERT

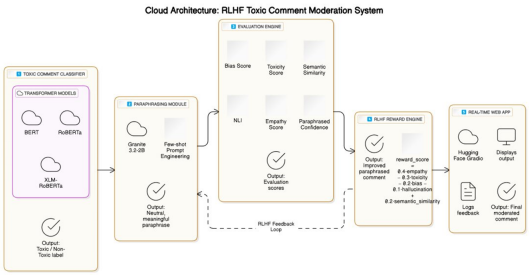

Fig. 1. Cloud Architecture

- **Multilingual Models:** XLM-RoBERTa, mBERT, mDistilBERT

Performance Metrics of Multilingual Models:

TABLE I
MULTILINGUAL MODELS TRAINING RESULTS

Model	Accuracy	F1	Precision	Recall	AUC
XLM-RoBERTa	0.9017	0.9023	0.9025	0.9021	0.9638
mBERT	0.9050	0.9043	0.9167	0.9021	0.9643
DistilBERT (Multi)	0.9132	0.9139	0.9135	0.9143	0.9553

TABLE II
CONFUSION MATRIX OF MULTILINGUAL MODELS

Model	TP	TN	FP	FN
XLM-RoBERTa	2082	2056	225	226
mBERT	2059	2094	187	249
DistilBERT (Multi)	2064	2035	246	244

Performance Metrics of English Models:

TABLE III
ENGLISH MODELS TRAINING RESULTS

Model	Accuracy	F1	Precision	Recall	AUC
BERT	0.9253	0.9279	0.9012	0.9562	0.9810
RoBERTa	0.9335	0.9340	0.9330	0.9350	0.9785
DistilBERT	0.9296	0.9311	0.9175	0.9450	0.9791

Insights: BERT demonstrated high sensitivity, making it effective for flagging edge cases. DistilBERT was more conservative and efficient for real-time processing. Among multilingual models, XLM-RoBERTa achieved balanced performance across languages.

2) Stage 2: Paraphrasing Toxic Comments: **Goal:** Rewrite flagged toxic content into emotionally neutral, non-toxic alternatives.

Method: Prompt engineering using few-shot examples. The implementation of paraphrasing of toxic comments is shown in Fig. 2.

Models Evaluated: GPT-J, GPT-NeoX, Claude 2, Flan-T5 XXL, Mixtral, Falcon, LLaMA 2, BLOOM, GPT-3.5 Turbo, Granite 3.2-2B

4.4.2.1 Prompt Engineering Strategy: To guide large language models in transforming toxic comments into emotionally neutral and constructive alternatives, prompt engineering was employed with few-shot examples.

TABLE IV
CONFUSION MATRIX OF ENGLISH MODELS

Model	TP	FP	FN	TN
BERT	2039	242	101	2207
RoBERTa	2126	155	150	2158
DistilBERT	2085	196	127	2181

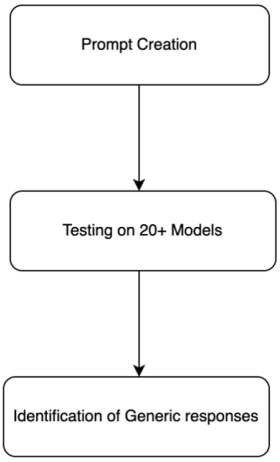

Fig. 2. Stage 2 implementation steps: Paraphrasing toxic comments using LLMs.

TABLE V
OBSERVATIONS OF LLM MODEL CHOSEN

Metric	Value	Description
BERT Toxicity Detection	30.33%	Initial toxicity score of the flagged comment as detected by a BERT-based classifier.
Toxicity Reduction	$\sim$0.45	Average reduction in toxicity score after paraphrasing the comment using each evaluated model.
Semantic Similarity	> 0.85	Ensures the paraphrased response retains the original meaning, measured via cosine similarity.
Empathy Improvement	Detected via classifier	The revised comment shows improved emotional sensitivity.
Final Model	Granite 3.2-2B	Best balance of reducing toxicity, preserving semantics, and improving tone.

