석 사 학 위 논 문
Master's Thesis

온라인 토론의 부정에 대한 피드백 메시지에 대한 사용자의 반응 이해

Understanding Users' Reactions to Feedback Messages about Incivility in Online Discussion

2020

나할 지번 (Naher, Jibon)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

석 사 학 위 논 문

온라인 토론의 부정에 대한 피드백 메시지에 대한 사용자의 반응 이해

2020

나할 지번

한 국 과 학 기 술 원

전산학부

온라인 토론의 부정에 대한 피드백 메시지에 대한 사용자의 반응 이해

나할 지번

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2020년 06월 16일

심사위원장 김주호 (인)

심 사 위 원 차미영 (인)

심 사 위 원 장정우 (인)

Understanding Users' Reactions to Feedback Messages about Incivility in Online Discussion

Jibon Naher

Advisor: Juho Kim

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Computer Science

Daejeon, Korea
June 16, 2020

Approved by

Juho Kim
Associate Professor in School of Computing

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MCS . 온라인 토론의 부정에 대한 피드백 메시지에 대한 사용자의 반응 이해. 전산학부 . 2020년. 26+iv 쪽. 지도교수: . (영문 논문)
Jibon Naher. Understanding Users' Reactions to Feedback Messages about Incivility in Online Discussion. School of Computing . 2020. 26+iv pages. Advisor: Juho Kim. (Text in English)

초 록

온라인 토론의 일반적인 문제는 괴롭힘이 정치적, 성적, 인종적 등일 수있는 괴롭힘입니다. 기존의 중재 방법은 괴롭힘 게시물을 탐지하고 적절한 경우 조치를 취하는 데 중점을 둡니다. 다양한 탐지 및 중재 기술을 사용할 수 있지만 온라인 괴롭힘은 쉽게 줄일 수있는 문제가 아닙니다. 기존 연구에서 대부분 누락 된 것은 커뮤니티의 규범 적 행동과 일치하는 행동으로 사용자를 독려하고 민첩한 컨텐츠를 줄이는 것입니다. 이 작업에서는 사용자 콘텐츠의 무의미한 존재에 대한 피드백 메시지를 사용자에게 제공하여 민중 댓글 작성의 잠재력을 탐구합니다. 피드백 메시지 후 사용자의 반응 및 향후 게시 동작을 조사합니다. 게시물에 잠재적인 부정적인 콘텐츠가 포함 된 ML 출력을 제공하는 경우 사용자는 어떻게 반응합니까? 피드백 메시지의 여러 구성 요소가 사용자의 반응에 어떤 영향을 줍니까? 이러한 피드백 메시지를 보내는 향후 주석 처리 동작에 긍정적 인 영향이 있습니까? 위의 질문에 대한 답변을 분석하면 온라인에서 괴롭힘 콘텐츠를 크게 줄일 수있는 새로운 기회를 제공 할 수 있습니다. 이 작업의 결과는이 측면에서 더 많은 것을 조사 할 유망한 가능성을 보여줍니다.

핵 심 낱 말 온라인 괴롭힘, 콘텐츠 조정, 피드백 메시지, 행동, 독성

Abstract

A common problem of online discussion is harassment where harassment may be political, sexual, or racial. Existing moderation methods focus on detecting harassment contents and taking action when appropriate. In spite of the availability of different detection and moderation techniques, online harassment is not an issue which can be reduced easily. Mostly missing in the existing research is designing to encourage users in the behavior which align with the normative behavior of the community, and reduce uncivil content. In this work, I explore the potential of promoting civil commenting, by providing feedback messages to the user regarding the existence of incivility in user generated content; and investigate user's reaction and future posting behavior after the feedback message. How do users react if providing the machine learning output of having potential negative content in their post? How do different components of the feedback message affect users' reactions? Is there any positive effect in the future commenting behavior of sending these feedback messages? Analyzing the answers of the above questions can offer new opportunities to significantly reduce harassment content online, by proactively notifying users about the potential bad contents in their post. The result from this work, shows promising potential, which needs to investigate more in this aspect.

Keywords Online harassment, Content moderation, Feedback message, Behavior, Toxicity

Contents

Contents i
List of Tables iii
List of Figures iv

Chapter 1. Introduction **1**

Chapter 2. Background **3**
2.1 Content moderation 3
2.2 Using AI/ML tools to assist moderation 3

Chapter 3. Method **5**
3.1 Feedback Message 5
3.1.1 Study 1 5
3.1.2 Study 2 5
3.2 Formulating Hypotheses 6
3.2.1 Study 1 6
3.2.2 Study 2 6
3.3 Study conditions 7
3.3.1 Study 1 7
3.3.2 Study 2 7
3.4 Study Variables 7
3.5 Study Materials 8
3.6 Study pipeline 9

Chapter 4. RESULT **11**
4.1 Analysis strategy 11
4.1.1 *Reaction* 11
4.1.2 *Acceptance* 11
4.1.3 *Intent* 11
4.1.4 *FutureCommenting* 12
4.2 Quantitative result 12
4.2.1 Study 1 12
4.2.2 Study 2 14
4.2.3 Validity of study design 15
4.3 Qualitative findings 16

Chapter 5. DISCUSSION **18**
5.1 Effect in study 1 18
5.2 Effect in study 2 18
5.3 Significance test explanation 19
5.3.1 *Reaction* 19
5.3.2 *Acceptance* 19
5.3.3 *Intent* 19
5.3.4 *FutureCommenting* 19

Chapter 6. LIMITATION **20**

Chapter 7. CONCLUSION **21**

Bibliography **22**

Acknowledgments **25**

Curriculum Vitae **26**

List of Tables

3.1 Study conditions for study 1 7
3.2 Study conditions for study 2 8
3.3 Control, independent variables for study 1 and 2, and dependent variables 8

