HateBuffer: Safeguarding Content Moderators' Mental Well-Being through Hate Speech Content Modification

SUBIN PARK*, School of Electrical Engineering, KAIST, Republic of Korea
JEONGHYUN KIM*, School of Computing, KAIST, Republic of Korea
JEANNE CHOI, School of Computing, KAIST, Republic of Korea
JOSEPH SEERING, School of Computing, KAIST, Republic of Korea
UICHIN LEE, School of Computing, KAIST, Republic of Korea
SUNG-JU LEE†, School of Electrical Engineering, KAIST, Republic of Korea

Hate speech remains a persistent and unresolved challenge in online platforms. Content moderators, working on the front lines to review user-generated content and shield viewers from hate speech, often find themselves unprotected from the mental burden as they continuously engage with offensive language. To safeguard moderators' mental well-being, we designed *HateBuffer*, which anonymizes targets of hate speech, paraphrases offensive expressions into less offensive forms, and shows the original expressions when moderators opt to see them. Our user study with 80 participants consisted of a simulated hate speech moderation task set on a fictional news platform, followed by semi-structured interviews. Although participants rated the hate severity of comments lower while using *HateBuffer*, contrary to our expectations, they did not experience improved emotion or reduced fatigue compared with the control group. In interviews, however, participants described *HateBuffer* as an effective *buffer* against emotional contagion and the normalization of biased opinions in hate speech. Notably, *HateBuffer* did not compromise moderation accuracy and even contributed to a slight increase in recall. We explore possible explanations for the discrepancy between the perceived benefits of *HateBuffer* and its measured impact on mental well-being. We also underscore the promise of text-based content modification techniques as tools for a healthier content moderation environment.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Content moderation, Content moderators, Hate speech, Mental health

ACM Reference Format:
Subin Park, Jeonghyun Kim, Jeanne Choi, Joseph Seering, Uichin Lee, and Sung-Ju Lee. 2025. *HateBuffer*: Safeguarding Content Moderators' Mental Well-Being through Hate Speech Content Modification. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW428 (November 2025), 39 pages. https://doi.org/10.1145/3757609

CONTENT WARNING: This paper contains hate speech examples, including pejorative terms, that readers may find disturbing.

*Equal contributions.
†Corresponding author.

Authors' Contact Information: Subin Park, subin.park@kaist.ac.kr, School of Electrical Engineering, KAIST, Daejeon, Republic of Korea; Jeonghyun Kim, jeonghyun.kim@kaist.ac.kr, School of Computing, KAIST, Daejeon, Republic of Korea; Jeanne Choi, jeanne.choi@kaist.ac.kr, School of Computing, KAIST, Daejeon, Republic of Korea; Joseph Seering, seering@kaist.ac.kr, School of Computing, KAIST, Daejeon, Republic of Korea; Uichin Lee, uclee@kaist.edu, School of Computing, KAIST, Daejeon, Republic of Korea; Sung-Ju Lee, profsj@kaist.ac.kr, School of Electrical Engineering, KAIST, Daejeon, Republic of Korea.

ACM 2573-0142/2025/11-ARTCSCW428
https://doi.org/10.1145/3757609

1 Introduction

As the volume of user-generated content continues to grow on social media [59], the prevalence of harmful content, particularly hate speech, has become a significant concern [133]. Hate speech is defined as any form of communication in speech or writing that attacks or discriminates against a person or group based on their identity (e.g., religion, ethnicity, nationality, race, gender, or other factors) [134], and is a long-established issue. For example, users on X reported 66.9M instances of hate speech [140], and Facebook took action on 14.6M instances of hate speech in the first half of 2024 [86].

Hate speech can elicit strong negative emotions [56], resulting in serious psychological effects (e.g., trauma [114, 126] or depression [131, 135]) for readers, particularly the target and onlookers who identify with the targets. Given the significant harm that hate speech can cause to individuals and society at large by reinforcing harmful stereotypes [14, 19], many platforms have employed commercial content moderators to review and moderate them [104]. Many efforts have been developed for automated hate speech moderation, mostly based on heuristic rules [87, 88, 130] and artificial intelligence (AI) [43, 89]. However, these tools often fail to grasp the context and subtle nuances of hate speech [2, 13, 80, 137]. In fact, X's 2024 transparency report indicates that of the over 2M removed hate speech posts, 99.75% required human moderators to address nuanced content, with only 0.25% managed solely by automated systems [140]. Given the persistent need for human judgment in hate speech moderation, platforms continue to recruit human moderators [4, 58, 144].

These content moderators are at risk for a variety of detrimental impacts to their mental well-being due to regular exposure to harsh and offensive content, creating a challenging and emotionally demanding work environment [105, 122]. Recent studies have shown that modifying the style of image- and video-based content that moderators review using approaches such as grayscaling, blurring, and adding cartooning [27, 67, 74] can reduce the emotional burden of moderating harmful content without compromising accuracy. However, the potential for similar modifications to text content remains underexplored. To safeguard the mental well-being of moderators who review hate speech, adapting content modification techniques for text content could be highly beneficial; however, unlike visual content, where immediate sensory stimuli often drive emotional impact, text-based hate speech derives its detrimental effect from the semantic meaning, making it challenging to directly apply visual content modification techniques. To address this, we designed *HateBuffer*, a text content modification system for hate speech moderation to alleviate moderators' mental burdens while preserving moderation performance.

HateBuffer consists of four features to modify the hate speech content. First, `target_anonymization` anonymizes the target group of potential hate speech to reduce the negative emotional impact on moderators caused by feeling attacked [25] or experiencing vicarious trauma [138]. Second, `paraphrasing_offensive` paraphrases offensive expressions to less offensive versions to prevent the emotional contagion from reviewing offensive language [21, 34, 51]. Lastly, `revealing_target` and `revealing_original` allow moderators to optionally view the original target and offensive expressions by clicking, providing control over whether to access the full content.

To investigate whether *HateBuffer* can reduce moderators' emotional burden without harming their performance, we aimed to answer the following research questions (RQs):

- **RQ1**: How does each feature of *HateBuffer* contribute to moderators' **mental well-being** during hate speech moderation?
- **RQ2**: How does each feature of *HateBuffer* influence **moderation strategies** and contribute to moderators' **performance** in hate speech moderation?

To address these RQs, we conducted a between-subjects study with 80 participants. We distributed them into four groups: the control group as the baseline, the anonymizing group using `target_anonymization`, the paraphrasing group using `paraphrasing_offensive`, and the revealing group using *HateBuffer* with all features. For the user study, we selected 100 comments from the K-HATERS dataset [99] and applied text content modification, utilizing a Large Language Model (LLM) for `paraphrasing_offensive`. Participants were assigned the role of moderators to perform simulated hate speech moderation for a fictional news platform. Through surveys and semi-structured interviews, we investigated *HateBuffer*'s impact on participants' mental well-being and moderation performance.

In contrast to our expectations, we did not detect a difference in post-study fatigue or negative emotion levels between the study conditions. However, participants in the text modification conditions rated the severity of hate speech lower than those in the control condition, and participants noted many positive aspects of the system in post-study interviews; they perceived *HateBuffer* as a buffer, providing time to prepare themselves to face the hateful content. Participants also noted that *HateBuffer* helped to protect them from normalizing biased and hateful opinions from the comments. In addition, despite *HateBuffer* modifying the comments by anonymizing targets and paraphrasing offensive expressions, the moderation accuracy remained similar at between 0.75–0.80 for all groups. Notably, the paraphrasing and revealing groups showed slightly higher moderation recall.

Building on this finding, we explore possible explanations for the discrepancy between perceived benefits and the actual impact on mental well-being. Additionally, we highlight how text content modification can provide positive friction to enable a more thoughtful moderation process, and we provide considerations for adopting text content modification in commercial settings. Finally, we discuss the importance of protecting moderators' mental well-being to support a more sustainable working environment.

2 Related Work

We review current practices in hate speech moderation, highlighting the complexity that necessitates the involvement of human moderators. We discuss the challenges human moderators face and existing approaches that address these challenges, focusing on mental well-being.

2.1 Hate Speech Moderation

Hate speech is a widespread issue in online communication [28, 37, 107, 133] and remains a persistent concern within HCI and CSCW. Although the exact definition of hate speech varies among countries and communities [64], it is generally defined as any form of communication, in speech or text, that attacks or discriminates against individuals or groups based on aspects of identity, such as religion, ethnicity, nationality, race, gender, or similar factors [79, 134].

Hate speech results in significant psychological and social harm to users within online communities [106, 117]. A large body of research has reported that hate speech targeting individuals who create content on platforms, such as YouTubers, live-streamers, or journalists, can lead to considerable emotional distress [46, 50, 56, 65, 108, 126]. This emotional harm can lead to long-term psychological consequences such as depression [131, 135] and trauma [114, 126]. The negative impact of hate speech extends beyond direct victims, affecting viewers who share the targeted identity [25]. Furthermore, prevalent hate speech could foster hostility and reinforce harmful stereotypes, carrying the potential to incite violence in offline settings [14, 19].

To address these negative impacts of hate speech, various efforts have been directed toward effective content moderation [42]. Content moderation is an organized practice of screening user-generated content posted to Internet sites, social media, and other online outlets [105]. Moderation

is usually shaped by community guidelines and policies [102, 141], and content moderation may include manual review by human moderators (sometimes volunteer users [115] but often contractors hired by platforms [105]), semi-automated review where automated tools assist moderators, and/or fully automated methods usually based on machine learning algorithms [42] or hashes [32].

Automated approaches have been extensively developed to facilitate large-scale content moderation. Word filtering, a traditional automated moderation technique that detects specific words or similar textual patterns violating community guidelines, is widely employed across various platforms (e.g., Twitch [130], Facebook [87], Instagram [88], etc.). However, word filtering often falls short due to high false positive and false negative rates, stemming from its inability to interpret context [62, 63, 72]. In response, advanced AI-based moderation systems incorporate contextual understanding to improve accuracy and reduce errors [5, 45, 112].

Although AI-based moderation models claim high accuracy, their real-world performance often falls short. Gordon et al. noted that evaluating moderation models using crowdsourced data can dramatically overstate their capabilities [44]. For example, by adjusting for intra-annotator consistency in the popular Jigsaw toxicity task, where the model initially achieved a reported ROC AUC of 0.95, they found the performance dropped to an ROC AUC of 0.73. The complex nature of hate speech adds further challenges to detection. As hate speech targets a specific individual or group, understanding related context or background is often needed to judge whether certain content is hate speech [2, 3, 80].

Commercial moderation models frequently struggle with limited contextual understanding, particularly for identity-based language, as in cases based on race and gender [52, 92]. This can lead to over-moderation, inadvertently flagging non-hateful content, such as counter-speech, which employs similar linguistic features to challenge or subvert discriminatory narratives [91]. Prior research has documented various classification biases [29], which further exacerbate these issues. Such biases complicate classifier adjustments, as reducing over- or under-moderation often worsens the other, particularly when moderation outcomes are unevenly distributed across identity groups [92]. Additionally, while LLMs show some potential for integration into content moderation processes, current efforts fall short in terms of accuracy and consistency [70, 77], showcasing the continued need for human moderators.

These limitations have led platforms such as YouTube and Spotify to take various approaches such as classifying comments into different categories – e.g., public, held for review, and likely spam – delegating final decisions on ambiguous content to human moderators [54, 84]. Similarly, X's 2024 transparency report revealed that among the posts flagged for hate speech, only 0.25% were handled automatically, with the vast majority requiring human moderators [140]. TrustLab's ModerateAI follows a similar approach, using AI to pre-process content and flag potential issues, while human moderators verify automated decisions to ensure policy alignment [128]. This underscores the limitations of automatic moderation and the ongoing essential role of human judgment in accurately identifying and moderating hate speech in online communities [3].

2.2 Mental Burden on Human Moderators: Challenges and Mitigation Strategies

A substantial body of research has highlighted the mental burden faced by human moderators [94, 105, 122, 138]. Regular exposure to highly offensive content, including sexism, racism, and various other abuses with varying degrees of intensity [139], places a significant psychological burden on moderators [122]. This burden can result in high levels of mental distress [27, 121], burnout [120], anxiety [138], and even post-traumatic stress disorder (PTSD) [7, 8, 40, 103, 139]. This mental strain can also drive moderators to leave their positions due to the cumulative impact of ongoing exposure [113].

Various tools and services have been developed to enhance scalable, efficient, and effective moderation processes, including technological support and improvements to the work environment to alleviate the mental burden of moderators [122]. For example, a recent approach provides visual cues that direct moderators' attention to potentially problematic content [111]. Highlighting offensive language or hate speech targets identified by machine learning models has supported faster moderation [17, 48, 53, 81]. Additionally, providing moderators with explanations about why the content violates the community's policies can reduce the time required for conducting the moderation task [17]. Expanding on this concept, an LLM-generated description of an implied social bias in content can enrich moderators' understanding of the underlying issues, improving moderation efficiency [143].

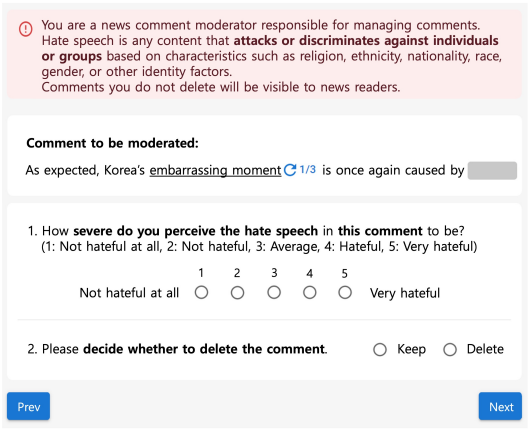
To further address the mental burden faced by moderators, various approaches have been taken to improve mental well-being and to promote moderators' workplace wellness. These include risk mitigation strategies (e.g., scheduling recess, showing relaxing images), providing clinical support, and fostering peer support connections to help manage their work's psychological impacts [95, 122]. For instance, offering mindfulness content, such as meditation videos, has shown potential benefits. Lee et al. found that providing positive videos (e.g., scenic landscapes, baby animals) during breaks in a car accident video annotation task reduced negative emotions among moderators [74]. In contrast, Cook et al. observed no positive effect from using similar stimuli (cute and relaxing images) during breaks from moderating text content related to sexism, racism, and threats, indicating that visual relief alone may not fully address moderators' emotional stress [24].

