Developing a User-Centric Stepwise Comment Moderation AI Service: Balancing Free Expression and Respectful Online Interaction

Jonghan Kim[1], Suhwan Jo[2], Woosung Jung[2], Manseo Kim[1], and Sung Park[3(✉)]

[1] Department of Human-Centered Artificial Intelligence, Sangmyung University, Seoul 03016, Republic of Korea

[2] Department of Electronic Engineering, Sangmyung University, Seoul 03016, Republic of Korea

[3] Department of Emotion Engineering, Sangmyung University, Seoul 03016, Republic of Korea

sjpark@smu.ac.kr

Abstract. The rise in online trolling, marked by intentionally posting offensive messages to provoke or disrupt, poses significant challenges to digital interactions. Trolling negatively impacts psychological well-being and disrupts online communities. Our research focuses on AI-based moderation, moving beyond traditional word censorship to accommodate user diversity and mental states in content moderation. We incorporated the discover, define, develop, and deliver stages to create a user-centered comments moderation service. This service was informed and validated through user feedback. Our research included conducting user surveys (n = 106) and in-depth interviews (IDI, n = 9), revealing a diversity in preferences for comment moderation in existing services. Notably, many users expressed a preference for replacing offensive words with alternatives rather than entirely blocking comments. In response to these findings, we designed and developed a stepwise moderation service using current large language models, such as ChatGPT. This service empowers users to choose their preferred level of comment moderation, striking a balance between free expression and a respectful online environment. We tested a working prototype with potential users (n = 9) to evaluate its effectiveness. The results highlight the wide range of user needs and preferences in comment moderation. While there was positive feedback towards progressive comment moderation, it was accompanied by concerns about potential over-moderation and the need to preserve genuine communication. These findings underscore the critical need for nuanced moderation approaches that respect user diversity.

Keywords: sentence purification · comments moderation · large language model · generative AI · AI design · user-centered design

S. Jin et al. (Eds.): IEA 2024, SSDI 56, pp. 299–305, 2025.
https://doi.org/10.1007/978-981-95-0289-9_44

1 Introduction

Trolling is when someone posts messages to upset people, start arguments, or draw others into pointless debates, either subtly or blatantly. This behavior includes sharing unconstructive content designed to elicit reactions, trap targets in pointless arguments, and disturb the normal functioning of online communities [1].

Recently, platforms like YouTube adopted automatic detection and removal of malicious comments, aiming to create a safer online environment. This method helps filter harmful content in public forums or news articles. However, this automatic censorship has drawn criticism from users who feel their freedom of expression is being stifled, leading some to switch platforms or find ways to bypass the censorship algorithms for continued trolling. The challenge lies in balancing the minimization of trolling with preserving user freedom and addressing the limitations of current systems. In light of these findings, we hypothesized that reactions to moderated comments are likely to be diverse and multifaceted. Understanding and addressing this diversity is crucial to ensure that AI moderation systems are inclusive, effective, and sensitive to the needs of all users. To thoroughly investigate this hypothesis and to gain a deeper understanding of the varying user needs and responses, we embarked on a service design and development process.

2 Related Works

Studies on handling malicious comments have explored a range of methods. Such as, censoring or filtering techniques aim to prevent user exposure to such comments by not displaying inappropriate content. Another strategy is the outright disabling of the comment feature, which blocks users from posting comments altogether [2]. More recent research has focused on AI- based moderation, which advances beyond mere word or sentence censorship. Notably, the use of generative AI, particularly large language models (LLMs), has opened new possibilities for analyzing and transforming the nature of malicious comments [3].

Several studies [4–6] have introduced innovative methods for modifying text styles through generative AI. While these techniques hold significant potential, they may not universally meet the diverse needs and preferences of all users. For instance, Park's research [7] demonstrates that users' perceptions and reactions to content can vary greatly based on their mental state and personal experiences. These variations in user experiences underscore the complexity of effectively moderating online content.

Considering these findings, we hypothesized that reactions to moderated comments are likely to be diverse and multifaceted. Understanding and addressing this diversity is crucial to ensure that AI moderation systems are inclusive, effective, and sensitive to the needs of all users. To thoroughly investigate this hypothesis and to gain a deeper understanding of the varying user needs and responses, we embarked on a service design and development process.

3 Methods

3.1 Survey and Interview

We collected subjective responses from participants, aged between 20 to 50 years old, with 89% in their 20 s and the remainder in their 30 s to 50 s. Among them, 65 (61%) were men and 41 (39%) were women, totaling 106 participants.

