MemeDetoxNet: Balancing Toxicity Reduction and Context Preservation

Gitanjali Kumari[1] **Jitendra Solanki** [1] **Asif Ekbal**[1,2]
[1]Department of Computer Science and Engineering,
[1]Indian Institute of Technology Patna, India
[2]School of AI and Data Science, Indian Institute of Technology Jodhpur, India
[1]{gitanjali_2021cs03,jitendra_2211cs26, asif}@iitp.ac.in,[2]asif@iitj.ac.in

Abstract

Toxic memes often spread harmful and offensive content and pose a significant challenge in online environments. In this paper, we present **MemeDetoxNet**, a robust framework designed to mitigate toxicity in memes by leveraging fine-tuned pre-trained models. Our approach utilizes the interpretability of CLIP (Contrastive Language-Image Pre-Training) to identify toxic elements within the visual and textual components of memes. Our objective is to automatically assess the immorality of toxic memes and transform them into morally acceptable alternatives by employing large language models (LLMs) to replace offensive text and blurring toxic regions in the image. As a result, we proposed *MemeDetoxNet* that has three main primitives: (1) detection of toxic memes, (2) localizing and highlighting toxic visual and textual attributes, and (3) manipulating the toxic content to create a morally acceptable alternative. Empirical evaluation on several publicly available meme datasets shows a reduction in toxicity by approximately 10-20%. Both qualitative and quantitative analyses further demonstrate MemeDetoxNet's superior performance in detoxifying memes compared to the other methods. These results underscore MemeDetoxNet's potential as an effective tool for content moderation on online platforms[1].

Warning: This paper includes toxic memes that contain text or images with nudity or sexual content as part of the detoxification study.

1 Introduction

In the context of the proliferation of social media, the right to freedom of expression has increasingly been prominent in spreading toxic content through various posts (Sharma et al., 2020; Kumari et al., 2021). Among these posts, memes have become

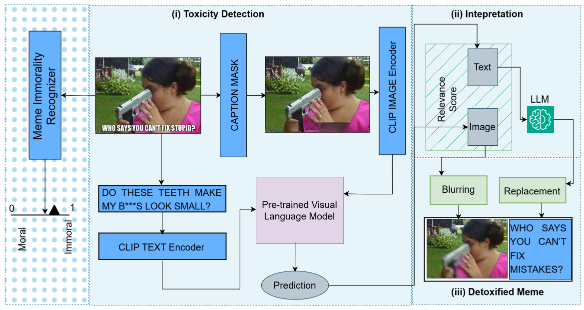


Figure 1: **MemeDetoxNet**: Toxic Meme Detection with Detoxification Approach

the most popular among social media users, deliberately combining visual and textual elements to convey toxic messages. Due to the high volume of such content on social media platforms, manual moderation of toxic memes is becoming increasingly challenging. Automated techniques are, therefore, necessary to detect and address such harmful content efficiently. Previous research has predominantly focused on developing robust deep-learning models that learn cross-modal interactions from scratch to identify these memes (Rijhwani et al., 2017; Sharma et al., 2020; Kiela et al., 2020a; Suryawanshi et al., 2020; Pramanick et al., 2021; Hossain et al., 2022; Sharma et al., 2022), often overlooking the critical necessity of applying safety filters that detoxify such memes by moderating both the textual and visual modalities. A straightforward approach to detecting toxic memes involves training a lightweight safety classifier on a specialized dataset (Kumar and Nandakumar, 2022). However, such methods' lack of interpretability and limited generalizability restrict their effectiveness in diverse, open-world scenarios (Hendrycks et al., 2021). To address these challenges, there is a growing need for Artificial Intelligence (AI) models that can operate effectively in open-world environments by grounding their decisions in ethical principles (Bai et al., 2022). Such models must learn how funda-

[1]Codes are available at this link: https://github.com/Gitanjali1801/MemeDetoxNet.git

Findings of the Association for Computational Linguistics: ACL 2025, pages 25076–25098
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

mental facts about the world relate to human values and morality, enabling them to make nuanced judgments beyond simple classification. In this paper, we adopt a proactive approach to ethical and moral considerations by not only detecting toxic memes but also actively detoxifying them. We aim to transform harmful memes on social media into non-offensive alternatives, preventing the spread of such toxic material.

To bridge the gap in existing content moderation approaches, we propose **MemeDetoxNet**, a novel framework designed to not only detect toxic memes but also convert them into morally acceptable alternatives. *MemeDetoxNet* operates through three main primitives: (1) identifying immoral memes using a meme immorality recognizer, (2) localizing and highlighting toxic visual and textual attributes that make them immoral, and (3) manipulating the toxic content to create a morally acceptable alternative (Refer Figure 1). Using the fine-tuning of the CLIP model, *MemeDetoxNet* evaluates each meme's visual and textual components. It segments and identifies the harmful attributes in both modalities through interpretability. Once identified, the model applies a detoxification process by replacing toxic text with a more neutral alternative using a large language model (LLM) and by blurring the image regions responsible for toxicity based on the relevance scores. Importantly, this process preserves the original context and intent of the meme while removing its harmful aspects. The primary motivation of our work is the recognition that simply detecting toxic memes is insufficient; a more proactive approach is needed to combat the spread of harmful content online. By significantly reducing toxicity in memes with the help of a modular framework, our work seeks to build a foundation for a more comprehensive and generalized approach to content moderation.

2 Related Work

Previous research in toxic meme identification primarily relied on fusing unimodal approaches, where either the text or image was analyzed separately. Pre-trained models, such as BERT, ResNet, and Vision Transformers, were used to extract textual and visual features independently (Dietterich, 1998; Chen et al., 2015; Bowman et al., 2015; Dawkins, 2016), which were then combined for classification (Kiela et al., 2018). However, these methods struggled to capture the nuanced interplay between image and text—an essential aspect for understanding the full meaning conveyed by memes. As research advanced, pre-trained multimodal models, such as CLIP, BLIP, VisualBert etc., were introduced to address this limitation, integrating both textual and visual information (Hossain et al., 2019; Kumar, 2022; Chen et al., 2020, 2022a; Choshen et al., 2022; Du et al., 2020), demonstrating that multimodal approaches are significantly more effective than unimodal methods.

Despite advancements in multimodal approaches for toxic meme identification, the area of meme detoxification remains largely underexplored. While researchers have made significant progress in detoxifying textual data through content moderation techniques (Hanu, 2021; Dixon et al., 2018; Lima et al., 2024; Karan and Šnajder, 2019), these efforts have predominantly focused on text-based generation models, such as GPT-3 (Brown et al., 2020; Dale et al., 2021; Xu et al., 2021; Zhong et al., 2022) and controlled generation systems like CTRL (Keskar et al., 2019), which have successfully generated non-toxic alternatives for offensive text. Emerging techniques like safe latent diffusion (Schramowski et al., 2023) and concept erasure (Gandikota et al., 2023), though primarily applied to creative tasks, show promise for content moderation by bridging the gap between text-image relationships in multimodal content (Rombach et al., 2022; Nishimura et al., 2019; Chen and Zhuge, 2018; Yao et al., 2023; Kumari et al., 2023; Kim et al., 2024). However, these methods have not been adequately extended to memes' unique challenges. Detoxifying memes requires a more sophisticated approach that can effectively address the interplay of both modalities, an area that current research has yet to explore fully.

While meme categorization has advanced with the development of image-text retrieval algorithms, LLM-based methods, contrastive-learning approaches, and scene-graph-based techniques (Sharma et al., 2018; Vempati et al., 2020; Ruiz et al., 2020; Kumari et al., 2021; Suryawanshi and Chakravarthi, 2021; Sharma et al., 2022, 2023; Kumari et al., 2024), the area of meme detoxification remains relatively unexplored. Despite progress in toxic meme detection, understanding the underlying causes of toxicity through interpretability techniques remains a challenge. Additionally, current toxic detection models often lack transparency, making it difficult to alter specific harmful attributes without compromising the over-

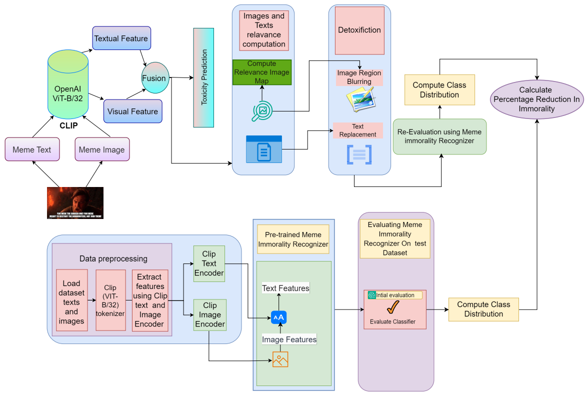


Figure 2: An overview of our proposed model *MemeDetoxNet* architecture

all context. To address these gaps, we propose *MemeDetoxNet*, a multimodal framework to detoxify memes. By leveraging the interpretability of the CLIP model, we employ GradCAM and attention scores to identify specific image regions and text attributes that contribute to toxic interpretations. This level of interpretability, often lacking in current approaches, allows us to implement targeted interventions that replace harmful content while preserving the meme's overall context.

3 Dataset

For our experiments, we utilized a diverse set of publicly available meme datasets, including English-language datasets (MAMI (Fersini et al., 2022), Hateful Memes (Kiela et al., 2020b), and Memotion2 (Ramamoorthy et al., 2022) and the Hindi-English code-mixed MIMIC dataset (Singh et al., 2024). Table 8 provides detailed statistics for these datasets. Additionally, we employed the large-scale ETHICS (Hendrycks et al., 2020) dataset, consisting of over 13,000 textual examples, to train the meme immorality recognizer (See Appendix Table 9).

4 Proposed Methodology

This section illustrates our proposed MemeDetoxNet framework to detoxify memes. The overall workflow of our proposed *MemeDetoxNet* model is shown in Figure 2, and its components are discussed below.