A complementary line of work for supporting moderation processes has explored methods for reducing harm from content exposure by proactively modifying the content itself. In image moderation, Karunakaran et al. applied grayscaling and blurring to images that moderators reviewed to reduce visual stimuli, finding that this adjustment led to more positive emotional responses without compromising moderation performance [67]. Building on this, Das et al. introduced an interactive blurring intervention that allowed moderators to selectively reveal content as needed, which further reduced emotional distress for moderators [27]. Recently, similar techniques have been extended to video content moderation, incorporating blurring, grayscaling, and cartoonizing effects [74]. While blurring and grayscaling maintained comparable moderation accuracy to the baseline, cartoonizing was perceived as an effective intervention for reducing negative emotions when dealing with provocative and unpleasant videos.

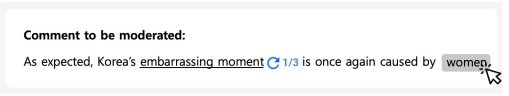
Inspired by the existing work on modifying the content to alleviate moderators' mental burdens, we explore how to realize that for user-generated textual content. In this work, we propose *HateBuffer*, a text-based content modification system for hate speech moderation. While previous research has focused on limiting moderators' exposure to visual stimuli by altering images or videos, our work focuses on textual content by introducing novel content modification techniques. Given the essential role that human moderators play in online spaces and the unique challenges they encounter, we explore the impact of *HateBuffer*'s features on their mental well-being and moderation effectiveness.

3 *HateBuffer*

We propose *HateBuffer*, a system designed to support moderators' mental well-being during hate speech moderation (Fig. 1). *HateBuffer* consists of four features: `target_anonymization`, `paraphrasing_offensive`, `revealing_target`, and `revealing_original`. We present each feature and its design rationale.



(a) Default view.



(b) `revealing_target`.

(c) `revealing_original`.

Fig. 1. Screenshots of *HateBuffer*. Instructions for the moderation task, a comment to moderate, and an interface for moderation are given. (a) By default, *HateBuffer* provides a modified comment with `target_anonymization` and `paraphrasing_offensive`. (b) By clicking the anonymized target, moderators can see the original target expression (`revealing_target`). (c) By clicking the paraphrased expression, moderators can see the original offensive expression (`revealing_original`).

Target Anonymization

Moderators frequently encounter hate speech that targets their identities or communities, which can have a negative emotional impact [25]. Even when the hate speech is not personally directed at them, offensive language aimed at their social group or identity can still evoke secondary trauma or PTSD [122]. To mitigate the emotional impact of moderating hate speech, we designed `target_anonymization`. As shown in Fig. 1a, it anonymizes the original target expression 'women' using a gray cover.

Paraphrasing Offensive Expressions

Reading offensive expressions can evoke negative sentiment through the behavioral phenomenon of emotional contagion [6, 41, 47, 55]. Research has shown that emotional contagion, usually examined through non-verbal cues such as facial expressions [55], can also occur through text [21, 34, 51]. These expressions not only cause immediate emotional responses but also tend to linger in memory, resulting in an enduring cognitive effect [114, 126]. To mitigate these impacts on content moderators, we designed a feature to reduce exposure to offensive language.

One approach to reducing the negative impact of text involves using euphemism, which replaces direct, potentially harmful language with softer alternatives to create emotional distance and reduce the shock factor [82]. Euphemisms have been observed to obscure the harshness of the original content, thereby lowering the emotional intensity of the message [15]. Prior research has shown that, even when the content remains the same, the tone of expression can significantly alter the reader's perception and emotional response [57].

With this in mind, we designed `paraphrasing_offensive`, which paraphrases offensive expressions within hate speech into less offensive versions. As shown in Fig. 1a, it paraphrases the

original offensive expression 'downfall' to a less offensive version, 'embarrassing moment.'[1] Since we replaced only specific expressions rather than entire sentences, the paraphrased terms may occasionally feel out of place in the surrounding context, even if the overall meaning is preserved. For better readability, *HateBuffer* provides up to three alternatives shown as '1/3' next to the underlined text, to ensure a smoother integration within the comment. We used an LLM to achieve this goal, and the detailed process of generating the paraphrased expressions is described in §4.2.

Revealing Targets or Original Offensive Expressions

Research on online harassment has shown that providing users with a high-level summary of negative responses to their posts allowed them to engage with the content that elicited negative emotions on an optional basis. Interestingly, even when users opted to read the feedback, the summary helped them mentally prepare for the negative content. Participants reported feeling less emotionally impacted when they could anticipate the criticism before fully engaging with it. This strategy highlights the importance of introducing a preparatory phase, giving users a moment to prepare before exposure to potentially harmful materials [68].

This approach aligns with the idea of positive friction, which refers to intentional interventions designed to momentarily interrupt automatic processes and encourage mindfulness and deliberate decision-making [20]. For instance, trigger warnings operate as positive friction by prompting users to decide whether they want to engage with sensitive content that might negatively affect their emotional state [49, 116].

Building on this idea, we developed two features: `revealing_target` and `revealing_original`. `revealing_target` shows the original target of the hateful content when the moderators click the anonymized target expression (Fig. 1b). Similarly, `revealing_original` displays the original offensive expression when the moderator clicks the paraphrased offensive expression (Fig. 1c). By offering these options, we create a psychological buffer that reduces moderators' emotional strain while enabling them to effectively carry out their moderation tasks.

We implemented *HateBuffer* as a web application using TypeScript and React.js, with real-time logging on Firebase Firestore [35].

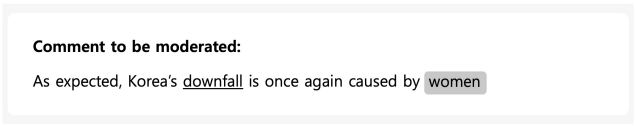
4 User Study

In this section we detail the user study methods, including data curation, evaluation metrics, study design, and analyses used to assess *HateBuffer*'s support for moderators' well-being and performance in hate speech moderation tasks.

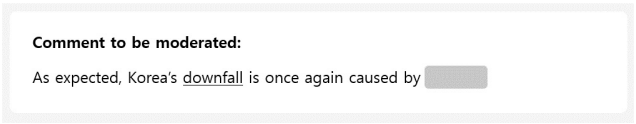
4.1 Study Setup

We designed a between-subjects user study that included a hate speech moderation experiment followed by semi-structured interviews to explore how *HateBuffer* supports moderators in protecting their mental well-being and moderation task performance. To clearly observe the effectiveness of `target_anonymization` and `paraphrasing_offensive`, we designed four study groups: the control group as the baseline, the anonymizing group using only `target_anonymization`, the paraphrasing group using only `paraphrasing_offensive`, and the revealing group using *HateBuffer*, where participants initially encounter anonymized targets and softened expressions but can reveal the target and original expressions using `revealing_target` and `revealing_original`.

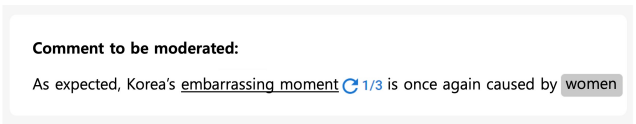
[1]Note that examples provided throughout this paper have been translated by the research team from their original Korean into English. We have endeavored to find English phrasings that capture the same sentiment, but this is difficult in some cases due to cultural differences. Participants in the study saw the phrases and their modifications in the original Korean.



(a) The control group.



(b) The anonymizing group.



(c) The paraphrasing group.

Fig. 2. Screenshots of the user interface with an example comment for (a) control, (b) anonymizing, and (c) paraphrasing groups.

For the control group, target expressions are highlighted in gray and offensive expressions are underlined (Fig. 2a). This design reflects the standard visual support typically employed in collaborative moderation support systems, both in real-world platforms [31, 118] and academic research [17, 22, 48, 53, 100, 119]. For the anonymizing group, only `target_anonymization` is given, where the target expression is anonymized with a gray cover (Fig. 2b). For the paraphrasing group, only `paraphrasing_offensive` is given, where the offensive expression is paraphrased as less offensive (Fig. 2c). Lastly, for the revealing group, the full set of features in *HateBuffer* is given.

4.2 Data Curation

To curate 100 comments for a moderation experiment, we reviewed Korean hate speech datasets: K-HATERS [99], K-MHaS [75], BEEP! [90], and KoLD [61]. From these, we selected the largest dataset that included sufficiently detailed labeling criteria and rich annotations (e.g., topic, target, and offensive rationales). K-HATERS is the largest Korean hate speech dataset with human-labeled hate classes (i.e., hate speech or normal) and labels for target and offensive expressions [99]. The dataset defines hate speech as *"words or phrases containing aggression or derogatory remarks directed at individuals or groups with specific attributes"* and categorized comments into 11 topics (e.g., insult, violence, sexual hate, and race) similar to hate speech policies used by popular social media platforms [85, 101, 141].

We began by reviewing randomly selected 300 hate speech comments and 300 normal comments with diverse topic labels (i.e., gender, politics, region, and job) from the entire dataset. The first and second authors individually reviewed and selected comments that provide sufficient context to be understood without reading the corresponding news article, as the dataset consists of news comments. We also verified that the hate class, target expressions, offensive expressions, and topics of the comments were correctly labeled. The final selection consisted of 50 hate speech comments and 50 normal comments, chosen based on agreements between at least two authors. Fleiss' Kappa score reached 0.88, indicating almost perfect agreement [73]. Additionally, to prevent participants from moderating hate speech without reading the comments using annotations of target and offensive expressions as cues, we added annotations for key subjects/objects (instead of

target expressions) and keywords (instead of offensive expressions) in the normal, non-hate speech comments through iterative discussions.

To apply `paraphrasing_offensive` intervention on curated comments, we paraphrased the offensive expressions using LLM. We used GPT-4o, chosen for its fluency in Korean and effectiveness in rephrasing [98]. We generated ten paraphrased versions of each comment by prompting the LLM as follows: (1) assigning the role of a text content moderator for Korean news comments, (2) explaining the definition of euphemism, (3) requesting the paraphrasing of only the annotated offensive expressions euphemistically, and (4) ensuring the original meaning remained unchanged (see Supplementary material A for the full prompt). We then filtered the paraphrased comments, retaining only those with a similarity score higher than 0.7 using cosine similarity of two text embeddings (OpenAI text-embedding-3-small [97]) by referring to the criteria used by LLM-based text data augmentation works [60, 127]. Afterward, the two lead authors selected three paraphrased comments for each original comment that best preserved the original meaning while still providing alternative expressions. We followed the same process for normal comments but paraphrased them into similar expressions.

In addition to the 100 comments to be used in the experiment, we selected eight more hate speech comments to be used to obtain the hate sensitivity level of each participant as part of the recruitment process, following the same process we used for the main data curation but without the paraphrasing process.

4.3 Measures

To investigate the effectiveness of *HateBuffer* in protecting hate speech moderators' mental well-being, we collected quantitative measures across two dimensions: mental well-being and moderation performance.

4.3.1 Mental Well-being. To understand how *HateBuffer* supports participants in preserving their mental well-being during hate speech moderation, we collected perceived hate severity, perceived effectiveness in mental well-being protection, Scale of Positive and Negative Experience (SPANE) score, and Multidimensional Fatigue Symptom Inventory (MFSI). The detailed explanations of each measure are as follows:

- **Perceived Hate Severity.** To understand whether *HateBuffer* reduces the offensiveness of comments, we observed how participants perceived the hate severity of the comments they saw. We asked them to evaluate the hate severity of each comment during the main task: "How severe do you perceive the hate speech in this comment to be?" in a 5-point Likert scale (1: Not hateful at all, 5: Very hateful).
- **Perceived Effectiveness in Mental Well-Being Protection.** We measured how participants perceived the effectiveness of *HateBuffer* in protecting their mental well-being. We asked them to rate the statement, "{feature} helped protect my mental well-being," on a 5-point Likert scale (1: Strongly disagree, 5: Strongly agree) after the experiment.
- **SPANE.** We examined how moderating hate speech with *HateBuffer* affected participants' emotions differently using the Scale of Positive and Negative Experience (SPANE) [30]. Participants responded to the adjusted question, "To what extent do you feel at this moment?", on six positive items (e.g., Pleasant, Happy) and six negative items (e.g., Unpleasant, Sad) on a 5-point Likert scale (1: Not at all, 2: A little, 3: Moderately, 4: Quite a bit, 5: Extremely) before and after the experiment. SPANE_B, the balance of positive and negative, is calculated by subtracting the sum of negative item scores from the sum of positive item scores, ranging

Table 1. Statistical summary of participants' demographics of each group. Moderators # represent the number of participants who have served as a moderator on the online platform (e.g., Facebook groups, YouTube channels, Naver Cafes, etc.).

Group	Num	Age			Gender			Moderators #	Hate Sensitivity
		Mean	Min	Max	Female	Male	No disclosure		
Control	20	23.05±2.35	18	27	8	12	0	3	3.94±0.70
Anonymizing	20	24.75±7.12	19	51	7	12	1	2	3.98±0.75
Paraphrasing	20	24.70±7.02	18	49	10	10	0	2	4.08±0.71
Revealing	20	24.55±4.01	19	33	8	11	1	4	4.05±0.72

from -24 to +24, indicating that a positive value represents participants feeling more positive than negative emotions.

- **MFSI.** Given the mentally demanding nature of hate speech moderation, we assessed the fatigue caused by moderating hate speech with *HateBuffer*. We used the Multidimensional Fatigue Symptom Inventory-Short Form (MFSI-SF) [123], focusing on emotional, mental, and vigor subscales before and after the experiment. We asked "Which of the following best describes how true each statement is for you at this moment?" on a 5-point Likert scale (1: Not at all, 2: A little, 3: Moderately, 4: Quite a bit, 5: Extremely). MFSI is calculated by subtracting the sum of emotional and mental scores from the sum of vigor scores, ranging from -28 to 54, with higher values indicating greater fatigue.

4.3.2 Moderation Performance. To evaluate *HateBuffer*'s feasibility as a moderation support tool, we collected perceived effectiveness in hate speech moderation, moderation accuracy, moderation recall, and task completion time. The detailed explanation of each measure is as follows:

- **Perceived Effectiveness in Hate Speech Moderation.** We assessed how participants perceived each part of *HateBuffer*'s effectiveness for moderating hate speech. We asked them to rate the statement, "{feature} helped perform the moderation task," on a 5-point Likert scale (1: Strongly disagree, 5: Strongly agree) after the experiment.
- **Moderation Accuracy and Recall**. We measured moderation accuracy and recall to compare how consistently participants moderated hate speech with *HateBuffer*. These statistics are considered here more as baselines for comparison rather than objective indicators of real-world performance. During the main task, participants made moderation decisions—either 'delete' or 'keep'—for each comment. We then calculated moderation accuracy and recall based on their responses. Moderation accuracy, defined as the proportion of participants' decisions consistent with the labeled dataset, captures how reliably participants identified what should or should not be deleted. Moderation recall refers to the ratio of deleted hate speech comments to the total number of true hate speech comments.
- **Task Completion Time.** We evaluated the time efficiency of using *HateBuffer* for moderating hate speech by collecting the task completion time for moderating 100 comments.

