Encouraging Self-Reflection in Online Conversations: A Comment Queuing Approach to Mitigating Toxicity and Enhancing Emotional Regulation

AKRITI VERMA*, Deakin University, Australia
SHAMA ISLAM, Deakin University, Australia
VALEH MOGHADDAM, Deakin University, Australia
ADNAN ANWAR, Deakin University, Australia

Online conversations are often disrupted by deliberate trolling, leading to emotional distress and conflict among users. Prior research has focused on identifying and moderating harmful content after it has been posted, but there is a lack of exploration of real-time strategies for managing emotions in digital interactions. Self-reflection has been shown to help attenuate emotion arousal for everyday emotion regulation. In this work, we propose a method to reduce the impact of trolls by implementing a delay in responses using a comment queuing approach. This delay is intended to encourage self-reflection among users, providing them with an opportunity to regulate their emotions before continuing to engage in the conversation. We evaluated the effectiveness of this approach by analysing 15K instances of user posts and interactions on Reddit, examining the effect of queuing on reducing negative emotional propagation in conversations. Preliminary analysis indicates that this framework can decrease the propagation of hate speech and anger by up to 15% during an active conversation, with only 4% of comments being temporarily withheld for up to 47 seconds on average. As the next phase of the study, we plan to assess user perceptions of the queuing mechanism's effectiveness through a targeted user survey.

Additional Key Words and Phrases: Digital Emotion Regulation (DER), Interpersonal Emotion Regulation (IER), Emotions in Social Media, Emotions Online, Human Computer Interaction (HCI), Affective Computing

ACM Reference Format:
Akriti Verma, Shama Islam, Valeh Moghaddam, and Adnan Anwar. 2024. Encouraging Self-Reflection in Online Conversations: A Comment Queuing Approach to Mitigating Toxicity and Enhancing Emotional Regulation. 1, 1 (November 2024), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The emergence of online platforms has transformed how people interact and communicate, creating new opportunities for social connection and community development [Parameswaran and Whinston 2007], [Wadley et al. 2020]. However, with these advancements come the difficulties of handling negative behaviours, particularly trolling, which can disrupt online discussions [Smith et al. 2022]. Trolling, involving deliberately provoking emotional responses from others, often results in conflict, emotional distress, and a decline in the quality of online conversations. As digital communication

Authors' Contact Information: Akriti Verma, vermaakr@deakin.edu.au, Deakin University, Australia; Shama Islam, Deakin University, Australia, shama.i@deakin.edu.au; Valeh Moghaddam, Deakin University, Australia, valeh.moghaddam@deakin.edu.au; Adnan Anwar, Deakin University, Australia, adnan.anwar@deakin.edu.au.

ACM XXXX-XXXX/2024/11-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

continues to grow, there is an urgent need for effective strategies to tackle and control these harmful interactions in real-time [Maarouf et al. 2022], [Goel et al. 2016].

Current approaches to managing online toxicity mainly focus on post-hoc moderation, where harmful content is identified and removed after it's been posted [Chandrasekharan et al. 2022]. While these methods are crucial for maintaining the integrity of online spaces, they often fail to prevent the initial emotional harm caused by trolling [Trujillo and Cresci 2022]. Moreover, the reactive nature of these approaches does little to promote emotional regulation among users, potentially allowing negative behaviours to persist or even escalate over time [Gongane et al. 2022].

In contrast, self-reflection has been acknowledged as a powerful tool for emotion regulation in daily life [Herwig et al. 2010]. Encouraging individuals to pause and reflect on their emotions can assist them in managing their responses effectively, reducing the likelihood of impulsive behaviour [Kiskola et al. 2021]. However, the application of self-reflection as a real-time strategy for managing emotions in digital interactions remains unexplored.

This research seeks to address this gap by proposing a novel approach to mitigating the impact of trolls through a comment queuing mechanism. By introducing a delay in the posting of potentially toxic comments, this method encourages users to engage in self-reflection, offering them the opportunity to regulate their emotions before their comment is made public. This approach shifts the focus from merely moderating content to fostering healthier emotional interactions online. This research aims to enhance digital emotion regulation by incorporating self-reflection, contributing to the development of real-time, user-centred strategies for managing online interactions.

