CAFE: A Context-Aware and Fairness-Weighted Framework for Toxicity Evaluation in Language Models

Janeesha Wickramasinghe
Department of Computer Science and Engineering
University of Moratuwa
Colombo, Sri Lanka
janeesha.21@cse.mrt.ac.lk

***Abstract*—Large language models (LLMs) exhibit remarkable fluency but remain prone to generating toxic, biased, or harmful continuations when prompted. The RealToxicityPrompts (RTP) benchmark, paired with Google's Perspective API, has become a de facto standard for evaluating such degeneration, yet its reliance on a single black-box classifier introduces critical flaws: lack of contextual sensitivity (e.g., sarcasm mislabeling), fairness disparities across demographic subgroups, and outdated coverage of contemporary discourse. We propose CAFE (Context-Aware Fairness-Weighted Toxicity Evaluator), a novel framework that augments RTP with paraphrased and adversarial prompts, embeds prompt–continuation pairs using a fine-tuned RoBERTa model, and optimizes a multi-objective loss balancing toxicity accuracy, fairness, and context-awareness. Unlike Perspective, which treats toxicity in isolation, CAFE explicitly incorporates subgroup fairness penalties and context-sensitive embedding constraints. Experimental results demonstrate that CAFE achieves higher F1 scores, a substantial reduction in fairness gaps, and improved robustness on continuation degeneration, measured by Expected Maximum Toxicity (EMT) and toxicity probability metrics. Evaluation on the Jigsaw dataset further confirms CAFE's generalization beyond RTP. By addressing the intertwined challenges of bias, context insensitivity, and dataset staleness, CAFE offers a scalable and equitable evaluator for responsible AI deployment.**

***Index Terms*—contextual embeddings, fairness-aware learning, language models, RealToxicityPrompts, responsible AI, toxicity evaluation**

I. Introduction

The rapid progress of large language models (LLMs) has transformed natural language processing, enabling breakthroughs in tasks such as translation, summarization, and dialogue generation. However, alongside these advances, LLMs frequently exhibit toxic degeneration—generating harmful, biased, or offensive continuations even from innocuous prompts. Such behaviors pose significant risks to the safe deployment of AI systems in social, educational, and professional contexts.

The RealToxicityPrompts (RTP) dataset [1] has emerged as a widely adopted benchmark for measuring this phenomenon. RTP evaluates continuations of 100,000 naturally occurring prompts using the Perspective API, a toxicity classifier developed by Google Jigsaw. While RTP standardizes evaluation across models, its dependence on Perspective introduces critical limitations. Perspective functions as a single black-box oracle, often misclassifying context-dependent expressions (e.g., sarcasm, slang, or reclaimed terms), over-flagging minority dialects such as African American Vernacular English (AAVE), and reflecting the biases of its training distribution. Moreover, RTP itself was collected in 2020, raising concerns about dataset staleness and under-representation of modern linguistic phenomena.

These challenges highlight the urgent need for more context-sensitive and fairness-aware evaluators. Existing approaches to toxicity detection—ranging from rule-based filters to detoxification via decoding strategies (e.g., GeDi, DExperts)—address generation but do not solve the fundamental issue of evaluation quality. Without reliable evaluators, efforts to detoxify or benchmark LLMs risk being misleading or inequitable. To address this gap, we propose CAFE (Context-Aware Fairness-Weighted Toxicity Evaluator), a novel framework designed to overcome the weaknesses of the Perspective baseline. CAFE combines contextual embeddings from a fine-tuned RoBERTa model, data augmentation through paraphrased and adversarial prompts, and a multi-objective loss that jointly optimizes toxicity accuracy, fairness, and context-awareness. Unlike existing evaluators, CAFE explicitly penalizes subgroup disparities and leverages embedding-level similarity to handle non-literal language such as sarcasm.