4.1 Average score comparison for *acceptance, intent* in study 1 13
4.2 Significance test results in study 1 14
4.3 Average score comparison for *acceptance, intent* in study 2 14
4.4 Significance test results in study 2 15

List of Figures

3.1 Feedback template and example message for study 1 5
3.2 Feedback template and example message for study 2 6
3.3 Simplified study pipeline 9

4.1 Percentage comparison of the characteristics of the dependent variables *reaction, acceptance,* and *intent*, for study 1 12
4.2 Comparison of incivility score between target and next comments of the users, and between next comments of the holistic and word-specific feedback type 13
4.3 Percentage comparison of the characteristics of the dependent variables *reaction, acceptance,* and *intent*, for study 2 15
4.4 Comparison of incivility score between target and next comments of the users, and between next comments of the low and high prompt question condition of study 2 16

Chapter 1. Introduction

Most online platforms prohibit obviously racist, homophobic, and hateful content. Still, the existence of abusive content is common across online platforms [1, 2]. To reduce potential damage caused by bad actors, different platforms adopt different techniques to moderate the contents [3]. These techniques take two primary forms: human moderation and human moderation augmented by automated techniques. In the former case, teams of human moderators including externally contracted workers, and/or a small number of selected users from the platform, manually go through the contents, and remove contents that violate the terms and conditions of the platform [4]. Users can also contribute in content moderation via voting or reporting mechanism. However, human moderators need time to filter content, and the constant exposure to disturbing content negatively and substantially affects the mental health of the moderators [5].

To speed up the moderation decision and to keep up with the immense volume of content generated by users everyday, online platforms are known to apply machine learning (ML) algorithms trained with large datasets of past moderation decisions on the platform [6] - [8]. Machine learning approaches are specially helpful in saving time and effort of human moderators by algorithmically triaging comments to review [9]. Moreover, bots are using largely to moderate content directly without any human intervention on several platforms [21, 22, 30]. Still, these moderation approaches are not enough, and online platforms are struggling in moderating the ample bad content every day [23, 2]. Mostly missing in the existing research is designing to encourage users in the behavior which align with the normative behavior of the community, and reduce uncivil contents. If users are encouraged in civil behavior, there is no uncivil content in the first place to moderate.

In this work, I explore the potential of promoting civil commenting using machine learning, by providing feedback messages to the user regarding the existence of incivility in users' content; and investigate users' reaction and future posting behavior after the feedback message. How do users react if providing the ML output of having potential negative content in their post? How do different components of the feedback message affect users' reactions? Is there any positive effect in future commenting behavior of sending these feedback messages? Analyzing the answers of the above questions can offer new opportunities to significantly reduce harassment content online, by proactively notifying users about the potential bad content in their post. This work adds support to the line of research that calls for taking an educational approach, rather than a punitive approach to content moderation [12].

Prior work analyzes end-users' perception of fairness about content removal decisions, and how it can affect their future posting tendency [10]. There is also a growing body of research on understanding bad actors online [13] - [18]. Coleman [17] and Phillips [18] conducted deep ethnographic investigations to understand the subculture of internet trolls. More recently, in several works in the context of reddit, Jhaver et al. pointed out the challenges of distinguishing sincere users from bad actors online [11, 13, 14]. This work adds to this research by analyzing the reaction and posting behavior of users on reddit after sending the feedback message.

This work is guided by the following four research questions -

RQs

- How do users react when sending a feedback message of having incivility in their comment using machine learning in online discussion?

- Do users accept the ML output of their comment?
- Do users intend to change the comment after sending the feedback message?
- What would be the effect in future posting behavior of sending feedback messages to users?

The feedback message has two key components and I did two separate studies in respect of the two parts of the feedback message, to investigate user's reaction to the feedback message, acceptance of ML output, intention to change the comment, and any effect on the future commenting behavior of user. The result shows some positive effect, which need further exploration to make a meaningful conclusion.

In the following sections, I first describe the background work related to this work. Then, the detailed design and development of both studies. After that, I show and discuss the findings from the studies. I conclude by the limitation and future plan to improve the findings of this work.

Chapter 2. Background

In background, I discuss the literature from two areas: content moderation in online space, and the use of artificial intelligence(AI)/machine learning(ML) to help in content moderation.

2.1 Content moderation

Different platforms adopt different techniques to moderate content in online space, those techniques take two primary forms: human moderation, and human moderation augmented by automated techniques.

Human moderation:

Online platforms employ the services of moderators (either paid or unpaid) who regulate content generated within the platform. Human moderation typically has two forms: centralized and distributed. In the centralized approach, teams of human moderators such as externally contracted workers, and/or a small number of power users from the platform, manually go through posts, and remove content with profane text or imagery [3]. In the distributed approach, users in the platform flag inappropriate submissions via voting or reporting mechanisms, which notify the moderators and they take action. However, the constant exposure to disturbing content negatively and substantially affects the mental health of moderators [4].

Human moderation augmented by automated techniques

As online communities grow larger, moderating content becomes increasingly difficult [19]. To keep up with the immense volume of content created by users, online platforms train and apply machine learning algorithms by compiling large datasets of past moderation decisions on the platform [6] - [8]. AI/ML tools, or moderator bots are common to use in content moderation to assist the human moderators in almost every platform [22].