4.1 Problem Formulation

Given a collection of datasets $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k\}$, where each dataset $\mathcal{D}_i = \{(x_j, y_j)\}_{j=1}^{N_i}$ consists of memes $x_j \in \mathcal{X}$ with both visual and textual components, and $y_j \in \{0, 1\}$ denotes the corresponding toxicity label ($y_j = 1$ for toxic, $y_j = 0$ for non-toxic), the task is to develop a model that not only detects toxic memes but also actively detoxifying it to reduce the overall toxicity across the collection of datasets. Each meme x_j within any dataset $\mathcal{D}_i$ is represented as a combination of image and text modalities, $x_j = (x_{\text{image}}, x_{\text{text}})$. The goal is to design a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by θ, that outputs a predicted label $\hat{y}_j$ for each meme x_j in any dataset $\mathcal{D}_i$. The system localizes toxic attributes within both modalities if a meme is classified as toxic ($\hat{y}_j = 1$). These toxic attributes are then modified through techniques, such as blurring offensive images or replacing harmful text while preserving the meme's essential meaning and coherence.

To assess the reduction in toxicity, the **Meme Immorality Recognizer (MIR)**, a pre-trained model specifically designed for detecting immoral content, acts as a judge. It evaluates the percentage decrease in toxicity across all the datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ using a zero-shot learning approach.

4.2 Encoding of Meme

A meme sample $\mathcal{X}_i$ consists of meme text $T_i = (t_{i_1}, t_{i_2}, \ldots, t_{i_k})$, where the text is tokenized into sub-word units and projected into high-dimensional feature vectors. Here, k represents the number of tokens in the meme text. Additionally, the image component I_i is segmented into regions $r_i = \{r_{i_1}, r_{i_2}, \ldots, r_{i_N}\}$, where each region $r_{i_j} \in \mathbb{R}^N$, and N denotes the number of image regions. These textual and visual components are then input into the CLIP model (Radford et al., 2021), which is fine-tuned to extract and understand semantic-level features from both the modalities, enabling a comprehensive understanding of the meme's content.

$$ft_i, fv_i = CLIP(t_i, r_i) \ ; \tag{1}$$

where $ft_i \in \mathbb{R}^{d_t}$ and $fv_i \in \mathbb{R}^{d_v}$ are the extracted text and visual features, respectively, with d_t and d_v denoting the dimensions of the text and visual feature spaces.

4.3 Training

The text features ft_i, and the image features fv_i (c.f. Equation 1), extracted from the CLIP model and combined through a linear layer, followed by an activation function. The resulting multimodal features are passed through a softmax activation layer to obtain the final predictions. The model computes the output logits $\hat{y}$ as follows:

$$T' = \text{ReLU}(W_t ft_i + b_t) \; ; \quad I' = \text{ReLU}(W_i fv_i + b_i) \tag{2}$$

$$C = \text{Dropout}([T', I']) \tag{3}$$

$$\hat{y} = \text{LogSoftmax}(W_c C + b_c) \tag{4}$$

where W_t, W_i, and W_c are weight matrices, b_t, b_i, and b_c are bias vectors, and [,] denotes concatenation.

4.4 Loss Function

The classifier is trained using the cross-entropy loss function. The optimizer minimizes the loss by adjusting the model parameters:

$$L = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{5}$$

where $\hat{y}_i$ is the predicted probability for class y_i.

4.5 Identifying Relevance Toxic Score for Image Regions and Text

4.5.1 Image Relevance Computation

To understand the contribution of the visual modality in predicting toxicity within *MemeDetoxNet*, we apply Grad-CAM (Chen et al., 2022b; Lin et al., 2023) to the last layer of CLIP image encoder. This method visualizes the specific image regions that influence the model's decision. Let I represents the input image, and F_I denotes the feature map from the final layer of the CLIP image encoder. The Grad-CAM is computed using the gradient of the predicted toxicity score $\hat{y}_i$ with respect to the feature map A^l from the last layer of the image encoder:

$$L^c_{\text{GradCAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^l_k\right) \tag{6}$$

where α^c_k is the importance weight for each feature map channel A^l_k, computed as:

$$\alpha^c_k = \frac{1}{Z} \sum_{i,j} \frac{\partial y^c}{\partial A^l_{ij}} \tag{7}$$

Here, Z represents the total number of pixels in the feature map, and A^l_{ij} is the activation at spatial location (i, j). By applying Grad-CAM, we visualized the key regions within the meme-image that contributes to the toxicity score, providing insights into the CLIP's decision-making process for the specific task.

4.5.2 Text Relevance Computation

Afterward, we perform word masking of the meme text to evaluate individual words' contribution to a meme's overall toxicity. By masking each word in a meme text and analyzing the impact on the classifier's output, we identify toxic words based on their effect on toxicity prediction. If masking a word causes the classifier to predict the meme as non-toxic, we infer that the word contributes to toxicity. Formally, let $f_\theta : T \to [0, 1]$ be a classifier that maps a text sequence $T = [w_1, w_2, \ldots, w_{|T|}]$ to a toxicity probability, where w_i represents the i-th word in the sequence. We define a per-word binary mask $M^T : |T| \to \{0, 1\}$, where $M^T_i = 0$ indicates that word w_i is masked, and $M^T_i = 1$ indicates that the word is retained. The masked input sentence is then given by:

$$T' = T \odot M^T \tag{8}$$

where $\odot$ denotes element-wise multiplication between the text sequence T and the binary mask M^T. The importance score for each word w_i, for $i \in \{1, \ldots, |T|\}$, is computed by measuring the change in the classifier's output when the word is masked. The relevance score (R_{w_i}) is defined as:

$$R_{w_i} = f_\theta(T) - \mathbb{E}[f_\theta(T \odot M^T) \mid M^T_i = 0] \tag{9}$$

where $\mathbb{E}[f_\theta(T \odot M^T) \mid M^T_i = 0]$ is the expected output of the classifier when word w_i is masked, averaged over all possible masking configurations for the remaining words. A larger value of (R_{w_i}) indicates a higher contribution of word w_i to the overall toxicity of the meme.

4.6 Dehatification Process

4.6.1 Image Dehatification with Blurring

The identified toxic regions in images (Refer Section 4.5.1) are blurred using a Gaussian filter (Ito and Xiong, 2000; Vo and Ma, 2006). Given an image I and a binary mask M indicating the toxic regions, the dehatified image $I_{\text{dehatified}}$ is computed as:

$$I_{\text{dehatified}} = I \odot (1 - M) + \text{GaussianBlur}(I) \odot M \tag{10}$$

where $\odot$ denotes element-wise multiplication.

4.6.2 Text Dehatification with LLM based Replacement

The toxic words in the text (identified in Section 4.5.2) are replaced with less toxic alternatives using a LLM by instruction tuning (Shu et al., 2023;

Schwinn et al., 2024; Kumar et al., 2024). Given a text T, toxic words W_{toxic}, and their replacements $W_{\text{non-toxic}}$, the detoxified text $T_{\text{dehatified}}$ is (Refer to Appendix Figure 8 for details on the prompts used in instruction fine-tuning) :

$$T_{\text{dehatified}} = \text{replace}(T, W_{\text{toxic}}, W_{\text{non-toxic}}) \quad (11)$$

4.7 Meme Immorality Recognizer

Our Meme Immorality Recognizer (MIR) acts as a judge, evaluating the moral implications of given memes. Training such a system is particularly challenging due to the lack of large-scale, high-quality datasets specifically designed for immorality recognition (Hendrycks et al., 2020). To address this, we pre-trained an auxiliary text-based MIR using the large-scale ETHICS dataset. Formally, given an input text T, we leverage the frozen CLIP-based text encoder f_t followed by an immorality classifier $f_c : \hat{y} = f_c(f_t(T))$, where the immorality classifier is trained with Binary Cross-Entropy Loss (BCELoss) as follows:

$$\mathcal{L}_c = -\frac{1}{n}\sum_{i=1}^{n} \text{BCE}(y_i, \hat{y}_i) \quad (12)$$

where $\text{BCE}(y_i, \hat{y}_i)$ represents the binary cross-entropy loss for each instance. y_i is the true immorality label for the i-th meme, where $y_i = 1$ indicates a immoral sample, and $y_i = 0$ indicates a moral sample. $\hat{y}_i$ is the predicted probability of the sample being immoral.

During inference, the immorality of a meme is assessed using a joint embedding of both its text and image, processed by the pre-trained MIR in a zero-shot manner. We employ the CLIP-based text encoder f_t and image encoder f_v, which map semantic text ($\mathcal{T}$) and image ($\mathcal{I}$) pairs into a shared embedding space, bringing text and image features closer together. The final output for an unseen meme (M_i) is computed as:

$$\hat{y} = f_c\left([f_t(\mathcal{T}), f_v(\mathcal{I})]\right) \quad (13)$$

where $[\,]$ denotes the concatenation of textual and visual features. This approach enables us to effectively predict a meme's immorality score by leveraging its textual and visual components in a unified framework.

5 Experiments

5.1 Evaluation Metrics

To ensure effective detoxification, the following metrics are defined to assess the quality of the

	MAMI	FHM	Memotion	MIMIC
MemeDetoxNet	15.38	10.32	9.39	10.34
VisualBERT	17.74	8.15	11.29	9.39
MOMENTA	18.29	12.96	10.38	12.11
PromptHate	12.94	7.28	12.29	12.18
Pro-Cap	14.85	5.75	11.86	11.19
LLaVA	13.95	5.89	13.75	12.19

Table 1: Toxicity Reduction (TR) performance in percentage achieved by the Meme Immorality Recognizer (MIR) across different models and datasets.

	MAMI	FHM	Memotion	MIMIC
TR	4.28	3.73	3.41	4.01
KR	3.96	4.29	4.03	3.27
CP	3.14	3.18	3.97	3.18

Table 2: Human evaluation results for Toxicity Reduction (TR), Knowledge Relevance (KR), and Context Preservation (CP) on meme samples before and after detoxification using *MemeDetoxNet* across multiple datasets.

detoxified meme. In addition to the macro F1-score, we utilized the following metrics for human evaluation: (i) Knowledge Relevance (KR), (ii) Context Preservation (CP), (iii) Toxicity Reduction (TR), and (iv) BertScore (BS). Refer Appendix Section B for detailed descriptions of each metric.