4.4 Participants

We recruited 80 participants by uploading the recruitment post to our institution's online communities and sending cold emails to moderators of 139 active Naver Cafes.[2] To be eligible for the study, participants had to be (1) over 18 years old and (2) fluent in Korean. We calculated each participant's hate sensitivity score by asking them to rate the severity of eight hate speech comments—including

[2]Naver Cafe (https://section.cafe.naver.com/ca-fe/home) is one of the most popular online community platforms in South Korea.

Table 2. Interview participants' demographic and moderator experience information. P represents participant number, G represents gender, and M represents moderator experience.

Control group				Anonymizing group				Paraphrasing group				Revealing group			
P	**Age**	**G**	**M**	**P**	**Age**	**G**	**M**	**P**	**Age**	**G**	**M**	**P**	**Age**	**G**	**M**
C01	18	F		A01	19	F		P01	20	F		R01	19	M	
C02	21	F		A02	20	M		P02	20	F	✓	R02	20	M	
C03	23	M		A03	21	F		P03	22	F		R03	21	M	✓
C04	23	M		A04	22	M		P04	23	F		R04	21	F	
C05	23	M	✓	A05	23	M		P05	23	F		R05	24	F	✓
C06	23	F		A06	24	M		P06	26	M	✓	R06	28	M	
C07	23	F		A07	25	F		P07	26	M		R07	28	F	
C08	24	M		A08	26	M	✓	P08	28	M		R08	29	F	✓
C09	24	M	✓	A09	28	M	✓	P09	32	M		R09	32	M	✓
C10	25	F	✓	A10	30	M		P10	49	F		R10	33	M	

Fig. 3. Overall user study procedure. Half of the participants from each group were interviewed.

two comments from each of four categories: gender, politics, region, and job—and then averaging their ratings to determine their overall hate sensitivity level. We utilized hate sensitivity scores to evenly distribute participants across groups. To minimize potential biases, we balanced groups based on participants' age and gender. Despite our best efforts, minor imbalances inevitably occurred due to unexpected factors, such as last-minute scheduling changes from participants. However, we ensured these differences were minimal and did not significantly affect group comparability. To confirm this, we conducted a Kruskal-Wallis test specifically on participants' hate sensitivity scores across the groups, which revealed no statistically significant difference (H=2.55, p=.466).

Table 1 presents the statistical summary of participants across the four groups. The average age was 24.26 years (min=18; max=51; std=5.45). Of the participants, 33 (41.25%) identified as female, 45 (56.25%) as male, and two chose not to disclose. Eleven participants (13.75%) had prior experience as content moderators for various platforms (e.g., Facebook groups, YouTube channels, Naver Cafes, etc.). Table 2 shows the detailed demographic and moderator experience of the interviewees in each group. To gather more meaningful insights from a moderator's perspective, we included those with moderation experience as interviewees, and we randomly selected additional participants from each group to ensure that half of the group participated in the interviews. Participants were compensated with 20,000 KRW (approximately USD 14.50) for the 1-hour user study and an additional 10,000 KRW (approximately 7.25 USD) if they participated in the subsequent half-hour interview.

4.5 Study Procedure

Fig. 3 shows the overall procedure of our user study, which was consistent across the four groups. First, we delivered definitions for hate speech in the introductory session and explained content moderators' roles and tasks during the experiment. To ensure participants understood how to use the features available in their assigned condition, we presented dummy comments (e.g., *"In the comment, the targets are highlighted in gray, and offensive expressions are underlined."*) and explained each feature in detail. After the explanation, participants were given the chance to ask clarifying questions. We guided participants to follow a one-minute meditation video to allow them to begin the study with a neutral emotional state. Next, participants responded to the pre-survey

with the SPANE and MFSI questionnaire and practiced with group-specific features on dummy comments introduced during the introductory session.

In a simulated moderation experiment for a fictional news platform, participants were assigned as moderators and asked to assess the perceived hate severity and moderate hate speech within a set of 100 comments. After completing the task, participants completed a post-survey, which asked SPANE, MFSI, and questions about the perceived effectiveness of *HateBuffer* in protecting mental well-being and moderating hate speech. Finally, we selectively interviewed participants to understand their experiences and perceptions of the content modification features. We interviewed participants, focusing on their perceptions of *HateBuffer* and each feature in supporting mental well-being and moderation performance, as well as their experiences during the simulated hate speech moderation experiment. The two lead authors conducted the user study online via Zoom. Up to four participants participated in each 1-hour session, with a maximum of two participants pre-selected during the scheduling phase for an additional half-hour one-on-one interview. This arrangement allowed two interviews to be conducted simultaneously using Zoom's breakout room feature. Since the study did not require face-to-face interaction, participants were allowed to turn off their cameras. The direct message function of the chat was used for participants to ask any help if needed during the study. Participants did not interact with each other at any point. We share the survey questionnaire and interview protocol in Supplementary material E and F, respectively.

4.6 Analysis

We used a mixed-methods approach, combining qualitative and quantitative data to comprehensively understand the participants' experience with content modification interventions.

4.6.1 Quantitative Analysis. We conducted a descriptive statistical analysis on the collected measures: perceived hate severity, SPANE_B, MFSI, moderation accuracy and recall, and task completion time. To account for personal bias in perceived hate severity, we applied z-score normalization. We confirmed normality using the Shapiro-Wilk test. If the data followed a normal distribution, we conducted a one-way ANOVA; otherwise, a Kruskal-Wallis test, to observe significant differences across groups. We performed two-tailed t-tests or Mann-Whitney U tests for pairwise comparisons. All between-subject analysis p-values were corrected with Bonferroni correction to control for multiple comparisons. We conducted Wilcoxon signed-rank tests for within-subject comparisons of SPANE_B and MFSI changes within each group.

4.6.2 Qualitative Analysis. We conducted inductive thematic analysis [12] on the participants' interview data to understand participants' perceptions of interventions and moderation experiences during the experiment. Korean speech-to-text services transcribed the recorded interviews.[3] Two lead authors independently open-coded each group's first three out of ten interviews, focusing on moderation experiences with *HateBuffer* and emotion change during the experiment. All authors then had discussion sessions to develop an initial codebook by discussing emerging themes, addressing inconsistencies, and resolving disagreements to reach a consensus. Based on the initial codebook, we coded the remaining interviews and had iterative discussion sessions to finalize the codebook.

4.7 Ethical Considerations

Our institution's Institutional Review Board (IRB) approved all study phases for ethical compliance. Each participant was pseudo-anonymized using a unique nickname throughout the research, including the user study, data analysis, and reporting. The participants were informed of their right

[3] https://clovanote.naver.com

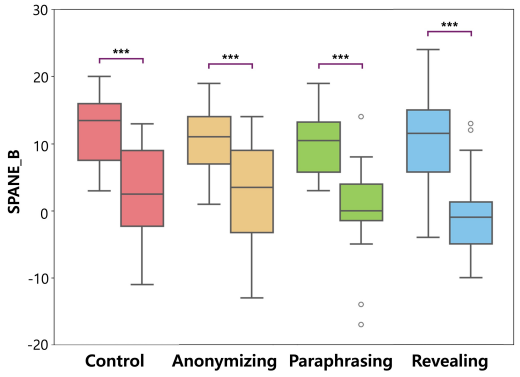


Fig. 4. Changes in participants' SPANE_B before and after the moderation task. *** represents p<.001 of Wilcoxon signed-rank test result.

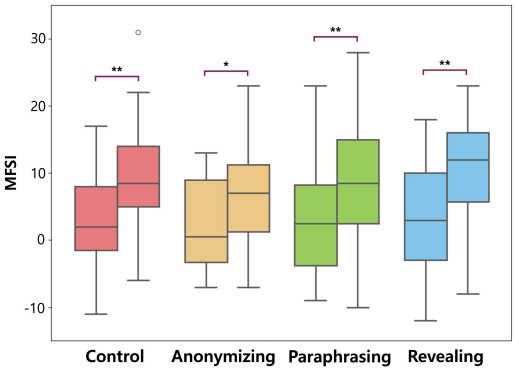


Fig. 5. Changes in participants' MFSI before and after the moderation task. * represents p<.05 and ** represents p<.01 of Wilcoxon signed-rank test result.

to decline any interview questions. We notified participants in advance of the study that they would engage with hate speech content, which could impact their mental well-being. We also notified them that they could withdraw their participation at any time.

5 Results

We assess *HateBuffer*'s feasibility as a moderator support tool by answering the following RQs: (1) How does each feature of *HateBuffer* contribute to moderators' mental well-being during hate speech moderation? (§5.1, §5.2) and (2) How does each feature of *HateBuffer* influence moderation strategies and contribute to moderators' performance in hate speech moderation? (§5.3, §5.4)

We present an overview of quantitative findings for each research question and contextualize our results with insights from the semi-structured interviews.

5.1 Quantitative Findings: *HateBuffer*'s Impact on Moderators' Mental Well-Being

We quantitatively evaluate how each feature of *HateBuffer* impacts moderators' mental well-being during hate speech moderation (RQ1). First, we analyze participant emotions and fatigue changes after the hate speech moderation task across different experimental conditions. We compare the perceived hate severity of comments modified by *HateBuffer* to that of unmodified comments. We then explore participants' self-reported perceptions of the effectiveness of *HateBuffer* in protecting mental well-being.

5.1.1 Impact of HateBuffer on Emotional State. We first examined how *HateBuffer* influenced participants' emotional state, focusing on changes in SPANE_B scores before and after the moderation task. Fig. 4 shows each group's changes in SPANE_B scores. The average SPANE_B score in the pre-survey was 12.00 for the control group, 10.65 for the anonymizing group, 10.35 for the paraphrasing group, and 10.35 for the revealing group. A one-way ANOVA test found no statistically significant difference in SPANE_B results across groups in the pre-survey (F=0.36, p=.785), indicating that participants started the moderation task with similar emotional states.

In the post-survey, the average SPANE_B score was 1.95 in the control group, 2.70 in the anonymizing group, 0.20 in the paraphrasing group, and -0.30 in the revealing group. A one-way ANOVA test indicates no statistically significant difference in SPANE_B results across groups in the post-survey (F=0.84, p=.477). This result shows that although participants' emotional states were significantly more negative after the moderation task, this change was roughly equal across

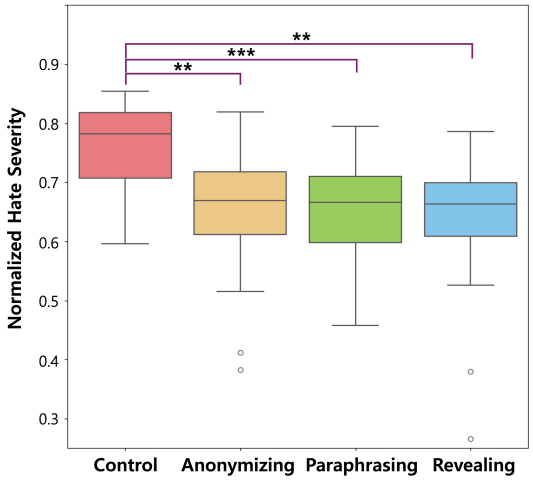


Fig. 6. Comparison of normalized hate severity scores of 50 hate speech comments across four groups. ** indicates p<.01 and *** indicates p<.001.

all experimental groups. The descriptive statistics and Wilcoxon signed-rank test results are in Supplementary material B and D.1, respectively.

5.1.2 Impact of HateBuffer on Fatigue. Next, we examined how *HateBuffer* affected participants' fatigue levels based on MFSI scores. Fig. 5 shows the changes in participants' MFSI scores before and after the moderation task for each group. The average MFSI score in the pre-survey was 2.60 for the control group, 2.25 for the anonymizing group, 2.80 for the paraphrasing group, and 3.75 for the revealing group. A one-way ANOVA test revealed no statistically significant difference in MFSI across groups in the pre-survey (F=0.36, p=.785), showing that participants started the moderation task with similar fatigue levels.

In the post-survey, the average MFSI score was 9.05 in the control group, 6.20 in the anonymizing group, 8.35 in the paraphrasing group, and 10.50 in the revealing group. The post-survey one-way ANOVA test indicated no statistically significant difference in fatigue scales across groups (F=0.90, p=.447). As with the SPANE_B results we analyzed previously, we found that participants in all groups felt significantly greater fatigue after the hate speech moderation based on within-subjects comparison of MFSI, but no experimental condition mitigated this increase in fatigue. We report the descriptive statistics and Wilcoxon signed-rank test results in Supplementary material C and D.2, respectively.

In summary, both emotional and fatigue scales showed similar trends: participants felt more negative emotions and greater fatigue after performing hate speech moderation, regardless of the assigned study group.

5.1.3 Perceived Hate Severity. Following our assessment of emotion and fatigue, we examined how participants rated the severity of hate in the comments across all groups. We found that the control group rated the severity of hate in the 50 hate speech comments as significantly higher (mean=3.94, std=0.58) than the anonymizing group (mean=3.56, std=0.75), the paraphrasing group (mean=3.66, std=0.65), and the revealing group (mean=3.58, std=0.80). To clearly observe the distribution of perceived hate severity across groups, independent of individual bias, we applied z-score normalization to the perceived hate severity scores of all 100 comments for each participant (Fig. 6). A one-way ANOVA test found a statistically significant difference in the normalized

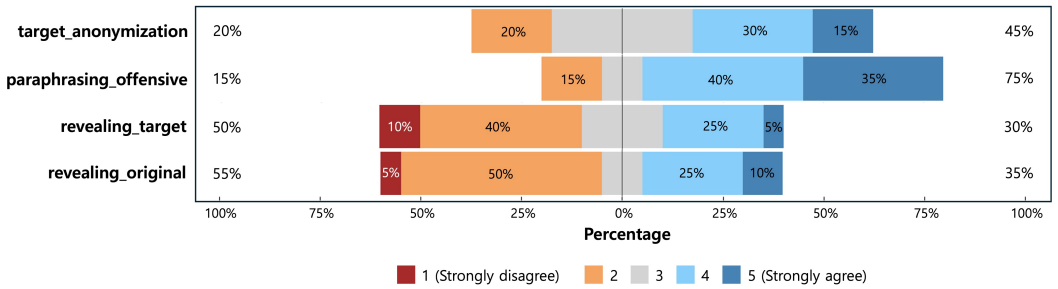


Fig. 7. Distribution of impact on the mental well-being of each feature. Responses range from 1 (Strongly Disagree) to 5 (Strongly Agree).

hate severity distribution across groups (F=6.53, p=.001). Pair-wise two-tailed t-tests revealed that the three experimental groups evaluated the comments as significantly less severe than the control group: the anonymizing group (U=3.54, p=.006), the paraphrasing group (U=4.27, p=.001), and the revealing group (U=3.94, p=.002). There was no significant difference among the experimental groups (i.e., anonymizing, paraphrasing, and revealing groups). This result indicates that `paraphrasing_offensive` and `target_anonymization` reduce the perceived hatefulness of hate speech. These effects appear to persist in *HateBuffer*, even when participants had the option to view the original expressions through `revealing_target` and `revealing_original`. As participants perceived comments as less severe with *HateBuffer*, we next examine how these perceptions contributed to *HateBuffer*'s perceived effectiveness in protecting mental well-being.