To evaluate the effectiveness of this self-reflection-based emotion regulation strategy, we employed a mixed-method research design. This began with an analysis of text-based social media data, focusing specifically on user interactions on Reddit, and will be followed by a user survey to gather feedback on the proposed framework. Preliminary results from the initial data analysis suggest that this approach can significantly reduce the propagation of negative emotions, such as hate speech and anger, during active conversations. Furthermore, implementing the queuing mechanism appears to minimally impact the overall flow of communication, with only a small percentage of comments being temporarily withheld for brief durations.

Therefore, this work makes the following research contributions:

- Emotion Regulation Approach: The research introduces a real-time emotion regulation method for online conversations by implementing a comment queuing system. Unlike traditional moderation techniques focusing on post-hoc content analysis, this approach proactively encourages users to self-reflect, reducing the likelihood of emotional escalation during active conversations.
- Integration of Self-Reflection in Digital Platforms: The study extends the use of self-reflection beyond personal emotional regulation to digital interactions, offering an approach to mitigate the spread of adverse emotions and online toxicity.
- Empirical Evidence on the Effectiveness of Comment Queuing: The study provides empirical evidence that implementing a comment queuing mechanism can reduce the spread of negative emotions and toxic comments in online conversations. Initial findings show a notable decrease in the spread of anger and hate speech. This will be further evaluated through user surveys.

The rest of this paper is structured as follows: It begins with a review of relevant literature, followed by a description of the proposed framework. Then, the results of the preliminary analysis and the evaluation plan are outlined.

2 Literature Review

The increasing interest in managing online toxicity and promoting positive digital interactions has attracted extensive research in recent years. Here, we focus on three main areas: the characteristics and effects of trolling in online discussions, current methods for regulating emotions in digital communication, and the potential of self-reflection as a tool for regulating emotions in online environments.

2.1 Trolling in Online Conversations

Trolling, which involves intentionally disrupting online conversations with provocative or offensive messages, has been extensively researched in the field of Human-Computer Interaction (HCI). Studies have revealed a range of motivations behind trolling, from seeking amusement and attention to more malicious intentions such as causing harm or manipulating discussions [Buckels et al. 2014]. The impact of trolling varies, often leading to heightened emotional arousal, conflict, and the breakdown of productive discourse in online communities [Kumar et al. 2017], [Cheng et al. 2017]. Trolling can also contribute to broader issues of online harassment and cyberbullying, with significant emotional and psychological consequences for victims [Jane 2020].

Despite the considerable body of work on the nature and impact of trolling, much of the existing research has focused on identifying and removing harmful content post-facto. Tools like automated moderation systems and community guidelines are commonly used to detect and mitigate trolling behaviours after they occur [Yin et al. 2009], [Chandrasekharan et al. 2017], [Chandrasekharan et al. 2022]. While these reactive measures are crucial for maintaining safe online environments, they do not address the root causes of trolling nor do they prevent the initial emotional damage that such behaviours can cause.

2.2 Emotion Regulation in Digital Communication

Emotion regulation, the process by which individuals influence their emotional states and expressions, is essential for maintaining positive interactions both offline and online. In digital communication, where the absence of non-verbal cues can make emotional interpretation challenging, effective emotion regulation is particularly important [Derks et al. 2008]. Research in this area has explored various strategies for regulating emotions in digital contexts, including cognitive reappraisal, expressive suppression, and the use of emoticons or emojis to clarify emotional intent [Holtzman et al. 2017].

Several studies have examined the role of digital platforms in supporting or hindering emotion regulation. For instance, social media platforms often amplify emotional content through algorithmic promotion of highly engaging (and often emotionally charged) posts, which can exacerbate emotional responses and contribute to online toxicity [Roberts 2016], [Kramer et al. 2014]. Conversely, some platforms have implemented features designed to support emotion regulation, such as content warnings, options to mute or block other users, and tools for taking breaks from online activity [Schoenebeck 2014].

However, there is a gap in the literature regarding real-time strategies for emotion regulation in digital interactions. Most existing approaches are either post-hoc or rely on user-initiated actions, such as choosing to mute a conversation. There is a limited exploration of how digital platforms themselves can be designed to promote emotion regulation at the moment, particularly through mechanisms that encourage users to pause and reflect before reacting.

