Reducing the Emotional Distress of Content Moderators through LLM-based Target Substitution in Implicit and Explicit Hate-Speech

Nazanin Jafari
University of Massachusetts Amherst
Amherst, USA
nazaninjafar@cs.umass.edu

James Allan
University of Massachusetts Amherst
Amherst, USA
allan@cs.umass.edu

Abstract

Exposure to hate speech and toxic content can lead to significant psychological harm, including increased stress and anxiety levels. Content moderators are particularly vulnerable due to their prolonged exposure to such harmful material. Hate speech presents unique challenges as it targets specific individuals and communities. To alleviate the mental burden associated with moderating harmful text, this work explores the effect of targeted content on emotional distress and investigates whether target substitution can reduce this burden and its effect in accuracy of the moderators. Our approach involves substituting the original target entity and all associated references in hate speech with a corresponding fictional character thereby implementing a de-identification process with the aim of retaining the semantic integrity of the content. We conduct both automated and human-based evaluations of this approach, assessing its emotional impact and moderation accuracy. Our findings show that target substitution significantly reduces emotional distress across all evaluated groups, though with a trade-off in accuracy. Additionally, we observe that moderators achieve the highest accuracy when the content's target aligns with their demographic background—a pattern that persists even after target substitution. Additionally, our study highlights the cumulative impact of prolonged exposure to hate speech, showing that moderators experience increased emotional distress over time, particularly in non-targeted scenarios. Despite this, target substitution consistently mitigates distress while maintaining moderation efficacy.

CCS Concepts

• **Human-centered computing → Collaborative and social computing design and evaluation methods**.

Keywords

hate speech, content moderation, LLMs

ACM Reference Format:

Nazanin Jafari and James Allan. 2018. Reducing the Emotional Distress of Content Moderators through LLM-based Target Substitution in Implicit and Explicit Hate-Speech. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

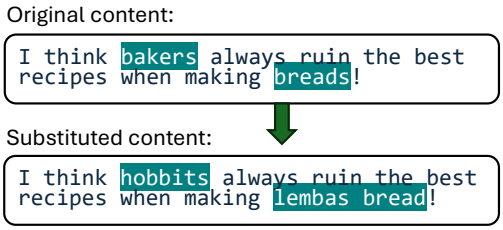

Figure 1: Example of the target substitution approach. Target and its references are substituted with a fictional character.

1 Introduction

The negative impact of hate speech has intensified in the digital era. Online hate speech has significant societal impacts, including the potential to incite violence and perpetuate discrimination against individuals and groups based on race, ethnicity, religion, nationality, socioeconomic status, gender, sexual orientation, disability, and etc [8]. The United Nations defines hate speech as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are" [35].

Exposure to hate speech and toxic content can lead to harmful psychological effects, including increase stress and anxiety [27]. This problem can be elevated in people who are exposed to harmful content for prolonged times such as content moderators and annotators. AlEmadi and Zaghouani [3] highlight that such exposure can lead to mental health challenges, including anxiety, depression, and symptoms akin to post-traumatic stress disorder (PTSD). A recent article in The Guardian [33] reported that over 140 Facebook content moderators have been diagnosed with severe PTSD.

Although automated approaches, such as supervised machine learning, deep learning methods [18, 42], and more recently, large language models (LLMs), have been employed to assist humans in the complex task of hate speech detection [20, 24, 30], human moderators remain essential. This necessity arises for several factors. First, the high accuracy requirements for automated systems [32], which current models often fail to meet, because hate speech detection, demands an understanding of nuanced context, sarcasm, humor, cultural subtleties [14, 19, 41], as well as implicit, subtly expressed, or indirect forms of hate speech [21]. Second, to train these models we still need human-annotated data.

To support human moderators in detecting hate speech, this paper explores whether the perceived harmfulness of the textual content can be mitigated by altering the target while still having a high quality moderation system. Our approach involves identifying target spans within hate speech and leveraging large language models (LLMs) to replace these spans with fictional characters (see Figure 1 for an example[1].). We study whether de-identification in the content can decrease the emotional burden on moderators by dissociating them from the content that is targeted towards real communities while retaining the necessary context for moderation.

To evaluate the effectiveness of target substitution, we first develop an automated moderation system leveraging LLMs and test its performance across a wide range of contents. Next, we assess its effects on human moderators via a user study. Finally, we compare the outputs of the automated system with those of human moderators to determine alignment and consistency.

We introduce two metrics to evaluate the emotional impact of content moderation, namely *Negative Intensity (NI)* and *Emotional Distress (ED)*. Negative Intensity (NI) refers to the perceived intensity in the content, encompassing the harshness of the language, the intent and severity of the message. Emotional Distress (ED) is a subjective measure in which individuals rate the level of distress caused by the given content. Additionally, we assess moderator accuracy in correctly identifying and flagging content as hate speech.

The moderation system operates in two settings: one with unmodified original content and another with substituted content, where the target is replaced with fictional characters. Both the automated system and human moderators are tasked with moderating the original content (control) and the substituted content (treatment). The study reveals that target substitution significantly reduces emotional distress and negative intensity across all evaluated groups, while maintaining moderation accuracy to a reasonable extent. However, a trade-off is observed, as accuracy declines in both automated and human moderation settings. Additionally, moderators demonstrate higher accuracy when the content aligns with their own demographic background, a pattern that persists even after substitution. Automated moderation systems also show reduced ED and NI when substitution is applied. However, they tend to predict higher distress compared to human moderators and lower NI than humans.

In summary, our contributions are as follows:

- We propose and evaluate textual obfuscation method based on a novel target substitution approach, that aims to mitigate the psychological burden on moderators while preserving content interpretability by replacing the original target of hate speech with fictional characters.
- We conduct both automated (LLM-based) and human-based evaluations, assessing the impact of target substitution on emotional distress (ED), negative intensity (NI), and moderation accuracy.
- Our results demonstrate that target substitution significantly reduces distress and perceived intensity across all evaluated groups, with a trade-off in moderation accuracy observed in both human and automated moderation settings.
- We examine the impact of prolonged exposure to harmful content, showing that moderators experience increased distress over time, particularly when moderating non-targeted content. However, substitution consistently mitigates this effect.