5.1.4 Perceived Effectiveness of HateBuffer in Protecting Mental Well-Being. For the final quantitative analysis for RQ1, Fig. 7 illustrates the distribution of the perceived effectiveness of *HateBuffer*'s features for protecting mental well-being based on analysis of a five-point Likert scale survey question. Participants perceived `target_anonymization`'s effectiveness in protecting mental well-being as moderate: 45% of participants agreed or strongly agreed that `target_anonymization` was effective in protecting their mental well-being, while 20% disagreed. Participants' perceptions of the effectiveness of `paraphrasing_offensive` in protecting them from hate speech were more positive, with 75% agreeing that it offered emotional protection.

In contrast to the positive evaluations of `target_anonymization` and `paraphrasing_offensive`, participants had mixed evaluations of the effectiveness of `revealing_target` and `revealing_original` features. While 30% of participants from the revealing group evaluated `revealing_target` as effective in protecting their mental well-being, 50% did not. Similarly, 35% evaluated `revealing_original` as effective in protecting their mental well-being, but 55% disagreed. These results indicate that although `target_anonymization` and `paraphrasing_offensive` are largely perceived as supportive of mental well-being, `revealing_target` and `revealing_original` had mixed evaluations.

Overall, participants felt more negative emotions and greater fatigue after moderating hate speech, regardless of the study condition. However, the three experimental groups rated the severity of hate speech in the comments they saw as less severe than the control group, suggesting that anonymizing targets and paraphrasing offensive expressions can make hate speech appear less hateful.

We now turn to qualitative insights to complement these mixed quantitative findings, exploring participants' experiences and perceptions of *HateBuffer*'s features in supporting their mental well-being. We also discuss potential explanations for these mixed results in § 6.1.

5.2 Qualitative Insights: *HateBuffer*'s Impact on Moderators' Mental Well-Being

Through qualitative analysis, we gain deeper insights into participants' experiences with *HateBuffer*'s features and their perceptions of its role in supporting mental well-being, including its effectiveness in preventing the normalization of hateful and biased opinions.

5.2.1 Aspects of HateBuffer that Were Perceived as Helpful for Mental Well-Being. In the interview, participants shared diverse perspectives on the value and impact of each feature. Participants expressed varying views on `target_anonymization`'s value. While some were **uncertain about anonymization's effectiveness**, others highlighted how it **helped when the participant shared an identity** with the target. For example, A01 noted, *"When there are hate speech comments directed at a certain gender, if I belong to that gender, it could hurt me more and make me feel more depressed than when the target is hidden."* In addition, some participants described `target_anonymization` as effectively preventing emotional contagion from pejorative terms. As A03 explained, *"I think hiding [targets] is much better for my emotions. If I'm repeatedly exposed to unpleasant words like 'Feminazi' [pejorative term for feminist] or 'Democrap' [pejorative term for Democratic Party], I feel like my emotions might get influenced by them."*

For `paraphrasing_offensive`, participants reported **a noticeable impact on emotional well-being**. P01 described, *"When I read the softened expressions, I don't really feel offended."* Some participants emphasized that, although they could infer the original expressions, the protection provided by the `paraphrasing_offensive` was still meaningful. P02 said *"Imagining is just imagining, so I told myself that I was over-interpreting the comments. Because of that, I don't think [moderating the hate speech] really had much of an emotional impact on me."*

When it came to the revealing features, some participants described how **encountering the target of hate speech was surprising**, saying *"When the target was hidden, it felt like vague or meaningless talk, but once I clicked and saw the real subject, it felt more emphasized and dramatic. It didn't make me feel worse or sad, but I was just surprised"* (R02). Participants also mentioned that inferring the original expressions and reading the original felt a bit different. R07 noted *"I had expected similar expressions to some extent, but there were still moments when I felt some discomfort. But it wasn't like my mental state was severely affected."* These reactions suggest that exposure to the original target and offensive expressions was the source of discomfort, and the hiding features gave participants control over whether to experience this.

Some participants explicitly described how they valued this type of control. They explained how `revealing_target` and `revealing_original` enabled a **phased approach that made them feel less offended** by hate speech. R08 explained how *"Going through the somewhat complicated process of checking the original made me feel less offended. So, whether the target was directed at me or my group, I think this more complex process helped reduce the impact on my mental state."* R04 also described how revealing allowed them time to prepare mentally, saying *"For any content that involves hate speech, I feel like I need time to mentally prepare before clicking to reveal it. If I were a moderator, I would probably encounter mostly hate speech, so instead of trying to encounter all the content right away, I think I'd prefer to know first that it's likely to contain violent or offensive language so I can be mentally prepared."*

To sum up, some participants noted how `target_anonymization` and `paraphrasing_offensive` could be effective in preventing their attitudes from being influenced by hate speech, and others noted how `revealing_target` and `revealing_original` worked as a form of positive friction. We next examine how these perceptions influenced their attitudes during the moderation process.

5.2.2 HateBuffer Safeguarded Against Normalization of Biased and Hateful Opinions. Participants in the control group reported concerns that exposure to hate speech not only triggered emotional

Table 3. Descriptive analysis of moderation accuracy and recall.

Group	Moderation accuracy						Moderation recall					
	Mean	Std	Min	Max	Shapiro-Wilk		Mean	Std	Min	Max	Shapiro-Wilk	
					W	p-value					W	p-value
Control	0.79	0.12	0.54	0.96	0.97	.749	0.62	0.28	0.05	0.98	0.95	.375
Anonymizing	0.75	0.13	0.52	0.92	0.94	.212	0.54	0.31	0.01	0.94	0.92	.089
Paraphrasing	0.80	0.09	0.68	0.91	0.97	.851	0.67	0.24	0.34	0.97	0.95	.299
Revealing	0.79	0.11	0.59	0.92	0.97	.752	0.71	0.26	0.27	1.00	0.94	.250

distress but also could lead to **normalization of biased and hateful opinions from the comments**. Per C07, *"... even though I don't personally think the baseball players did anything wrong, reading the comments made me wonder, 'Did they really do something wrong?' ... It seems like these comments make me think more negatively, and I don't feel good about that."*

In contrast, the experimental groups felt that anonymizing specific targets of hate speech or paraphrasing offensive expressions **safeguarded them from the normalization of biased or hateful opinions** from hate speech, creating a needed distance between their perspectives and the negative opinions they encountered during the experiment. For instance, A01 reflected on the role of `target_anonymization` in containing negative emotional contagion: *"When it's hidden, it feels like I'm just dealing with my own negative emotions [about the comments]. If I read comments about a specific country without any hiding, I end up reading all the criticisms directed at that country. This makes me feel like I might start to resonate with those negative emotions and end up disliking that country as well."* Similarly, P05 reflected on the experience of reading hate speech in the recruitment form and said, *"When I read hate speech in the recruitment form, I had the feeling that if I kept seeing these kinds of comments, I might start to think like the people who wrote them. But today, I didn't feel like my emotions became extremely negative, and I was less worried that doing this kind of work continuously would negatively affect my mental state."* These reflections suggest that, while reviewing hate speech can influence participants' attitudes, *HateBuffer* offered a protective layer for moderators' mental well-being by supporting them in maintaining emotional distance from biased and hateful opinions.

5.3 Quantitative Findings: *HateBuffer*'s Impact on Moderation Performance

We explore how altering comment text through the features of *HateBuffer* affects moderation performance (RQ2) by analyzing moderation accuracy and recall, task completion time, and perceived effectiveness for hate speech moderation.

5.3.1 Moderation Accuracy and Recall. Table 3 presents a descriptive analysis of each group's moderation accuracy and recall. Overall, the average accuracy scores were similar across all groups. The average accuracy was 0.79 for the control group, 0.75 for the anonymizing group, 0.80 for the paraphrasing group, and 0.79 for the revealing group. A one-way ANOVA test was conducted, as the Shapiro-Wilk test confirmed a normal accuracy distribution in each group, and no significant differences were found across the groups (H=1.05, p=.377). We also investigated moderation recall, the ratio of deleted hate speech comments to the ground truth hate speech. The average recall was 0.62 for the control group, 0.54 for the anonymizing group, 0.67 for the paraphrasing group, and 0.71 for the revealing group. A one-way ANOVA test was conducted, as the Shapiro-Wilk test confirmed a normal distribution of recall in each group, and no significant differences were found across the groups (H=1.71, p=.173).

To summarize, there was no statistically significant difference in moderation accuracy and recall across groups despite `target_anonymization` and `paraphrasing_offensive` limiting the

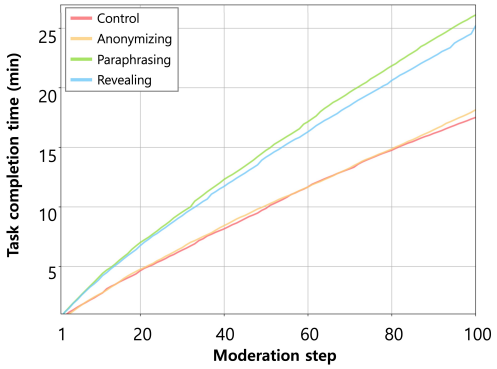


Fig. 8. Cumulative task completion time by moderation step.

Table 4. Descriptive analysis of task completion time in minutes. * indicates p<.05.

Group	Mean	Std	Min	Max	Shapiro-Wilk	
					W	p-value
Control	18.07	5.08	10.45	29.13	0.96	.125
Anonymizing	18.13	7.30	9.60	34.95	0.87	.010*
Paraphrasing	26.10	9.15	12.12	48.47	0.95	.326
Revealing	25.17	9.66	10.72	48.12	0.96	.630

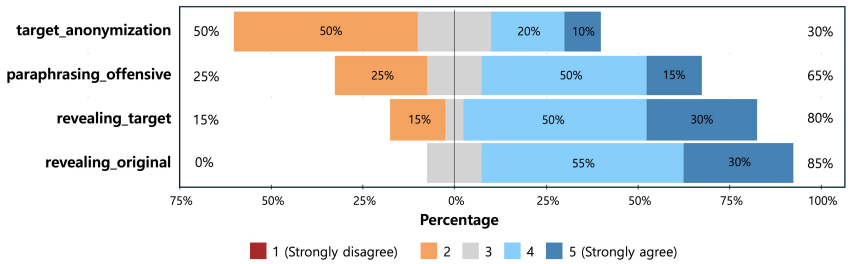


Fig. 9. Distribution of perceived effectiveness of each feature for hate speech moderation. Responses range from 1 (Strongly disagree) to 5 (Strongly agree).

information participants received about the comments. Still, the paraphrasing group demonstrated higher moderation recall than the control group, despite the offensive expressions being paraphrased as less aggressive. The revealing group achieved the highest moderation recall. We explore what enables participants to moderate accurately and sensitively despite content modification in § 5.4.2.

5.3.2 Task Completion Time. Fig. 8 plots the cumulative task completion time across moderation tasks, and Table 4 provides descriptive statistics for the task completion time of each group. Overall, the control and anonymizing groups displayed a similar trend, completing the moderation of 100 comments in an average of 18.07 minutes and 18.13 minutes, respectively. The paraphrasing group took the longest to complete, averaging 26.10 minutes. While the revealing group showed a similar trend to the paraphrasing group, they completed the task slightly faster, with an average time of 25.17 minutes, despite the availability of more interactive features.

Kruskal-Wallis test revealed a statistically significant difference across the groups (H=15.38, p=.002), indicating that at least one group differs significantly from the others. There were significant differences in task completion time between the paraphrasing and control groups (U=84.5, p=.002), paraphrasing and anonymizing groups (U=94.5, p=.005), revealing and control groups (U=106.0, p=.011), and revealing and anonymizing groups (U=110.0, p=.015). We discuss the friction introduced by *HateBuffer*, which led to increased task completion times in hate speech moderation, in §6.2.

5.3.3 Perceived Effectiveness for Hate Speech Moderation. Fig. 9 illustrates the distribution of perceived effectiveness for each *HateBuffer* feature in hate speech moderation. In the anonymizing group, 30% of participants agreed or strongly agreed that `target_anonymization` was helpful for

the moderation task, while 50% disagreed. In the paraphrasing group, 65% of participants agreed or strongly agreed that `paraphrasing_offensive` helped moderate hate speech, while 25% disagreed. In the revealing group, 80% of participants agreed or strongly agreed that `revealing_target` was helpful for the moderation task, while 15% disagreed. Finally, 85% of the participants agreed or strongly agreed that `revealing_original` was beneficial for performing moderation tasks, and none disagreed.

To summarize, we explored the impact of *HateBuffer* on moderation performance and found that, despite `target_anonymization` and `paraphrasing_offensive` modifying the original comments, overall accuracy was not negatively affected. Notably, both the paraphrasing and revealing groups demonstrated increased moderation recall, with `paraphrasing_offensive` providing subtle cues in paraphrased expressions. For the task completion time, the paraphrasing and revealing groups took longer than the others. Overall, participants considered `paraphrasing_offensive`, `revealing_target`, and `revealing_original` features helpful for moderating hate speech, despite `paraphrasing_offensive` modifying the original expressions. Building on these quantitative findings, we next explore qualitative insights into participants' experiences and perceptions of how *HateBuffer*'s features support moderation performance.

5.4 Qualitative Insights: *HateBuffer*'s Impact on Moderation Performance

In this section, we explore participants' strategies in applying *HateBuffer*'s features and examine how these approaches influenced moderation accuracy and recall, shedding light on the benefits and challenges of hate speech moderation with textual content modifications.

5.4.1 Strategies to Moderate Hate Speech with HateBuffer. We first describe participants' strategies for moderating hate speech using *HateBuffer*: focusing on explicit, offensive language, evaluating overall content, and selectively revealing the original expressions. The revealing group varied in their use of revealing features, balancing accuracy in moderation with protecting their emotions.