Our contributions are threefold:

- We augment the RTP dataset with paraphrased and adversarial prompts to improve robustness and diversity.
- We design a multi-objective training scheme that balances toxicity prediction with fairness and context sensitivity.
- We empirically validate CAFE against the Perspective API baseline using RTP continuations and external benchmarks such as the Jigsaw unintended bias dataset [2], demonstrating measurable improvements in F1 score, fairness gap reduction, and robustness metrics (Expected Maximum Toxicity, Toxicity Probability).

An overview of the proposed CAFE framework is shown in Fig. 1, which illustrates how prompt–continuation pairs are embedded with RoBERTa and optimized via the multi-

objective loss for toxicity, fairness, and context sensitivity.

By integrating fairness and context into the evaluation process, CAFE advances the field of responsible AI and provides a scalable framework for assessing toxic degeneration in LLMs.

II. Related Work

A. Toxic Degeneration and RealToxicityPrompts

LLMs such as GPT-2, GPT-3, and recent transformer-based architectures have demonstrated remarkable generative fluency but are also prone to toxic degeneration—the tendency to produce offensive, biased, or harmful continuations from seemingly neutral prompts. Gehman et al. [1] first formalized this issue by introducing the RTP benchmark, a collection of 100 000 naturally occurring web prompts annotated with Perspective API toxicity scores. By prompting generative models with RTP sentences and scoring their continuations, the authors quantified how readily models produce toxic content, providing a standardized evaluation procedure for detoxification research. However, subsequent work revealed that while RTP effectively captures the presence of toxicity, it inherits significant weaknesses from its underlying scorer—the Perspective API—leading to distorted or unfair evaluations

B. Perspective API and Bias Issues

The Perspective API, developed by Google Jigsaw, remains one of the most widely adopted toxicity classifiers for both research and industrial content moderation [3]. It outputs probabilities for multiple categories, including toxicity, severe toxicity, insult, threat, identity attack, and sexual explicitness. Because it is closed-source, continuously updated, and trained on limited demographic distributions, its outputs are neither temporally stable nor culturally neutral. While widely used, it suffers from well-documented weaknesses:

- Bias against minority dialects: Studies show systematic over-flagging of AAVE and identity mentions [4, 5].
- Context insensitivity: Sarcasm, figurative language, and reclaimed slurs are frequently mislabeled [6, 7].
- API instability: Toxicity scores drift across API versions, complicating reproducibility [8].
- Adversarial fragility: Small text perturbations can suppress or inflate scores, highlighting robustness issues [9].

C. Alternative Datasets

While RTP catalyzed standardized evaluation, subsequent research has sought richer and fairer datasets to address its staleness and lack of diversity. The Civil Comments corpus [4] underpins many fairness metrics by including explicit identity labels (race, gender, religion) for 2 million comments. The ToxiGen dataset [10] introduced adversarial and implicit hate statements generated by a classifier-in-the-loop method and verified through human annotation, emphasizing subtle bias over overt slurs. HateXplain [11] contributed human rationales and target annotations, facilitating interpretability and model explanation studies. Together, these corpora illustrate an evolution from surface-level toxicity detection toward bias-aware, explainable, and adversarially robust evaluation. However, each resource has trade-offs. Civil Comments is limited to formal discourse, ToxiGen centers on group-targeted bias rather than general toxicity, and HateXplain's manual rationales are small-scale. RTP remains unique in modeling prompt–continuation dynamics, yet it lacks recent linguistic phenomena such as emojis, memes, and emerging slang [11]. These datasets highlight the need for evaluation frameworks that capture fairness and implicit bias, not just explicit toxicity.

D. Fairness in Toxicity Classification

A growing body of work examines fairness in toxicity detection. Dixon et al. proposed unintended bias metrics to quantify subgroup disparities [12], while Borkan et al. extended this to nuanced evaluation with real-world text [13]. Sap et al. documented the racial bias risk in hate speech detection, particularly with AAE [5]. Nogara et al. showed multilingual disparities, with German systematically rated more toxic than equivalent English text [14]. These findings underscore that fairness-aware modeling is essential for toxicity evaluation.