TABLE VI
PROMPT EXAMPLE AND MODEL JUSTIFICATION

Component	Content
Prompt Type	Few-shot prompt with toxic-to-neutral transformations
Toxic Input	"This is the dumbest thing I've ever seen."
Neutral Output	"I don't think this idea works well."
Purpose	Showcases how the model rewrites aggressive language into respectful feedback while preserving meaning.
Outcome	Granite 3.2-2B was chosen for its effectiveness in: - Reducing inflammatory language - Preserving intent - Generating emotionally intelligent responses

The prompt was designed to reflect real-world moderation goals by incorporating clear guidelines:

- Remove explicit hate speech, personal attacks, or offensive language.
- Keep the response neutral and professional.
- Ensure the rewritten comment retains the original intent but in a constructive tone.
- Match the length and brevity of the original toxic comment whenever possible.

Few-shot Examples:

- Toxic: "You're so dumb! You never understand anything!"
- Neutral: "You might be misunderstanding this."

Outcome: Granite 3.2-2B was selected due to its consistent performance across tone control, semantic preservation, and brevity.

3) Stage 3: Evaluation Metrics: **Goal:** This stage focuses on quantitatively and qualitatively evaluating the paraphrased outputs to ensure that toxic content is not only neutralized but also retains its original intent and tone in a more constructive manner. A comprehensive evaluation pipeline was implemented using a set of seven metrics covering toxicity, bias, semantic integrity, and emotional intelligence.

Evaluation Workflow: We began by analyzing the original comments for toxicity, bias, and classification confidence. Toxic comments were then paraphrased using the best-performing model (Granite 3.2-2B). The new outputs were evaluated to confirm they were non-toxic, unbiased, semantically consistent with the original, and conveyed a more empathetic tone. This guaranteed that transformations were not only linguistic but also ethical and emotionally savvy. See Table for evaluation metrics. VII.

TABLE VII
LLM EVALUATION METRICS

Metric	Description
Prediction & Confidence	Determines if the original comment is toxic and shows model confidence.
Toxicity Score	Measures the toxicity level of the original comment (scale: 0 to 1).
Bias Score	Detects any bias or discriminatory language in the original comment.
Paraphrased Output & Prediction	Displays the rephrased version and re-evaluates its toxicity.
Paraphrased Toxicity & Bias Scores	Confirms the new version is free from toxicity and bias.
Semantic Similarity	Assesses how well the paraphrased comment retains the original meaning.
Empathy Score	Measures emotional tone, ensuring the new comment is empathetic and constructive.

4) Stage 4: RLHF: Refining with Human Feedback and Reinforcement Learning: **Goal:** To iteratively enhance the quality of paraphrased toxic comments by aligning model outputs more closely with human expectations for empathy, fairness, and semantic integrity using a reward-based optimization approach.

Refinement Process Overview: To simulate RLHF in a controlled, reproducible setting, we used proxy-labeled reward

scores rather than live human feedback for every iteration. This proxy-labeled feedback was constructed based on a custom-designed reward function combining five core evaluation metrics.

Reward Function:

$$\begin{aligned}\text{reward\_score} = {} & 0.4 \times \text{empathy} - 0.3 \times \text{toxicity} \\ & - 0.2 \times \text{bias} - 0.1 \times \text{hallucination} \\ & + 0.2 \times \text{semantic\_similarity}\end{aligned} \tag{1}$$

Each model output was scored using this formula, and higher scores reflected better alignment with emotionally intelligent and socially appropriate behavior. The reward signal served as feedback to fine-tune the model iteratively, simulating the reinforcement learning process with human preferences embedded indirectly via proxy metrics.

Steps Implemented:

- Defined a custom reward function combining empathy, toxicity, bias, hallucination, and semantic similarity.
- Scored each paraphrased comment using the function after Stage 3 evaluation.
- Tracked reward scores across multiple generations and models.
- Identified top-performing outputs and used them to fine-tune the paraphrasing behavior.
- Measured progress over iterations using cumulative reward, correlation with human scores, and convergence rate.

The observations from this refinement process are shown in Table VIII.

TABLE VIII
RLHF METRICS OBSERVATION

Metric Tracked	Observation
Cumulative Reward Improvement	Reward scores increased with each refinement cycle, indicating successful tuning.
Reward-Human Correlation (Pearson)	Moderate-to-strong correlation observed, validating the reward function design.
Convergence over Iterations	Reward scores began to stabilize after a few iterations, showing learning saturation.