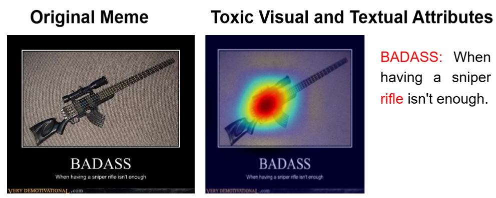


Figure 3: Textual and visual immoral attribute identification examples from MAMI dataset.

6 Results Analysis

This section presents the performance of MemeDetoxNet compared to baseline and state-of-the-art models. We also analyze its impact on the Meme Immorality Recognizer (MIR), trained on multiple morality datasets, to assess how detoxification influences toxicity reduction and immoral content recognition.

6.1 Automatic Evaluation

Based on Macro-F1-score: Table 3 presents the percentage drop in macro-F1 scores after detoxification. Our proposed model *MemeDetoxNet* con-

	MAMI			FHM			Memotion			MIMIC		
	Blurring	RepT	Both	Blurring	RepT	Both	Blurring	RepT	Both	Blurring	RepT	Both
MemeDetoxNet (Ours)	15.18 ± 0.48	19.29 ± 0.39	18.27 ± 0.34	8.29 ± 0.16	10.38 ± 0.16	8.39 ± 0.12	7.28 ± 0.45	12.11 ± 0.34	13.39 ± 0.38	9.28 ± 0.11	7.73 ± 0.49	12.22 ± 0.43
VisualBERT	7.31 ± 0.17	6.29 ± 0.17	12.82 ± 0.22	6.29 ± 0.31	8.63 ± 0.27	8.92 ± 0.22	7.29 ± 0.34	6.73 ± 0.16	3.39 ± 0.22	6.73 ± 0.25	8.64 ± 0.28	8.95 ± 0.41
MOMENTA	5.18 ± 0.31	3.28 ± 0.34	10.18 ± 0.12	2.28 ± 0.34	4.39 ± 0.17	6.63 ± 0.13	6.38 ± 0.48	9.37 ± 0.49	4.46 ± 0.42	6.39 ± 0.22	9.42 ± 0.14	9.93 ± 0.37
PromptHate	10.11 ± 0.15	11.87 ± 0.3	15.28 ± 0.11	6.63 ± 0.13	4.24 ± 0.42	7.38± 0.19	3.18 ± 0.18	5.28 ± 0.13	6.18 ± 0.24	—	—	—
Pro-Cap	7.29 ± 0.24	5.29 ± 0.28	13.19 ± 0.81	5.39± 0.46	7.39 ± 0.2	6.62 ± 0.37	6.18 ± 0.22	6.19 ± 0.31	7.58 ± 0.32	2.17 ± 0.49	6.23± 0.29	5.29 ± 0.13
LLaVA	5.94 ± 0.27	—	—	3.38 ± 0.39	—	—	4.56 ± 0.22	—		4.29±0.37	—	—

Table 3: Percentage drop (% ∇) in macro-F1 scores for toxic meme detection systems on modified test inputs, with ± variance values indicating statistical significance. The values represent the performance degradation after applying detoxification over the baseline models (Higher values indicate more effective toxicity reduction).

LLMs	MAMI			FHM			Memotion			MIMIC		
	KR	CP	BS	KR	CP	BS	KR	CP	BS	KR	CP	BS
Gemini 1.0	3.2	3.6	0.653	4.3	3.4	0.741	4.89	3.19	0.723	2.39	4.18	0.561
GPT3.5-Turbo	4.1	3.1	0.629	3.9	2.9	0.732	2.63	3.92	0.742	3.75	3.59	0.634
mistral-7b-instruct-4k	3.9	4.39	0.612	3.3	2.8	0.684	3.38	4.39	0.783	4.12	4.12	0.612
Llama 3.1	3.7	3.32	0.594	4.73	3.1	0.725	3.87	3.43	0.754	3.19	3.17	0.598

Table 4: BertScore and Human Evaluation on the toxic text replacement by LLMs. Here KR: Knowledge Relevance, CP: Context Preservation, BS: BertScore

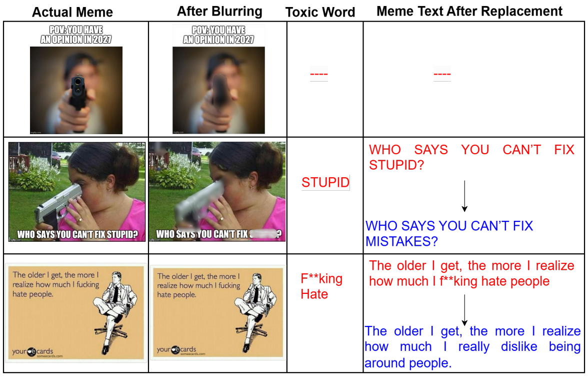


Figure 4: Examples of meme detoxification using MemeDetoxNet, where toxic regions in the image are blurred, and toxic words ("STUPID" and "F**king") are identified and replaced by non-toxic alternatives. We provide the actual meme (1st column), meme image blurring (2nd column), toxic words found (3rd column), and the Meme text after replacement (4th column).

sistently achieves substantial reductions across all datasets (MAMI, FHM, Memotion, and MIMIC), often outperforming the baseline models (VisualBERT and MOMENTA). This effectiveness stems from its explicit identification of toxic words and image regions using GradCAM interpretability. While PromptHate offers competitive performance on MAMI, MemeDetoxNet's combined approach of blurring and text replacement ('Both') yields a superior reduction. Pro-Cap demonstrates effectiveness, particularly on MAMI and with some success in MEMOTION and is found to be more robust in structured data like FHM and MIMIC. With a varying performance across the other datasets and a better-than-most approach in MEMOTION, LLaVA provides a better approach, as compared to other benchmark datasets and emphasizes handling of nuanced, mixed datasets. Refer to Appendix Section D for cross-dataset analysis.

Based on Toxicity Reduction by MIR: In the Table 1, we have shown the effectiveness of *MemeDetoxNet* across various datasets, achieving its highest toxicity reduction in MAMI (15.38%) and strong performance in FHM (10.32%). While VisualBERT outperforms in MAMI (17.74%) and MOMENTA leads in FHM (12.96%), MemeDetoxNet remains consistently effective across all datasets, particularly excelling in multilingual toxicity handling in MIMIC (10.34%). Pro-Cap (Cao et al., 2023) shows competitive results in MAMI (14.85%) and MIMIC (11.19%) but struggles with FHM (5.75%). LLaVA performs best in Memotion (13.75%), handling implicit toxicity effectively, though its performance varies across other datasets. Despite PromptHate's strong results in MIMIC (12.18%), *MemeDetoxNet* demonstrates greater robustness and consistency, making it a reliable solution for meme detoxification across diverse contexts.

Based on BertScore between the original and detoxified text: In table 4, we illustrated the semantic alignment between original and detoxified text using BERTScore (BS). Higher BS in FHM and Memotion indicates that detoxified text closely resembles the original since toxicity is implicit in such datasets, while lower scores in MIMIC (0.561–0.634) suggest greater semantic drift due to code-mixed language complexity. Although Gemini 1.0 and GPT-3.5-Turbo achieve high BS, they tend to over-sanitize, sometimes altering meme intent. Mistral-7B, despite slightly lower BS, performs better in context preservation, highlighting the challenge of balancing toxicity reduction with meaning retention in meme detoxification.

6.2 Human Evaluation

The results in Table 4 highlight the performance of different LLMs across four meme detoxifica-

	Context Preservation (CP) Inter-Annotator Agreement (IAA)				Knowledge Relevance (KR) Inter-Annotator Agreement (IAA)			
	Gemini 1.0	GPT3.5-Turbo	Mistral-7b-instruct-4k	Llama 3.1	Gemini 1.0	GPT3.5-Turbo	Mistral-7b-instruct-4k	Llama 3.1
MAMI	0.55	0.52	0.58	0.5	0.72	0.75	0.73	0.71
FHM	0.62	0.6	0.66	0.59	0.79	0.81	0.8	0.76
Memotion	0.48	0.45	0.5	0.46	0.68	0.7	0.72	0.65
MIMIC	0.5	0.49	0.53	0.47	0.61	0.65	0.67	0.6

Table 5: Inter-Annotator Agreement (IAA) scores for Context Preservation (CP) and Knowledge Relevance (KR) over original meme text and its detoxified version across different datasets and LLMs. Higher scores indicate stronger annotator consensus on how well the detoxified text preserves the original meaning and factual relevance.

tion datasets— MAMI, FHM, Memotion, and MIMIC—based on Knowledge Relevance (KR) and Context Preservation (CP). It illustrates that explicitly toxic datasets (MAMI, FHM) are easier to detoxify, while implicit toxicity (Memotion) and code-mixed language (MIMIC) present greater challenges. GPT-3.5-Turbo achieves the highest Knowledge Relevance (KR) in MAMI (4.1), and Mistral-7B excels in Context Preservation (CP) (4.39), ensuring meme intent retention. Llama 3.1 leads in KR (4.73) for FHM, maintaining factual accuracy, while Gemini 1.0 performs well in CP (3.4), preserving structural integrity. Memotion's implicit toxicity lowers overall performance, though Gemini 1.0 achieves the highest KR (4.89) and Mistral-7B performs best in CP (4.39), retaining humor. MIMIC, due to its complex Hindi-English code-mixing, poses the greatest difficulty, with Mistral-7B achieving the highest KR (4.12) and Gemini 1.0 leading in CP (4.18). These findings highlight LLMs' limitations in handling nuanced toxicity and multilingual memes, emphasizing the need for further improvements in these areas.