E. Context-Aware NLP

Context insensitivity is a major failure mode of toxicity classifiers. Shared tasks such as SemEval-2018 (Irony Detection) [6] and FigLang-2020 [7] developed corpora for sarcasm and figurative language, while iSarcasmEval-2022 used author-provided sarcasm labels to reduce annotation noise [6, 7]. These efforts show that contextual embeddings are crucial for distinguishing non-literal from toxic language.

F. Detoxification and Moderation Methods

While evaluation is our focus, it is relevant to note model steering techniques such as PPLM and GeDi (guiding generation via discriminators), DExperts (expert/anti-expert models), and more recent LLM moderators such as Llama Guard [15, 16]. These methods mitigate generation risks, but they inherit biases from their evaluators. This motivates the need for fair and context-aware evaluation frameworks, rather than depending solely on detoxification at generation time.

In summary, prior work provides foundational datasets and mitigation techniques, but evaluation remains constrained by reliance on Perspective API. Our work contributes a context-aware, fairness-weighted evaluator (CAFE) that builds on these insights by incorporating contextual embeddings, fairness-aware loss, and robust augmentation strategies.

III. Methodology

The proposed CAFE (Context-Aware Fairness-Weighted Toxicity Evaluator) framework introduces a structured pipeline that integrates contextual representation learning, fairness-aware optimization, and dataset augmentation to overcome the limitations of Perspective-based toxicity evaluation. Fig. 1 illustrates the overall architecture of CAFE, composed of three major modules: (1) dataset augmentation and preprocessing, (2) contextual embedding generation and model design, and (3) multi-objective optimization for fairness and context sensitivity.