2.3 Self-Reflection as a Tool for Emotion Regulation

The process of self-reflection, which involves examining and understanding one's thoughts, feelings, and behaviours, is widely recognised as an effective strategy for regulating emotions in psychological research [Herwig et al. 2010]. Research studies have demonstrated that self-reflection can provide individuals with a better understanding of their emotions, reduce impulsive reactions, and lead to more intentional decision-making [Gross 2002], [Upadhyaya 2020]. In the context of conflict resolution, self-reflection has been associated with decreased aggression and improved interpersonal outcomes [Kross and Ayduk 2008], [Ayduk and Kross 2010].

Despite its potential, the integration of self-reflection into digital platforms has been relatively limited, especially as a real-time tool for managing emotional responses [Kiskola et al. 2021], [Kiskola et al. 2022]. Some studies have explored the incorporation of reflective practices into digital experiences, such as mindfulness apps that encourage users to pause and consider their emotional state [Howells et al. 2016], [Ruckenstein and Turunen 2020], [Torre and Lieberman 2018]. However, the application of self-reflection to the particular challenge of managing online trolling has received little attention.

We aim to address this gap by proposing a new approach that utilises self-reflection to regulate emotions in online conversations. Through the implementation of a comment queuing mechanism that introduces a delay before potentially toxic content is posted, we intend to prompt users to reflect on their emotional state and the potential impact of their comments. This approach not only tackles the immediate issue of trolling but also fosters a more deliberate form of digital communication.

2.4 Gaps in the literature

The key gaps that exist in the literature on Self-Reflection as a Tool for Emotion Regulation:

- Lack of Real-Time Emotion Regulation Strategies: Current approaches to addressing negative online behaviour, such as trolling, focus on post-facto moderation, identifying and responding to trolling after it occurs rather than managing it in real-time. There's a lack of research on proactive approaches that can help control emotional reactions as they happen in online conversations.
- Under-utilisation of Self-Reflection in Digital Platforms: Self-reflection is a well-established tool in psychological research for emotion regulation. However, its application in digital environments, especially in real-time, is limited. There is a need for exploring how self-reflection can be integrated into digital communication platforms to help users regulate their emotions before reacting impulsively.
- Limited Focus on the Root Causes of Trolling: Current research on trolling focuses on removing or moderating toxic (or potentially toxic) content, rather than addressing the underlying causes of trolling behaviour, such as emotional arousal and impulsivity. There is a need for strategies that directly target these factors by encouraging users to pause and reflect before posting.

3 Methodology

In this section, we present a strategy for implementing a comment queuing system aimed at promoting self-reflection and moderating emotional reactions in online discussions.

3.1 Data Collection

The process of collecting data begins with gathering text-based social media data. Reddit is a highly popular platform where users engage in in-depth discussions, offer support, and share

their views on current events [Manikonda et al. 2018]. Our approach involves obtaining a diverse dataset that includes text-based interactions from Reddit, which we acquired using the Reddit API (PRAW). Specifically, we targeted data from r/politics, r/worldpolitics, r/politicaldiscussions, and r/politicaldebates, as these subreddits are recognised for facilitating extensive discussions and often elicit strong emotional responses to both national and international political issues.

To ensure a diverse range of conversation samples, we categorized posts based on comment volume, including those with at least 300, 500, and 1000 comments. This stratification allowed for an analysis of emotional dynamics across conversations of varying scales, from moderate to highly active threads, thereby enhancing the generalizability of our findings. By filtering posts, we aimed to capture high-impact conversations where emotional exchanges tend to be more intense and varied. We gathered a total of 65 conversations for this analysis, covering the period from August 2023 to August 2024.

3.2 Data Preprocessing

Our dataset consisted solely of posts containing text and emoticons, facilitating the capture of nuanced emotional expressions. We limited our analysis to English-language content and excluded all other languages. We only consider posts containing text and emoticons, and the analysis focuses solely on English-language content while excluding all other languages. CSV files were used to save the data, with one file per user. Afterwards, the data was cleaned to remove extraneous characters, links, and special symbols. The emojis in the tweets were replaced with vector representations generated by Gensim using the Emojinal library [Barry et al. 2021] after which the text was broken down into individual words or tokens for easier analysis [Bird et al. 2009], [Verma et al. 2023].