2.2 Using AI/ML tools to assist moderation

To keep up with the immense volume of content created by users, using AI/ML tools to assist in moderation is very common. Machine learning approaches are helpful in saving time and effort of human moderators by algorithmically triaging comments to review [9]. Moreover, bots are using largely to moderate content directly without any human intervention on several platforms [21, 22, 30]. However, deploying these algorithms without any human oversight can sometimes be problematic; for example, in 2018, Tumblr launched a new anti-porn algorithm to flag pornography, but it was accused of creating chaos by flagging random, nonsexual posts [24]. Nonetheless, machine learning approaches can be especially helpful for algorithmically triaging contents for human moderators to review. For example, from April, 2019 to June, 2019, 99.3% of comments on Youtube were removed after flagging from the automatic detection [7]. The number of reports on Twitter had decreased from 868,349 in January, 2018

to approximately 504,259 in June, 2018, after it introduced technology to proactively identify offensive content, which was able to flag 97% of users account for violation before any report [8].

Although adopting these moderation techniques help in reducing the existing uncivil contents in the discussion, still, these moderation techniques are not enough. The generation of uncivil contents is increasing every day, and online platforms are struggling in moderating the content [23]. If users can be encouraged in normative behavior, uncivil contents can be potentially reduced.

In this work, I investigate the potential of promoting a user's civil commenting behavior by exploring the user's reaction to the feedback message. I also discuss the future implications and consideration in designing for civil commenting, from the analysis of the study findings.

Chapter 3. Method

In this chapter, I describe the details of the design and development of the two studies. I discuss the details (feedback message, study conditions, hypotheses, study variables) for study 1 and 2 separately, except three parts: study materials, study pipeline, and participants, which are the same for both studies.

3.1 Feedback Message

To design the feedback message, I took inspiration from the feedback intervention theory (FIT) [20]. The feedback message has two key components: feedback about the ML output, and prompt questions for critical thinking. According to FIT, feedback is processed hierarchically and there are two level of the hierarchy in simpler form: task process feedback (lower level), tells about what the feedback is about; meta task process feedback (higher level), focuses on the user's self-belief, self-goal or self-perception about the task. In study 1 and 2, I investigate the effect of lower level and higher level of feedback message on user's reaction, and commenting behavior separately.

3.1.1 Study 1

In study 1, I focused on the lower level of the feedback message. I designed two versions of message in the lower level, to design the ML output. I designed holistic vs word-specific feedback; holistic feedback only says the comment is uncivil, and word-specific feedback gives some additional information (bad words responsible for making the comment uncivil). I investigated how the two versions of this message affect users' reactions and commenting behavior in online discussion, also what is the potential of sending this message in reducing uncivil commenting. I kept the high prompt questions for the higher level of the feedback hierarchy in study 1. The message template has five components.

Message Template:

1. [Base message part 1: greetings+indicating the user's comment+self-intro]
2. [Task process feedback, ML output: holistic vs word_specific]
3. [Feedback req]
4. [Prompt questions - high]
5. [Ending greeting + Sender's name]

(a) Feedback template

Hi receiver,
I just saw one of your comments in r/AskReddit.
"Comment text Comment text Comment text Comment text Comment text Comment text Comment text Comment text Comment text Comment text"
I am a researcher working on a project on detecting toxicity/incivility in text, and our algorithm found the comment above as toxic.

Could you provide feedback about the algorithmic output by answering the questions below? Your valuable feedback will help us improve the algorithm.

1. How acceptable is the algorithmic output? (on a 10-point scale: 1 - not acceptable at all, 10 - completely acceptable)
2. If you think it is acceptable, why do you think so?
3. To what extent do you intend to change the comment to make it less toxic? (on a 10-point scale: 1 - not intended at all, 10 - very much intended)
4. If you think it is not acceptable, why do you think the algorithm may predict it as toxic?

Thanks!
Sender

(b) Example message

Figure 3.1: Feedback template and example message for study 1

3.1.2 Study 2

For the higher level of feedback hierarchy, I designed prompt questions about the ML output. In study 2, I designed low and high prompt questions about the ML output in respect of the critical thinking: what questions in the lower level, and why questions (higher level) on top of it in the higher level. For

study 2, I kept the holistic message as a lower level message of the feedback hierarchy. After conducting study 1, I added part 5 in the message of study 2, as I found out this has the potential of reducing the negative tone of the feedback message. The message template for study 2 has six components.

Figure 3.1 and 3.2 shows the message template and example message for study 1 and 2, respectively.

Message Template:

1. [Base message part 1: greetings+indicating the user's comment+self-intro]
2. [ML output - holistic]
3. [Feedback req]
4. [Prompt question: low vs high]
5. [Liability]
6. [Ending greeting + Sender's name]

(a) Feedback template

Hi receiver,
I just saw one of your comments in r/AskReddit.
I am a researcher working on a project on detecting toxicity/incivility in text, and our algorithm found the comment as toxic.

Could you provide feedback about the algorithmic output by answering the questions below:

1. As the author of the comment, how acceptable is the algorithmic output to you? (on a 10-point scale: 1 - not acceptable at all, 10 - completely acceptable)
2. To what extent do you intend to change the comment to make it less toxic? (on a 10-point scale: 1 - not intended at all, 10 - very much intended)
3. If you think it is acceptable, why do you think so?
4. If you think it is not acceptable, why do you think the algorithm may predict it as toxic?

[N.B: Hope this message does not bother you. Your valuable feedback will help us improve the algorithm.]

Thanks!
Sender

(b) Example message

Figure 3.2: Feedback template and example message for study 2

3.2 Formulating Hypotheses

The effect of feedback intervention (FI) depends on the focus of the message. If the message is task-focused than self-focused, especially in the higher level of the feedback, it improves the task performance, which is the focus of FI [20]. The lower level feedback is already task-oriented. To design task-oriented feedback in higher level, I designed prompt questions about the ML output; thus, focusing attention to the task, not on the self.