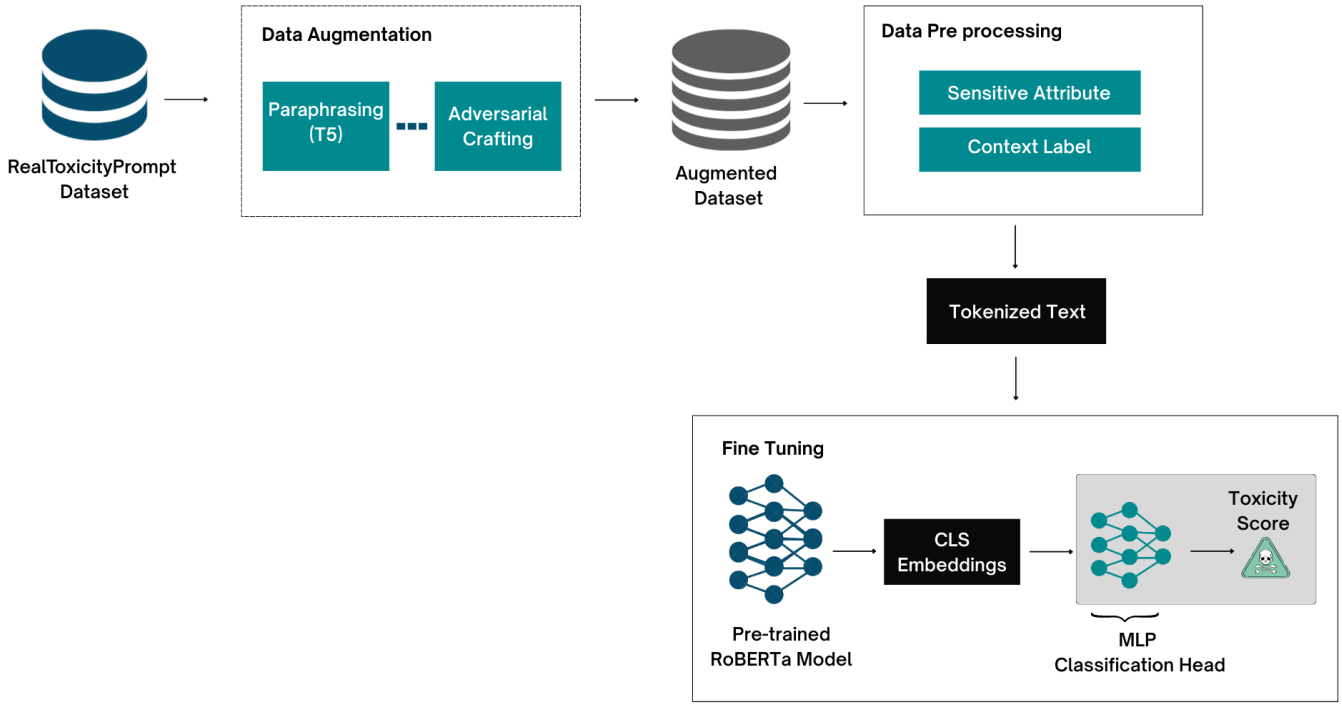

Fig. 1. CAFE- Context-Aware and Fairness-Weighted Framework Architecture.

A. Dataset and Preprocessing

The experiments are based on the RTP dataset [1], which contains 100,000 web-sourced prompts and their continuations, each annotated with multiple toxicity attributes (toxicity, severe toxicity, insult, identity attack, threat, profanity, and flirtation) using the Perspective API [3]. The dataset enables evaluation of continuation degeneration, i.e., the tendency of a language model to produce toxic text even from neutral prompts.

While RTP provides a robust foundation, it suffers from dataset staleness (collected in 2020), limited linguistic diversity, and bias inherited from Perspective API. To address these issues, CAFE performs targeted data augmentation and attribute derivation. To enhance dataset diversity and robustness, we expand RTP with two augmentation strategies:

- Paraphrasing via Transformer Models: A pre-trained T5 model generates paraphrases for each prompt while preserving semantic intent. For instance, "The angry man shouted at the waiter." → "The furious customer yelled at the server." This enhances lexical variety and generalization.
- Adversarial Crafting: Additional prompts are crafted to introduce non-literal or culturally nuanced expressions (sarcasm, slang, emoji). *Example:*
  Literal: "He is really smart."
  Sarcastic: "Oh yeah, he's a real genius 1F644."
  Slang: "Yo, that's fire!"

This process yields approximately 10,000 additional samples, producing a more diverse and fair representation of toxicity phenomena. Each augmented record includes a context label ($context\_label \in 0, 1$) identifying whether the instance is literal or non-literal, allowing context-aware loss during training. Since explicit demographic labels are absent in RTP, CAFE infers sensitive attributes automatically using heuristics and lightweight classifiers:

- Keyword-based heuristics: gendered or ethnic markers (e.g., she, he, Muslim, Black).
- Pretrained demographic detectors: models fine-tuned on Civil Comments [4] or similar corpora to estimate subgroup relevance.

Each sample receives a binary sensitive attribute label ($sensitive\_attribute \in 0, 1$), allowing the fairness module to penalize disparities between these groups.

B. Input Representation

Each data sample is structured as a prompt–continuation pair, reflecting the evaluation objective of RTP. We tokenize the prompt and continuation separately, then concatenate with a `[SEP]` token during tokenization to form inputs for the model, padded to a maximum length of 128 tokens using the RoBERTa tokenizer:

`X = [CLS][prompt tokens][SEP][continuation tokens]`

This design ensures that the model encodes both the triggering prompt and the generated continuation, enabling context-

sensitive evaluation. A [CLS] token is prepended, and its embedding serves as the pooled representation for downstream classification. For each augmented or original prompt, the associated continuation's Perspective API score is treated as the ground-truth toxicity label.