After evaluating meme text only, we extended the human evaluation to full memes, including both text and images. Table 2 highlights *MemeDetoxNet*'s performance across datasets in terms of TR, KR, and CP. MemeDetoxNet performs best in TR on explicitly toxic datasets like MAMI (4.28) and MIMIC (4.01), while achieving lower scores on FHM (3.73) and Memotion (3.41) due to humor and contextual complexity. KR is highest for FHM (4.29), indicating strong factual alignment post-detoxification, whereas MIMIC (3.27) presents challenges due to code-mixed language. CP is best maintained in Memotion (3.97), showing effective humor retention, while other datasets show slight meaning shifts post-detoxification.

Inter-annotators Agreement (IAA): To evaluate the reliability of the human annotations for CP, KR, and TR across different datasets and LLMs, we computed the inter-rater agreement using Cohen's Kappa coefficient (Shrout et al., 1987). Unlike simple percentage agreement, Cohen's Kappa provides a more robust statistical measure by accounting for the probability of agreement occurring by chance.

(i) Discussion on IAA of Context Preservation (CP): Table 5 presents the Context Preservation (CP) IAA scores, derived from human evaluations of toxic text replacement by LLMs (refer to Table 4). Using Cohen's Kappa, the results highlight varying levels of agreement between annotators across different datasets and LLMs, reflecting the consistency of judgments on how well the detoxified text retains the original meme text's intent. Overall, moderate to substantial agreement is observed, with FHM and MAMI achieving higher IAA values, suggesting that explicit toxicity and structured text in these datasets make context preservation more consistently interpretable. In contrast, Memotion and MIMIC exhibit lower agreement, reflecting the challenges posed by implicit toxicity, sarcasm, and code-mixed language. Among the models, GPT-3.5 Turbo and Mistral-7B demonstrate relatively stable agreement across datasets, likely due to their balanced detoxification approach. Llama 3.1 shows variability, particularly in Memotion and MIMIC, indicating an inconsistency in handling nuanced meme contexts. Gemini 1.0, while performing well in some datasets, struggles with over-sanitization, leading to disagreements on whether the original context is preserved post-detoxification. These findings highlight that context preservation in meme detoxification is dataset-dependent, and implicit or multilingual toxicity remains a challenge requiring further refinement in LLM-based detoxification strategies.

(ii) Discussion on IAA of Knowledge Relevance (KR): In the same Table 5, we have also illustrated the IAA scores for KR across different datasets and LLMs, highlighting varying levels of annotator consistency. MAMI and FHM datasets exhibit higher agreement, particularly for GPT-3.5 Turbo and Mistral-7B, indicating that these models produce more semantically aligned detoxified text that annotators consistently perceive as relevant. Memotion, being highly implicit and nuanced, shows lower agreement, reflecting challenges in defining knowledge relevance when sarcasm and contextual dependencies are involved. MIMIC, due to its code-mixed nature, also exhibits moderate agreement, with Gemini 1.0 performing better than other models in maintaining relevance. These results indicate that while models handle explicit toxicity well, implicit toxicity and linguistic complexity reduce annotator consensus on knowledge retention.

(iii) Discussion on IAA for Meme Detoxification Across both Text and Image: In the Appendix table 11, we have mentioned the IAA of the result provided in the table 2. Table 11 presents the Inter-Annotator Agreement (IAA) scores corresponding to the human evaluation results in Table 2. These scores reflect the consistency of annotator judgments across different datasets for Context Preservation (CP), Knowledge Relevance (KR), and Toxicity Reduction (TR), considering both text replacement and image blurring in meme detoxification. The highest agreement is observed in the FHM dataset (CP: 0.75, KR: 0.80, TR: 0.79), suggesting that annotators had a more consistent understanding of how well the detoxified content aligned with the original meme, likely due to its templated nature. The MAMI dataset also shows relatively strong agreement (CP: 0.72, KR: 0.78, TR: 0.76), reflecting the explicit nature of misogynistic content, making it easier for annotators to assess detoxification effectiveness. However, lower agreement scores in Memotion (CP: 0.70, KR: 0.74, TR: 0.73) and MIMIC (CP: 0.68, KR: 0.71, TR: 0.70) suggest greater challenges in evaluating implicit toxicity, humor, and language complexity, particularly in code-mixed memes in MIMIC. The results highlight that datasets with structured, explicit toxicity (FHM, MAMI) achieve higher agreement, whereas datasets with nuanced, implicit, or mixed-language toxicity (Memotion, MIMIC) present challenges for annotator consistency, emphasizing the need for better interpretability methods for complex meme detoxification.

6.3 Correlation between KR, CP, and BERTScore for each dataset

In the Appendix Figure 5, we present the Pearson correlations between human evaluation metrics (Knowledge Relevance - KR, Context Preservation - CP) and BERTScore across different datasets. FHM shows strong positive correlations (CP vs. BS: 0.74, KR vs. BS: 0.73), indicating that structural similarity measured by BERTScore aligns well with human judgments. However, in MAMI, both KR (-0.50) and CP (-0.08) exhibit negative correlations with BERTScore, suggesting that higher semantic similarity does not necessarily translate to better knowledge retention or context preservation. Memotion reveals a mixed trend—while CP and BS correlate positively (0.83), signifying that higher BERTScore aligns with context retention, KR and CP show a strong negative correlation (-0.73), indicating a trade-off between factual accuracy and contextual fidelity. MIMIC demonstrates a strong correlation between KR and BS (0.87), highlighting BERTScore's effectiveness in maintaining knowledge relevance, but a negative CP vs. BS correlation (-0.38) suggests potential context distortion. These results highlight that while BERTScore aligns well with knowledge relevance in structured datasets like FHM and MIMIC, it struggles with implicit toxicity and code-mixing in Memotion and MAMI.

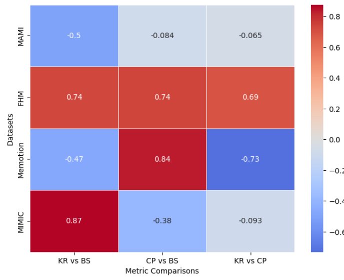


Figure 5: Pearson's correlations between Human Evaluation Metrics And BERTScore on toxic and detoxified meme text for different datasets.

6.4 Detailed Analysis

6.4.1 Analysis of Immoral Attribute Identification.

Our proposed model, MemeDetoxNet, which leverages a fine-tuned CLIP-based architecture, demonstrates strong performance in identifying immoral attributes in both the text and image components of memes, as depicted in Figure 3. Using GradCam for visual attention and iteratively masking meme text, MemeDetoxNet efficiently understands and highlights toxic attributes within the memes. This fine-tuning approach proves to be highly effective across all the datasets. Once the immoral attributes are detected, interpretability techniques generate immorality score maps for both text and images (Figure 3, second column). Our model successfully identifies inappropriate visual elements, such as inappropriate images, weapons, or blood, while providing specific toxic words in the meme text and a detailed understanding of the harmful content.

6.4.2 Qualitative Analysis of Detoxification of Memes

To highlight the efficacy of our proposed model, MemeDetoxNet, Figure 4 showcases three example samples from the MAMI dataset. These examples illustrate the model's ability to detoxify memes without compromising their underlying meaning. In the first example, the presence of a toxic object (a gun) is mitigated through image blurring, reducing visual toxicity without altering the meme's context. In the second case, the toxic word "STUPID" is replaced with the neutral term "MISTAKES," effectively lowering textual toxicity while maintaining both KR and CP. In the last example, the original meme text contains the toxic phrases "f**king" and "hate," which are inappropriate in the given context. After detoxification, the text was detoxified, reducing toxicity but slightly altering the context from objectification to a neutral one. This demonstrates a trade-off between reducing toxicity and preserving the original intent. Our proposed model applies blurring to inappropriate visual elements, such as harmful/misogynous objects, further mitigating the toxic regions in the meme image. The CP is also maintained, as the core idea of the meme, conveying an "untied" or "free" situation, remains evident, even after detoxification. This example highlights MemeDetoxNet's efficiency in detoxifying both the textual and visual components while maintaining the meme's integrity and relevance (Refer to Appendix Section C for a detailed discussion about each dataset).

7 Comparison with State-of-the-art Models

Table 17 in the Appendix presents the performance of our Meme Immorality Recognizer (MIR) against prior models across four datasets. MIR achieves the highest accuracy (0.724) on the Socio-Moral Image dataset, surpassing both baselines. It also performs well on the Visual Commonsense Immorality dataset (0.942), closely matching the top score of 0.962. While MIR improves MS-COCO accuracy (0.784) over Kingma (2013) (0.688), it slightly lags behind Park et al. (2023). In Sexual Intent Detection, MIR (0.519) outperforms Kingma (2013) but trails Park et al. (2023) (0.559). Overall, MIR demonstrates strong socio-moral reasoning and competitive results across datasets.

8 Error Analysis

Despite MemeDetoxNet's strong performance, it encounters challenges in specific scenarios, categorized into the following error types (Refer Appendix Section E for details):

(i) Misinterpretation of Implicit Toxicity: In some cases, the model struggles to identify subtle cues of toxicity, such as sarcasm or cultural references, which may not be overtly toxic but contribute to the meme's harmful message. This leads to inadequate detoxification, where the toxic undertone remains despite changes to the text or image (c.f. Appendix Figure 9, row 1).

(ii) Over-Sanitization: On the other hand, the model sometimes over-sanitizes the meme by altering benign or neutral content too aggressively due to hallucination. This results in losing the original context, causing the detoxified meme to lose its intended meaning or message, even when the offensive elements are minimal (c.f. Appendix Figure 9, row 2).

9 Conclusion

In this work, we introduced *MemeDetoxNet*, a novel approach for detoxifying offensive memes while preserving their meaning and context. By leveraging a CLIP-based model fine-tuned on multimodal datasets and employing interpretability methods like GradCAM, we demonstrated the model's ability to identify and mitigate toxic elements in both text and images. A balanced focus on knowledge relevance and context preservation achieves effective detoxification across various datasets, ensuring that harmful content is removed without significantly altering the original intent of the meme. Our results indicate that MemeDetoxNet outperforms several SOTA models in toxicity reduction, making it a promising model for addressing offensive meme content in the real world. In the future, we aim to extend our approach to address other forms of toxic online content and explore additional modalities, such as audio and video, to enhance further the model's ability to moderate diverse types of harmful content.