5) Stage 5: Deployment Phase: Following evaluation, *ToxiFix* was deployed as a live, real-time moderation demo.

Environment:

- **Frontend:** Built using Gradio interface on Hugging Face Spaces for user input and display.
- **Backend:** Python pipeline with Granite 3.2-2B, Streamlit UI, and integrated evaluation module.
- **Deployment Method:** All components uploaded to a Hugging Face public repository for hosting.

Features:

- Enables real-time toxic comment moderation and paraphrasing.
- Provides instant evaluation scores for toxicity, bias, empathy, and similarity.
- Displays side-by-side comparison of original and paraphrased comments.
- Architecture Flow: User input → Toxicity Classification → Paraphrasing → Evaluation → Output (rewritten + scores).

Deployment KPIs:

TABLE IX
DEPLOYMENT SUMMARY

Metric	Observation
Response Time	3–5 seconds end-to-end latency.
Detox Success Rate	90% toxic inputs successfully neutralized.
Semantic Fidelity	Paraphrased output maintains >0.85 similarity.
Empathy Improvement	Verified using an empathy classifier.
User Feedback	Positive during live demo sessions.

TABLE X
DEPLOYMENT TECH STACK

Component	Tools/Frameworks Used
Language	Python 3.8+
Model Deployment	Hugging Face Spaces + Gradio
Toxicity Classifiers	BERT, RoBERTa, XLM-RoBERTa
Paraphrasing	Granite 3.2-2B with prompt engineering
Evaluation Metrics	Toxicity, Sentence-BERT, Emotion-BERT, RoBERTa NLI
Reward Modeling (RLHF)	Custom scoring with human feedback
Visualization & Logging	CSV logging, web interface display

Challenges:

- Faced Hugging Face memory and runtime constraints due to model size.
- No API integration with external platforms like Coursera, Outlook, or Gmail (marked as future scope).

VI. DATA VISUALIZATION AND MODEL TESTING RESULTS SUMMARY

We evaluated how well the model detects toxic comments using key metrics: F1 Score, Recall, Accuracy, Precision, and AUC-ROC. Precision shows how many flagged toxic remarks were correct, while recall indicates the model's capacity to spot all toxic comments.

A. Dataset Details

A real-time, web-scraped YouTube comment dataset containing approximately 20,000 cleaned comments was used to evaluate performance across English and multilingual transformer models. The results of the English model toxicification test and the results of the multilingual model toxicification test are shown in Tables XI and XII, respectively.

TABLE XI
ENGLISH MODELS TOXICITY CLASSIFICATION TEST RESULTS

Model	Total Comments	Toxic Comments	Percentage Toxic
BERT	19,301	5,854	30.33%
DistilBERT	19,301	3,194	16.55%
RoBERTa	19,301	4,832	25.04%

TABLE XII
MULTILINGUAL MODELS TOXICITY CLASSIFICATION TEST RESULTS

Model	Total Comments	Toxic Comments	Percentage Toxic
XLM-RoBERTa	19,296	3,756	19.47%
mDistilBERT	19,296	5,001	25.92%
mBERT	19,296	4,697	24.35%

VII. RESULTS, DISCUSSIONS AND CONCLUSIONS

BERT had the highest toxicity detection rate (30.33%), indicating greater sensitivity but possibly more false positives.

DistilBERT showed the lowest rate (16.55%), suggesting a more conservative and precise classification approach.

RoBERTa balanced sensitivity and specificity with a moderate detection rate (25.04%).

Toxicity detection rates of English models are displayed in Fig. 3.

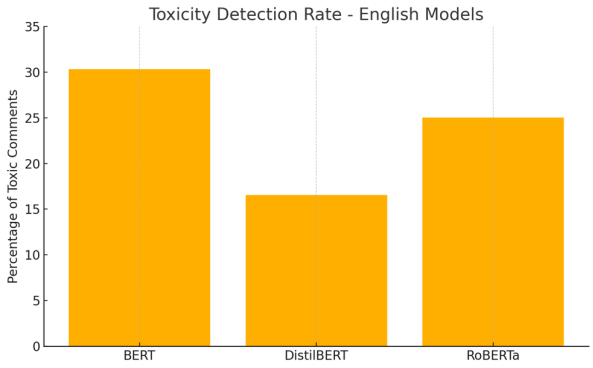

Fig. 3. English Models Toxicity Detection Rate

See Fig. 3, BERT had the highest detection rate (30%) and DistilBERT the lowest.