[1]The examples are not an actual hate speech

2 Related Work

Hate speech and toxic content moderation present significant challenges, requiring a balance between protecting individuals and vulnerable communities from harmful content while upholding the principles of freedom of speech [34].

2.1 Automatic Hate Speech Detection

Automated approaches have been developed to assist in content moderation by identifying and filtering harmful speech and reducing human exposure to distressing content. To this end, various machine learning and deep learning-based methods have been explored for textual hate speech detection, including LSTMs [25, 36], CNNs [9], and Transformer-based architectures [37]. Specialized models such as HateBERT [4] for abusive language detection, as well as GPT-2-based [40] and LLM-driven approaches [20, 24, 26, 30], have further advanced automated moderation systems.

While these models improve moderation efficiency, they remain imperfect and cannot fully replace human moderators due to several limitations: (1) difficulties in capturing cultural nuances [16, 17, 21] and contextual subtleties [23]; (2) lack of interpretability and explainability, limiting the trust in automated decisions [18]; (3) biases in ML models and LLMs due to biased training data [12]; and sensitivity of LLM outputs to prompt instructions [20]. Moreover, automated systems still require human-annotated training datasets for supervised learning [13], necessitating human oversight.

Given the inherent limitations of automated moderation, human moderators remain essential for addressing the performance gaps in automated approaches. Recognizing the psychological burden placed on human moderators, recent research has explored solutions to mitigate their distress and improve working conditions.

2.2 Mitigating Harms on Moderators

Existing literature explores various approaches to reduce the emotional burden associated with content moderation, which can be broadly categorized into two strategies: *supporting moderators post-exposure* and *mitigating exposure risks*.

2.2.1 Post-Exposure Support Strategies. Several studies focus on alleviating the psychological harm moderators experience after exposure to harmful content. Spence et al. [31] conducted interviews with 11 moderators and identified key factors that help mitigate distress, including social support, fostering supportive relationships, normalizing reactions, and reducing feelings of isolation. Similarly, Scott et al. [29] advocate for trauma-informed design principles in social media and online platforms to minimize long-term harm and promote psychological recovery.

2.2.2 Reducing Exposure to Harmful Content. Other approaches aim to minimize direct exposure to distressing content while moderating. Several studies have explored methods to reduce the psychological burden for visual content moderators by obfuscating harmful imagery [6, 7].By blurring images. Das et al. [7] demonstrate that

image blurring effectively reduces distress without significantly affecting moderation accuracy. Additionally, Lee et al. [16] examine six intervention techniques designed to mitigate negative emotions among video content moderators. Beyond obfuscation, Cook et al. [5] investigate the use of positive emotional stimuli between moderation sessions, finding that an important factor in negative emotions is the prolonged exposure to harmful content, and that inserting positive stimuli, may induce compassion fatigue due to its contrastive nature with negative content.

While these studies primarily focus on mitigating distress for image and video content moderators, our study is the first to explore an approach for obfuscating textual hate speech to reduce emotional distress. Our approach is similar in concept to image blurring, but for text. By "blurring" the target (substituting it with a fictional character) we reduce moderator stress but still allow them to identify harmful content. Crucially, our method preserves the content's integrity and ensures it remains within the boundaries of free speech.

3 Methodology

In this section, we present an overview of the target substitution approach, which consists of two key components: *identifying the targets* and *substituting them* with fictional entities.

3.1 Target Identification

This step involves identifying text spans within content that reference a targeted individual or community. For example:

> I think bakers always ruin the best recipes when making bread.

The identified target spans are "bakers" and making "bread", as they explicitly reference a specific group and an associated activity or characteristic related to that group. Identifying both the entity and its contextual attributes is essential for effective substitution while preserving the overall meaning of the content. Manual identification of target spans is both time-consuming and impractical. To address this, we leverage a semi-automatically labeled dataset from TargetDetect [15]. This dataset employs a pooling methodology to extract target spans, combining the capabilities of large language models (LLMs) with human evaluation. It builds upon publicly available implicit hate speech datasets, including DynaHate [38], IHC [11], and SBIC [28].

3.2 Target Substitution

Following the identification of target spans in the previous step, we replace these target spans with fictional entities and characteristics. To achieve this, we utilize gpt4o-mini[2] LLM and employ a few-shot prompting strategy to instruct the LLM to substitute the input spans with fictional characters.

To ensure the modifications maintain the content's integrity, we follow two key principles: 1) Avoid overly altering the text to the point where it becomes unmoderatable. 2) Prevent replacing targeted communities with potentially offensive characters. 3) Replaced fictional characters have some similarity to the target of the content for better moderation practice. To uphold these principles, we carefully tune our prompts to guide the LLM in generating substitutions that align with the original content. Below is an example of the prompt used for target substitution:

Prompt: *Given the post and the targets in the post (provided in the order they appear in the text), replace the exact matches of the targets with semantically similar, fictional characteristics. Only replace the specified targets and leave the rest of the text unchanged.*

4 Experimental Setup

We evaluate the effectiveness of a moderation system by examining the impact of applying *target substitution* to the content under review. The goal is to determine whether this smoothing approach improves the moderation outcomes.

To achieve this, we first define the *evaluation metrics* used to assess the system's performance. Following this, we provide detailed descriptions of both the *LLM-Based Moderation System* and the *Human-Based Moderation System*. Next, we introduce the *Datasets* and finally *the models used in automated moderation system*.

4.1 Evaluation Metrics

4.1.1 Accuracy measures the ability of humans or LLMs to correctly classify a given piece of content as *hate speech* or *not hate speech*. Specifically, we compare the classifications provided by humans or LLMs against ground-truth labels from previously annotated datasets i.e., Accuracy = $\frac{|\text{Correctly classified}|}{|\text{Contents}|}$.