3.2.1 Study 1

Conflicting arguments exist about how much information should provide about an algorithm output. Although providing more information can increase user's trust on the algorithm, previous work showed too much information can have a negative effect on users' reaction about the algorithm output [31, 32]. In this work, I designed holistic vs word-specific messages for providing the ML output; holistic feedback only says the comment is uncivil, and word-specific feedback gives additional information (bad words responsible for making the comment uncivil). In accordance with the previous work, I expect that users getting holistic messages will react more positively than users getting word-specific messages. My hypotheses in study 1 are follows -

S1H1: Users getting holistic messages will show more positive effect on user's reaction and commenting behavior than users getting word-specific messages.

3.2.2 Study 2

In study 2, the focus of interest was the effect of the higher level of feedback on user's reaction. Prompt questions are an effective way of triggering critical thinking [27]. Explicitly prompting self-explanations improves comprehension of information, even a simple prompt can trigger critical thinking and self-explanation [26, 27, 29]. I expected to see a higher effect on user's reaction and commenting behavior in the high prompt questions than low prompt questions. My hypothesis in this study is -

S2H1: High prompt questions will have more positive effect on user's reaction and commenting behavior than low prompt questions.

3.3 Study conditions

In both the studies, I have two versions of message of the lower and higher level of feedback message, and two conditions corresponding to these two versions.

3.3.1 Study 1

I have two versions of message for providing the ML output to the user, and two conditions corresponding to these two versions. Table 3.1 shows the study conditions for study 1. Since the focus of this study is to understand the user's reaction after sending a feedback message, I do not have any "no message" condition.

Conditions	Explanation	Example feedback message part
Condition 1: Holistic	Holistic feedback message about ML output	The detection algorithm found the comment as toxic.
Condition 2: Word-specific	Word-specific feedback message about ML output	The detection algorithm found the comment as toxic, and words which contributes most are - "asshole, stupid"

Table 3.1: Study conditions for study 1

3.3.2 Study 2

In study 2, I have two versions of prompt questions about the ML output to the user, and two conditions of this study corresponding to these two versions.Table 3.2 shows the study conditions for study 1. Similar to study 1, I do not have any "no message" condition.

3.4 Study Variables

The control and dependent variables in both studies were the same, only the independent variables differ depending on the focus of the specific study. In both the studies, I took four dependent variables to investigate: user's reaction to the feedback message, acceptance of the ML output, intention of changing the target comment (the comment for which the message is sending), and future commenting behavior after sending the feedback message (FM); I refer these dependent variables as *reaction, acceptance, intent* and *futureCommenting*.

I took two variables as control variables, which I suspected and observed from an initial pilot study that they are likely to have an effect on the dependent variables: incivility measure, initiator user. Incivility measure is related to the corresponding comment, I use ML and bad word counts to measure it, and select the comment having a score greater than 70% incivility. The second control variable, initiator user, checks whether the user is the initiator of uncivil comments in a thread, which provokes

Conditions	Explanation	Example feedback message part
Condition 1: Low prompt question	Low level prompt question about the ML output	1. How acceptable is the algorithmic output to you? (on a 10-point scale: 1 - not acceptable at all, 10 - completely acceptable) 2. If you think it is acceptable, to what extent do you intend to change the comment to make it less toxic? (on a 10-point scale: 1 - not intended at all, 10 - very much intended)
Condition 2: High prompt question	High level prompt question about the ML output	1. How acceptable is the algorithmic output? (on a 10-point scale: 1 - not acceptable at all, 10 - completely acceptable) 2. To what extent do you intend to change the comment to make it less toxic? (on a 10-point scale: 1 - not intended at all, 10 - very much intended) 3. If you think it is acceptable, why do you think so? 4. If you think it is not acceptable, why do you think the algorithm may predict it as toxic?

Table 3.2: Study conditions for study 2

more uncivil comments in the same thread. I observed sometimes there is a comment thread of uncivil comments, in such cases I sent FM if the target user is the uncivil initiator user in the thread. Otherwise, if there is no such thread of uncivil comment, I sent the FM to the target comment author.

Independent variables for study 1 is the lower level of the feedback hierarchy message, and for study 2 it is the higher level of the feedback hierarchy message. Table 3.3 shows all the variables.

Control variables	Independent variables	Dependent variables
1. Incivility measure 2. Initiator user	Study 1: lower level of feedback: Holistic vs. Word-specific Study 2: higher level of feedback: Low prompt question vs. High prompt question	1. User's reaction to FM: *reaction* 2. Acceptance of the ML output: *acceptance* 3. Intention of changing the target comment: *intent* 4. Future commenting behavior after FM: *FutureCommenting*

Table 3.3: Control, independent variables for study 1 and 2, and dependent variables

3.5 Study Materials

To calculate the first control variable (incivility measure), I combined two measures: the incivility score from perspective API[1], and ratio of bad words among total words in the comment. Perspective API gives a high score irrespective of having one or multiple bad words in the comment. With the additional measure of bad word ratio, I can filter comments which are blatantly offensive, and which have bad words as part of a long text. I also needed to filter bad words to add in the feedback message, which may have directed the high incivility score for the word-specific feedback message version. I used harassment lexicon from a previous work [28], and modify the list based on the occurrence of words

[1]https://www.perspectiveapi.com/

in previous reddit comments [1]. I removed words which did not occur at all, and added some words which were not in the list, but occured frequently in uncivil manner in the comments (e.g., modified or misspelled version of some bad words). The final list contains 164 bad words.

I used the toxicity score provided by perspective API. There are total eight types of score one can get from the perspective API. Among them, six are in experimental stage, and the other two are toxicity and severe toxicity respectively, where severe toxicity ignore usage of bad words in small scale. I used the toxicity score, which is the general score of incivility level of a text.