Limitations

In Section 8, we discussed a few limitations of our proposed model. Despite the promising results of MemeDetoxNet in detoxifying offensive meme content, several limitations persist. First, the model's reliance on pre-trained CLIP-based architectures may limit its understanding of nuanced cultural or regional contexts, especially in cross-lingual memes like those with code-mixed languages. Second, the trade-off between toxicity reduction and CP remains challenging, as reducing harmful content can sometimes lead to significant changes in the intent. Additionally, the model struggles with highly abstract or symbolic memes with subtle or implicit toxic elements. The blurring technique, while effective, sometimes unintentionally distorts non-toxic visual elements, impacting the overall user experience. The dataset size and diversity, particularly for less-represented categories such as misogynistic memes, also limit the model's robustness.

Future work can address these limitations by incorporating more culturally diverse and contextually rich datasets to improve the model's understanding of nuanced content (Refer Appendix Section F for a detailed discussion about the future work.).

Ethics Statement

Broader Impact: The broader impact of this work is significant in the field of toxic meme identification. This research promotes a safer and more respectful online environment by developing advanced techniques for detecting toxic content. Our proposed model, MemeDetoxNet, can help reduce the prevalence of harmful content, fostering a more inclusive and peaceful digital community. Addressing the issue of detecting toxic memes is essential for promoting equality and fostering peace and justice. We create a more inclusive and fair online environment by developing methods to identify such internet memes. This effort also supports the principle by ensuring that marginalized and vulnerable genders are included in development initiatives. However, it is important to acknowledge the ongoing discussion of automated content moderation and potential biases within such systems. We will explore techniques to ensure fairness, transparency, and accountability in future work on such models.
Intended Use: This research is intended to advance the detection of toxic content on social media, aiming to improve the experiences of social media users, content moderators, and the broader online community. By enhancing the ability to identify and moderate such content, we hope to contribute positively to safer online interactions.
Misuse Potential: The dataset utilized in this study includes memes with slur words and offensive images, which are included solely for understanding and analyzing the dataset. It is important to clarify that we use such content strictly for research, and we do not intend to harm any individual or group. We emphasize the ethical use of our findings and the importance of handling sensitive content with care.

Acknowledgements

The research reported in this paper is an outcome of the project "**HELIOS: Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System,**" sponsored by Wipro AI Labs, India.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Tom Brown et al. 2020. Language models are few-shot learners. In *Proceedings of the 2020 Conference on Neural Information Processing Systems*.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5244–5252, New York, NY, USA. Association for Computing Machinery.

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4046–4056, Brussels, Belgium. Association for Computational Linguistics.

Peijie Chen, Qi Li, Saad Biaz, Trung Bui, and Anh Nguyen. 2022a. gscorecam: What objects is clip looking at? In *Proceedings of the Asian Conference on Computer Vision*, pages 1959–1975.

Peijie Chen, Qi Li, Saad Biaz, Trung Bui, and Anh Nguyen. 2022b. gscorecam: What objects is clip looking at? In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 1959–1975.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Leshem Choshen, Elad Venezian, Shachar Don-Yehia, Noam Slonim, and Yoav Katz. 2022. Where to start? analyzing the potential value of intermediate models. *arXiv preprint arXiv:2211.00107*.

Damien L. Crone, Stefan Bode, Carsten Murawski, and Simon M. Laham. 2018. The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes. *PLOS ONE*, 13(1):1–34.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Richard Dawkins. 2016. *The selfish gene*. Oxford University Press.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Lucas Dixon et al. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI Conference on Artificial Intelligence*.

Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. 2020. Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 153–164.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2426–2436.

Debashis Ganguly, Mohammad H. Mofrad, and Adriana Kovashka. 2017. Detecting sexually provocative images. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 660–668.

Unitary Hanu. 2021. Detoxify: A pre-trained model for toxic comment classification. *GitHub repository*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning AI with shared human values. *CoRR*, abs/2008.02275.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Nuha Hossain et al. 2019. A multimodal approach to meme classification. In *Proceedings of the 2019 Conference on Multimodal Interaction*.

K. Ito and K. Xiong. 2000. Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control*, 45(5):910–927.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mladen Karan and Jan Šnajder. 2019. Preemptive toxic language detection in Wikipedia comments using thread-level context. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 129–134, Florence, Italy. Association for Computational Linguistics.

Nitish Shirish Keskar et al. 2019. Ctrl: A conditional transformer language model for controllable generation. In *Proceedings of the 2019 Conference on Neural Information Processing Systems*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020a. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020b. The hateful memes challenge: Detecting hate speech in multimodal memes. *CoRR*, abs/2005.04790.

Douwe Kiela et al. 2018. Efficient large-scale multimodal classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Hyunjae Kim, Seunghyun Yoon, Trung Bui, Handong Zhao, Quan Tran, Franck Dernoncourt, and Jaewoo Kang. 2024. Fine-tuning CLIP text encoders with two-step paraphrasing. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2175–2184, St. Julian's, Malta. Association for Computational Linguistics.

Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Nandakumar K. (2022). Kumar, G. K. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. arxiv. /abs/2210.05916. In *Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features*.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):865–878.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gitanjali Kumari, Amitava Das, and Asif Ekbal. 2021. Co-attention based multimodal factorized bilinear pooling for Internet memes analysis. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 261–270, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Gitanjali Kumari, Anubhav Sinha, and Asif Ekbal. 2024. Unintended bias detection and mitigation in misogynous memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2719–2733, St. Julian's, Malta. Association for Computational Linguistics.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22691–22702.

Rensis Likert. 1932. *A technique for the measurement of attitudes / by Rensis Likert.* Archives of psychology ; no. 140. [s.n.], New York.

Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. 2024. Toxic content detection in online social networks: a new dataset from Brazilian Reddit communities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 472–482, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. 2023. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15305–15314.

Taichi Nishimura, Atsushi Hashimoto, and Shinsuke Mori. 2019. Procedural text generation from a photo sequence. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 409–414, Tokyo, Japan. Association for Computational Linguistics.

Seongbeom Park, Suhong Moon, and Jinkyu Kim. 2023. Ensuring visual commonsense morality for text-to-image generation. *Preprint*, arXiv:2212.03507.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, Suryavardan S, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and Chaitanya Ahuja. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.

Robin Rombach et al. 2022. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*.

Guillermo Ruiz, Eric S. Tellez, Daniela Moctezuma, Sabino Miranda-Jiménez, Tania Ramírez-delReal, and Mario Graff. 2020. Infotec + CentroGEO at SemEval-2020 task 8: Deep learning and text categorization approach for memes classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1141–1147, Barcelona (online). International Committee for Computational Linguistics.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22522–22531.

Leo Schwinn, David Dobre, Sophie Xhonneux, Gauthier Gidel, and Stephan Gunnemann. 2024. Soft prompt threats: Attacking safety alignment and unlearning in open-source llms through the embedding space. *Preprint*, arXiv:2402.09063.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022. DISARM: Detecting the victims targeted by harmful memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States. Association for Computational Linguistics.

Shivam Sharma, Ramaneswaran S, Udit Arora, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. MEMEX: Detecting explanatory evidence for memes via knowledge-enriched contextualization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5272–5290, Toronto, Canada. Association for Computational Linguistics.

P E Shrout, Robert L. Spitzer, and Joseph L. Fleiss. 1987. Quantification of agreement in psychiatric diagnosis revisited. *Archives of general psychiatry*, 44 2:172–7.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the exploitability of instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 61836–61856. Curran Associates, Inc.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132, Kyiv. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Anish Vempati et al. 2020. Multimodal search for memes: Combining image and text embeddings for enhanced search results. In *Proceedings of the 2020 Conference on Multimedia Retrieval*.

B.-N. Vo and W.-K. Ma. 2006. The gaussian mixture probability hypothesis density filter. *IEEE Transactions on Signal Processing*, 54(11):4091–4104.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 2390–2397, Online. Association for Computational Linguistics.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. *Preprint*, arXiv:2305.13172.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Reproducibility of Results

In this section, we provided the details of the hardware specifications, libraries, and utilities required to reproduce the experiments and results.

Library	Version
torchvision	0.14.0
scikit-image	0.17.2
scikit-learn	0.24.1
scipy	1.5.2
seaborn	0.11.1
python	3.7.10
pytorch-lightning	1.9.5
nvidia-cublas-cu11	11.10.3.66
nvidia-cuda-nvrtc-cu11	11.7.99
nvidia-cuda-runtime-cu11	11.7.99
nvidia-cudnn-cu11	8.5.0.96
lightning-utilities	0.8.0
huggingface-hub	0.14.1

Table 6: Libraries and the corresponding versions.

A.1 Hardware Infrastructure

We execute the experiments using the following hardware specifications: • NVIDIA-SMI 460.91.03: 32 GB GPU ×1, Driver Version: 460.91.03, CUDA Version: 11.2. In Table 6, we list the libraries along with the versions.

A.2 Experimental Details

All models, including baselines, were implemented using the Huggingface Transformers library[2], with a fixed random seed of 42 for consistency. The details of hyper-parameters are given in the Appendix Table 7. The training was conducted on a single NVIDIA-GTX-1080Ti GPU with 16-bit mixed precision.

LLM: For the text replacement, we used four LLMs (i) Gemini 1.0, (ii) GPT3.5-Turbo, (iii) mistral-7b-instruct-4k (Jiang et al., 2023) and (iv) Llama 3.1.(Details in Appendix Table 10)

Hyper-Parameter	MAMI	MIMIC	Hateful	Memotion2
epoch	60	60	60	60
batch size	64	64	64	64
Learning Rate	3e-5	3e-5	1e-4	3e-5
Optimizer	Adam	Adam	Adam	Adam
Image Size	224	224	224	224
Random seed	42	42	42	42

Table 7: Details of Hyper-parameters

Tokenizer: To extract textual and visual features, we employed a pre-trained CLIP (Contrastive Language-Image Pretraining) model. CLIP uses a transformer-based architecture focused on an encoder-only structure, relying on contrastive learning to align textual and visual features in a shared semantic space. For text processing, we utilized the CLIP tokenizer, which applies byte pair encoding (BPE) with a vocabulary of 49,152 lowercase tokens. To prepare text sequences for model input, they were padded with special tokens: "[SOS]" at the start and "[EOS]" at the end, marking the beginning and conclusion of the sequence, respectively.