Policy Definition	Levels	NI Rating
Content includes literal killing, death/eliminatyion, physical harm, or violence against individuals or groups based on their characteristics, or promotes or glorifies harm or violence against individuals or groups.	Tier 1	Extremely Intense
Content targets individuals or groups based on their characteristics by dehumanizing them, such as comparing them to animals or other degrading things, and by nonviolent characterizations using derogatory terms, slurs, and insults.	Tier 2	Intense
Content portrays individuals or groups negatively through nonviolent stereotypes or characterizations based on their characteristics, expresses disagreement with their right to exist, or challenges claims about attempts to change their existence.	Tier 3	Slightly Intense

Table 1: Policy Tier and Negative Intensity Rating. A content that violates policy tier 1 should be rated as extremely intense.

4.1.2 Negative Intensity (NI) refers to the measure the perceived intensity of a given content. This metric aims to minimize subjectivity by rating the content itself—its language, tone, and overall harshness. NI serves as an indicator of the harmful nature of textual content, capturing the presence of derogatory and dehumanizing language, offensive terms, or other negative attributes.

NI is rated on a scale from slightly intense to extremely intense. For assigning NI ratings, Facebook hate speech policy[3] is used as a reference, which categorizes hate speech into three tiers—Tier

[2]https://platform.openai.com/docs/models#gpt-4o-mini

[3]https://transparency.meta.com/policies/community-standards/hate-speech/

1, Tier 2, and Tier 3—according to their level of harmfulness. This means that to rate the NI, first we check which policy tier it violates then we can map the tiers to intensity levels (shown in Table 1.) For instance, content classified under Tier 1—which includes the most severe forms of hate speech—receives the highest NI rating (e.g., extremely intense). In contrast, content falling under Tier 3, which is comparatively less harmful, is rated as slightly intense.

4.1.3 Emotional Distress (ED) This is a subjective metric designed to capture the perceived distress caused by content to the individual moderator. Moderators are asked to rate their own emotional distress elicited by the content on a scale from **0 to 10**, where: **0** indicates *"no distress at all"* and **10** indicates *"extremely distressing"*. This measure enables the analysis of whether specific types of content evoke differing levels of emotional distress based on individual moderator characteristics or the predicted sensitivity in the automated approach.

4.2 Automated Moderation

To simulate human moderation at a larger scale, we design an automated moderation system using large language models (LLMs) to replicate human decision-making in content moderation. This system operates in two distinct settings: *Targeted* and *Not Targeted*.

In the *Not Targeted* setting, the LLM performs the moderation task without any additional contextual information about the target. In contrast, the *Targeted* setting provides the LLM with contextual details regarding the characteristics of the target referenced in the content. To incorporate this contextual information, the prompt is enhanced using the *original target spans* and structured as follows: *"How would a person with [target] characteristics perform the following tasks?"*

This instruction guides the LLM to evaluate content from the perspective of an individual in the referenced target group.

The automatic moderation system comprises three key tasks: 1) **Hate Speech Classification:** To evaluate overall *accuracy*. 2) **Measuring Negative Intensity (NI):** To assess the severity of the harmful content. 3) **Measuring Emotional Distress (ED):** To gauge the emotional impact of the content on moderators.

4.2.1 Hate Speech Classification This task is asked to classify given content as either hate speech or not. Given the content and a model, the prompt is structured as follows: *Determine if the [content] constitutes hate speech or not.* The output generated by the model is used as the predicted annotation.

4.2.2 Negative Intensity (NI) Measurement To measure Negative Intensity (NI), we deploy an approach inspired by the Chain-of-Thought (CoT) framework [39], utilizing step-by-step reasoning. The process is outlined as follows:

(1) **Policy Mapping**: The LLM is first instructed to analyze the given content and, if classified as hate speech in the previous task, assign it to one of the three policy tiers (defined in Section 4.1.2), denoted as $\tau \in \{1, 2, 3\}$. For content classified as non-hateful, Tier 0 is automatically assigned, indicating no negative intensity. Additionally, the LLM generates a policy explanation $\mathcal{E}$, providing the rationale for the selected tier τ.
(2) **Negative Intensity Rating**: The prompt is augmented with the policy tier τ and the explanation $\mathcal{E}$ obtained from the policy mapping step. This enhanced prompt instructs the LLM to evaluate the negative intensity of the content, taking into account both the policy tier τ and the rationale $\mathcal{E}$. The LLM then assigns a numerical rating to the content, reflecting the degree of negativity or harmfulness it perceives.

Final **prompt** for Negative Intensity rating is as follows:

Given the [content], the policy tier [τ], and the explanation [$\mathcal{E}$], output the intensity of the post using following guideline: For policy tier 1 output "extremely intense", for tier 2 output "intense" and for tier 3 output "slightly intense".

Our preliminary results demonstrated that the Chain-of-Thought methodology produces higher quality and more reliable outputs compared to directly prompting the LLM to rate the negative intensity of the content. This structured approach yields more nuanced evaluations, reduces ambiguity, and ensures better alignment with human assessments.

4.2.3 Emotional Distress (ED) measurement To assess Emotional Distress (ED), the LLM is prompted to evaluate the emotional impact of a given content. The model assigns a rating on a scale from 0 to 10, where 0 indicates no distress and 10 represents extreme ED.

The final **prompt** for ED rating is as follows: *Given the [content], output the rate of emotional distress caused by it to humans on a scale of 0 to 10.*

4.3 Human Moderation

For the human moderation, we recruited participants from 3 different protected communities. These communities are Women, LGBTQ community, and Muslims. The age range is between 18-34 and the participants are university students. Each participant completed 40 tasks and were paid $10 US dollars for completing the study. Participants completed tasks across four sessions, each consisting of 10 tasks. Two of these sessions were *targeted*, featuring content tailored to the participants' community, while the other two were *not targeted*, presenting random content from various communities excluding their own community. To study the effect of "target substitution", the *targeted* and *not targeted* sessions are grouped into substituted (treatment group) and not substituted targets (control group). The order of the sessions was randomized for each participant to eliminate ordering effects. Participants in the same community are divided into two equal groups. Each group received the same content related to their community; however, the substitution of targets is flipped between the groups i.e., if the first group views content in its substituted form, the second group views the same content in its non-substituted form. This design ensures a balanced, between-subjects comparison to evaluate the effects of target substitution. Figure 2 provides an overview of the procedure.