For this study, we will utilise a dataset containing 15,000 instances of user posts and interactions from Reddit. The dataset will encompass the following main components:

- Original Posts: These will act as the foundational nodes in conversation graphs.
- Replies and Comments: These will serve as nodes that will be examined for their emotional content.
- Timestamps: We will leverage timestamps to monitor the timing of each comment's submission.

In the dataset, each comment is categorised with an emotion based on its content. We achieve this by matching the emotion and sentiment of specific words in the user's comments with the dictionary from the NRC Word-Emotion Association Lexicon [Mohammad and Turney 2013]. The NRC Emotion Lexicon comprises words associated with 8 basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) as well as 2 sentiments (negative and positive). Each comment is assigned an emotion intensity score between 0.1 and 1.0, which reflects the strength of the conveyed emotion.

3.3 Graph-based Conversation Analysis

We use directed acyclic graphs (DAGs) to represent conversations, with nodes being individual comments or posts, and edges denoting the reply relationships between comments. In this structure, an edge is directed from the replying comment to the comment it is replying to, resulting in a hierarchical framework where the original post acts as the root node. This graph-based structure enables us to compute the impact of each comment on the overall emotional tone of the conversation. The influence of each comment is determined using a combination of metrics:

- Number of Replies: Comments that receive more replies are considered more influential.
- Distance from Root Node: Comments closer to the original post are deemed more influential due to their proximity to the root.

- PageRank: A measure of a comment's importance based on its position within the graph.
- Emotion Intensity: The strength of the expressed emotion is factored into the influence score. The emotional influence (E_m) of each comment on the root node is given by:
  The impact of nodes in ***G = (V, E, A)*** *on the root node* ***R*** *is given by:*

$$\forall V \in G - \{R\}, E_m(R) = \sum E_m(V1, V2, V3....Vn) \quad (1)$$

where:

$E_m(Vi)$ = f(Ai)