Multilingual models showed consistent performance across languages, with toxicity rates between 19.47% and 25.92%. Multilingual Models Toxicity Detection Rate is shown in Fig. 4.

As per Fig. 4, moderate detection is observed across XLM-RoBERTa, mDistilBERT, and mBERT. A small difference in total comment count (5 comments) across models is due to non-English or malformed entries being skipped during multilingual processing.

The dataset maintained structural integrity, aligning with the 19,300 entries in the cleaned input dataset.

A. Application Screenshots

Toxic Comment Classifier System in Hugging Face, Toxic Prediction Application Outputs, and Non-Toxic Prediction Application Outputs are shown in Figs. 5, 6, and 7, respectively.

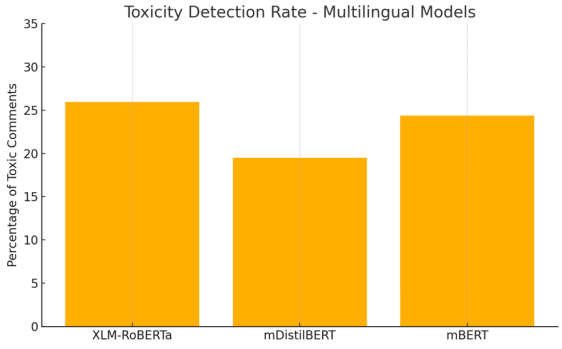

Fig. 4. Multilingual Models Toxicity Detection Rate

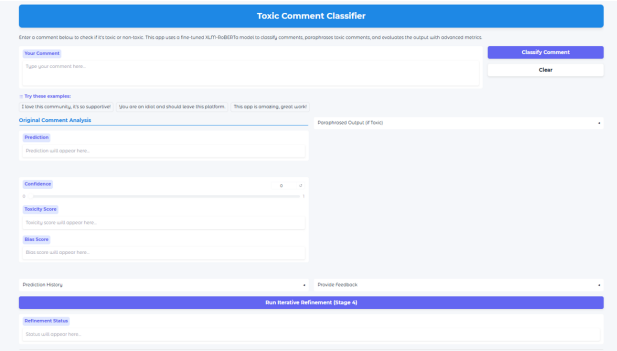

Fig. 5. Toxic Comment Classifier System in Hugging Face

VIII. CONCLUSION

This research dives into how different deep learning methods handle toxic comment classification. We dug deep into how both balanced and unbalanced datasets affect model results. We compared the suggested methods, focusing on computational complexity. A key highlight is our framework's skill in extracting and classifying text from images tied to online messages.

AI and machine learning are revolutionizing digital forensics, enhancing tool capabilities. Yet, scalability is still key. With improvements in computing, data handling, and cutting-