Each participant were first assigned to a group based on their background. Participants were then asked to rate their emotional distress before beginning the moderation. Next, participants completed four sessions presented in random order. At the end of each session, they were asked to rate their emotional distress again. Each participant viewed two sessions from the *Control* and two sessions from the *Treatment* group, which featured substituted content.

As previously mentioned, each session comprises 10 tasks. Each task presents a sample of content from publicly available, pre-annotated hate speech datasets and asks participants to answer

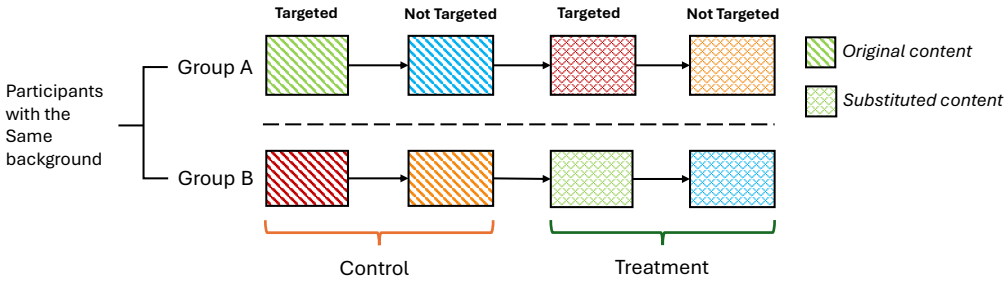

Figure 2: The procedure for human moderation system. Participants from the same background are divided into two groups: Group A receives the substituted content, while Group B receives the original content. Identical colors indicate the same content.

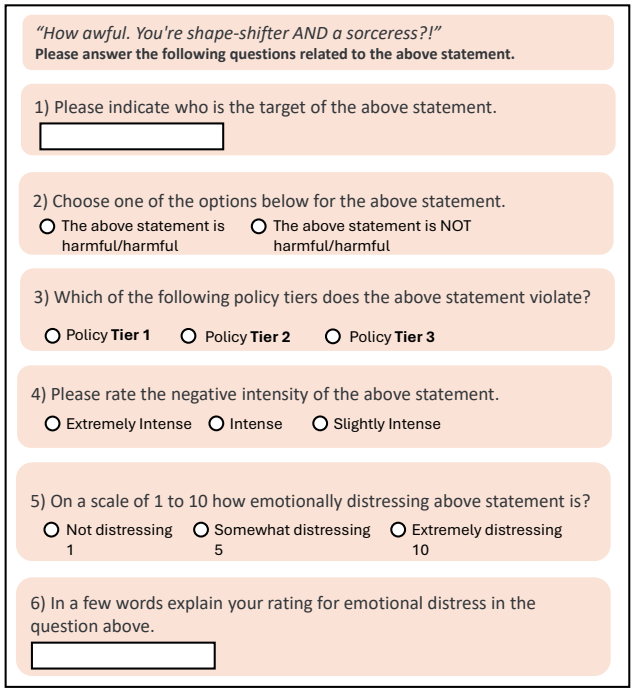

Figure 3: An example of a single task performed by a participant.

6 questions. Details of the questions and a prototype for a single task are shown in Figure 3.

4.4 Datasets

Here we provide details of the datasets used in our experiments. We utilize the *TargetDetect* dataset introduced by Jafari et al. [15], which builds upon and extends existing datasets, including the *IHC*, *DynaHate*, and *SBIC*. The TargetDetect dataset enhances these datasets by incorporating target spans through a pooling-based approach. These extended versions enable us to leverage the annotated target spans effectively for the *target substitution*. We utilize this dataset in full test-set scale for automated moderation system and its subsample for human moderation system.

For the automated system, all the test sets from the *TargetDetect* dataset is used. In total we have 6161 contents across three test sets and 220 unique targeted communities. The number of targets is derived by mapping the target spans from the *TargetDetect* dataset to their corresponding target groups. We removed redundant target names and standardized all target group representations.

For the human moderation system, we create a subsample of the *TargetDetect* dataset, focusing on specific targeted groups. From the dataset, we randomly select 20 samples for each of the following three protected groups: *Woman*, *LGBTQ*, and *Muslim*. Each group's 20 samples are evenly divided into two categories of Substituted Content: 10 samples where target spans have been substituted and Non-Substituted Content: 10 samples with original, unmodified text. Within each of these 10 samples, 8 are labeled as hate speech, targeting the respective community, and 2 are labeled as not hate speech, mentioning the community but without hateful content.

Each moderator annotates 40 samples, including 20 targeting their background and 20 evenly subsampled from the other four categories. Among these additional samples, 10 are substituted for target spans, while the remaining 10 retain their original content. We ensure same label distribution in that set (8 for hate speech and 2 for not hate speech).

4.5 LLMs used in Automated Moderation

For the automated moderation system, we evaluate five large language models (LLMs):

Llama-3.1-8B-Inst [10]: An 8-billion parameter instruction tuned Llama3.1 optimized for general NLP tasks.

Llama-3.1-70B-Inst [10]: A larger 70-billion parameter instruction tuned Llama3.1 offering improved contextual understanding.

Mistral-8B-Inst [1]: A fine-tuned 8-billion parameter Mistral LLM designed for instruction-based tasks.

Mistral-Nemo-Inst [2]: A 12-billion parameter Mistral fine-tuned for dialogue and task-specific instructions.

gpt4-o-mini [22]A lightweight variant of GPT-4o optimized for efficiency and language understanding.