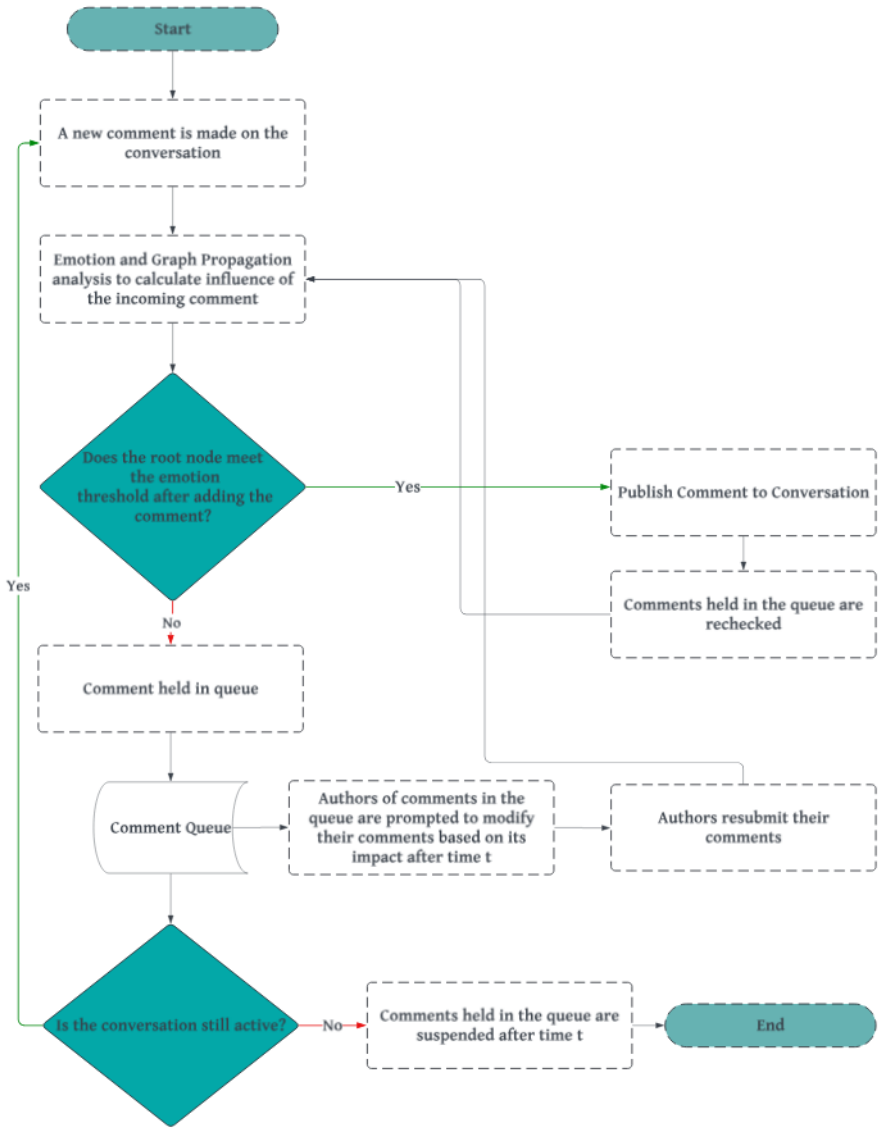

Fig. 1. Proposed Comment Queuing to encourage Self-reflection

These influence scores are used to update the emotion board of the root node, which aggregates the emotional influence of all comments in the conversation.

3.4 Proposed Comment Queuing to encourage Self-reflection

To prevent the escalation of negative emotions, we introduce a comment queuing mechanism that temporarily holds comments before they are added to the conversation graph, as shown in Fig. 1.

As new comments are added, their potential impact on the root node's emotion board is evaluated. The emotion board aggregates the emotional influence of all comments in the discussion, with specific thresholds set for each emotion (e.g., Anger > 50%, Fear > 60%). If a comment's emotional influence surpasses any of these thresholds, it is flagged as toxic and temporarily withheld from the conversation. Toxic comments are placed in a queue rather than being immediately added to the conversation graph, where they are re-evaluated each time a new comment is posted.

The toxicity thresholds are dynamically adjusted using an algorithm that considers the rate of comments, the overall emotional distribution within the conversation, and recent changes in emotional intensity. For instance, in particularly active discussions where many comments reflect high-intensity emotions, the thresholds for anger or fear may be raised temporarily to reduce the frequency of queuing. Conversely, during quieter periods, thresholds may be slightly lowered to facilitate stricter moderation and prevent the potential escalation of intense emotions. This adaptive mechanism ensures that thresholds remain contextually relevant and responsive to the evolving emotional tone of the conversation.

Additionally, the system utilises a sliding window approach, focusing on the most recent comments within a specific time-frame to maintain the relevance of the emotion board to the current conversation context. If a queued comment no longer causes the emotion board to exceed established thresholds due to these ongoing adjustments, it is reintegrated into the conversation.

Comments that remain in the queue after all others have been processed prompt the author to revise their input. Once modified, the comment undergoes re-evaluation, considering any changes to the emotion board and thresholds. If the revised comment's emotional impact falls within acceptable limits, it is added back to the conversation. Conversely, if it still surpasses toxicity thresholds or the user chooses not to revise, the comment will be suspended to prevent further emotional escalation in the discussion. This approach ensures effective moderation of potentially inflammatory content while fostering constructive participation.

In accordance with the framework established by [Slovak et al. 2023] for classifying technology-based emotion regulation interventions, the proposed queuing mechanism in our study can be positioned within its three core dimensions:

- Theoretical Component: This queue-based approach is designed to foster self-reflection prior to response, specifically in line with Gross's Process Model of Emotion Regulation [Gross 2008]. It encompasses response modulation and attentional deployment, prompting users to reconsider or modify their comments if flagged for exceeding toxicity thresholds.
- Strategic Component: The queuing mechanism employs an experiential learning strategy by introducing an on-spot intervention during the comment-posting process. It offers real-time feedback regarding the emotional tone of the conversation, by dynamically adjusting the queuing thresholds based on the conversation's emotion levels.
- Practical Component: In practical terms, this design utilises both implicit feedback mechanisms (queue time acting as a moderating factor) and explicit prompts (requests for comment revision) to encourage self-reflection and reduce the escalation of negative emotions.