I conducted the study on r/AskReddit[2] subreddit. I decided to select this subreddit based on three properties: high subscriber number, high number of removed comments, general subreddit (not political or religious stance). r/AskReddit is one of the top subreddit in respect to the number of subscribed users[3]. It also has the highest number of removed comments other than political subreddits [1]. This subreddit has high responsiveness in respect to a post, and the environment of the subreddit is average in case of toxicity [34]. Since, this is a subreddit to ask any kind of questions, people may ask questions which has adult content understanding. Normally, these posts are tagged as NSFW, but not always. In NSFW posts, comments may exhibit toxicity, but this is contextual. Other than this, the environment is not highly toxic. As per the rules[4] of the subreddit, toxic comments are strictly forbidden.

3.6 Study pipeline

I monitored the comment stream of subreddit r/AskReddit using a reddit bot, and checked the comment's civility score using Google's perspective API, also the bad words in the comment. If I found a comment having a high incivility score, I did additional filtering (NSFW tagged post, bad words in a quotation etc.). Then, I assigned a study condition for the user, balancing the two conditions of a study, and sent a message to the user. Finally, saved the target comment and user information in files for later reference. Figure 3.3 shows the simplified pipeline of the study.

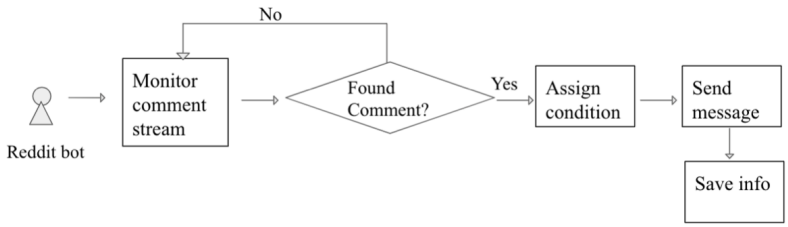


Figure 3.3: Simplified study pipeline

I excluded user if the comment was made in a post tagged as 'NSFW', which means 'Not safe for work'. NSFW posts contain adult content, and the comments made in NSFW posts naturally have contextual offensive contents[5]. I also excluded users if the bad words found were inside a quotation, which used to quote from another person or text. Moreover, if there was no activity (no reply to my FM or no comments by the user) of the user after sending the message, I considered that the user did not

[2]https://www.reddit.com/r/AskReddit/
[3]http://redditlist.com/
[4]https://bit.ly/3i2VptY
[5]https://bit.ly/2AfVJ7y

see the message, and excluded him from the user list. Also, I only considered the user whose account was at least one month old, to ensure the user was familiar with the community environment.

Finally, there was concern of being reported or flagged as spam by other users while doing the study. Fortunately, the occurrence of such incidents were less than I expected. There was one such occurrence, one user reported the feedback message as spam. I got a warning message from reddit admin about the reported message. Other than that, no such occurrence happened during the studies.

Chapter 4. RESULT

4.1 Analysis strategy

I used a similar strategy to analyze the dependent variables for both study 1 and study 2, which is described in this section.

4.1.1 *Reaction*

After observing the responses I got from the users, I decided on five characteristics for *reaction* as positive: whether the user give reply, whether the user give any feedback about the questions asked, whether the feedback is in the expected format, whether the reply has civil tone, whether the user give explanation of the possible reason behind the ML output. I annotated these characteristics of the replies, and compared the percentage of users in each characteristics under both conditions. Since, the last characteristics (whether the user gives explanation of the possible reason behind the ML output) was expected only in high prompt questions in study 2, but not in the low prompt questions, I discarded this one and annotated the first four characteristics for *reaction* in study 2. As the number of replies was different in the two conditions in both studies, I decided to use percentage to compare, but not the actual count of users.

4.1.2 *Acceptance*

For measuring the DV Acceptance, I asked the user's acceptance (out of score 10) of the ML output regarding the comment as the prompt question in the feedback message. I observed two types of responses from the replies, some users specifically answer this question with a score, as I asked, while some users gave text responses (e.g., *yes, this can be considered as toxic*). I annotated two characteristics for *acceptance*: whether users give answers to the acceptance question, and whether users accept the ML output; and compared the percentages of these characteristics under both conditions. I also compared the average of exact acceptance scores (given by the user) under both conditions.

4.1.3 *Intent*

I asked the user's likelihood of intention of changing the comment (out of score 10). I found three types of responses should be considered for *intent*: two are similar to the acceptance response (based on exact score given by users, based on text reply from users), and another one is by checking the actual changes in the target comments. Similar to *acceptance*, I annotated two characteristics for *intent*: whether users give answers to the intent question, and whether users intend to change the comments; and compared the percentages of these characteristics under both conditions. Also, compared the average of exact *intent* scores (given by the user) under both conditions.

4.1.4 *FutureCommenting*

I collected users' next comment after the target comment, and compared the incivility score between target and next comments of each condition, and also compared the incivility score between next comments of the two conditions.