5 Results

In this section, we present the results for two moderation settings: an automatic system and a human-based system. For the automatic moderation system, we used five LLMs. For the human-based system, we recruited 25 participants from diverse three targeted groups. Among the participants, 9 belonged to the "women" targeted group, 7 to the "Muslims," 9 to the LGBTQ community.

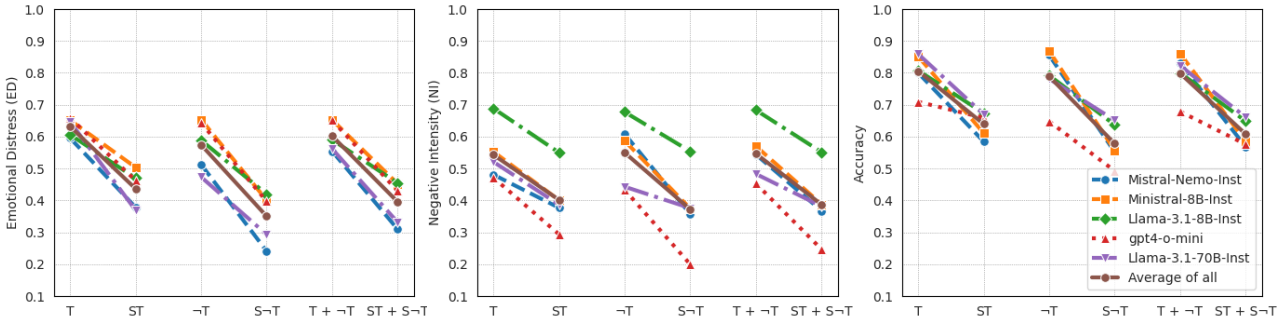

Figure 4: Results for the LLM based Moderation System reported for three metrics of ED, NI and Accuracy.

The results report the three metrics introduced in Section 4.1. ED and NI scores are normalized to a range between 0 and 1 to ensure consistency across evaluations. Additionally, we provide a comparative analysis of the performance of the automatic and human-based moderation systems using identical samples. The results are categorized into two settings: *Targeted* (T) and *Not Targeted* ($\neg T$). Substitution was applied to these settings, resulting in the *Substituted Targeted* (ST) and *Substituted Not Targeted* ($S\neg T$) groups. We compare the *Targeted* and *Not Targeted* scenarios (control group) with their substituted counterparts (treatment group) to evaluate the impact of substitution.

5.1 Automatic Moderation System

Figure 4 presents the results of the automatic moderation system across all 6,161 test set samples, encompassing 220 unique targets. The results are presented as average score for each metric in 4 settings (T, ST, $\neg T$ and $S\neg T$).

5.1.1 Emotional Distress (ED) The findings demonstrate that applying substitution consistently reduces the average predicted emotional distress (ED) rating across all five LLMs. Overall, the ED rating decreases from 0.6 to 0.39, representing a substantial 35% reduction ($T + \neg T$ vs. $ST + S\neg T$).

Among the five models, *Mistral-Nemo-Inst* exhibits the highest reduction, with a 45% decrease in ED between the control and treatment groups, while *Llama-3.1-8B-Inst* shows the smallest reduction at 23%. On average, the LLMs predict an ED rating of 0.63 for targeted (T), which drops significantly to 0.43 for substituted targeted (ST), with a decrease of 33%. Similarly, for non-targeted ($\neg T$), the average rating of 0.57 declines to 0.35 with substitution ($S\neg T$). These results indicate that LLMs consistently assign slightly higher ED ratings to targeted content compared to non-targeted content. Statistical analysis further supports these observations. Welch's t-test between the control and treatment groups revealed statistically significant differences in all scenarios, including the transitions from T to ST and $\neg T$ to $S\neg T$, with a p-value of < 0.05.

5.1.2 Negative Intensity (NI) A A similar trend is observed in the Negative Intensity (NI) ratings, where applying substitution results in significantly lower NI predictions across all models, with statistical significance (p-value < 0.05). On average, NI decreases from 0.54 to 0.40 for targeted content (a 25% reduction) and from 0.55 to 0.37 for non-targeted content (a 31% reduction).

The smaller decrease in NI ratings vs. ED suggests NI is more influenced by overall language, which remains unchanged. However, the significant NI drop after substitution indicates that depersonalizing the target reduces perceived intensity, even without altering tone or structure.

5.1.3 Accuracy Applying substitution also results in a drop in accuracy of LLMs for hate speech classification. We observe that the overall accuracy decreases by 22%, with a 21% drop from T to ST and a 27% drop from $\neg T$ to $S\neg T$. Interestingly, the average accuracy of the five LLMs is higher (0.8) for targeted scenarios (T). When the same data is presented without the targeted context (i.e., $\neg T$), the average accuracy slightly decreases to 0.78. For substituted targeted (ST) scenarios, the accuracy further drops to 0.64, and for substituted non-targeted ($S\neg T$) scenarios, it is 0.57. These results suggest that adding a contextual prompt such as *"How would a person of <target> characteristics determine if the content is hate speech or not?"* enables the LLMs to perform better overall.

Welch's t-test conducted between control ($T + \neg T$) and treatment groups (vs. $ST + S\neg T$) for all five LLMs in all three metrics yielded statistically significant results (p-value < 0.05).

In summary, based on the automated moderation system, we observe that LLMs consistently generate lower ED and NI scores after substitution with a trade-off in drop in accuracy.

Metrics	Tested Groups	p-value
ED	$T \longrightarrow ST$	**0.019**
	$\neg T \longrightarrow S\neg T$	**0.003**
	$T + \neg T \longrightarrow ST + S\neg T$	**0.000**
NI	$T \longrightarrow ST$	**0.019**
	$\neg T \longrightarrow S\neg T$	**0.000**
	$T + \neg T \longrightarrow ST + S\neg T$	**0.000**
Acc	$T \longrightarrow ST$	**0.000**
	$\neg T \longrightarrow S\neg T$	0.121
	$T + \neg T \longrightarrow ST + S\neg T$	**0.008**

Table 2: Statistical significance test based on Welch's t-test for human moderation. P-value< 0.05 suggests a statistically significant difference which is shown in bold.