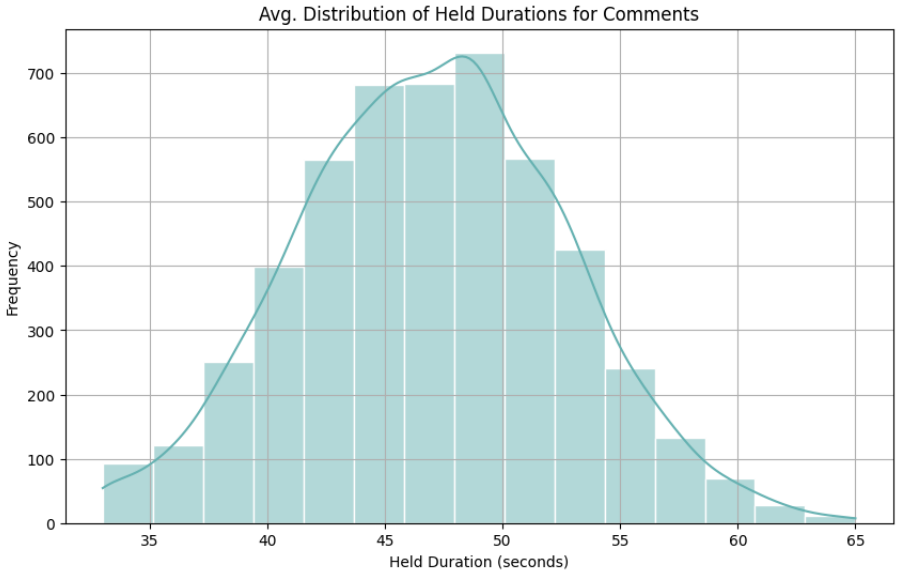

Fig. 2. Average Distribution of Held Durations for Comments when using the proposed queue approach

4 Results

The preliminary analysis of the comment queuing mechanism, as illustrated in Fig. 2, presents a histogram of the frequency distribution of the time comments are held in the queue before being added to the conversation graph.

The average range of held durations varies from approximately 33 to 65 seconds for conversations with up to 5000 comments, demonstrating that the queuing mechanism accommodates a range of response times based on the nature of the conversation. The majority of comments are held for durations ranging from 40 to 55 seconds, with a significant peak at 47 seconds, indicating that this time frame is frequently sufficient for evaluating the emotional impact of a comment before deciding whether to include it in the conversation. The smooth curve on the histogram illustrates the trend of held durations.

As comments are released from the queue and added to the conversation graph, the emotional intensity of various emotions in the root node of the conversation is consistently changing. When comparing emotion boards for scenarios with and without the use of the queue, we observe a balanced influence across emotions when the queue is employed. The emotion board consistently maintains balanced emotional levels, preventing any single emotion from dominating the conversation as shown in Fig. 3. Conversely, when the queue is not utilised, the conversation is repeatedly dominated by anger and fear.

When the queue is used, the changes in emotion levels occur gradually, indicating that the comment queuing mechanism effectively mitigates spikes in negative emotions, by holding potentially inflammatory comments until they can be integrated into the conversation in a manner that preserves emotional balance. This process helps to limit the spread of negative emotions and ensures a more constructive and less emotionally charged discussion.

Initial findings suggest that the comment queuing approach effectively delays the posting of potentially toxic comments, giving users time to reconsider and revise their responses. With only 4%