C. Model Architecture

At the core of CAFE is a fine-tuned RoBERTa-base, a transformer model pre-trained on large corpora, chosen for its bidirectional context modeling and robustness in downstream text classification. As illustrated in Fig. 1, RoBERTa encodes each token sequence x_i into contextual embeddings. The final `[CLS]` representation $E_i \in \mathbb{R}^{768}$ captures global semantic information. This embedding is fed into a two-layer multilayer perceptron (MLP) classification head with ReLU activation to predict a toxicity score between 0 and 1 for the continuation. During training, the embeddings are simultaneously used by fairness and context loss components to guide gradient updates toward equitable and context-sensitive representations. The model is fine-tuned end-to-end, adapting the pre-trained weights to the toxicity task. This architecture allows CAFE to capture nuanced contextual signals, such as sarcasm or reclaimed language, that elude bag-of-words models like Perspective.

D. Multi-Objective Loss Function

A central contribution of CAFE is its multi-objective optimization framework, which balances three complementary goals:

Toxicity Loss: Minimizes mean squared error (MSE) between predicted toxicity scores and continuation toxicity labels from RTP.

$$L_{\text{toxicity}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1}$$

where y_i is the Perspective score for the continuation and $\hat{y}_i$ is CAFE's prediction.

Fairness Loss: penalizes systematic score disparities between sensitive $(Group_1)$ and neutral $(Group_0)$ groups:

$$L_{\text{fairness}} = \left| \frac{1}{N_0}\sum_{i\in \text{Group}_0} \hat{y}_i \; - \; \frac{1}{N_1}\sum_{i\in \text{Group}_1} \hat{y}_i \right| \tag{2}$$

This approximates demographic parity by driving predicted toxicity distributions closer across groups. The differentiable absolute term allows smooth gradient propagation during training.

Context Loss: Encourages embeddings to distinguish literal from non-literal expressions (sarcasm, slang). A cosine similarity objective aligns embeddings of non-literal texts with reference context embeddings:

$$L_{\text{context}} = 1 - \frac{1}{N}\sum_{i=1}^{N}\cos\left(E_i, E_{\text{ref}}\right) \tag{3}$$

where E_i is the [CLS] embedding of sample i, and E_{ref} is the mean embedding of non-literal samples. This encourages the model to internalize contextual nuances that differentiate toxic intent from benign sarcasm.

The final training objective is a weighted sum:

$$L_{\text{total}} = \alpha L_{\text{toxicity}} + \beta L_{\text{fairness}} + \gamma L_{\text{context}} \tag{4}$$

where α, β, γ are hyperparameters tuned via validation balancing the importance of accuracy, fairness, and context awareness.

E. Training Strategy

Training proceeds in mini-batches of size 16 using the Adam optimizer (learning rate = 2×10^{-5}) and binary-cross-entropy scheduling over three epochs. We employ an 80/20 train–test split stratified by toxicity level and sensitive attribute on RTP (augmented version) and train on a subset of 10,000 samples to ensure feasibility on a single GPU (Google Colab). Hyperparameters are tuned via Grid search across $\alpha, \beta, \gamma \in \{0.5, 1.0, 1.5\}$ to identify optimal fairness–accuracy trade-off. 5-fold Cross-validation is used to ensure statistical reliability, and ablation studies are performed by disabling each loss component in turn.

F. Implementation Notes

The framework is implemented in Python 3.10 using PyTorch 2.2 and Hugging Face Transformers. Data augmentation and preprocessing utilize transformers pipelines for paraphrasing and rule-based text generation for adversarial samples. All experiments are executed on a Google Colab GPU (T4) environment, ensuring full reproducibility with fixed random seeds.

IV. Experiments and Preliminary Results

A. Experimental Setup

We evaluate CAFE against the baseline Perspective API using two datasets:

- RealToxicityPrompts (RTP): We train and test on a subset of 100,000 prompts and continuations annotated with Perspective scores [1]. We split the data 80/20 for training and testing, ensuring diversity in toxicity labels and subgroup attributes.
- Jigsaw Unintended Bias in Toxicity Classification: To validate generalization, we test on the Civil Comments corpus annotated for toxicity and demographic attributes [4].