Questionnaire items were diverse, aimed at gathering user evaluations on existing malicious comments detection methods and comment moderation assessments. Additionally, we conducted in-depth interviews with 11 individuals selected based on their susceptibility to comment influence compared to other survey participants. This diverse group included individuals like a composer who regularly shares their creations online.

3.2 Understanding User Needs

Through surveys and interviews, we have identified user requirements and perceptions in three key areas. Firstly, users express dissatisfaction with the current detection methods, highlighting their negative impact and the need for improvement. Many participants reported feeling discomfort and hostility when encountering online comments, raising concerns about the inadequacy of existing strategies.

Secondly, diverse preferences exist regarding comment moderation. While some users preferred filtering out only profanity, others favored replacing offensive words with alternatives. There was also a preference for modifying the overall tone of comments to make them less aggressive.

Lastly, some users express interest in reading the original malicious comments. For instance, one participant, a professional composer, shared a nuanced perspective on comment moderation, stating, *"I am in favor of moderating malicious comments. However, I also value seeing unfiltered feedback on my work, as such comments, even if critical, often motivate me to strive harder in my creative endeavors."*

3.3 Designing The Service Concept

We designed and developed a stepwise moderation service using current large language models, such as ChatGPT to address the user needs defined earlier (see Fig. 1).

Each moderation level is characterized by its unique prompt instructions. When a sentence is input, it undergoes pre-processing, which includes removing special characters and normalizing words. After pre-processing, the sentence is sent to the corresponding model section based on the assigned moderation level. In this stage, the model moderates the sentence and generates a revised version. The moderated sentence then undergoes error detection in the post-processing section before being stored in the database. Simultaneously, it is pre-processed for the next moderation level. This cycle continues, with moderated sentences from levels 1 to 3 being sequentially processed and stored in the database. The moderation levels were defined based on user feedback gathered from surveys and interviews, which provided insights into effective moderation of malicious comments (refer to Table 1).

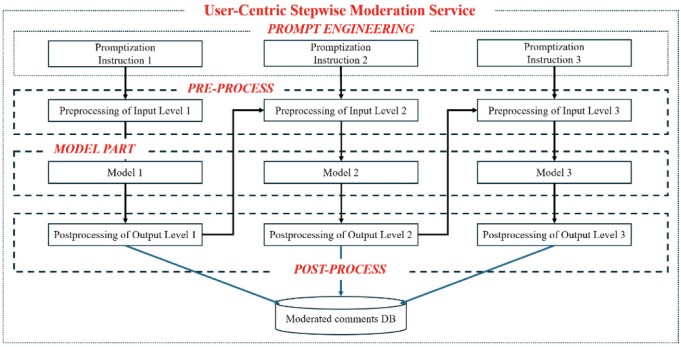

Fig. 1. The streamlined comment generation process of a stepwise moderation service

Table 1 Examples of sentence based on a stepwise moderation service

Moderation levels	Example	
Original comment	You're just a good for nothing jerk!	
Level 1	You're just a good for nothing @@@!	
Level 2	You're just a good for nothing bro!	
Level 3	You're like a sloth. Put in more effort, please.	

At the "Original Comment" stage, the comment is displayed in its unaltered form. At "Level 1," slang and swear words are filtered out and replaced with "@@" symbols. Moving to "Level 2," the sentences containing "@@" symbols from the previous level are supplemented with appropriate words. Finally, at "Level 3," an AI rephrases the sentences from Level 2, aiming to soften any aggressive or provocative expressions while also adding positive elements to enhance the overall tone. Therefore, users can adjust the moderation level from Level 1 to Level 3 according to their preference. The primary aim of our service is to mitigate psychological harm by moderating harmful content such as criticism, profanity, and offensive language, without resorting to complete removal. This service empowers users to choose their preferred level of comment moderation, striking a balance between free expression and a respectful online environment.

4 User Evaluations

To gather user feedback on our service, we developed a community website that serves as a representative platform. On this site, users went through several steps as part of their interaction with our service. Initially, they read the main articles along with the existing comments. Subsequently, they had the option to adjust the moderation settings from level 1 to level 3, thereby observing the changes in the comments at each level. After interacting with the service, we presented users with a set of questions designed to gather feedback on their experience.

To enhance user engagement and provide a realistic experience, we populated the website with actual articles and comments sourced from the internet. This approach encouraged users to naturally engage with the content, immersing themselves in the experience. Importantly, it allowed users to experiment with different moderation levels, observe how comments were altered at each stage, and determine which level best suited their preferences.