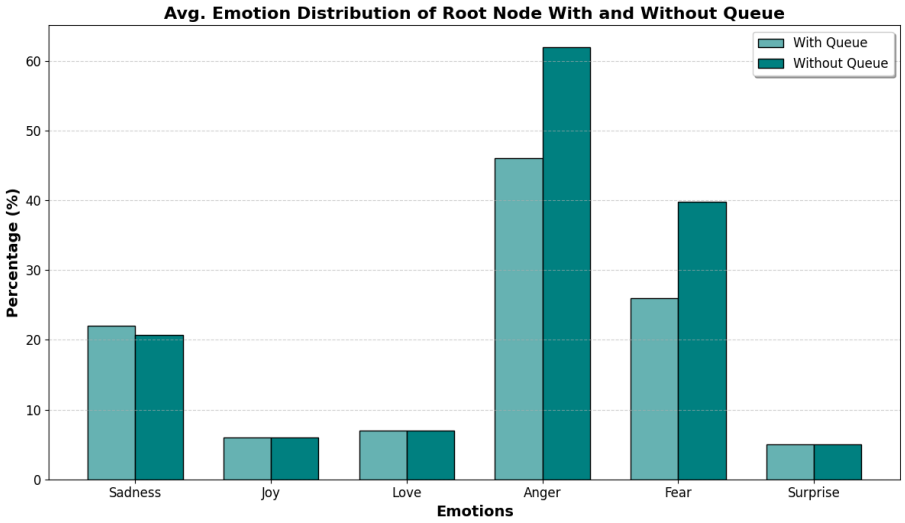

Fig. 3. Average Emotion Distribution of Root Node With and Without Queue

of comments being temporarily held for an average duration of 47 seconds, the system successfully manages the need for moderation while allowing smooth conversation flow.

Furthermore, there is an observed avg. 15% reduction in the spread of emotions such as anger and fear when comparing the emotional board of conversations using the queue versus those without it. This indicates the potential of this system to foster a more positive and less hostile online environment. The queuing mechanism maintains an active conversation while minimising emotional escalation often associated with trolling and other disruptive behaviours.

5 Conclusion

This research introduces an innovative method for moderating online conversations. It involves using a comment queuing system that encourages users to engage in self-reflection, ultimately curbing the spread of toxic and emotionally charged content. By introducing a brief delay in comment publication, we observed a reduction in the dissemination of anger and fear in the conversation. Our preliminary analysis indicates that the queuing mechanism can mitigate negative emotions, with only a small fraction of comments being temporarily withheld. We plan to evaluate this approach through user surveys.

6 Future Work

6.1 User Feedback and Evaluation

To assess the potential effectiveness and user perception of this queuing mechanism, we will conduct the following evaluations:

6.1.1 User Surveys. We plan to recruit participants to engage with prototype interfaces of the new system. They will participate in simulated conversations using the comment queue and will be asked to provide feedback on their experience, specifically in their emotional responses and any reflections they may have during the comment delay.

6.1.2 Surveys with HCI Experts. We plan to conduct surveys with Human-Computer Interaction (HCI) experts to gather their insights on the design and feasibility of the comment queuing approach.

This will involve discussions on the broader implications of integrating such mechanisms into digital platforms.

During the drafting of this paper, [Grammarly 2024] was used to check and enhance the grammar and writing style of this document.

References

Özlem Ayduk and Ethan Kross. 2010. From a distance: implications of spontaneous self-distancing for adaptive self-reflection. *Journal of personality and social psychology* 98, 5 (2010), 809.

Elena Barry, Shoaib Jameel, and Haider Raza. 2021. Emojional: Emoji Embeddings. In *UK Workshop on Computational Intelligence*. Springer, 312–324.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. 2014. Trolls just want to have fun. *Personality and individual Differences* 67 (2014), 97–102.

Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction* 1, CSCW (2017), 1–22.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1217–1230.

Daantje Derks, Agneta H Fischer, and Arjan ER Bos. 2008. The role of emotion in computer-mediated communication: A review. *Computers in human behavior* 24, 3 (2008), 766–785.

Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. 2016. The structural virality of online diffusion. *Management Science* 62, 1 (2016), 180–196.

Vaishali U Gongane, Mousami V Munot, and Alwin D Anuse. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining* 12, 1 (2022), 129.

Grammarly. 2024. Grammarly. https://www.grammarly.com. Accessed: 2024-06-20.

James J Gross. 2002. Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology* 39, 3 (2002), 281–291.

James J Gross. 2008. Emotion regulation. *Handbook of emotions* 3, 3 (2008), 497–513.

Uwe Herwig, Tina Kaffenberger, Lutz Jäncke, and Annette B Brühl. 2010. Self-related awareness and emotion regulation. *NeuroImage* 50, 2 (2010), 734–741.