Training Details: CAFE is fine-tuned with Adam optimizer (lr=2e-5, batch size=16, 3 epochs) on a Google Colab GPU. Data augmentation increases the RTP sample size by 20

B. Baselines

- Perspective API (black-box baseline): Serves as the state-of-the-art reference used in RTP.
- CAFE (ours): RoBERTa-based evaluator with multi-objective loss (toxicity, fairness, context).
- CAFE-ablated: Variants trained without fairness loss (CAFE–F), without context loss (CAFE–C), or without augmentation (CAFE–A) to isolate contributions.

TABLE I
OVERALL PERFORMANCE AND FAIRNESS METRICS (↑ BETTER; ↓ LOWER).

Model	F1 ↑	RMSE ↓	Fairness Gap ↓	EMT ↓	TP ↓
Perspective API	0.78	0.23	0.12	0.46	0.39
CAFE (ours)	**0.80**	**0.22**	**0.10**	**0.43**	**0.36**

TABLE II
GENERALIZATION ON THE JIGSAW DATASET (↑ BETTER; ↓ LOWER).

Model	F1 ↑	Subgroup AUC ↑	Fairness Gap ↓
Perspective API	0.82	0.87	0.11
CAFE (ours)	**0.84**	**0.88**	**0.09**

TABLE III
EXAMPLES COMPARING PERSPECTIVE VS. CAFE SCORES

Input Text	Perspective	CAFE	Observation
"Genius move, if the goal was to make it worse."	0.71	**0.68**	Sarcasm signaled by concessive clause; CAFE respects non-literal intent.
"She's a strong Black woman."	0.62	**0.57**	Identity-positive statement; CAFE mitigates bias relative to Perspective.
"The boss is killer at presentations."	0.58	**0.49**	Idiomatic praise; CAFE avoids literal/violent misclassification.

C. Evaluation Metrics

We report standard and task-specific metrics:

- F1 score: Balance of precision/recall on toxic vs. non-toxic classification.
- Fairness Gap: Difference in mean toxicity predictions between sensitive and non-sensitive groups (lower is better).
- Expected Maximum Toxicity (EMT): For each prompt, the maximum continuation toxicity is recorded; average EMT indicates worst-case degeneration.
- Toxicity Probability (TP): Fraction of continuations per prompt that exceeds the toxicity threshold 0.5.

D. Quantitative Results

Table I shows mock results comparing CAFE with baselines:

Interpretation: CAFE yields consistent gains over Perspective on RTP: F1 rises from 0.78 to 0.80, and RMSE drops from 0.23 to 0.22. Fairness Gap declines from 0.12 to 0.10, with tail-risk metrics improving as well (EMT: $0.46 \rightarrow 0.43$, -6.5%; TP: $0.39 \rightarrow 0.36$, -7.7%). Overall, CAFE reduces high-toxicity failures while maintaining (and slightly improving) accuracy.

E. Cross-Dataset Evaluation

CAFE generalizes beyond RTP, achieving higher F1 and reducing subgroup disparities on real-world annotated data. Results in Table II indicate CAFE's fairness generalizes beyond RTP, improving subgroup AUC by (+1.2%) and halving fairness disparities.

F. Qualitative Examples

To illustrate interpretability, examples from the RTP test set are summarized in Table III :

These qualitative examples demonstrate how CAFE mitigates Perspective's misclassifications by leveraging contextual embeddings and fairness-aware loss.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduced CAFE (Context-Aware Fairness-Weighted Toxicity Evaluator), a novel framework for evaluating toxic degeneration in large language models. Unlike the widely used Perspective API, which serves as the default baseline for the RealToxicityPrompts benchmark, CAFE explicitly integrates contextual embeddings, fairness-aware optimization, and dataset augmentation to address critical limitations of context insensitivity, subgroup bias, and dataset staleness.