5.2 Human Moderation System

The results from the human moderation are presented in Figure 5. We report aggregated results for all participants, and a breakdown by the targeted groups: *women*, *LGBTQ*, and *Muslims*. The results are presented as the average score for each metric in 4 sessions.

5.2.1 Emotional Distress (ED) As shown in the left panel of Figure 5, applying substitution leads to a consistent and notable reduction in emotional distress across all groups. The aggregated ED score decreases from 0.54 in the targeted (T) scenario to 0.46 in the substituted targeted (ST) scenario, representing a 15% reduction. For the non-targeted scenario ($\neg T$), ED drops from 0.57 to 0.42 (26% reduction) after substitution ($S\neg T$). These results indicate that substitution significantly mitigates emotional distress for all groups and scenarios. Among the individual groups, *Women* reported the highest ED for T and $\neg T$, while *LGBTQ* reported the lowest ED overall. It is important to note that the participants in this study are not exclusively in one target group meaning that, they can identify with several targeted communities. For example among 9 individuals in *LGBTQ* community, 8 of them also identify as *woman*. Therefore, higher emotional distress and negative intensity observed in the *LGBTQ* community for $\neg T$ may indicate the intersectionality of individuals' identities and their ratings, as several of these examples also align with the targets associated with other groups.

5.2.2 Negative Intensity (NI). The middle panel of Figure 5 shows a similar trend for negative intensity. The aggregated NI scores decrease from 0.59 ($T + \neg T$) to 0.45 ($ST + S\neg T$), with a 24% reduction. It decreases from 0.57 to 0.47 for Targeted scenario and for non-Targeted scenario, the NI score drops from 0.62 ($\neg T$) to 0.43 ($S\neg T$), representing a 30% reduction. Across all groups, substitution effectively reduces NI, with the *LGBTQ* group experiencing the largest decrease. These findings re-instates the impact of substitution in reducing the perceived negativity of hate speech content.

5.2.3 Accuracy The right panel of Figure 5 illustrates the impact of substitution on participant accuracy. Overall, accuracy decreases from 0.83 to 0.72, representing a 13% reduction. For targeted content (T), accuracy decreases by 10%, while for non-targeted content, accuracy drops from 0.82 ($\neg T$) to 0.69 ($S\neg T$), reflecting a 16% reduction. Despite this decline, participants consistently demonstrated higher accuracy for targeted scenarios (T) compared to non-targeted ones ($\neg T$). This pattern persists even after substitution, indicating that participants perform better when content is aligned with their own community's context. Among the groups, *Women* exhibited the highest accuracy when annotating content related to their own group. Even after substitution (ST), their accuracy remained higher compared to non-targeted scenarios ($\neg T$).

5.2.4 Statistical Analysis We performed Welch's T-test on the human-based moderation system, and the p-values are reported in Table 2. The t-test was conducted between the control groups (T, $\neg T$) and the treatment groups (ST, $S\neg T$) for all three metrics. As shown in the table, statistically significant results are observed in all cases except for accuracy between $\neg T$ and $S\neg T$.

This suggests that while target substitution has a significant impact on reducing emotional distress (ED) and negative intensity (NI) across both targeted and non-targeted scenarios, its effect on accuracy is less pronounced in non-targeted cases. The lack of significance in accuracy between $\neg T$ and $S\neg T$ indicates that participants are less affected by substitution in non-targeted scenarios and were able to perform the annotation task with similar accuracy for non-substituted and substituted content, while experiencing significantly lower emotional distress.

5.2.5 Impact of Prolonged Exposure to Harmful Content Here, we report the overall impact of prolonged exposure to harmful content on participants' emotional distress (ED).

To evaluate the effect of prolonged exposure, we calculated the *difference* in ED levels between the beginning and after each session. Specifically, we measured participants' ED ratings at two key points: *before* starting the moderation and *after* completing each session, where they were exposed to 10 harmful contents per session. For instance, to assess the ED after completing the T session, we computed the change as $\Delta = T - S_0$, where S_0 represents the participant's ED at the beginning. This analysis includes responses only from participants who successfully completed all tasks and the end-of-session questions. As a result, the analysis is based on data from 12 participants.

The results of this analysis are presented in Figure 6, illustrating the Δ in ED levels after each session. The box plot reveals that participants experienced an **overall increase** in emotional distress after engaging with the moderation. Notably, exposure to original, unsubstituted content (T and $\neg T$) resulted in higher increases in ED with a median of +2 increase compared to the substituted content (ST and $S\neg T$) with a median of +1. Additionally, the variance in ED is higher for $\neg T$, as indicated by the larger interquartile range and whiskers, suggesting that non-targeted harmful content elicited more varied responses among participants. Outliers in T and $\neg T$ indicate heightened impact on some individuals from prolonged exposure to harmful content.

Overall, these results suggest that target substitution helps mitigate the emotional burden associated with prolonged exposure to harmful content, as participants exposed to substituted content reported lower increases in distress. This reinforces the potential of substitution techniques in reducing the psychological impact of content moderation tasks especially for prolonged exposure.

5.3 Comparison between Automatic and Human-Based Moderation Systems

Finally, we compare human outputs with LLM outputs using the same content samples as in the user study rather than the 6,161 used previously. This means that the inputs provided to the LLM are identical to those used for human evaluation across each targeted category. The results are reported as aggregated outputs for both the LLM and human moderators for three targeted groups in Figure 7.