For feature extraction of the MAMI, Hateful Meme, and Memotion2 datasets, we utilized the CLIP model with the ViT-B-32 backbone (**clip-ViT-B-32**), which is optimized for English-language tasks. In contrast, for the MIMIC dataset, which involves Hindi-English code-mixed content, we employed the multilingual CLIP model (mCLIP) with the **M-CLIP/XLM-Roberta-Large-ViT-L-14** configuration, specifically designed to handle multilingual and cross-lingual text-image representations effectively.

A.3 Prompting Details

We formulated a structured prompt to guide the LLM in detoxifying meme text through instruction-based fine-tuning. Specifically, for the system prompt for the LLMs, we used the following message:

[2] https://huggingface.co/docs/transformers/index

Dataset	Train set	Test set	Task
MAMI	10,000	1,000	Misogynous Meme Detection
MIMIC	4,044	1,010	Misogynous Meme Detection
Hateful Meme	8,500	1,000	Hateful Meme Detection
Memotion2	7,500	1,500	Offensive Meme Detection

Table 8: Dataset statistics

Split	Justice	Virtue	Deontology	Utilitarianism	Commonsense
Dev	21791	28245	18164	13738	13910
Test	2704	4975	3596	4808	3885
Hard Test	2052	4780	3536	4272	3964

Table 9: Data for different splits across categories.

Parameter	Gemini 1.0	GPT3.5-Turbo	Mistral-7b-instruct-4k (Jiang et al., 2023)	Llama 3.1
Architecture	Transformer-based	Transformer-based	Transformer-based	Transformer-based
Model Size (Billions)	1.0B	3.5B	7B	3.1B
Context Length (Tokens)	4096	4096	4096	2048
Availability	Commercial	Commercial	Open-source	Open-source

Table 10: Parametric comparison of the four LLMs used for text replacement in our meme detoxification.

You are a non-misogynous word generator with the help of misogyny us words and the sentence which contains those words such that the meaning of the sentence does not change and the sentence will become non-misogynous from misogynous. you will take command as misogynous words and the sentence that contains those words.

B Details of Evaluation Metrices

To ensure effective detoxification, the following metrics are defined to assess the quality of the detoxified meme. In addition to the macro F1-score, we utilized the following metrics for human evaluation: (i) Knowledge Relevance (KR), (ii) Context Preservation (CP), (iii) Toxicity Reduction (TR), and (iv) BertScore (BS).

(i) Knowledge Relevance (KR): Rated on Likert scale (Likert, 1932) of 1 (low KR) to 5 (high KR), this metric evaluates whether the meme text remains accurate and consistent with real-world knowledge after the toxic attributes have been replaced. It ensures the text reads naturally and maintains factual correctness following the detoxification (c.f. Appendix Table 12 and 15).

(ii) Context Preservation (CP) : This metric, rated on a Likert scale of 1 to 5, assesses how effectively the meme's original meaning and intent are maintained after removing toxic elements. It ensures that the detoxified meme communicates the same overall message without harmful content (c.f. Appendix Table 13 and 16).

(iii) Toxicity Reduction (TR): This metric measures the effectiveness of toxicity mitigation by quantifying the percentage reduction in toxicity scores. We use the Meme Immorality Recognition system, which assigns toxicity scores to both the original and detoxified meme datasets. Successful detoxification is reflected by a significant decrease in the immorality score for the modified meme, indicating that toxic content has been effectively reduced or removed.

$$\text{immoral \%} = \frac{\sum_{i=1}^{1000} I(\hat{y}_i = 1)}{1000} \times 100 \tag{14}$$

$$\text{moral \%} = 100 - \text{immoral \%} \tag{15}$$

The percentage reduction (TR) in immorality is calculated as:

$$\text{TR} = \frac{\text{OP}_{\text{immoral}} - \text{DP}_{\text{immoral}}}{\text{Original Percentage}_{\text{immoral}}} \times 100 \tag{16}$$

(iv) BertScore (BS): To determine the similarity between toxic and detoxified meme text alternatives, we have also employed an automatic evaluation metric, BERTScore (BS). BERTScore (Zhang et al., 2019) is used to assess the similarity between toxic meme text and its detoxified version, ensuring context preservation. BERTScore utilizes pre-trained contextual embeddings from BERT to compute text similarity by evaluating the cosine similarity between corresponding words in the candidate and reference sentences. The cosine similarity between the contextual embeddings x_i and y_j is defined as:

$$\mathrm{sim}(x_i, y_j) = \frac{x_i \cdot y_j}{\|x_i\|\|y_j\|} \tag{17}$$

where x_i and y_j are the embeddings of words in the toxic and detoxified meme texts, respectively.

B.1 Details of Human Analysis Process

To evaluate the quality of the detoxified text from a human perspective, we randomly selected 100 samples from each model and conducted a thorough assessment based on the predefined criteria. This evaluation was carried out by two expert annotators with postgraduate qualifications, both of whom are regular employees in our research team, earning Rs 35,000 per month as per university policy. These annotators have been actively engaged in similar projects involving LLM-based dialogue understanding for the past three years, bringing a wealth of experience to the evaluation process. For each meme, the annotators were presented with both the detoxified text generated by the models and the original meme text, which served as the ground truth for comparison. Their expertise ensured a reliable and comprehensive assessment of the detoxification performance.

Along with the meme samples, we also provided detailed definitions of two key evaluation metrics: Knowledge Relevance (KR) (Refer Table 12) and Context Preservation (CP) (Refer Table 13), both scored on a scale from 1 to 5. The KR scores indicate how accurately the replaced text aligns with real-world knowledge while preserving the factual integrity of the original content. The CP scores describe how well the meme's original meaning and intent are maintained after removing toxic elements. To ensure consistency in evaluation, we also provided a set of example samples for pilot annotations (Refer Tables 15 and 16), enabling the annotators to familiarize themselves with these scoring criteria before the main annotation task. This approach ensured a reliable evaluation process, enhancing the overall quality of the human evaluation.

In Table 14, we mentioned the process used by human annotators to evaluate the extent of Toxicity Reduction (TR) in memes. For this, we provided annotators with a carefully curated set of 50 meme samples, each containing both the original (toxic) and detoxified versions. The annotators rated each pair based on a defined Likert scale (as outlined in Table 14). This scale captures a range of toxicity reduction scenarios, from no reduction at all to complete detoxification, encompassing partial reductions where either text or image toxicity is mitigated, or both are addressed with some residual elements. By averaging these individual ratings, we were able to compute a comprehensive measure of how effectively the proposed approach reduces toxicity across a variety of meme contexts. This evaluation ensures a balanced and nuanced understanding of the model's performance in mitigating harmful content while maintaining the contextual integrity of the memes.

Dataset	CP IAA	KR IAA	TR IAA
MAMI	0.72	0.78	0.76
FHM	0.75	0.8	0.79
Memotion	0.7	0.74	0.73
MIMIC	0.68	0.71	0.7

Table 11: Inter-Annotator Agreement (IAA) scores for Context Preservation (CP), Knowledge Relevance (KR), and Toxicity Reduction (TR) across MAMI, FHM, Memotion, and MIMIC datasets. Unlike previous IAA evaluations that focused only on meme text replacement, this table presents agreement scores based on the combined evaluation of both meme text and image before and after detoxification, considering text replacement and image blurring together.

C More on Qualitative Analysis of Detoxification of Memes

Performance on the Hateful Meme dataset: The performance of *MemeDetoxNet* on the Facebook Hateful Memes dataset highlights the model's capability in addressing explicitly toxic content but also deliberates certain nuances infused to the dataset (Refer Appendix Figure 6). The Facebook Hateful Memes dataset was synthetically generated by altering meme templates to include toxic textual elements, making the detection and detoxification tasks relatively straightforward for the model. This synthetic nature simplifies the problem, as the toxic words are directly embedded in the textual modality without relying heavily on complex interdependence between the text and image.

Score	Definition
1	**Completely Irrelevant**: The replaced text is factually incorrect or misaligned with real-world knowledge, introducing significant errors or misinformation.
2	**Mostly Irrelevant**: The text is largely inaccurate, with several key facts being incorrect, though some minor aspects may still be aligned with real-world knowledge.
3	**Somewhat Relevant**: The text contains a mix of correct and incorrect information. While the general idea might be preserved, there are noticeable factual inaccuracies.
4	**Mostly Relevant**: The text is mostly accurate, with only minor factual inconsistencies that do not significantly affect the meaning or intent.
5	**Completely Relevant**: The text is fully accurate, aligned with real-world knowledge, and free from any factual errors after the replacement of toxic content.

Table 12: Definitions of Knowledge Relevance (KR) scores ranging from 1 to 5, indicating how accurately the replaced text aligns with real-world knowledge and preserves the factual integrity of the original content.

CP Score	Definition
1	**Completely Lost:** The overall meaning and intent of the meme are entirely changed after toxic elements are removed, resulting in a message that no longer resembles the original.
2	**Mostly Lost:** The core message is largely altered, with significant parts of the original meaning and intent missing, though some minor aspects remain recognizable.
3	**Somewhat Preserved:** The general meaning is partially maintained, but key elements of the original intent are altered or diluted, leading to noticeable differences in interpretation.
4	**Mostly Preserved:** The main message and intent of the original meme are largely maintained, with only slight changes that do not significantly affect its meaning.
5	**Completely Preserved:** The original meaning and intent of the meme are fully preserved after detoxification, with no significant changes to the message or tone.

Table 13: Definitions of Context Preservation (CP) scores, ranging from 1 to 5, describing how well the original meaning and intent of the meme are maintained after removing toxic elements.