In the study groups where original offensive expressions were visible (i.e., control, anonymizing, and revealing groups), participants frequently **relied on the presence of explicit expressions** to determine whether a comment constituted hate speech. Many participants reported removing comments with raw and acute expressions. A01 noted, *"I believe that even if the content is the same, the choice of words can determine whether a comment should be kept. ... Some people phrase things more gently, in a way that helps other people improve, while others use provocative words to hurt someone intentionally, and that's the real issue. ... So, I focused on the underlined words [offensive expressions] to see whether they used hateful language."*

In particular, some participants mentioned removing comments containing pejoratives, as these words tended to ridicule certain entities. R06 explained, *"There are cases where the noun itself carries hate speech. For example, when I revealed the target like 'Democratic Party,' I found nouns ['Democrap'] that people use mockingly. In such cases, when I sensed mockery or discrimination, especially toward race, gender, or nationality, I decided to delete the comment if the expression seemed overly aggressive."*

On the other hand, some participants, especially those exposed to milder expressions through `paraphrasing_offensive` (i.e., paraphrasing and revealing groups), made moderation decisions by **assessing the detailed context**, such as keeping logical criticism or removing offensive comments without valid reasoning. P10 described, *"I think I deleted more comments based on the overall content rather than specific words. If the content was fake news or targeted certain regions, countries, or races with hate, I tended to delete those comments first."*

Participants in the revealing group mentioned that they **adopted a selective approach** when using `revealing_target` and `revealing_original`. R04 explained that `revealing_target` was used where determining whether a comment constituted hate speech depended on the target.

Referring to a comment in Fig. 1, she noted, *"If the target here had been something like, 'It's criminals who bring embarrassing moment or downfall to Korea,' I wouldn't have considered it hate speech. So, in cases like this, I checked the target."* R02 explained he used `revealing_original` only when the flow of altered comments felt awkward, saying *"I only checked the original expression when the sentence didn't match the previous or following sentences. Otherwise, I could generally understand the comment, so I didn't often check the original expressions."*

The reasons for the selective use of revealing features were clustered into two purposes: accurate moderation and mental well-being protection. Half of the interviewed participants from the revealing group said they frequently used `revealing_target` and `revealing_original`, as they found it challenging to determine whether a comment should be deleted or kept based solely on the anonymized or paraphrased expressions. R01 described *"When I saw only the softened expression, I couldn't really tell whether the comment was something that should be deleted, so I immediately checked the original version."* This frequent use of revealing functions was driven by their sense of responsibility as moderators, emphasizing accuracy in their moderation decisions. R02 noted *"... In the end, it was my role to make a judgment, whether this is too hateful, so I checked everything to review the (original) content."*

In contrast to the frequent use of `revealing_target` and `revealing_original`, a few participants expressed concerns on the potential emotional impact of confronting offensive expressions. R09 explained, *"I felt it was less stressful not to click and reveal everything, so I often just skipped over comments with that kind of (offensive) tone without even looking at the original version. ... I tried to avoid looking at the original as much as possible"*. This participant ended up using the `revealing_original` feature only twice.

The observed strategies illustrate how participants engaged with *HateBuffer*'s content modification features to moderate hate speech effectively with less mental burden. Building on these insights, the next sections illustrate how participants described the specific impact of *HateBuffer*'s features on moderation accuracy and recall, examining how each feature contributed to or hindered participants' ability to make precise moderation decisions.

5.4.2 Impact of HateBuffer on Precise Moderation Decisions. The anonymizing group often mentioned that they could *"infer the target based on the revealed information, though targets were anonymized"* (A08). However, they also reported that **`target_anonymization` hindered accurate moderation** in certain cases, especially when a comment could be interpreted as either an ordinary opinion or hate speech, depending on the target. A04 noted, *"I found it difficult to judge whether the comment should be deleted. Once the anonymized target is revealed, it feels like they might not be genuine hate toward a specific group, but more like fake news."* Another participant, A06, expanded on the issue, pointing out that certain words could describe factual observations but might still be perceived as hate speech depending on the context: *"Words like 'incompetent' are negative, but depending on the target, someone might think it's a fair opinion. For example, saying someone is incompetent could be based on actual data; in that case, deleting the comment doesn't seem right."* This uncertainty around `target_anonymization` was also reflected in the lower perceived effectiveness of `target_anonymization`, as reported in § 5.3.3.

In the paraphrasing group, participants explained they could still sense the nuance of the offensive content. P09 noted, *"Even if an offensive expression is toned down, it still feels like it demeans this group. From the overall context, even if the aggressive wording is reduced, the sentence itself still demeans or disparages a certain group."* Some participants mentioned that they **could infer the original offensive expressions**, saying, *"You know, if it's something like 'a person with foggy mind,' you can pretty much guess the kind of insult that's implied"* (P08). Understanding the overall context and inferring the original offensive expressions behind `paraphrasing_offensive` could

make participants more sensitive to hate speech, requiring careful consideration. Moreover, P02 mentioned that she "*took action to delete based on imagining the worst-case scenario in many cases.*"

Meanwhile, both anonymizing and paraphrasing groups expressed that they wanted to delete some comments that had been kept while reviewing their moderated comment lists during the interview. The comments they referred to often contained pejorative terms, which were anonymized by the `target_anonymization` or replaced with neutral targets by the `paraphrasing_offensive`. For instance, terms such as 'Feminazi,' 'Ching chong' (a pejorative term for Chinese), and 'Nip' (a pejorative term for Japanese) were mentioned by participants. Participants from the revealing group, who initially encountered modified comments through `target_anonymization` and `paraphrasing_offensive` but had the option to view the original expressions, were **able to moderate comments containing these pejoratives**, which might explain their higher recall than the other groups. R06 noted, *"For example, when the target is anonymized, technically, it's not considered hate speech. But in cases like calling Japanese people 'monkeys,' the noun itself becomes hate speech, and seeing these terms made me decide to delete them."*

In summary, participants adopted different strategies depending on the visibility of original expressions, either focusing on explicit expressions or evaluating the broader context. In addition, participants exhibited a bifurcated approach to using the revealing features: some prioritized accuracy in moderation, while others engaged the features selectively to protect their emotional well-being. While `target_anonymization` may compromise accurate moderation by limiting information, `paraphrasing_offensive` encourages a more complex moderation practice, prompting moderators to infer the original expressions and interpret the overall contents carefully. Furthermore, `revealing_target` and `revealing_original` facilitated precise moderation by allowing them to verify pejoratives.

6 Discussion

We evaluated *HateBuffer*, a text content modification system designed to support hate speech moderation while protecting moderators' mental well-being. *HateBuffer* offers four main features: `target_anonymization`, which anonymizes the target of hate speech; `paraphrasing_offensive`, which paraphrases offensive expressions into less harmful language; and `revealing_target` and `revealing_original`, which allows users to reveal the target and original expressions with a click. Contrary to our expectations, we did not observe a significant improvement in emotional state or a reduction in fatigue after moderation when comparing the experimental groups with the control group. However, the experimental groups considered modified comments less severe and perceived *HateBuffer* to effectively protect their mental well-being.

Furthermore, the moderation accuracy remained similar despite *HateBuffer* modifying the comments by anonymizing the target and paraphrasing offensive expressions into less offensive ones. Notably, the participants who used `paraphrasing_offensive` showed slightly higher moderation recall. In interviews, our participants described the revealing features of *HateBuffer* as a type of *buffer*, providing time for them to prepare to face the offensive original expressions. They also noted that *HateBuffer* prevented them from normalizing biased and hateful opinions from the comments by anonymizing targets and paraphrasing offensive expressions. Despite such positive aspects, we did not find clear evidence that *HateBuffer* can better safeguard a moderator's overall emotion and fatigue after moderation.

In this section, we discuss possible explanations for the mixed findings between perceived benefits and actual impacts of *HateBuffer* on mental well-being. We then examine in depth how hate speech moderation can still be accurate despite textual modifications, highlighting further considerations for content modification in text content moderation. Additionally, we shed light

on possible strategies for protecting moderators' mental well-being for a sustainable working environment.

6.1 Understanding the Discrepancy Between Perceived Benefits and Actual Impact on Mental Well-Being

Our findings revealed a discrepancy between measured outcomes and participants' opinions: while participants perceived comments modified by *HateBuffer* as less hateful and positively viewed the effectiveness of `target_anonymization` and `paraphrasing_offensive`, there was no significant difference in the emotion scale or fatigue levels between experimental groups and the control group. Several factors may contribute to this gap between perceived benefits and actual impact on mental well-being.

One potential explanation is that while the paraphrasing and revealing groups evaluated the modified comments as less hateful than the control group, they also took significantly longer to process each comment owing to the uncertainty induced by paraphrasing offensive phrases. Although these features helped lower the perceived severity of each comment, our participants spent a longer time on moderation, resulting in longer cognitive engagement with each potentially hateful comment. This might have contributed to significant changes in emotion and fatigue scales after the experiment. In other words, *HateBuffer* reduced the intensity of hate speech, but the longer exposure may have had a negative impact, potentially diminishing the intended emotional protective effect.

Additionally, the cognitive load associated with inference generation could cause negative emotion and fatigue at the end of the experiment. Our participants indicated that as the target of hate speech was anonymized and the offensive expression was paraphrased, they often inferred and imagined the original expression to understand the full context of a comment and make a moderation decision. This inference process, which requires considerable cognitive effort [83], may have affected participants in the experimental groups, resulting in changes in emotion and fatigue levels similar to those observed in the control group.

Another consideration is the short duration of hate speech exposure in our study. Participants moderated 100 comments, with only 50 containing hate speech, and the control group took an average of 18.07 minutes in total moderation time, with the shortest session lasting only 10.45 minutes. Given this limited exposure, it is possible that the system's protective benefits were not fully realized in such a short session. In contrast, commercial moderators typically work extended hours each day [105]. A longer moderation session using *HateBuffer* could better reveal its protective effects, not only in perceived benefits but also in measurable outcomes for mental well-being.

The results of this study show that text content modification systems such as *HateBuffer* have potential, but underlying complexities impact their real-world effectiveness in protecting moderators' mental well-being. Future studies might examine how these factors (e.g., exposure duration, perceived severity, and cognitive load) interact over longer moderation sessions to better understand how text content modification tools can sustainably support moderators' mental well-being.

6.2 Text Content Modification in the Context of Content Moderation

We found that using `target_anonymization` and `paraphrasing_offensive` maintained comparable moderation accuracy to a control condition, even with anonymized targets and paraphrased offensive expressions. This resilience may stem from *inference generation* processes, where readers actively use their knowledge and context to fill gaps in a text to create logical conclusions [83]. Our participants explained that they inferred the original hateful meaning to understand and make moderation decisions, mentally reconstructing the likely intent of paraphrased or concealed

expressions. This ability to *fill in* the modified elements of hate speech, even when information was limited, allowed participants to maintain accuracy comparable to that of the control group.

Another unexpected finding was that participants in the paraphrasing and revealing groups, who moderated comments using `paraphrasing_offensive`, exhibited slightly higher recall than the control group. This may also be attributed to the inference generation process, as participants took more time to reconstruct the possible offensive intent behind paraphrased comments, resulting in more careful judgments. This aligns with the dual process theory of human decision making [96], which is also known as the concept of *thinking fast and slow*—system 2's slow thinking involves a deeper and more thoughtful evaluation of statements and their implications, whereas system 1's fast thinking lacks such deliberate reasoning [23, 66].

Anonymizing targets or paraphrasing offensive phrases fosters a reflective approach to moderation, as it introduces uncertainty that requires moderators to actively infer concealed targets or offensive meanings. We posit that this uncertainty in user interactions facilitates mindful interaction, similar to traditional 'interaction lockout' or 'friction' designs, which restrict user interactions for safety or prevent human errors [26]. Our moderation features can introduce a lightweight interaction lockout that slows cognitive processing and encourages deliberate thought.

While uncertainty or ambiguity in HCI literature has historically been explored to inspire design and enrich hedonic interactions [39], its role is expanding. Traditionally, designers introduced elements of ambiguity or uncertainty (e.g., information, context, and relationship) to foster curiosity, enhance engagement, and encourage self-reflection, particularly in creative and playful systems. Uncertainty also plays an important role in pragmatic applications that involve data contextualization and sense-making [69]. In content moderation contexts, such elements of uncertainty not only create a psychological buffer for content moderation but also possibly improve the performance of content moderation (e.g., higher recall in hate speech moderation).

While our study underscores the potential of text content modification for hate speech moderation, additional directions remain for exploration. One possible consideration is the question of authorship in modified user-generated content. Given that *HateBuffer* presents content in its modified form for evaluation, users might argue that moderation decisions do not accurately reflect the original intent or expression behind what they had written. Authorship of LLM-based paraphrased content has been actively discussed in AI and Human-AI Interaction fields [127, 142], where authorship typically depends on two factors: *content*, representing the subject matter of thematic focus, and *style*, the distinctive manner of expression [110]. Recent research has argued that LLM-based paraphrasing retains the core *content* while altering the style [127]. Traditional views on authorship often emphasize the author's unique ideas, concepts, or thoughts, aligning more closely with *content* [36, 109]. Given that *HateBuffer* preserves core *content* (i.e., hateful intent) while modifying *style* to reduce offensive language, moderation decisions based on the modified content could reasonably translate to judgments on the original version. However, further investigation is needed to understand how users perceive having their content assessed in a modified form and how they might react to decisions based on these modifications.

Additionally, when users request a reconsideration of decisions they find incorrect or unfair, the appeals process generally allows them to explain in free text why they believe the decision was incorrect [129]. However, text content modification through *HateBuffer* introduces an information asymmetry: moderators view the modified text, while appealers present arguments based on their original content. Further research exploring the impact of this information disparity could be valuable in guiding the development of a fair and transparent appeal process for situations where text content modification techniques are used for moderation.

6.3 Toward a Sustainable Career in Content Moderation

Our qualitative findings highlight the potential for text content modification in hate speech moderation to support moderators' mental well-being. Participants discussed how *HateBuffer* protected their personal viewpoints from being shaped by biased and hateful opinions from the comments. Unlike image and video moderation, where explicit visual stimuli (e.g., violent or sexual content) can trigger immediate, sensory-based reactions [132], text-based hate speech presents a distinct set of concerns via repeatedly exposing moderators to biased or extreme opinions. Moderating this type of content requires individuals to engage in a complex cognitive process to interpret and reconstruct the underlying message, as discussed in §6.2, which may have long-term effects on their personal perspectives.

Repeated and prolonged exposure to such violent, negatively stimulating content places moderators at risk of longer-term psychological damage, including emotional desensitization — characterized by a gradual numbing of emotional reactions and reduced empathy toward real-world situations [71, 78] — or, conversely, emotional sensitization, wherein repeated exposure heightens their responsiveness to stressful stimuli [10]. Recent work has documented instances where experienced moderators exhibited adverse reactions even to traditionally positive emotional stimuli intended to alleviate stress, underscoring the complexities of sensitization and its challenges for effective emotional support [24]. Although our study did not directly measure sensitization or desensitization, participants indicated that content modification, anonymizing target expressions, and/or paraphrasing offensive expressions could potentially reduce their mental burden. Such support may help moderators sustain their roles over a longer period, creating a more stable working environment and reducing turnover rates commonly observed in moderation work [18].