Overall we see results in line with what we observed so far: substitution both in humans and LLMs results in reduced ED and NI and a drop in accuracy. Moreover we see that LLM and human moderators exhibit similar alignment in terms of accuracy. However, the LLM tends to produce higher ratings for Emotional Distress (ED) compared to human evaluations in targeted and non-targeted scenarios. LLM also show more dramatic drop in ED with 30%, when the target is substituted whereas humans show more subtle drop

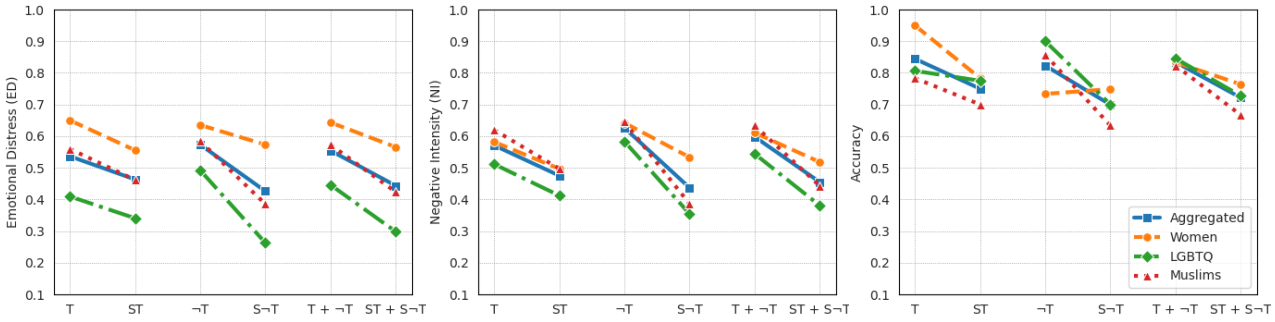

Figure 5: Results for the Human based Moderation System reported for three metrics of ED, NI and Accuracy for Aggregated and different target groups.

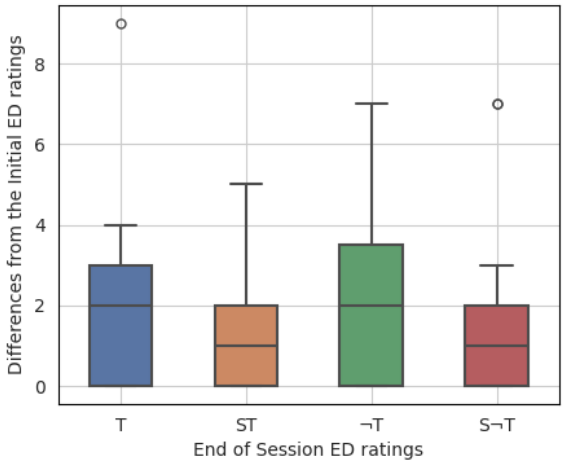

Figure 6: The results are reported for the Δ of 12 participant between S_0 and corresponding session.

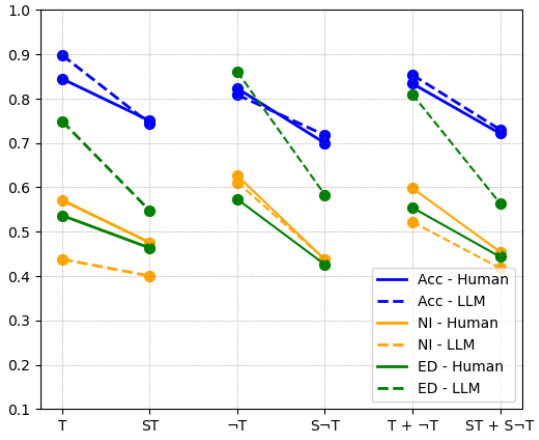

Figure 7: Comparison between LLM and Human evaluation on same samples. We used Llama-3.1-70B-Inst model for this experiment.

18%. These discrepancies highlight the differences in human emotion and perceived emotion by LLMs in performing such human oriented tasks (emotional distress judgement). We also observe a compatibility between LLM and human NI scores in the untargeted scenario ($\neg T$ and $S\neg T$). However, in the targeted scenario (T and ST), the LLM exhibits lower NI ratings compared to human evaluators. This discrepancy may be due to the LLM's tendency to rely more on explicit textual cues rather than implicit contextual understanding, making it less sensitive to nuanced meaning shifts that humans naturally perceive. Additionally, human evaluators may incorporate real-world knowledge and infer intent more effectively, contributing to higher NI scores in the targeted scenario.

6 Conclusion

In this study, we investigated the impact of target substitution in hate speech moderation across both automated and human-based moderation systems. Our results demonstrate that target substitution effectively reduces emotional distress (ED) and negative intensity (NI) across all targeted groups. Additionally, human moderators reported higher ED for non-targeted content compared to targeted content on an individual basis. However, by the end of the session, distress levels were similar across conditions, suggesting a cumulative effect of prolonged exposure to hate speech. Moderation accuracy declined by 22% in automated systems and 13% in humans after substitution, indicating a trade-off between emotional relief and detection performance. However, accuracy remained higher when moderating content related to one's community, emphasizing the role of contextual familiarity.

These findings underscore the need to balance emotional well-being and moderation quality in content moderation. While target substitution effectively reduces emotional distress, further refinements are needed to minimize its impact on accuracy. Optimizing this approach can lead to more sustainable and ethical moderation practices that prioritize both well-being and performance.

References

[1] Mistral AI. 2024. Ministral-8B-Instruct Model. https://huggingface.co/mistralai/Ministral-8B-Instruct-2410.
[2] Mistral AI. 2024. Mistral-Nemo-Instruct-2407 Model. https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407.
[3] Maryam M AlEmadi and Wajdi Zaghouani. 2024. Emotional Toll and Coping Strategies: Navigating the Effects of Annotating Hate Speech Data. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies@ LREC-COLING 2024*. 66–72.

[4] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Association for Computational Linguistics, Online, 17–25. https://doi.org/10.18653/v1/2021.woah-1.3

[5] Christine L Cook, Jie Cai, and Donghee Yvette Wohn. 2022. Awe versus aww: The effectiveness of two kinds of positive emotional stimulation on stress reduction for online content moderators. *Proceedings of the ACM on human-computer interaction* 6, CSCW2 (2022), 1–19.