Susan Holtzman, Drew DeClerck, Kara Turcotte, Diana Lisi, and Michael Woodworth. 2017. Emotional support during times of stress: Can text messaging compete with in-person interactions? *Computers in Human Behavior* 71 (2017), 130–139.

Annika Howells, Itai Ivtzan, and Francisco Jose Eiroa-Orosa. 2016. Putting the 'app'in happiness: a randomised controlled trial of a smartphone-based mindfulness intervention to enhance wellbeing. *Journal of happiness studies* 17 (2016), 163–185.

Emma A Jane. 2020. Online abuse and harassment. *The international encyclopedia of gender, media, and communication* 116 (2020).

Joel Kiskola, Thomas Olsson, Aleksi H Syrjämäki, Anna Rantasila, Mirja Ilves, Poika Isokoski, and Veikko Surakka. 2022. Online Survey on Novel Designs for Supporting Self-Reflection and Emotion Regulation in Online News Commenting. In *Proceedings of the 25th International Academic Mindtrek Conference*. 278–312.

Joel Kiskola, Thomas Olsson, Heli Väätäjä, Aleksi H. Syrjämäki, Anna Rantasila, Poika Isokoski, Mirja Ilves, and Veikko Surakka. 2021. Applying critical voice in design of user interfaces for supporting self-reflection and emotion regulation in online news commenting. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National academy of Sciences of the United States of America* 111, 24 (2014), 8788.

Ethan Kross and Ozlem Ayduk. 2008. Facilitating adaptive emotional analysis: Distinguishing distanced-analysis of depressive experiences from immersed-analysis and distraction. *Personality and Social Psychology Bulletin* 34, 7 (2008), 924–938.

Srijan Kumar, Justin Cheng, and Jure Leskovec. 2017. Antisocial behavior on the web: Characterization and detection. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 947–950.

Abdurahman Maarouf, Nicolas Pröllochs, and Stefan Feuerriegel. 2022. The Virality of Hate Speech on Social Media. *arXiv preprint arXiv:2210.13770* (2022).

Lydia Manikonda, Ghazaleh Beigi, Huan Liu, and Subbarao Kambhampati. 2018. Twitter for sparking a movement, reddit for sharing the moment:# metoo through the lens of social media. *arXiv preprint arXiv:1803.08022* (2018).

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada* 2 (2013), 234.

Manoj Parameswaran and Andrew B Whinston. 2007. Social computing: An overview. *Communications of the association for Information Systems* 19, 1 (2007), 37.

Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. (2016).

Minna Ruckenstein and Linda Lisa Maria Turunen. 2020. Re-humanizing the platform: Content moderators and the logic of care. *New media & society* 22, 6 (2020), 1026–1042.

Sarita Yardi Schoenebeck. 2014. Giving up Twitter for Lent: how and why we take breaks from social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 773–782.

Petr Slovak, Alissa Antle, Nikki Theofanopoulou, Claudia Daudén Roquet, James Gross, and Katherine Isbister. 2023. Designing for emotion regulation interventions: an agenda for HCI theory and research. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–51.

Wally Smith, Greg Wadley, Sarah Webber, Benjamin Tag, Vassilis Kostakos, Peter Koval, and James J Gross. 2022. Digital Emotion Regulation in Everyday Life. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

Jared B Torre and Matthew D Lieberman. 2018. Putting feelings into words: Affect labeling as implicit emotion regulation. *Emotion Review* 10, 2 (2018), 116–124.

Amaury Trujillo and Stefano Cresci. 2022. Make reddit great again: assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW2 (2022), 1–28.

Geetanjali Upadhyaya. 2020. Richard H. Thaler and Cass R. Sunstein (2009). Nudge: Improving Decisions about Health, Wealth, and Happiness. *Economic Journal of Nepal* 43, 1-2 (2020), 96–97.

Akriti Verma, Shama Islam, Valeh Moghaddam, and Adnan Anwar. 2023. Encouraging Emotion Regulation in Social Media Conversations through Self-Reflection. *arXiv preprint arXiv:2303.00884* (2023).

Greg Wadley, Wally Smith, Peter Koval, and James J Gross. 2020. Digital emotion regulation. *Current Directions in Psychological Science* 29, 4 (2020), 412–418.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, Lynne Edwards, et al. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2, 0 (2009), 1–7.