We curated a set of questions aimed at capturing comprehensive user experiences and opinions about our service. We received insightful responses from nine individuals who interacted with the service on our community website.

Following their experience with the service, two distinct response categories emerged. While the majority had a positive reaction to the concept, some expressed negative views, particularly regarding the third level of moderation, which involved softening any aggressive or provocative expressions and adding positive elements to enhance the overall tone. A few examples of participants' responses are as follows:

> "Functionally, using it was smooth. I was quite amazed by the step-by-step moderation feature and how well it worked in practice." (Participant #6).

> "I like that I can choose the level of moderation for malicious comments, allowing me to select different levels based on my purpose." (Participant #1).

> "I think moderation was done well, but there were instances in Level 3 where it changed to praise, altering the meaning significantly." (Participant #4).

When considering potential utilization in daily life, participants primarily focused on the service's suitability for children and creators. A notable number of respondents (n = 5) believed that the service could be beneficial in contexts involving children, helping to shield them from inappropriate language in the content they view.

> "It would be great if this could be used for YouTube comments that are suitable for children to watch." (Participant #8).

> "I think that parents with children are often concerned that their kids might learn inappropriate language when using the internet. Having a service like this could be helpful in addressing such issues." (Participant #9).

Overall, while there was enthusiasm for the innovative approach of our moderation service, there were concerns about over-moderation and the preservation of authentic communication. These insights are invaluable for refining the service, ensuring it strikes the right balance between protecting users and maintaining the integrity of online discourse.

5 Discussion

Most feedback indicated a positive reception towards the concept of progressive comment moderation, suggesting our service could play a role in mitigating the prevalent issue of aggressive online culture. Our approach, which transcends simple censorship, considers both the author's right to expression and the reader's desire for a positive online experience. This method allows users to adjust the moderation level to their preference, thereby preserving freedom of expression while minimizing the harm caused by malicious comments.

Nonetheless, our study faced two primary limitations, leading to some negative feedback. The first was related to inaccuracies in translations and interpretations, particularly at Level 3 moderation, where AI sometimes produced awkward or misunderstood sentences. This issue was compounded by the lack of context, as isolated comments could be misinterpreted without the accompanying main article. For instance, sarcastic comments could be misconstrued as genuine praise. The second limitation concerned the inability to stay current with the latest internet slang and terminology. The AI model, with a cutoff point for its data, failed to recognize newly coined words and phrases, resulting in awkward or inaccurate moderations. This was particularly evident with trending yet mildly toxic words or memes, which were sometimes incorrectly classified and overly moderated, stripping away the original humor or creativity.

Future research could explore measuring the similarity between original and moderated comments to assess effectiveness more accurately. Additionally, further refining the moderation levels could offer more personalized experiences. Future studies might also test the effectiveness of this service in preventing the amplification of malicious comments, known as the echo chamber effect [8]. Despite the challenges, our study provides a foundation for developing more nuanced and effective online comment moderation solutions.

Acknowledgement. This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF2022R1I1A1A01066657).

References

1. Bishop, J.: The psychology of trolling and lurking: The role of defriending and gamification for increasing participation in online communities using seductive narratives. In: Li, H. (ed.) Virtual Community Participation and Motivation: Cross-Disciplinary Theories, pp. 160–176. IGI Global, Hershey, USA (2012). https://doi.org/10.4018/978-1-4666-0312-7.ch010
2. Lee, S.H.: Biased artificial intelligence: Analyzing the types of hate speech classified by 'cleanbot', NAVER AI for detecting malicious comments. J. Cybercom. Acad. Soc. **38**, 33–75 (2021)
3. Axelsen, H., Jensen, J. R., Axelsen, S., Licht, V., Ross, O.: Can AI Moder-ate Online Communities? Work in Progress
4. Atwell, K., Hassan, S., Alikhani, M.: APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations (2022)

5. Nogueira dos Santos, C., Melnyk, I., Padhi, I.: Fighting offensive language on social media with unsupervised text style transfer. (2018)
6. Wang, K., Hua, H., Wan, X.: Controllable unsupervised text attribute transfer via editing entangled latent representation. (2019)
7. Park, M., Mcdonald, D. W., Cha, M.: Perception differences between the depressed and non-depressed users in Twitter (2013). www.aaai.org
8. Cinelli, M., et al.: The echo chamber effect on social media (2021). https://doi.org/10.1073/pnas.2023301118/-/DCSupplemental.y