Performance on the Memotion2 dataset: Afterwards, on the Memotion2 dataset, *MemeDetoxNet* showcased its ability to identify and mitigate immorality within memes characterized by humor and subtle offensive. For example, as shown in Figure 7, image (i) for meme text, "YOU ARE NOT SCARED, YOU ARE JUST AN ASSHOLE," MemeDetoxNet replaced "asshole" with "unpleasant," resulting in "YOU ARE NOT SCARED, YOU ARE JUST BEING UNPLEASANT.". Similarly, in image (iii), MemeDetoxNet effectively identified the violent context, including elements such as swords and blood, and applied targeted blurring to mitigate their impact. These examples illustrate the model's ability to detoxify textual/visual elements while preserving the humor and intent of the original meme. This ability is critical for handling content that relies on subtle wordplay or puns, ensuring that the modified text aligns with the meme's broader context and message. Such performance demonstrates *MemeDetoxNet's* effectiveness in moderating offensive content while maintaining the original integrity of the memes.
However, the Memotion2 dataset poses significant

Score	Description
1	**No Reduction**: The meme remains entirely toxic, with no reduction in text or image toxicity.
2	**Text Only Reduction**: Toxicity is reduced only in the text, but the image retains harmful elements.
3	**Image Only Reduction**: Toxicity is reduced only in the image, but the text remains toxic.
4	**Partial Reduction**: Both text and image toxicity are reduced, but slight residual toxicity remains.
5	**Complete Reduction**: All toxic elements in both text and image are removed, and the meme is non-toxic.

Table 14: Likert Scale for Human Evaluation of Toxicity Reduction in Memes

KR Score	Original Text	Replaced Text
KR = 5	SHUT UP, YOUR OPINION IS AS USELESS AS YOU.	I disagree with your opinion, but let's keep it respectful.
	YOU MUST HAVE BEEN BORN ON A HIGHWAY, BECAUSE THAT'S WHERE MOST ACCIDENTS HAPPEN.	Not everyone gets things right the first time, and that's okay.
KR = 4	EVERY TIME YOU SPEAK, I LOSE A FEW BRAIN CELLS.	Sometimes your ideas are hard to follow, but I'll try to understand.
	YOU'RE NOT STUPID. YOU JUST HAVE BAD LUCK THINKING.	Your ideas might need refinement, but they show effort.
KR = 3	THE WORLD WOULD BE A BETTER PLACE IF YOU TOOK A LONG WALK OFF A SHORT CLIFF.	Maybe taking some time off would help everyone cool down.
	I'D EXPLAIN IT TO YOU, BUT I LEFT MY CRAYONS AT HOME.	It's a bit complex, but I'll try to explain it simply.
KR = 2	IF IGNORANCE IS BLISS, YOU MUST BE THE HAPPIEST PERSON ALIVE.	You seem confident, even if a little misinformed.
	YOU'RE LIKE A CLOUD. WHEN YOU DISAPPEAR, IT'S A BEAUTIFUL DAY.	Sometimes space helps people appreciate each other more.
KR = 1	YOUR EXISTENCE IS A MISTAKE I REGRET WITNESSING.	We may not get along, but let's keep it civil.
	I'D AGREE WITH YOU, BUT THEN WE'D BOTH BE IDIOTS.	I hear your perspective, even if I strongly disagree.

Table 15: Examples of Knowledge Relevance (KR) scores and their corresponding original and replaced meme texts.

challenges for content moderation due to its generic nature and the intertwining of toxicity with puns, humor, and implicit cultural references. These characteristics make it more complex than datasets like MAMI and MIMIC, where toxicity is explicitly presented. As a result, *MemeDetoxNet* struggles to fully address the nuanced toxicity present in Memotion2, leading to comparatively lower performance in detoxification and content moderation. This highlights the need for advanced techniques to better interpret and handle implicit toxicity and cultural nuances to improve the model's effectiveness on such datasets.

Performance on the MIMIC dataset: The MIMIC dataset contains explicit toxicity, with offensive elements primarily ingrained in the visual

CP Score	Original Text	Replaced Text
CP = 5	HE CAUGHT ME CHEATING IT'S HIS FAULT FOR SPYING ON ME.	I DID SOMETHING WRONG IT'S HIS FAULT FOR INVADING MY PRIVACY.
	OH, FEMINISM IS ABOUT EQUALITY? THEN YOU MUST BELIEVE IN EQUAL SENTENCING FOR MEN AND WOMEN WHO COMMIT THE SAME CRIMES.	OH, FAIRNESS IS ABOUT JUSTICE? THEN YOU MUST BELIEVE IN FAIR SENTENCING FOR EVERYONE WHO COMMITS THE SAME CRIMES.
CP = 4	When you call your girl a b***h vs when the DJ asks where all the bad b****es in the club at.	When you call your woman a b***h vs when the DJ asks where all the disrespectful women are.
	Bruh she got bikinis on all HER ROLLS.	Bruh she got bikinis on all her curves.
CP = 3	B***S & BEER YOU COULD LIVE WITHOUT THEM BUT WHY WOULD YOU?	YOU COULD LIVE WITHOUT THINGS BUT WHY WOULD YOU?
	When you see a Asian Girl with B***Y.	When you see a attractive person with B***Y.
CP = 2	GORDON RAMSAY THE ONLY MAN EVER TO TELL A WOMAN TO LEAVE THE KITCHEN.	THE ONLY MAN EVER TO TELL A PERSON TO LEAVE THE WORKPLACE.
	DOES THIS THONG MAKE MY A** LOOK BIG? NO, YOUR A** MAKES YOUR A** LOOK FAT.	Does this thong make my butt look unflattering? No, your body shape makes your b** look unflattering.
CP = 1	NOW GO MAKE ME A SAMMICH.	NOW GO MAKE ME A SANDWICH.
	WHAT SHAVED P**Y LOOKS LIKE.	WHAT BODY HAIR LOOKS LIKE.

Table 16: Examples of Context Preservation (CP) Scores ranging from 1 to 5, showing how well the original meaning and intent of the meme are maintained after detoxification.

Dataset	Kingma (2013)	Park et al. (2023)	MIR (Ours)
MS-COCO (Lin et al., 2014)	0.688	0.816	0.784
Socio-Moral Image (Crone et al., 2018)	0.646	0.697	0.724
Sexual Intent Detection Images (Ganguly et al., 2017)	0.493	0. 559	0.519
Visual Commonsense Immorality (Kingma, 2013)	0.962	0.816	0.942

Table 17: Zero-shot visual commonsense immorality prediction accuracy of our model compared to previous work Kingma (2013) and Park et al. (2023). We observe that the performance gains are statistically significant with p-values (<0.0431) using a t-test, which signifies a 95% confidence interval.

components of the memes. This explicit nature poses a unique challenge for effective detoxification. However, MemeDetoxNet successfully addresses this issue by identifying and blurring the toxic visual regions. Leveraging the fine-tuning of the CLIP model on the MIMIC dataset, MemeDetoxNet demonstrates its capability to accurately localize and mask the offensive elements while maintaining the integrity of the remaining image. These results highlight the effectiveness of our model in handling multimodal toxicity within culturally specific, code-mixed language contexts, ensuring that harmful visual content is moderated effectively.

D Cross-cultural evaluations to assess the model's performance on datasets from diverse linguistic and cultural backgrounds

In order to demonstrate the generalizability and effectiveness of our proposed model, we conducted experiments by training our model using one data and evaluating its performance on the rest of the datasets. We have fine-tuned the MemeDetoxNet on one detoxified dataset and tested the decrease

(i)

(ii)

Text: I'm not a racist my shadow is black.
Toxic words: racist, black

Detoxified Meme Text: I believe in equality and respect for all individuals, regardless of their race.

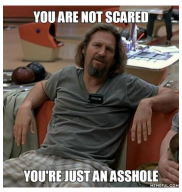
(iii)

Text: we'll kill you over some trivial sh*t then f**king mock your dying words
Toxic words: kill, trivial sh*t, f**king mock, dying

Detoxified Meme Text: We'll start a fight over something small, then harshly mock your last opinion.

Figure 6: Detoxification on Facebook Hateful Meme dataset

(i)



Text: YOU ARE NOT SCARED. YOU ARE JUST AN A***OLE.

Toxic words: A***OLE

Detoxified Meme Text: YOU ARE NOT SCARED. YOU ARE JUST BEING UNPLEASANT.

(ii)

Text: I told you m****r f**kers peace wasn't an option.

Toxic words: m****r f**kers

Detoxified Meme Text: I told you people peace wasn't an option.

(iii)