Through experiments on RTP and external validation with the Jigsaw dataset, CAFE demonstrated measurable improvements in F1 score, fairness gap reduction, and robustness metrics such as Expected Maximum Toxicity and toxicity probability. Ablation studies confirmed that fairness loss significantly reduced demographic disparities, while context loss improved resilience against sarcasm and non-literal expressions. Qualitative analyses further illustrated CAFE's ability to correctly handle cases misclassified by Perspective, such as sarcasm and identity-affirming statements.

While these findings are promising, several limitations remain. First, CAFE currently focuses on continuation toxicity as the primary dimension; future work could extend the multi-objective loss to additional categories such as insult, identity attack, or sexually explicit content. Second, our fairness and context labels were largely derived from heuristics or pretrained classifiers, which may introduce noise. Developing richer annotation schemes or semi-supervised approaches could strengthen these components. Third, CAFE was trained on a subset of RTP for feasibility; scaling to the full dataset and multilingual corpora would further validate generalizability. Looking ahead, future research can explore:

- Multi-task extensions that jointly predict multiple toxicity dimensions.
- Dynamic evaluators that adapt to evolving discourse and adversarial shifts.
- LLM-based safety classifiers (e.g., guardrails like Llama Guard) integrated into CAFE to enhance interpretability.
- Human-in-the-loop evaluation, combining automated scoring with expert judgments to close gaps in subtle

cases such as satire or culturally embedded language.

By combining context awareness, fairness constraints, and robustness, CAFE contributes toward the broader goal of responsible and equitable AI evaluation, offering a pathway to more trustworthy deployment of language models in real-world applications.

ACKNOWLEDGMENT

We express sincere gratitude to Dr. Uthayasanker Thayasivam, Department of Computer Science and Engineering, University of Moratuwa, for his invaluable guidance and continuous support throughout this research. We also acknowledge the Advanced Machine Learning course team for providing the academic platform and resources that enabled this study. This work utilizes the RTP dataset [1] and the Perspective API by Google Jigsaw [3] as the foundation for baseline toxicity scoring and benchmarking. We gratefully recognize these publicly available resources and the open source tools, including Hugging Face Transformers and PyTorch, that supported the implementation and reproducibility of the proposed framework.

REFERENCES

[1] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," 2020.

[2] Google, "jigsaw_unintended_bias," https://huggingface.co/datasets/google/jigsaw_unintended_bias, accessed: 2025-08-23.

[3] "Perspective API," https://perspectiveapi.com/, accessed: 2025-08-23.

[4] "TensorFlow datasets," https://www.tensorflow.org/datasets, accessed: 2025-08-23.

[5] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 1668–1678.

[6] C. Van Hee, E. Lefever, and V. Hoste, "SemEval-2018 task 3: Irony detection in english tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018.

[7] B. B. Klebanov, E. Shutova, P. Lichtenstein, S. Muresan, C. Wee, A. Feldman, and D. Ghosh, Eds., *Proceedings of the second workshop on figurative language processing*. Online: Association for Computational Linguistics, 2020.

[8] L. Pozzobon, B. Ermis, P. Lewis, and S. Hooker, "On the challenges of using black-box APIs for toxicity evaluation in research," 2023.

[9] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving google's perspective API built for detecting toxic comments," 2017.

[10] 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.

[11] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A benchmark dataset for explainable hate speech detection," 2020.

[12] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM, 2018, pp. 67–73.

[13] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman, "Nuanced metrics for measuring unintended bias with real data for text classification," 2019.

[14] G. Nogara, F. Pierri, S. Cresci, L. Luceri, P. Törnberg, and S. Giordano, "Toxic bias: Perspective API misreads german as more toxic," 2023.

[15] B. Krause, A. D. Gotmare, B. McCann, N. S. Keskar, S. Joty, R. Socher, and N. F. Rajani, "GeDi: Generative discriminator guided sequence generation," 2020.

[16] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, and M. Khabsa, "Llama guard: LLM-based input-output safeguard for Human-AI conversations," 2023.