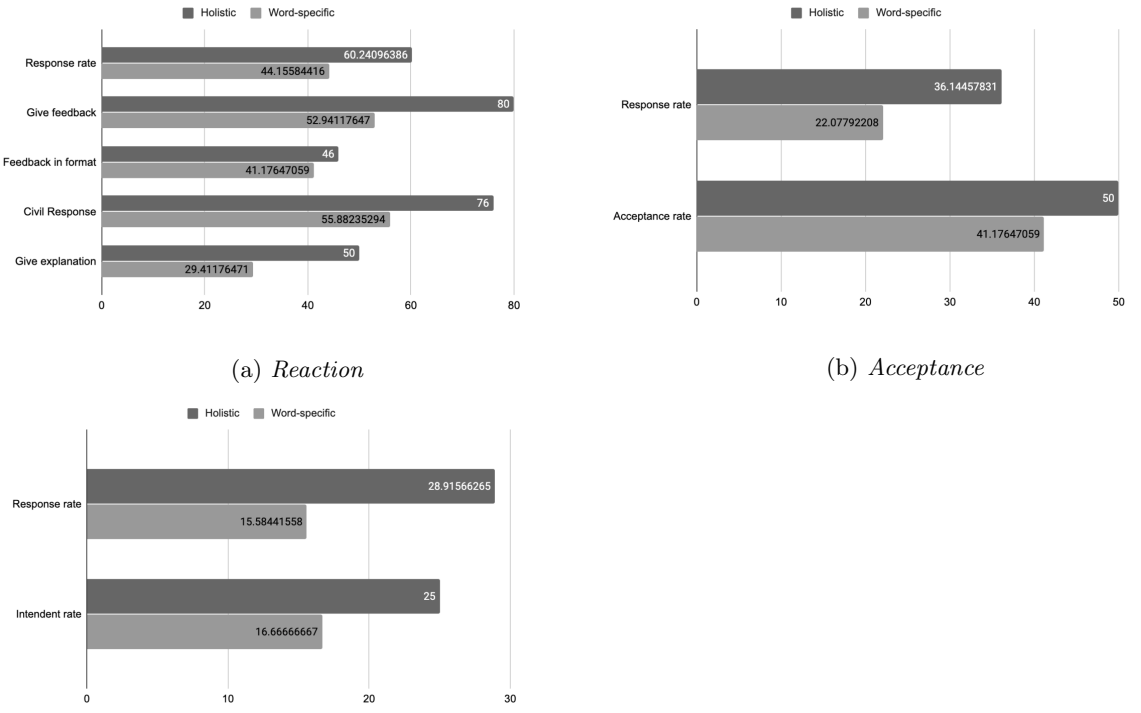
4.2 Quantitative result

In this section, I show the quantitative findings from study 1 and 2 separately.

4.2.1 Study 1

Percentage comparison of the annotated characteristics of *reaction, acceptance, intent*:

I annotated and compared the percentage of five characteristics of replies under both conditions for *reaction, acceptance,* and *intent*. Figure 4.1 shows the graph of the comparison.



(a) *Reaction*

(b) *Acceptance*

(c) *Intent*

Figure 4.1: Percentage comparison of the characteristics of the dependent variables *reaction, acceptance,* and *intent*, for study 1

Score comparison of *acceptance* and *intent*:

I asked the user's *acceptance* and *intention* of changing the target comment on a scale of 10 (1 - not acceptable/not intended at all, 10 - completely acceptable/very much intended). For *acceptance*, 21 out of 50 users who replied, gave a score in condition 1, and 15 out of 34 users gave in condition 2. For

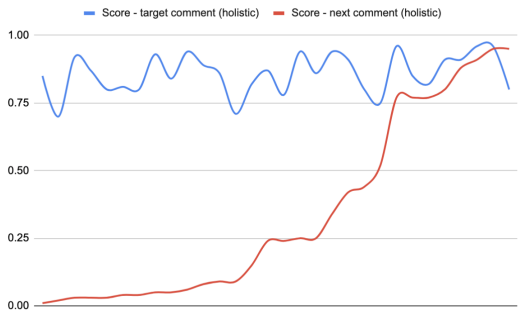
	Condition 1 - Holistic feedback	Condition 2 - Word-specific feedback
acceptance	4.77	3.9
intent	1.1333	1

Table 4.1: Average score comparison for *acceptance, intent* in study 1

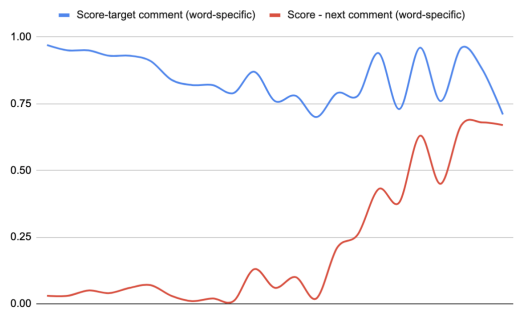
intent, users who gave a score are 15 and 11 respectively. I calculated the average of the values in both conditions shown in Table 4.1

***FutureCommenting* comparison:**

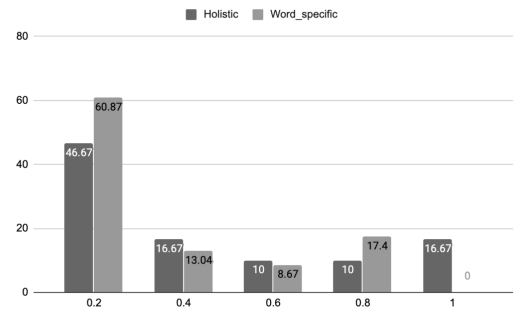
I compared the incivility score between target comment and next comment of a user, under each condition. I also compared the percentages of users in the next comments under each condition (holistic and word-specific feedback) of the study, by grouping into 5 bins. Since the number of users under each condition was different, I used percentage to compare the next comments. Figure 4.2 shows the result of the comparison.



(a) target and next comments in holistic feedback



(b) target and next comments in word-specific feedback



(c) next comments in holistic and word-specific

Figure 4.2: Comparison of incivility score between target and next comments of the users, and between next comments of the holistic and word-specific feedback type

Significant test results

To test the significance of the conditions, I did one-way ANOVA test for *reaction*, chi-square test for *acceptance, intent*, and Mann-Whitney Test for *futureCommenting*. Table 4.2 shows the result.