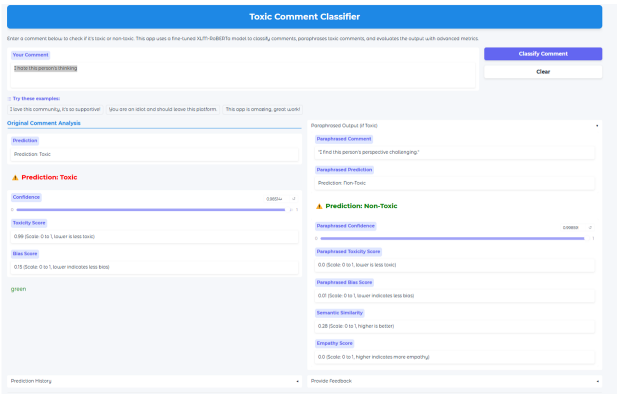

Fig. 6. Toxic Prediction Application Outputs

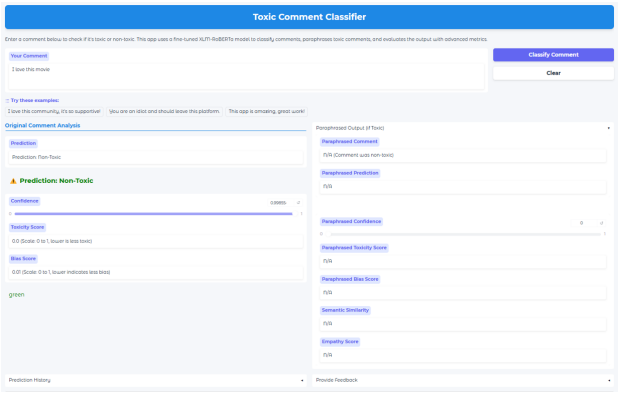

Fig. 7. Non-Toxic Prediction Application Outputs

edge algorithms, automating the detection of toxic comments is now more achievable.

ToxiFix demonstrated the viability of building a human-aligned, emotionally intelligent moderation system. By combining transformer models, prompt-tuned generative AI, and Reinforcement Learning from Human Feedback (RLHF), the system effectively detected and rephrased toxic comments in real-time. Deployment on Hugging Face validated both technical feasibility and ethical alignment, enabling live paraphrasing with evaluation transparency.

Compared to existing systems such as DistilRoBERTa classifiers, traditional ML models, and expensive explainability pipelines, ToxiFix provided a superior solution—real-time, multilingual, and semantically consistent moderation. Its unique ability to transform toxic content via reward-optimized paraphrasing positions it as a leading-edge framework.

IX. Limitations and Future Scope

A. Limitations

While ToxiFix offers a solid framework, some limitations were observed:

- The system primarily focused on English data despite including multilingual models, limiting cultural and linguistic inclusivity.
- Due to resource constraints, only Granite 3.2-2B was used for generation, restricting experimentation with more powerful generative models.
- The RLHF loop relied on simulated human feedback rather than real-time, limiting precision in alignment.
- Toxicity, bias, and empathy assessments used third-party classifiers, which may introduce inherited biases.
- The system has not yet been tested in production environments like Coursera, Gmail, or YouTube.

B. Future Work

Future research will aim to:

- Integrate real-time human feedback to enhance the RLHF loop and improve personalization.
- Deploy as a Chrome extension or API integration for platforms like Gmail and YouTube to enable broader adoption.
- Extend multilingual support to promote inclusivity and culturally nuanced moderation.
- Perform fairness audits and evaluate performance under real-world content loads.
- Expand the system into a fully automated content moderation suite with end-to-end monitoring and emotional intelligence.

RLHF Applications and Future Vision: RLHF has the potential to revolutionize AI by aligning systems with human expectations across fields like content moderation, recommender systems, robotics, and self-driving cars. However, RLHF also faces challenges:

- Human feedback is often noisy, subjective, and hard to generalize.
- Collecting and processing feedback is complex and resource-intensive.
- Scaling RLHF effectively while maintaining ethical responsibility is non-trivial.

Nevertheless, RLHF's ability to connect AI with human values holds transformative promise. Progress in generalization, ethical design, and feedback integration will help RLHF drive breakthroughs in business, healthcare, finance, and more. The path forward must prioritize trust, transparency, and innovation—turning AI into a socially responsible partner for solving complex human challenges.

X. Individual Contributions

Jahnavi Chintakindi: Led Stage 1 (Toxicity Classification) and contributed to all stages, focusing on model fine-tuning, prompt design, and evaluation metric integration.
Designed the end-to-end NLP pipeline, developed the reward function for RLHF, and conducted semantic, bias, and empathy-based analysis.

Poojitha Ganta: Led Stages 2 and 3, handling paraphrasing model evaluation and multi-metric performance tracking including hallucination, empathy, and similarity.
Benchmarked LLMs using few-shot prompts and constructed the evaluation framework that aligned model outputs with human preference scores.

Ramya Rangaraju: Led Stages 4 and 5, documenting the RLHF tuning process and planning real-time deployment on Hugging Face with web-based moderation flow.
Contributed to the literature review, architecture explanation, and research communication for RLHF application and multilingual deployment.