The revealing features of our study allowed participants to selectively view the targets and original offensive expressions, providing a sense of control. Facebook's Global Resiliency Team has noted that *"shielding moderators from harm begins with giving them more control over what they see and how they see it"* [125]. Given that a lower sense of control in high-stress workplaces can increase stress levels [1], tools such as *HateBuffer* could enhance moderators' agency over their exposure to potentially hateful text-based content. In addition, our participants found revealing features helpful in providing them time to mentally prepare before encountering potentially distressing language, consistent with findings of previous work [68]. Considering the repetitive and emotionally demanding nature of moderation [105, 122], this sense of control supports self-efficacy and promotes healthier engagement with the work.

With that said, implementing content modification tools such as *HateBuffer* in real-world moderation settings involves navigating a tension between performance and moderator well-being. Our findings show that reviewing modified content, designed to reduce the offensiveness of hate speech, led to longer task completion time while maintaining quantitative fatigue levels. These outcomes are consistent with observations with commercial moderators, who often prefer reviewing a smaller number of severe cases over a high volume of mildly offensive content, citing the increased cognitive load and time demands of the latter [124]. In such settings, especially where strict performance quotas are in place [16, 92, 139], the need to interpret softened expressions more carefully may contribute to slower moderation. However, given that commercial platforms typically provide far more detailed moderation guidelines than those used in our study [92], it remains unclear how such content modifications would affect moderators' speed and experience in practice. Future work should examine how tool use, time constraints, and policy design interact to shape moderation outcomes and moderator well-being in real-world environments.

In addition, prior research in content moderation has emphasized the importance of providing appropriate rest, which can benefit moderators' mental well-being [24, 33, 122]; structured breaks

are essential in maintaining workplace well-being and reducing fatigue [9]. While a standard recommendation for break duration is known as a 7.5-minute break after 50 minutes of content review [11], break scheduling could be further adjusted based on the frequency of exposure to original, unmodified content, especially when content modification techniques are used in text moderation. Future research should explore the optimal balance of modified and original content exposure and break frequency and duration to effectively support moderators' mental well-being.

6.4 Limitations and Future Work

In this paper, we present insights from a mixed-methods approach, combining quantitative analysis with in-depth qualitative insights, which allows us to explore the potential for text content modification to support moderators' mental well-being in hate speech moderation. In addition, our user study was conducted using a simulated moderation task with pre-curated data. This controlled setup enabled us to protect participants from encountering unexpected harmful content, but it may not fully capture the range and complexity of real-world hate speech moderation scenarios. Implementing a full pipeline that includes AI-based detection of targets and offensive expression through techniques such as entity recognition [76] and sentiment analysis [136], followed by paraphrasing via LLMs could facilitate field studies with real-world data. Moreover, evaluating the actual moderation accuracy with *HateBuffer* through large-scale studies in real-world settings could provide deeper insights into the practical applicability of content modification. Such studies could offer more comprehensive insights into *HateBuffer*'s effectiveness within content moderation's dynamic, unpredictable nature.

Additionally, our simulated moderation experiment utilized the K-HATERS dataset [99], which consists of news article comments that may lack full context. Although we made efforts to select comments that seemed understandable independently of their original news articles, there remains a possibility that participants might not fully grasp certain context-dependent meanings. However, context dependency is not unique to news comments; moderation of content from other platforms, such as X's threads, Reddit's discussions, or YouTube's comments, may similarly face challenges if moderators review comments without the original root content. Therefore, investigating how moderation performance and moderators' experiences with *HateBuffer* vary across diverse content types and platforms, including different degrees of contextual availability, would be a valuable direction for future research.

Furthermore, since the K-HATERS dataset includes topics specifically relevant to Korean society, such as gender issues and internal regional discrimination, our findings may have limited generalizability across different cultural contexts. Hate speech varies significantly across cultures, reflecting unique social, political, and historical factors [38]. For example, prevalent topics of hate speech in other cultures, such as gun control and immigration, were not present in our dataset. Even for the same topic, the targeted race or event may differ [19]. To gain a broader understanding of text content modification systems' impact on hate speech moderation, it would be valuable to investigate their effects across various cultural contexts, expanding our observations to account for global dynamic social issues. Moreover, while *HateBuffer*'s `paraphrasing_offensive` was designed to paraphrase hate speech effectively, a key challenge remains in how accurately LLMs can interpret and respond to cultural nuances. As the types, targets, and subtleties of hate speech vary widely across cultures, incorporating a culturally adaptive or localized model [93] could enhance *HateBuffer*'s ability to moderate contextually relevant hate speech and increase its applicability across diverse cultural backgrounds.

7 Conclusion

We designed *HateBuffer*, a text-based content modification system for hate speech moderation, to safeguard moderators' mental well-being while preserving moderation performance. We conducted a user study with 80 participants who were assigned the role of moderators to perform simulated hate speech moderation from a fictional news platform and observed qualitative insights through semi-structured interviews. In contrast to our expectation, we could not observe any improvement in emotion and fatigue after hate speech moderation with *HateBuffer* compared with the control group. However, the perceived hate severity of comments was significantly lower when *HateBuffer* is used, and participants recognized *HateBuffer* as an effective *buffer* for emotional contagion and normalization of biased opinions from hate speech. Notably, *HateBuffer* did not hinder the moderation accuracy, even enabling slightly higher recall. Building on these findings, we explored possible explanations for the discrepancy between perceived benefits and the actual impact of *HateBuffer* on mental well-being. We highlighted the potential of the content modification technique in the text as a content moderation tool and mental well-being protection tool, fostering a more sustainable working environment for content moderators.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-02263169, Detection and Prediction of Emerging and Undiscovered Voice Phishing), funded by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2022-II220064), and funded by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00348993).

References

[1] Jafar Akbari, Rouhollah Akbari, Mahnaz Shakerian, and Behzad Mahaki. 2017. Job demand-control and job stress at work: A cross-sectional study among prison staff. *Journal of education and health promotion* 6, 1 (2017), 15.

[2] Badr AlKhamissi, Faisal Ladhak, Srini Iyer, Ves Stoyanov, Zornitsa Kozareva, Xian Li, Pascale Fung, Lambert Mathias, Asli Celikyilmaz, and Mona Diab. 2023. ToKen: Task Decomposition and Knowledge Infusion for Few-Shot Hate Speech Detection. arXiv:2205.12495 [cs.CL] https://arxiv.org/abs/2205.12495

[3] Ali Alkhatib and Michael Bernstein. 2019. Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, 1–13. https://doi.org/10.1145/3290605.3300760

[4] Amazon Jobs. 2025. Content Reviewer, Amazon. https://amazon.jobs/en/search?base_query=content+reviewer&loc_query=&latitude=&longitude=&loc_group_id=&invalid_location=false&country=&city=&region=&county=

[5] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 759–760. https://doi.org/10.1145/3041021.3054223

[6] John Bargh, Mark Chen, and Lara Burrows. 1996. Automaticity of social behaviour: direct effects of trait construct and stereotype priming on action. *Journal of personality and social psychology* 71 (08 1996), 230–44. https://doi.org/10.1037/0022-3514.71.2.230

[7] Paul Barrett. 2020. *Who Moderates the Social Media Giants?* NYU Center for Business and Human Rights. https://bhr.stern.nyu.edu/wp-content/uploads/2024/02/NYUContentModerationReport_FINALVERSION.pdf

[8] Roukaya Benjelloun and Yassine Otheman. 2020. Psychological distress in a social media content moderator: A case report. *Archives of Psychiatry and Mental Health* 4, 1 (2020), 10.

[9] Gerhard Blasche, Sanja Pasalic, Verena-Maria Bauböck, Daniela Haluza, and Rudolf Schoberberger. 2017. Effects of rest-break intention on rest-break frequency and work-related fatigue. *Human factors* 59, 2 (2017), 289–298.

[10] Daniel T Blumstein. 2016. Habituation and sensitization: new thoughts about old ideas. *Animal behaviour* 120 (2016), 255–262.

[11] Wolfram Boucsein and Michael Thum. 1997. Design of work/rest schedules for computer work based on psychophysiological recovery measures. *International Journal of Industrial Ergonomics* 20, 1 (1997), 51–57.

[12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[13] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 1664–1674. https://doi.org/10.18653/v1/D19-1176

[14] Catherine Buerger. 2021. Speech as a driver of intergroup violence: A literature review. *Available at SSRN 4066876* (2021).

[15] Armin Burkhardt. 2010. Euphemism and truth. *Tropical truth (s): The epistemology of metaphor and other tropes* (2010), 355–372.

[16] Jie Cai, Donghee Yvette Wohn, and Mashael Almoqbel. 2021. Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities. In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences* (Virtual Event, USA) *(IMX '21)*. Association for Computing Machinery, New York, NY, USA, 61–72. https://doi.org/10.1145/3452918.3458796

[17] Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten W Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. 2024. Explainability and Hate Speech: Structured Explanations Make Social Media Moderators Faster. *arXiv preprint arXiv:2406.04106* (2024).

[18] Caroline Kimeu. 2024. 'The work damanged me': ex-Facebook moderators describe effect of horrific content. https://www.theguardian.com/technology/2024/dec/18/ex-facebook-moderators-describe-effect-of-horrific-content?utm_source=chatgpt.com Accessed: August 20, 2025.

[19] Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. 2021. Internet, social media and online hate speech. Systematic review. *Aggression and violent behavior* 58 (2021), 101608.

[20] Zeya Chen and Ruth Schmidt. 2024. Exploring a Behavioral Model of "Positive Friction" in Human-AI Interaction. In *Design, User Experience, and Usability*, Aaron Marcus, Elizabeth Rosenzweig, and Marcelo M. Soares (Eds.). Springer Nature Switzerland, Cham, 3–22.

[21] Arik Cheshin, Anat Rafaeli, and Nathan Bos. 2011. Anger and happiness in virtual teams: Emotional influences of text and behavior on others' affect in the absence of non-verbal cues. *Organizational Behavior and Human Decision*

Processes 116, 1 (2011), 2–16. https://doi.org/10.1016/j.obhdp.2011.06.002

[22] Frederick Choi, Tanvi Bajpai, Sowmya Pratipati, and Eshwar Chandrasekharan. 2023. ConvEx: A Visual Conversation Exploration System for Discord Moderators. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–30.

[23] Ryuhaerang Choi, Subin Park, Sujin Han, and Sung-Ju Lee. 2024. FoodCensor: Promoting Mindful Digital Food Content Consumption for People with Eating Disorders. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[24] 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. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 277 (Nov. 2022), 19 pages. https://doi.org/10.1145/3555168

[25] Gloria Cowan and Cyndi Hodge. 1996. Judgments of Hate Speech: The Effects of Target Group, Publicness, and Behavioral Responses of the Target. *Journal of Applied Social Psychology* 26 (02 1996), 355 – 374. https://doi.org/10.1111/j.1559-1816.1996.tb01854.x

[26] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1389–1397. https://doi.org/10.1145/2851581.2892410

[27] Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8 (10 2020), 33–42. https://doi.org/10.1609/hcomp.v8i1.7461

[28] Mithun Das, Binny Mathew, Punyajoy Saha, Pawan Goyal, and Animesh Mukherjee. 2020. Hate speech in online social media. *SIGWEB Newsl.* 2020, Autumn, Article 4 (Nov. 2020), 8 pages. https://doi.org/10.1145/3427478.3427482

[29] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. arXiv:1703.04009 [cs.CL] https://arxiv.org/abs/1703.04009

[30] Ed Diener, Derrick Wirtz, William Tov, Chu Kim-Prieto, Dong-won Choi, Shigehiro Oishi, and Robert Biswas-Diener. 2010. New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social indicators research* 97 (2010), 143–156.

[31] Discord. 2022. Auto Moderation in Discord. https://discord.com/safety/auto-moderation-in-discord Accessed: August 20, 2025.

[32] Hany Farid. 2021. An Overview of Perceptual Hashing. *Journal of Online Trust and Safety* 1, 1 (Oct. 2021). https://doi.org/10.54501/jots.v1i1.24

[33] Julia Faucett, James Meyers, John Miles, Ira Janowitz, and Fadi Fathallah. 2007. Rest break interventions in stoop labor tasks. *Applied ergonomics* 38, 2 (2007), 219–226.

[34] Emilio Ferrara and Zeyao Yang. 2015. Measuring Emotional Contagion in Social Media. *PLOS ONE* 10, 11 (11 2015), 1–14. https://doi.org/10.1371/journal.pone.0142390

[35] Google Firebase. 2024. Firestore. https://firebase.google.com. Accessed: August 20, 2025.

[36] Michel Foucault. 2003. What is an Author? In *Reading architectural history*. Routledge, 71–81.

[37] Imogen Foulkes Fracnes Mao. 2024. Crack down on racist hate speech, UN tells UK. https://www.bbc.com/news/articles/cgl21053rdzo. Accessed: August 20, 2025.

[38] Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems* 41, 8 (2024), e13562.

[39] William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a resource for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/642611.642653

[40] T. Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press. https://books.google.co.kr/books?id=cOJgDwAAQBAJ

[41] Amit Goldenberg and James Gross. 2019. Digital Emotion Contagion. https://doi.org/10.31219/osf.io/53bdu

[42] Vaishali Gongane, Mousami Munot, and Devidas Anuse. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining* 12 (09 2022). https://doi.org/10.1007/s13278-022-00951-3

[43] Google. n.d.. Google Jigsaw. https://jigsaw.google.com/. Accessed: August 20, 2025.