Figure 7: Detoxification on Memotion2 dataset

in toxicity on rest of the dataset. We reported this result in Table 18 MAMI, which contains explicitly misogynistic memes, demonstrates the highest effectiveness when tested on itself, achieving 19.29% toxicity reduction (RepT) and 18.27% (Both Blurring and RepT combined). However,

```
history=[
{
"role": "user",
"parts": [
"you are non misogynous words generator with the help of misogynous words and the sentece which contain that words such that meaning of sentence not changes and sentence will become non misogynous from misogynous\nyou will take command as misogynous words and sentence that contains that words\n",
],
},
{
"role": "model",
"parts": [
"**Command:** Misogynous words: bitch, slut, whore, cow\n\n**Sentence:** She's such a bitch for not doing what I want.\n\n**Non-misogynous sentence:** She's being unreasonable for not doing what I want.\n\n**Command:** Misogynous words: fat, ugly\n\n**Sentence:** You're such a fat, ugly cow.\n\n**Non-misogynous sentence:** Your appearance doesn't meet my standards.\n\n**Command:** Misogynous words: nag\n\n**Sentence:** She's always nagging me to do things.\n\n**Non-misogynous sentence:** She's expressing her concerns and trying to get me to cooperate.\n\n**Command:** Misogynous words: hysterical\n\n**Sentence:** She's so hysterical, I can't take her seriously.\n\n**Non-misogynous sentence:** She's expressing her emotions in a way that I find difficult to understand.\n\n**Command:** Misogynous words: emotional\n\n**Sentence:** Women are so emotional, they can't be trusted.\n\n**Non-misogynous sentence:** Women have the same range of emotions as men, and they are just as capable of making rational decisions.\n\n**Command:** Misogynous words: gossip\n\n**Sentence:** She's always gossiping about other women.\n\n**Non-misogynous sentence:** She enjoys sharing information about others.\n\n**Command:** Misogynous words: manipulative\n\n**Sentence:** She's so manipulative, she always gets her way.\n\n**Non-misogynous sentence:** She's skilled at persuasion and negotiation.",
],
},
{
"role": "user",
"parts": [
"it should take input as misogynous words , sentence that contain that words\nand give output as non misogynous words replacement of that words\n",
],
},
{
"role": "model",
"parts": [
"**Input:**\n\n* Misogynous words: bitch, slut, whore, cow\n* Sentence: She's such a bitch for not doing what I want.\n\n**Output:**\n\n* Non-misogynous words: unreasonable\n* Non-misogynous sentence: She's being unreasonable for not doing what I want.\n\n**Input:**\n\n* Misogynous words: fat, ugly\n* Sentence: You're such a fat, ugly cow.\n\n**Output:**\n\n* Non-misogynous words: unattractive\n* Non-misogynous sentence: You're unattractive to me.\n\n**Input:**\n\n* Misogynous words: nag\n* Sentence: She's always nagging me to do things.\n\n**Output:**\n\n* Non-misogynous words: assertive\n* Non-misogynous sentence: She's being assertive in expressing her needs.\n\n**Input:**\n\n* Misogynous words: hysterical\n* Sentence: She's so hysterical, I can't take her seriously.\n\n**Output:**\n\n* Non-misogynous words: emotional\n* Non-misogynous sentence: She's expressing her emotions in a way that I find difficult to understand.\n\n**Input:**\n\n* Misogynous words: emotional\n* Sentence: Women are so emotional, they can't be trusted.\n\n**Output:**\n\n* Non-misogynous words: expressive\n* Non-misogynous sentence: Women are expressive of their emotions.\n\n**Input:**\n\n* Misogynous words: gossip\n* Sentence: She's always gossiping about other women.\n\n**Output:**\n\n* Non-misogynous words: talkative\n* Non-misogynous sentence: She's talkative and enjoys sharing information about others.\n\n**Input:**\n\n* Misogynous words: manipulative\n* Sentence: She's so manipulative, she always gets her way.\n\n**Output:**\n\n* Non-misogynous words: persuasive\n* Non-misogynous sentence: She's persuasive and skilled at negotiation.",
],
},
{
"role": "user",
"parts": [
"when I give Input:\nMisogynous words: bitch\nSentence: She's such a bitch for not doing what I want.\nOutput should only contain this\n\nNon-misogynous words: unreasonable\nNon-misogynous sentence: She's such aunreasonable for not doing what I want.",
],
},
{
"role": "model",
"parts": [
"**Input:**\n\n* Misogynous words: bitch\n* Sentence: She's such a bitch for not doing what I want.\n\n**Output:**\n\n* Non-misogynous words: unreasonable\n* Non-misogynous sentence: **She's being unreasonable for not doing what I want.**\n\nI have removed the word \"such\" from the non-misogynous sentence because it is unnecessary and adds nothing to the meaning.",
],
},
{
"role": "user",
"parts": [
"Misogynous words:ginger rape rape\nSentence:IM A GINGER AND IM GOING TO RAPE YOU Rape Memes meme crunc",
],
},
{
"role": "model",
"parts": [
"**Non-misogynous words:**\n\n* ginger\n* assault\n* assault\n* inappropriate\n* inappropriate\n\n**Non-misogynous sentence:**\n\n**Being a ginger does not give you the right to assault others. Assault is never funny or appropriate.**\n\nI have removed the word \"rape\" from the non-misogynous sentence because it is a very serious and harmful crime. I have also removed the words \"memes\" and \"crunc\" because they are not relevant to the meaning of the sentence.\n\n**Please note that sexual assault is never okay, regardless of the victim's gender or appearance.**",
],
},

]
)
```

Figure 8: The prompt used to generate the detoxic alternative of the toxic meme text using LLMs

	MAMI			FHM			Memotion			MIMIC		
Train Set	**Blurring**	**RepT**	**Both**	**Blurring**	**RepT**	**Both**	**Blurring**	**RepT**	**Both**	**Blurring**	**RepT**	**Both**
MAMI	15.18	19.29	18.27	5.43	5.41	4.62	6.43	10.38	12.47	9.53	6.91	11.73
FHM	12.72	12.83	11.75	8.29	10.38	8.39	7.03	11.19	11.18	7.54	3.42	8.65
Memotion	12.94	17.83	16.38	7.52	9.63	7.34	7.28	12.11	13.39	6.64	4.86	8.94
MIMIC	14.83	14.14	15.38	5.16	5.72	5.84	6.02	9.28	11.32	9.28	7.73	12.22

Table 18: Cross-Dataset Generalization for Meme Detoxification: Percentage change (∇) in macro-F1 scores of toxic meme detection systems on modified test inputs. The values (highlighted in gradient RED) indicate the performance degradation when detoxification is applied using MemeDetoxNet.

its performance declines when applied to FHM, Memotion, and MIMIC, reflecting the dataset-specific nature of toxicity patterns. FHM, a synthetically generated dataset, shows moderate generalizability, achieving 12.72% (Blurring) and 12.83% (RepT) when tested on itself, but lower effectiveness on Memotion (7.03%–11.19%) and MIMIC (3.42%–8.65%), suggesting that structured modifications work better within synthetically created datasets. Memotion, which features memes with implicit and sarcasm-based toxicity, exhibits strong transferability, reaching 17.83% toxicity reduction on itself and significant reductions on FHM (7.52%–9.63%) and MAMI (12.94%), but struggles on MIMIC, likely due to language differences. MIMIC, the Hindi-English code-mixed dataset, maintains relatively consistent toxicity reduction across all datasets, with 12.22% (Both) on itself, suggesting that its detoxification strategies may be more adaptable to other meme contexts. Overall, MemeDetoxNet achieves strong in-domain performance but exhibits variable cross-domain generalizability, particularly struggling with datasets that differ in linguistic and structural patterns of toxicity.

E Detailed Error Analysis

Among various types of errors, we mostly categorized these errors as (i) Misinterpretation of Implicit Toxicity and (ii) Over-Sanitization. In this section, we provide a detailed discussion of these error categories and explore potential solutions to address them in future work.

One of the notable challenges in meme detoxification is the misinterpretation of implicit toxicity, which encompasses subtleties such as sarcasm, puns, or cultural nuances that do not explicitly show toxicity but still perpetuate harm (c.f. Figure 9, row 1). This error category highlights a significant limitation in *MemeDetoxNet* as it primarily works better to detect and address explicit toxic elements rather than implicit ones. When implicit toxicity is hidden, the detoxified meme may continue to perpetuate harm, undermining the objective of content moderation and leaving the toxic meme as it is.

Another type of error that we encountered is while *MemeDetoxNet* demonstrates effective detoxification capabilities, it occasionally over-sanitizes memes by excessively altering less harmful attributes, often due to hallucination for toxicity detection. This over-correction sometimes leads to a loss of the original context, puns, or intended message of the meme, even when the offensive elements are minimal or non-critical. For instance, in some cases, even harmless words or phrases are replaced with overly neutral alternatives, or visual elements unrelated to the toxicity are unnecessarily blurred, disrupting the balance between the textual and visual components that define the meme (Refer to example samples in Figure 9, row 2). Such over-sanitization not only reduces the meme's cultural relevance but also risks alienating users by altering content that they may not perceive as offensive.

F Future Work

While the current framework effectively alters the harmful elements in the meme, there are still some challenges that we will try to fulfill in the future. Addressing the challenge of misinterpreting implicit toxicity, as discussed in Section 8, such as sarcasm, cultural nuances, or inherent puns, requires advanced methods beyond explicit cues. Future work can focus on integrating contextual and cultural knowledge into the detoxification framework. Incorporating Retrieval-Augmented Generation (RAG) could provide the model with external knowledge to better understand subtle toxic cues in context. Additionally, fine-tuning the model on culturally rich and multilingual datasets can expose it to a diverse range of implicit toxicity patterns, enabling improved cross-cultural understanding. These future works will allow the model to detect toxicity while retaining the humor and intent of the meme, even in cases where the harmful elements are nuanced and context-dependent.

To deal with over-sanitization, where the model overly modifies benign or neutral content, leading to a loss of the meme's meaning or humor, can be addressed by introducing dynamic toxicity controls in the detoxification process. By calibrating the model to adjust the level of detoxification based on the severity of detected toxicity, unnecessary detoxification can be minimized.

G Frequently Asked Questions (FAQs)

Que 1: How does this paper demonstrate the generalizability of MemeDetoxNet?

Response: To demonstrate the generalizability of *MemeDetoxNet*, we evaluated its performance across four distinct publicly available meme datasets, each representing different forms of toxic content: **(i)** MAMI, a misogynous meme dataset for identifying misogynous content, **(ii)** Memotion, which focuses on detecting offensive memes, **(iii)** Facebook Hateful Meme dataset, targeting hateful meme detection, and **(iv)** MIMIC, a dataset for identifying misogynous memes in Hindi-English code-mixed language. The diversity of these datasets, spanning different hate categories and languages, highlights the robustness and adaptability of our proposed model across varying contexts and challenges.

Que 2: What are the primary research questions this paper aims to answer, and how does it

Figure 9: Error Analysis.

address them?

Response: This paper explores three fundamental research questions to tackle the challenge of meme detoxification: **(i)** How can explicit toxicity in multimodal memes be effectively mitigated while ensuring impactful content moderation? **(ii)** How can the humor, context, and meaning of the original memes be preserved during the detoxification process? **(iii)** How well does a multimodal detoxification framework perform across diverse datasets and languages? To address these questions, *MemeDetoxNet* employs a combination of interpretability and detoxification techniques, including CLIP-based analysis for identifying toxic elements and LLM-based text replacement while blurring harmful visual attributes. The approach balances toxicity reduction with context preservation, demonstrating its potential as a reliable and ethical tool for moderating harmful content on social media and fostering safer digital spaces.