References

Abbasi, A., & Javed, A. R. (2022). Deep learning for religious and continent-based toxic content detection and classification. *Scientific Reports*. https://doi.org/10.1038/s41598-022-22523-3

Agushaka, J. O., Ezugwu, A. E., & Abualigah, L. (2023). Gazelle optimization algorithm: A novel nature-inspired metaheuristic optimizer. *Neural Computing and Applications*, *35*(5), 4099–4131. https://doi.org/10.1007/s00521-022-07854-6

Bonetti, A., & Martínez-Sober, M. (2023). Comparison between machine learning and deep learning approaches for the detection of toxic comments on social networks. *Applied Sciences*, *13*(10), 6038. https://doi.org/10.3390/app13106038

Dutta, S., & Neoga, M. (2024). Assamese toxic comment detection on social media using machine learning methods. *IEEE Xplore*. https://doi.org/10.1109/icETITE58242.2024.10493331

Giridhar, P., & Singh, S. H. (2023). Exploring the efficacy of deep learning models for multiclass toxic comment classification in social media using natural language processing. *IEEE Xplore*. https://doi.org/10.1109/ACCAI58221.2023.10199737

Glazkova, V., & Morozov, D. A. (2023). Applying transformer-based text summarization for keyphrase generation. *Lobachevskii Journal of Mathematics*, *44*(1), 123–136. https://doi.org/10.1134/S1995080223010134

Jessica, A., & Sugiarto, M. S. (2024). A hybrid deep learning techniques using bert and cnn for toxic comments classification. *IEEE Xplore*. https://doi.org/10.1109/ICIMTech63123.2024.10780934

Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, *244*, 122778. https://doi.org/10.1016/j.eswa.2023.122778

Kim, Y., Park, S., Namgoong, Y., & Han, Y.-S. (2023). Conprompt: Pre-training a language model with machine-generated data for implicit hate speech detection. https://github.com/youngwook06/ConPrompt

Lin, S.-Y., & Chien, S.-Y. (2024). Combating online malicious behavior: Integrating machine learning and deep learning methods for harmful news and toxic comments. *Information Systems Frontiers*.

Maity, A., & More, R. (2024). Toxic comment detection using bidirectional sequence classifiers. *IEEE Xplore*. https://doi.org/10.1109/IDCIoT59759.2024.10467922

Mezghani, A., & Elleuch, M. (2024). Toward arabic social networks unmasking toxicity using machine learning and deep learning models. *International Journal of Intelligent Systems Technologies and Applications*, *22*(3), 260–280. https://doi.org/10.1504/IJISTA.2024.140948

Neoga, M., & Baruah, N. (2024). A hybrid deep learning approach assamese toxic comment detection in social media. *Procedia Computer Science*, *235*, 2297–2306. https://doi.org/10.1016/j.procs.2024.04.218

Onan, A. (2023). Gtr-ga: Harnessing the power of graph-based neural networks and genetic algorithms for text augmentation. *Expert Systems with Applications*, *232*, 120908. https://doi.org/10.1016/j.eswa.2023.120908

Patel, D., & Pramanik, P. K. D. (2025). Detecting toxic comments on social media: An extensive evaluation of machine learning techniques. *Journal of Computational Science*, *8*, 20.

Poojitha, K., & Charish, A. S. (2023). Classification of social media toxic comments using machine learning models.

Prabha, S., & Yadav, A. (2025). Toxic comments classification using machine learning and word embedding techniques. *2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*. https://doi.org/10.1109/CICTN64563.2025.10932491

Shahid, M., & Umair, M. (2024). Leveraging deep learning for toxic comment detection in cursive languages. *PeerJ Computer Science*, *10*, e2486. https://doi.org/10.7717/peerj-cs.2486

Shukla, A., & Arora, D. (2023). Deep learning model for identification and classification of web based toxic comments. *IEEE Xplore*. https://doi.org/10.1109/APSIT58554.2023.10201794

Ünver, H. A. (2023). Emerging technologies and automated fact-checking: Tools, techniques and algorithms. https://doi.org/10.13140/RG.2.2.20514.20165

Zaheri, S., Leath, J., & Stroud, D. (2023). Toxic comment classification. *SMU Data Science Review*, *3*(1). https://scholar.smu.edu/datasciencereview/vol3/iss1/13

Zhang, Y., Hangya, V., & Fraser, A. (2024). A study of the class imbalance problem in abusive language detection. *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH)*.