[44] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 388, 14 pages. https://doi.org/10.1145/3411764.3445423

[45] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7 (02 2020), 205395171989794.

https://doi.org/10.1177/2053951719897945

[46] Nitesh Goyal, Leslie Park, and Lucy Vasserman. 2022. "You have to prove the threat is real": Understanding the needs of Female Journalists and Activists to Document and Report Online Harassment. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 242, 17 pages. https://doi.org/10.1145/3491102.3517517

[47] Jamie Guillory, Jason Spiegel, Molly Drislane, Benjamin Weiss, Walter Donner, and Jeffrey Hancock. 2011. Upset now? emotion contagion in distributed groups. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 745–748. https://doi.org/10.1145/1978942.1979049

[48] Uma Gunturi, Xiaohan Ding, and Eugenia H. Rho. 2023. ToxVis: Enabling Interpretability of Implicit vs. Explicit Toxicity Detection Models with Interactive Visualization. arXiv:2303.09402 [cs.CL] https://arxiv.org/abs/2303.09402

[49] Oliver L. Haimson, Justin Buss, Zu Weinger, Denny L. Starks, Dykee Gorrell, and Briar Sweetbriar Baron. 2020. Trans Time: Safety, Privacy, and Content Warnings on a Transgender-Specific Social Media Site. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 124 (Oct. 2020), 27 pages. https://doi.org/10.1145/3415195

[50] Catherine Han, Joseph Seering, Deepak Kumar, Jeffrey T. Hancock, and Zakir Durumeric. 2023. Hate Raids on Twitch: Echoes of the Past, New Modalities, and Implications for Platform Governance. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 133 (April 2023), 28 pages. https://doi.org/10.1145/3579609

[51] Jeffrey T. Hancock, Kailyn Gee, Kevin Ciaccio, and Jennifer Mae-Hwah Lin. 2008. I'm sad you're sad: emotional contagion in CMC. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA, USA) *(CSCW '08)*. Association for Computing Machinery, New York, NY, USA, 295–298. https://doi.org/10.1145/1460563.1460611

[52] David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025. Lost in Moderation: How Commercial Content Moderation APIs Over-and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations. *arXiv preprint arXiv:2503.01623* (2025).

[53] Ming Shan Hee, Karandeep Singh, Charlotte Ng Si Min, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. Brinjal: A Web-Plugin for Collaborative Hate Speech Detection. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1063–1066. https://doi.org/10.1145/3589335.3651250

[54] YouTube Help. 2024. Manage your community & comments: Choose comment settings. https://support.google.com/youtube/answer/9482556?hl=en. Accessed: August 20, 2025.

[55] Carolina Herrando and Efthymios Constantinides. 2021. Emotional Contagion: A Brief Overview and Future Directions. *Frontiers in Psychology* 12 (2021). https://doi.org/10.3389/fpsyg.2021.712606

[56] Sharon Heung, Lucy Jiang, Shiri Azenkot, and Aditya Vashistha. 2024. "Vulnerable, Victimized, and Objectified": Understanding Ableist Hate and Harassment Experienced by Disabled Content Creators on Social Media. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 744, 19 pages. https://doi.org/10.1145/3613904.3641949

[57] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch your heart: A tone-aware chatbot for customer care on social media. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[58] Indeed. 2025. Trust & Safety – Content Reviewer. https://www.indeed.com/q-content-reviewer-jobs.html?vjk=a67aa177858229c0

[59] Influencer Marketing Hub. 2024. The Rise of User-Generated Content (UGC). https://influencermarketinghub.com/rise-of-ugc/ [Accessed: August 20, 2025].

[60] Lasal Jayawardena and Prasan Yapa. 2024. ParaFusion: A Large-Scale LLM-Driven English Paraphrase Dataset Infused with High-Quality Lexical and Syntactic Diversity. *arXiv preprint arXiv:2404.12010* (2024).

[61] Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean Offensive Language Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 10818–10833. https://doi.org/10.18653/v1/2022.emnlp-main.744

[62] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. 26, 5, Article 31 (July 2019), 35 pages. https://doi.org/10.1145/3338243

[63] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X. Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 205, 21 pages. https://doi.org/10.1145/3491102.3517505

[64] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PLOS ONE* 16, 8 (08 2021), 1–22. https://doi.org/10.1371/journal.

pone.0256762
[65] Mark Johnson. 2018. Inclusion and exclusion in the digital economy: disability and mental health as a live streamer on Twitch.tv. *Information, Communication & Society* 22 (05 2018), 1–15. https://doi.org/10.1080/1369118X.2018.1476575
[66] Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux* (2011).
[67] Sowmya Karunakaran and Rashmi Ramakrishan. 2019. Testing stylistic interventions to reduce emotional impact of content moderation workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 50–58.
[68] Haesoo Kim, Juhoon Lee, Jeong-Woo Jang, and Juho Kim. 2024. ReSPect: Enabling Active and Scalable Responses to Networked Online Harassment. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 117 (April 2024), 30 pages. https://doi.org/10.1145/3637394
[69] Inyeop Kim and Uichin Lee. 2024. Navigating User-System Gaps: Understanding User-Interactions in User-Centric Context-Aware Systems for Digital Well-being Intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 249, 15 pages. https://doi.org/10.1145/3613904.3641979
[70] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. LLM-Mod: Can Large Language Models Assist Content Moderation?. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 217, 8 pages. https://doi.org/10.1145/3613905.3650828
[71] Barbara Krahé, Ingrid Möller, L Rowell Huesmann, Lucyna Kirwil, Juliane Felber, and Anja Berger. 2011. Desensitization to media violence: links with habitual media violence exposure, aggressive cognitions, and aggressive behavior. *Journal of personality and social psychology* 100, 4 (2011), 630.
[72] Tina Kuo, Alicia Hernani, and Jens Grossklags. 2023. The Unsung Heroes of Facebook Groups Moderation: A Case Study of Moderation Practices and Tools. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 97 (April 2023), 38 pages. https://doi.org/10.1145/3579530
[73] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. http://www.jstor.org/stable/2529310
[74] 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* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 1701–1721. https://doi.org/10.1145/3643834.3660708
[75] Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022. K-MHaS: A Multi-label Hate Speech Detection Dataset in Korean Online News Comment. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 3530–3538. https://aclanthology.org/2022.coling-1.311/
[76] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering* 34, 1 (2020), 50–70.
[77] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "HOT" ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media. *ACM Trans. Web* 18, 2, Article 30 (March 2024), 36 pages. https://doi.org/10.1145/3643829
[78] Daniel Linz, Edward Donnerstein, and Steven Penrod. 1984. The effects of multiple exposures to filmed violence against women. *Journal of Communication* 34, 3 (1984), 130–147.
[79] Yonatan Lupu, Richard Sear, Nicolas Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Beth Goldberg, and Neil F Johnson. 2023. Offline events and online hate. *PLoS one* 18, 1 (2023), e0278511.
[80] Sarah Masud, Subhabrata Dutta, Sakshi Makkar, Chhavi Jain, Vikram Goyal, Amitava Das, and Tanmoy Chakraborty. 2021. Hate is the New Infodemic: A Topic-aware Modeling of Hate Speech Diffusion on Twitter. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. 504–515. https://doi.org/10.1109/ICDE51399.2021.00050
[81] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 14867–14875.
[82] Michael Mayor. 2009. *Longman dictionary of contemporary English*. Pearson Education India.
[83] Gail McKoon and Roger Ratcliff. 1992. Inference during reading. *Psychological review* 99, 3 (1992), 440.
[84] Ivan Mehta. 2024. Spotify adds auto-moderation tool to help podcasters manage comments. https://techcrunch.com/2024/10/24/spotify-adds-auto-moderation-tool-to-help-podcasters-manage-comments/. Accessed: August 20, 2025.
[85] Meta. 2025. Hateful Conduct. https://transparency.meta.com/en-us/policies/community-standards/hateful-conduct/ Accessed: August 20, 2025.
[86] Meta (Facebook). 2024. Community Standards Enforcement Report - Hate Speech. https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook/ [Accessed: August 20, 2025].

[87] Meta (Facebook). n.d. How do I block certain words from appearing in comments on my Facebook Page? https://www.facebook.com/help/131671940241729?helpref=faq_content [Accessed: August 20, 2025].
[88] Meta (Instagram). n.d. Hide comments or message requests you don't want to see on Instagram. https://help.instagram.com/700284123459336?cms_id=700284123459336 [Accessed: August 20, 2025].
[89] Microsoft. n.d.. Microsoft Azure AI Content Safety. https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety. Accessed: August 20, 2025.
[90] Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, Lun-Wei Ku and Cheng-Te Li (Eds.). Association for Computational Linguistics, Online, 25–31. https://doi.org/10.18653/v1/2020.socialnlp-1.4
[91] Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 742, 22 pages. https://doi.org/10.1145/3613904.3642025
[92] Arvind Narayanan and Sayash Kapoor. 2024. AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference. In *AI Snake Oil*. Princeton University Press.
[93] Rachna Narula and Poonam Chaudhary. 2024. A comprehensive review on detection of hate speech for multi-lingual data. *Social Network Analysis and Mining* 14, 1 (2024), 1–35.
[94] Casey Newton. 2019. The trauma floor: the secret lives of Facebook moderators in America. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona. Accessed: August 20, 2025.
[95] Casey Newton. 2019. The Trauma Floor: The secret lives of Facebook moderators in America. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona. Accessed: August 20, 2025.
[96] DA Norman. 1986. Attention to action: Willed and automatic control of behavior. *Consciousness and self-regulation: Advances in research and theory* 4 (1986).
[97] OpenAI. 2024. Embeddings. https://platform.openai.com/docs/guides/embeddings. Accessed: August 20, 2025.
[98] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: August 20, 2025.
[99] Chaewon Park, Soohwan Kim, Kyubyong Park, and Kunwoo Park. 2023. K-HATERS: A Hate Speech Detection Corpus in Korean with Target-Specific Ratings. *arXiv preprint arXiv:2310.15439* (2023).
[100] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1114–1125.
[101] Reddit. 2025. Promoting Hate Based on Identity or Vulnerability. https://support.reddithelp.com/hc/en-us/articles/360045715951-Promoting-Hate-Based-on-Identity-or-Vulnerability Accessed: August 20, 2025.
[102] Reddit. n.d. Reddit Content Policy. https://redditinc.com/policies/content-policy [Accessed: August 20, 2025].
[103] Martin J. Riedl, Gina M. Masullo, and Kelsey N. Whipple. 2020. The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior* 107 (2020), 106262. https://doi.org/10.1016/j.chb.2020.106262
[104] Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. (2016).
[105] Sarah T Roberts. 2019. *Behind the screen*. Yale University Press.
[106] 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* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 255–264. https://doi.org/10.1145/3292522.3326032
[107] Punyajoy Saha, Kiran Garimella, Narla Komal Kalyan, Saurabh Kumar Pandey, Pauras Mangesh Meher, Binny Mathew, and Animesh Mukherjee. 2023. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2212270120. https://doi.org/10.1073/pnas.2212270120
[108] Patrawat Samermit, Anna Turner, Patrick Gage Kelley, Tara Matthews, Vanessia Wu, Sunny Consolvo, and Kurt Thomas. 2023. "Millions of people are watching you": understanding the digital-safety needs and practices of creators. In *Proceedings of the 32nd USENIX Conference on Security Symposium* (Anaheim, CA, USA) *(SEC '23)*. USENIX Association, USA, Article 315, 17 pages.
[109] Edward Samuels. 1988. The idea-expression dichotomy in copyright law. *Tenn. L. Rev.* 56 (1988), 321.
[110] Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th international conference on computational linguistics*. 343–353.
[111] Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proc. ACM Hum.-Comput. Interact.*

6, CSCW2, Article 370 (Nov. 2022), 27 pages. https://doi.org/10.1145/3555095

[112] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Lun-Wei Ku and Cheng-Te Li (Eds.). Association for Computational Linguistics, Valencia, Spain, 1–10. https://doi.org/10.18653/v1/W17-1101

[113] Angela M. Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2024. Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors. *New Media & Society* 26, 10 (2024), 5677–5701. https://doi.org/10.1177/14614448221138529

[114] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 341, 20 pages. https://doi.org/10.1145/3544548.3581512

[115] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443. https://doi.org/10.1177/1461444818821316 arXiv:https://doi.org/10.1177/1461444818821316

[116] Yukari Seko and Stephen P Lewis. 2018. The self-harmed, visualized, and reblogged: Remaking of self-injury narratives on Tumblr. *New Media & Society* 20, 1 (2018), 180–198. https://doi.org/10.1177/1461444816660783 arXiv:https://doi.org/10.1177/1461444816660783

[117] Alexandra A. Siegel. 2020. *Online Hate Speech*. Cambridge University Press, 56–88.

[118] Sightengine. 2019. Text Moderation API: Detect and filter inappropriate content in any text. https://sightengine.com/text-moderation-api Accessed: August 20, 2025.

[119] Jean Y Song, Sangwook Lee, Jisoo Lee, Mina Kim, and Juho Kim. 2023. ModSandbox: Facilitating Online Community Moderation Through Error Prediction and Improvement of Automated Rules. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[120] Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 17, 4 (Sep. 2023), Article 8. https://doi.org/10.5817/CP2023-4-8

[121] Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2024. Content Moderator Mental Health, Secondary Trauma, and Well-being: A Cross-Sectional Study. *Cyberpsychology, Behavior, and Social Networking* 27, 2 (2024), 149–155. https://doi.org/10.1089/cyber.2023.0298 PMID: 38153846.

[122] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 341, 14 pages. https://doi.org/10.1145/3411764.3445092

[123] Kevin D Stein, Paul B Jacobsen, Chris M Blanchard, and Christina Thors. 2004. Further validation of the multidimensional fatigue symptom inventory-short form. *Journal of pain and symptom management* 27, 1 (2004), 14–23.

[124] Dimitra Eleftheria Strongylou, Marlyn Thomas Savio, Miriah Steiger, Timir Bharucha, Wilfredo R Torralba Manuel III, Xieyining Huang, and Rachel Lutz Guevara. 2024. Perceptions and experiences of severe content in content moderation teams: A qualitative study. In *International Congress on Information and Communication Technology*. Springer Nature Singapore Singapore, 1–12.

[125] Mark Sullivan. 2019. Facebook is expanding its tools to make content moderation less toxic. https://www.fastcompany.com/90367858/facebook-is-expanding-its-tools-to-make-content-moderation-less-toxic. Accessed: August 20, 2025.

[126] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. 2022. "It's common and a part of being a content creator": Understanding How Creators Experience and Cope with Hate and Harassment Online. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 121, 15 pages. https://doi.org/10.1145/3491102.3501879

[127] Nafis Irtiza Tripto, Saranya Venkatraman, Dominik Macko, Robert Moro, Ivan Srba, Adaku Uchendu, Thai Le, and Dongwon Lee. 2024. A Ship of Theseus: Curious Cases of Paraphrasing in LLM-Generated Texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 6608–6625. https://doi.org/10.18653/v1/2024.acl-long.357

[128] TrustLab. 2025. How ModreateAI Works. https://www.trustlab.com/moderateai Accessed: August 20, 2025.

[129] Trust & Safety Professional Association (TSPA). 2024. User Appeals. https://www.tspa.org/curriculum/ts-fundamentals/content-moderation-and-operations/user-appeals/. Accessed: August 20, 2025.

[130] Twitch. n.d. How to Use AutoMod. https://help.twitch.tv/s/article/how-to-use-automod?language=en_US [Accessed: August 20, 2025].