[6] Brandon Dang, Martin Johannes Riedl, and Matthew Lease. 2018. But Who Protects the Moderators? The Case of Crowdsourced Image Moderation. *arXiv: Human-Computer Interaction* (2018). https://api.semanticscholar.org/CorpusID:47016766

[7] Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 33–42.

[8] Talita Dias. 2022. Tackling online hate speech through content moderation: The legal framework under the International Covenant on Civil and Political Rights. *Countering online hate and its offline consequences in conflict-fragile settings (Forthcoming)* (2022).

[9] Ehsan Doostmohammadi, Hossein Sameti, and Ali Saffar. 2019. Ghmerti at SemEval-2019 Task 6: A Deep Word- and Character-based Approach to Offensive Language Identification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, USA, 617–621. https://doi.org/10.18653/v1/S19-2110

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[11] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 345–363. https://aclanthology.org/2021.emnlp-main.29

[12] Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An Investigation of Large Language Models for Real-World Hate Speech Detection. *2023 International Conference on Machine Learning and Applications (ICMLA)* (2023), 1568–1573. https://doi.org/10.1109/ICMLA58977.2023.00237

[13] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12. https://doi.org/10.1109/MIS.2009.36

[14] Ming Shan Hee, Rui Cao, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Understanding (Dark) Humour with Internet Meme Analysis. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1276–1279. https://doi.org/10.1145/3589335.3641249

[15] Nazanin Jafari, James Allan, and Sheikh Muhammad Sarwar. 2024. Target Span Detection for Implicit Harmful Content. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval* (Washington DC, USA) *(ICTIR '24)*. Association for Computing Machinery, New York, NY, USA, 117–122. https://doi.org/10.1145/3664190.3672525

[16] Dokyun Lee, Sangeun Seo, Chanwoo Park, Sunjun Kim, Buru Chang, and Jean Y Song. 2024. Exploring Intervention Techniques to Alleviate Negative Emotions during Video Content Moderation Tasks as a Worker-centered Task Design. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 1701–1721.

[17] Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate Speech Classifiers are Culturally Insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 35–46. https://doi.org/10.18653/v1/2023.c3nlp-1.5

[18] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and O. Frieder. 2019. Hate speech detection: Challenges and solutions. *PLoS ONE* 14 (2019). https://doi.org/10.1371/journal.pone.0221152

[19] Seema Nagar, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2023. Towards more robust hate speech detection: using social context and user data. *Social Network Analysis and Mining* 13, 1 (2023), 47.

[20] Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards Interpretable Hate Speech Detection using Large Language Model-extracted Rationales. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, Yi-Ling Chung, Zeerak Talat, Debora Nozza, Flor Miriam Plaza-del Arco, Paul Röttger, Aida Mostafazadeh Davani, and Agostina Calabrese (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 223–233. https://doi.org/10.18653/v1/2024.woah-1.17

[21] Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 1997–2013. https://doi.org/10.18653/v1/2023.eacl-main.147

[22] OpenAI. 2024. GPT-4o-mini Model. https://platform.openai.com/docs/models#gpt-4o-mini. Accessed: January 30, 2025.

[23] Juan Manuel Pérez, Franco M Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, et al. 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access* 11 (2023), 30575–30590.

[24] Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or Toxic? Using Zero-Shot Learning with Language Models to Detect Hate Speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, Yi-ling Chung, Paul R{\"ottger}, Debora Nozza, Zeerak Talat, and Aida Mostafazadeh Davani (Eds.). Association for Computational Linguistics, Toronto, Canada, 60–68. https://doi.org/10.18653/v1/2023.woah-1.6

[25] Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Hierarchical CVAE for Fine-Grained Hate Speech Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 3550–3559. https://doi.org/10.18653/v1/D18-1391

[26] Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6116–6128. https://doi.org/10.18653/v1/2023.findings-emnlp.407

[27] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM conference on web science*. 255–264.

[28] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5477–5490. https://doi.org/10.18653/v1/2020.acl-main.486

[29] Carol F Scott, Gabriela Marcu, Riana Elyse Anderson, Mark W Newman, and Sarita Schoenebeck. 2023. Trauma-informed social media: Towards solutions for reducing and healing online harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[30] Tanmay Sen, Ansuman Das, and Mrinmay Sen. 2024. HateTinyLLM: Hate Speech Detection Using Tiny Large Language Models. *arXiv preprint arXiv:2405.01577* (2024).

[31] Ruth Spence, Amy Harrison, Paula Bradbury, Paul Bleakley, Elena Martellozzo, and Jeffrey DeMarco. 2023. Content moderators' strategies for coping with the stress of moderating content online. *Journal of Online Trust and Safety* 1, 5 (2023).

[32] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[33] The Guardian. 2024. Kenya Facebook moderators sue after diagnoses of severe PTSD. *The Guardian* (18 December 2024). https://www.theguardian.com/media/2024/dec/18/kenya-facebook-moderators-sue-after-diagnoses-of-severe-ptsd

[34] Jeremiah Thuku Thuku and Margaret Mbaaro. 2022. Hate Speech Legislation and Freedom of Expression: Finding the Balance. *Interdisciplinary Studies in Society, Law, and Politics* (2022). https://doi.org/10.61838/kman.isslp.1.2.5

[35] United Nations. 2024. What is hate speech? https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

[36] Neeraj Vashistha and Arkaitz Zubiaga. 2020. Online multilingual hate speech detection: experimenting with Hindi and English social media. *Information* 12, 1 (2020), 5.

[37] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[38] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1667–1682. https://doi.org/10.18653/v1/2021.acl-long.132

[39] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.

[40] Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Towards Hate Speech Detection at Large via Deep Generative Modeling. *IEEE Internet Computing* 25, 2 (2021), 48–57. https://doi.org/10.1109/MIC.2020.3033161

[41] Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 5918–5930. https://doi.org/10.18653/v1/2022.naacl-main.433

[42] Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving Hate Speech Detection with Deep Learning Ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1404

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009