DV	Test	Result	Interpretation
Reaction	One-way ANOVA	F value, Pr(¿F) : 4.203, 0.0435	Bartlett's test did not show a violation of homogeneity of variances (p = 0.3303). With one-way ANOVA, we found a significant effect of Group on Value, F=4.203, p = 0.0435
Acceptance	Chi-square	X-squared = 0.077454, df = 1, p-value = 0.7808	Not significant
Intent	Chi-square	X-squared = 0.87891, df = 1, p-value = 0.3485	Not significant
FutureCommenting	Mann-Whitney Test	W = 426.5, p-value = 0.1455	Not significant

Table 4.2: Significance test results in study 1

4.2.2 Study 2

Percentage comparison of the annotated characteristics of *reaction, acceptance, intent*:

I annotated and compared the percentage of four characteristics of replies under both conditions for *reaction, acceptance,* and *intent*. Figure 4.3 shows the graph of the comparison.

Score comparison of *acceptance* and *intent*:

	Condition 1 - Low prompt question message	Condition 2 - High prompt question message
acceptance	4.0	4.41
intent	1.13	1.83

Table 4.3: Average score comparison for *acceptance, intent* in study 2

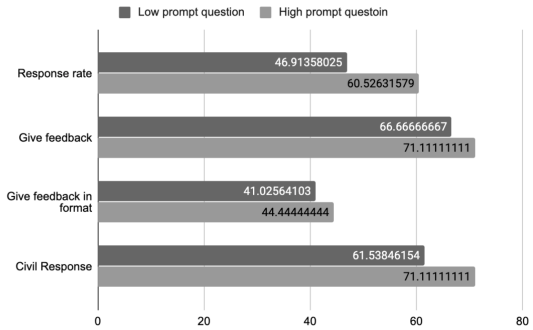
I did similar analysis for the score comparison of *acceptance* and *intention* for study 2, the value is shown in Table 4.3. In study 2, 14 out of 38 and 23 out of 46 users gave a score for the acceptance question for condition 1 and 2 respectively. And, for *intent*, users who gave a score are 14 and 13 respectively.

***FutureCommenting* comparison:**

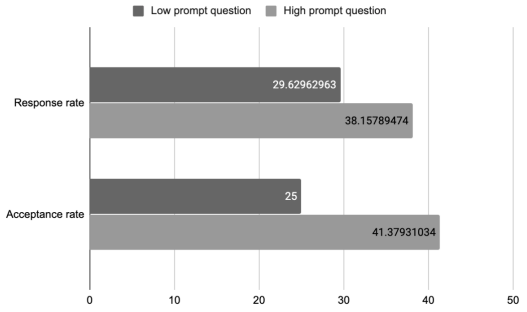
I did similar analysis for *futureCommenting* behavior of study 2 as study 1, shown in Figure 4.4

Significant test results

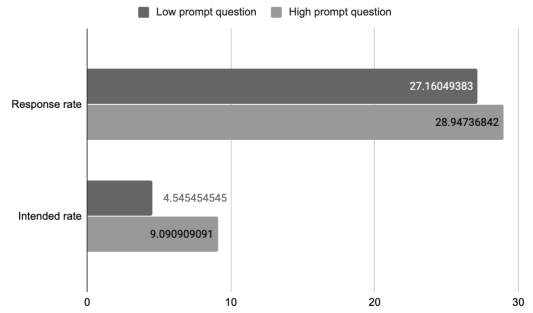
To test the significance of the conditions, I did similar significance test as study 1. Table 4.4 shows the result.



(a) *Reaction*



(b) *Acceptance*



(c) *Intent*

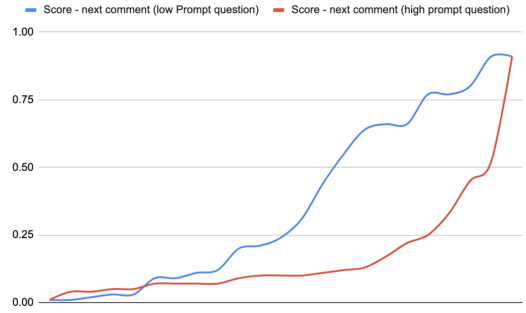
Figure 4.3: Percentage comparison of the characteristics of the dependent variables *reaction, acceptance,* and *intent*, for study 2

DV	Test	Result	Interpretation
Reaction	One-way ANOVA	F value, Pr(¿F) : 0.415, 0.521	Not significant
Acceptance	Chi-square	X-squared = 0.10041, df = 1, p-value = 0.7513	Not significant
Intent	Chi-square	X-squared = 0, df = 1, p-value = 1	Not significant
FutureCommenting	Mann-Whitney Test	W = 341, p-value = 0.4053	Not significant

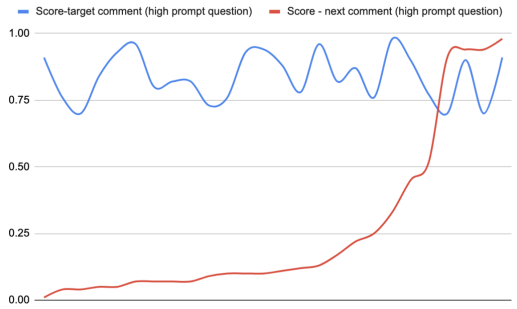
Table 4.4: Significance test results in study 2

4.2.3 Validity of study design

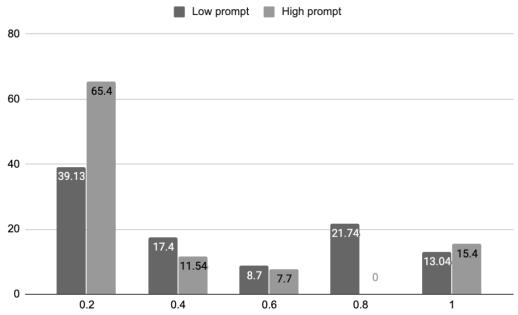
Since I varied one part of the feedback message keeping the other part fixed in study 1 and 2, I have a common condition from both studies, condition 1 in study 1: Holistic+High prompt question, and condition 2 in study 2: Holistic+High prompt questions. To test the validity of the study design, I compared the common condition from both study in the *reaction* DV. I did one way ANOVA test of these two conditions, and the result do not show significant difference between the two conditions, in other way, the condition can be considered same.