[131] Brendesha M. Tynes, Michael T. Giang, David R. Williams, and Geneene N. Thompson. 2008. Online Racial Discrimination and Psychological Adjustment Among Adolescents. *Journal of Adolescent Health* 43, 6 (2008), 565–569. https://doi.org/10.1016/j.jadohealth.2008.08.021

[132] Meike K Uhrig, Nadine Trautmann, Ulf Baumgärtner, Rolf-Detlef Treede, Florian Henrich, Wolfgang Hiller, and Susanne Marschall. 2016. Emotion elicitation: A comparison of pictures and films. *Frontiers in psychology* 7 (2016), 180.

[133] UNESCO. 2023. World insights on public opinion. https://www.unesco.org/sites/default/files/medias/fichiers/2023/11/unesco_ipsos_survey.pdf Accessed: August 20, 2025.

[134] United Nations. 2019. United Nations Strategy and Plan of Action on Hate Speech. https://digitallibrary.un.org/record/3889290?v=pdf Accessed: August 20, 2025.

[135] Sebastian Wachs, Manuel Gámez-Guadix, and Michelle Wright. 2022. Online Hate Speech Victimization and Depressive Symptoms Among Adolescents: The Protective Role of Resilience. *Cyberpsychology, Behavior, and Social Networking* (04 2022). https://doi.org/10.1089/cyber.2022.0009

[136] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review* 55, 7 (2022), 5731–5780.

[137] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*. 19–26.

[138] Tim Winchcomb. 2019. Use of AI in Online Content Moderation. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/other/cambridge-consultants-ai-content-moderation.pdf?v=324081. Accessed: August 20, 2025.

[139] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300390

[140] X (formerly Twitter). 2024. Transparency Report. https://transparency.x.com/en [Accessed: August 20, 2025].

[141] X Help Center. n.d. The X Rules. https://help.x.com/en/rules-and-policies/x-rules [Accessed: August 20, 2025].

[142] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105

[143] Yiming Zhang, Sravani Nanduri, Liwei Jiang, Tongshuang Wu, and Maarten Sap. 2023. Biasx:" thinking slow" in toxic content moderation with explanations of implied social biases. *arXiv preprint arXiv:2305.13589* (2023).

[144] ZipRecruiter. 2025. Content Review – Independetn Safety Contractor. https://www.ziprecruiter.com/Jobs/Content-Reviewer?lvk=miUutPyuUZSi_hHYzLLmIw.--Nk8mkyBIg

A Full Prompt for Paraphrasing Comments

A.1 Paraphrasing Hate Comments

You are a text content moderator for Korean news comments. Your task is to paraphrase the original comment into a euphemistic comment while maintaining the original meaning of the comment. For your information, Euphemism is an innocuous word or expression used in place of one that is deemed offensive or suggests something unpleasant. From a given Korean comment on the News, output 10 distinct Euphemistic paraphrased comments by paraphrasing only words between '※'. Do not change the original meaning and separate each output with '$'.

A.2 Paraphrasing Normal Comments

You are a text content moderator for Korean news comments. Your task is to paraphrase a comment. From a given Korean comment on the News, output 10 distinct paraphrased comments by paraphrasing only words between '※'. Do not change the original meaning and separate each output with '$'.

B Shapiro-Wilk Test for SPANE_B Scores: Pre- and Post-Experiment

B.1 Pre-experiment

Table 5. Shapiro-Wilk Test Results of Pre-experiment scores for SPANE_B

Group	Mean	Std	Min	Max	Shapiro-Wilk	
					W	p-value
Control	12.00	5.38	3.00	20.00	0.91	.625
Anonymizing	10.65	4.86	1.00	19.00	0.97	.659
Paraphrasing	10.35	5.03	3.00	19.00	0.94	.256
Revealing	10.35	7.84	-4.00	24.00	0.96	.587

B.2 Post-experiment

Table 6. Shapiro-Wilk Test Results of Post-experiment scores for SPANE_B

Group	Mean	Std	Min	Max	Shapiro-Wilk	
					W	p-value
Control	1.95	7.17	-11.00	13.00	0.96	.478
Anonymizing	2.70	7.19	-13.00	14.00	0.95	.420
Paraphrasing	0.20	7.12	-17.00	14.00	0.92	.096
Revealing	-0.30	6.20	-10.00	13.00	0.92	.109

C Shapiro-Wilk Test for MFSI Scores: Pre- and Post-Experiment

C.1 Pre-experiment

Table 7. Shapiro-Wilk Test Results of pre-experiment scores for MFSI

Group	Mean	Std	Min	Max	Shapiro-Wilk	
					W	p-value
Control	2.60	8.12	17.00	-11.00	0.96	.478
Anonymizing	2.25	6.67	13.00	-7.00	0.91	.072
Paraphrasing	2.80	8.32	23.00	-9.00	0.96	.482
Revealing	3.75	7.91	18.00	-12.00	0.97	.840

C.2 Post-experiment

Table 8. Shapiro-Wilk Test Results of post-experiment scores for MFSI

Group	Mean	Std	Min	Max	Shapiro-Wilk	
					W	p-value
Control	9.05	9.02	31.00	-6.00	0.96	.540
Anonymizing	6.20	7.18	23.00	-7.00	0.97	.777
Paraphrasing	8.35	9.32	28.00	-10.00	0.99	.988
Revealing	10.50	8.13	23.00	-8.00	0.96	.607

D Wilcoxon Signed-Rank Test Results comparing Pre- and Post-experiment for SPANE and MFSI

D.1 Pre- and Post-experiment for SPANE

Table 9. Wilcoxon Signed-Rank Test Results comparing pre- and post-experiment scores for SPANE questionnaire items.

Comparison	Statistics	p-value
Control	0.00	<.001***
Anonymizing	8.50	<.001***
Paraphrasing	0.00	<.001***
Revealing	3.00	<.001***

D.2 Pre- and Post-experiment for MFSI

Table 10. Wilcoxon Signed-Rank Test Results comparing pre- and post-experiment scores for MFSI questionnaire items.

Comparison	Statistics	p-value
Control	22.50	.002**
Anonymizing	28.00	.012*
Paraphrasing	17.50	.002**
Revealing	21.50	.002**

E Survey Questionnaire

This survey contains the following two sections.

(1) Korean Scale of Positive and Negative Experience (pre- and post-survey)
(2) Multidimensional Fatigue Symptom Inventory (pre- and post-survey)
(3) System Evaluation Questionnaire (post-survey)

Please answer all questions sincerely.

Korean Scale of Positive and Negative Experience

This scale consists of a number of words that describe different feelings and emotions. Read each item and indicate to what extent you feel at this moment before you have started the experiment (at this moment after you have finished the experiment). [1: Not at All, 2: A Little, 3: Moderately, 4: Quite a Bit, 5: Extremely]

(1) Positive
(2) Negative
(3) Good
(4) Bad
(5) Pleasant
(6) Unpleasant
(7) Happy
(8) Sad
(9) Afraid
(10) Joyful
(11) Angry
(12) Contented

Multidimensional Fatigue Symptom Inventory

Below is a list of statements that describe how people sometimes feel. Please read each item carefully, then mark the one number which best describes how true each statement is for you at this moment before you have started the experiment (at this moment after you have finished the experiment). [1: Not at All, 2: A Little, 3: Moderately, 4: Quite a Bit, 5: Extremely]

(1) I have trouble remembering things
(2) I feel upset
(3) I feel cheerful
(4) I feel lively
(5) I feel nervous
(6) I feel relaxed
(7) I am confused
(8) I feel sad
(9) I have trouble paying attention
(10) I am unable to concentrate
(11) I feel depressed
(12) I feel refreshed
(13) I feel tense
(14) I feel energetic
(15) I make more mistakes than usual
(16) I am forgetful
(17) I feel calm
(18) I am distressed

System Evaluation Questionnaire

Based on your experience using the system during the experiment, please indicate how much you agree or disagree with the following statements, and provide answers to the open-ended questions about your overall comment moderation experience. [1: Not at All, 2: A Little, 3: Moderately, 4: Quite a Bit, 5: Extremely]

- *Control group*

(1) I had no trouble understanding the meaning of the comments.
(2) Please describe the overall process you went through when reading the comments and making deletion decisions. How did you feel when reading the comments? What factors did you consider when deciding whether to delete them? [*open-ended*]

- *Anonymizing group*

(1) The fact that the targets were hidden was helpful in performing the comment moderation task.
(2) The fact that the targets were hidden was helpful in maintaining my mental well-being. [*open-ended*]
(3) Please describe the overall process you went through when reading the comments with hidden targets and making deletion decisions. How did you feel when reading these comments? What factors did you consider when deciding whether to delete them? [*open-ended*]

- *Paraphrasing group*

(1) The following questions pertain to the mitigated offensive expressions in the comments.
  (a) The fact that the expressions were mitigated was helpful in performing the comment moderation task.

(b) The fact that the expressions were mitigated was helpful in maintaining my mental well-being.

(2) Please describe the overall process you went through when reading the mitigated comments and making deletion decisions. How did you feel when reading these comments? What factors did you consider when deciding whether to delete them? [*open-ended*]

- *Revealing group*

(1) The following questions pertain to the target hiding feature in comments.
   (a) The "view original target" feature was helpful in performing the comment moderation task.
   (b) The "view original target" feature was helpful in maintaining my mental well-being.
(2) The following questions pertain to the mitigated offensive expressions in comments.
   (a) The "view original expression" feature was helpful in performing the comment moderation task.
   (b) The "view original expression" feature was helpful in maintaining my mental well-being.
(3) Please describe the overall process you went through when reading the modified comments and making deletion decisions. How did you feel when reading these comments? What factors did you consider when deciding whether to delete them? [*open-ended*]

F Interview Protocol

We will now begin the interview about the experiment you participated in today. If there are any questions you don't wish to answer during the interview, feel free to refuse to answer.

Warm-up and Overall Experience

First, I will ask some questions about your overall experience with comment moderation.

(1) What was the general intensity of hate speech you felt in the comments you reviewed during the experiment?
   (a) (If negative) You mentioned that the atmosphere of the comments was what the participant described. How did moderating such comments affect your emotions or mental well-being?
   (b) If you had to moderate these comments daily as a comment moderator, how would it impact your emotions or mental well-being?
(2) (Questions asked again about any parts the participant did not elaborate on in the survey)
   (a) How did you feel when reading the comments?
   (b) What are your criteria for determining hate speech?
   (c) Were there any difficult instances in deciding whether to delete a comment? Please explain the situation and the reason.

About the System

I will now ask some questions regarding the system you used today and comment moderation.

- *Control group*

  The system you used today annotated words that could be considered targets of hate speech and offensive expressions in the comments.

(1) Did the fact that targets were annotated help you determine hate speech? Why or why not?
(2) Did the fact that offensive expressions were annotated help you moderate hate speech? Why or why not?
(3) Did the fact that targets were annotated affect your mental well-being? If it did, was the effect positive or negative? Please explain the reason.

(4) Did the fact that offensive expressions were annotated affect your mental well-being? If it did, was the effect positive or negative? Please explain the reason.

- *Anonymizing group*
  The system you used today concealed words that could be considered targets of hate speech and underlined offensive expressions in the comments.

(1) Would you make the same moderation decisions even if the targets were not concealed? Why or why not?
(2) Do you think reviewing and moderating comments with concealed targets affects your mental well-being differently compared to seeing the original comments? Please explain the reason.
(3) If there is a feature to reveal the concealed targets, how would you use it in the decision-making process for comment moderation? If you think you wouldn't use it, please explain why.
(4) Did the fact that offensive expressions were annotated help you moderate hate speech? Why or why not?
(5) Did the fact that offensive expressions were annotated affect your mental well-being? If it did, was the effect positive or negative? Please explain the reason.

- *Paraphrasing group*
  The system you used today paraphrased offensive expressions in less offensive and annotated words that could be considered targets of hate speech.

(1) Would you make the same moderation decisions if you saw the original, unmodified comments? Why or why not?
(2) Have you changed your moderation decision after reading another version of the expression using the refresh feature? Please describe the situation and explain why you changed your decision.
(3) Would reviewing and moderating comments with the original offensive expressions affect your mental well-being differently than their paraphrased versions? Please explain the reason as well.
(4) Suppose there is a feature that allows you to see the original versions of the paraphrased offensive expressions. How would you use it in the decision-making process for comment moderation? If you feel you wouldn't use it, please explain why.
(5) Did the fact that targets were annotated help you determine hate speech? Why or why not?
(6) Did the fact that targets were annotated affect your mental well-being? If it did, was the effect positive or negative? Please explain the reason.

- *Revealing group*
  The system you used today concealed words that could be considered targets of hate speech and paraphrased offensive expressions into less offensive ones. It also allowed you to view the original comments or different versions of paraphrased expressions when needed.

(1) First, I will ask questions related to your moderation decision.
  (a) Was there a situation where you changed your moderation decision after checking the concealed target? Please explain the situation and why you changed your decision.
  (b) Was there a situation where you changed your decision to delete a comment after checking the original version of the mitigated offensive expression? Please explain the situation and why you changed your decision.

(c) Was there a situation where you used the refresh feature to switch to a different paraphrased version of a comment and subsequently changed your decision to moderate it? Please describe the situation and explain why you changed your decision.
(d) Among the features—target hiding, offensive expression paraphrasing, refresh, and view original—which was the most helpful in comment moderation, and why?

(2) Next, I will ask questions related to your mental well-being.
(a) Did checking the concealed targets affect your mental well-being? If so, how did it change, and why?
(b) Did checking the original version of the mitigated offensive expressions affect your mental well-being? If so, how did it change, and why?
(c) Did using the refresh feature to change the paraphrased versions of the comments affect your mental well-being? If so, how did it change, and why?
(d) Among the features—target hiding, offensive expression paraphrasing, refresh, and view original—which was the most helpful in protecting your mental well-being, and why?

Wrap-up

We've covered almost all the questions about today's experiment, but I'd like to ask you a simple question before we finish.

- (*Control group*) It is known that moderators are often exposed to content that negatively impacts mental health, potentially leading to vicarious trauma or PTSD. Considering the negative effects of moderating comments, are there any features you think should be added to a moderation system?
- (*Target, Offensive, Revealing group*) It is known that moderators are often exposed to content that negatively impacts mental health, potentially leading to vicarious trauma or PTSD. The system you used today was designed to help protect moderators. Are there any features you would like to see improved? Or are there any new features you would propose to better preserve the moderator's mental well-being?

Thank you for answering all of our questions thoroughly. We've asked everything we intended to cover. Before we conclude, is there anything you'd like to share with us or any responses you couldn't fully express during the interview?

Thank you. This concludes today's experiment. We will follow up with additional documents via email for your compensation.

Thank you again.

Received October 2024; revised April 2025; accepted August 2025