(a) target and next comments in low prompt question condition



(b) target and next comments in high prompt question condition



(c) next comments in low and high prompt question condition

Figure 4.4: Comparison of incivility score between target and next comments of the users, and between next comments of the low and high prompt question condition of study 2

4.3 Qualitative findings

From the reply of the users, I reported four qualitative findings: types of user's responses to the feedback message, user's perception about the ML output, user's justification of posting toxic comments, user's personal motivation for not changing the toxic comments. These qualitative findings give indication of the reason behind the lacking of strong differences from the significance test in both studies. I found these four qualitative findings from the responses of both study 1 and 2. Thus, in this section, I discuss the qualitative findings together for both studies, by mentioning study 1 or study 2 when needed specific information from a study.

User's responses to the feedback message:

I got four levels of responses from the user in respect to the ML output regarding the target comment: (1) taking the ML output as appropriate, these users agree with the ML output of being the target comment as toxic, without any confusion. (2) second level of users who accept the ML output having some conditions (e.g., without context). Even though they accept the ML output, but they also mentioned their lacking confidence of the certainty of the ML output. (3) third level of users are those who did not give a specific acceptance answer, but said that they do not care about the ML output of being toxic in the comments. (4) the final level is people who said specifically that they do not accept the ML output as appropriate.

Users' perception about the ML output:

Among the users who answered the acceptance question, 50% and 41.2% users in study 1 agree with the ML output as the target comment being toxic, in holistic and word-specific feedback respectively. For study 2, it is 33.33% and 41.37% respectively for low and high prompt questions. However, the majority of the users mentioned that the algorithm should consider the context of the comment. 25.5% users in study 1, and 22.6% users in study 2 mentioned the word "context" to explain their perception of acceptance of the ML output, suggesting that the ML is working only with the target comment, without any context.

Justification of posting toxic comments:

Users have their own justification for posting toxic comments. The most common form of justification is joking as mentioned by many users. 21.3% users in study 1, and 11.3% users in study 2 mentioned the word "joke" (either joke or joking) to explain why they made the comments, which might be considered as toxic. Other reasons mentioned by users were the hostile reddit environment, having different stance in political or religious perspectives, or opinion conflict with the other users made the users think that the comment created is justified whether it is toxic or not.

Reason for not intending to change the comment:

In study 1, 48% and 35.3% users answered the intent question (on a 10 point scale, how much intended the user is to make the comment less toxic) in holistic and word-specific feedback respectively. Among them, 25% and 16.7% of users showed the intention of changing the comment (either answering yes or actually making the changes). In study 2, the percentages were 4.5 and 9, among the total answered users of 56.4% and 47.8% for low and high prompt questions respectively. Majority of the users mentioned their willingness to stand for what they said as the reason for not intending to change the comment. In specific, 16.7% users in study 1, and 18.2% users in study 2 mentioned the word "stand" to explain why they did not intend to change the comment to a less toxic comment. Moreover, user pointed out the importance of moderator message for the feedback, which give user the motivation of changing the comment. As the feedback was from another user, they did not have intention of changing the target comment.

Chapter 5. DISCUSSION

From the result, the annotated characteristics comparison for *reaction, acceptance, intent* and *futureCommenting* shows positive effect in study 1 and 2: users getting holistic messages will show more positive effect on user's reaction and commenting behavior than users getting word-specific messages, and high prompt questions will have more positive effect on user's reaction and commenting behavior than low prompt questions. However, the significance test shows no strong differences between the conditions as shown in Table 4.2 and 4.4 for all the DVs, except *reaction* in study 1. In this chapter, I first discuss the result from the previous chapter and then, explain the result of significance test from the qualitative findings.

5.1 Effect in study 1

Figure 4.1, 4.2, and Table 4.1 show the quantitative results of the effect on *reaction, acceptance, intent* and *futureCommenting* with respect to the holistic and word-specific feedback type. From Figure 4.1, the percentages of positive responses of *reaction, acceptance* and *intent* in holistic feedback are higher than word-specific feedback. Also, the comparison of the average of exact values given by the users for *acceptance* and *intent* in Table 4.1 shows higher value in holistic feedback than word-specific feedback.

The *futureCommenting* behavior shown in Figure 4.2a, and 4.2b show that, both in holistic and word-specific feedback follows the same pattern between target and next comments; toxicity score for next comments is lower for majority of user. Figure 4.2c compare the percentages of users in each toxicity score range (group as 0.0, 0.2, 0.4, 0.6, 0.8, 1.0). The comparison shows no strong difference between the two condition.

5.2 Effect in study 2

For study 2, Figure 4.3, 4.4, and Table 4.3 show the quantitative results of the effect on *reaction, acceptance, intent* and *futureCommenting* with respect to the low and high prompt questions feedback type. From Figure 4.3, the percentages of positive responses of *reaction, acceptance* and *intent* in high prompt questions are higher than low prompt questions. Also, the comparison of exact value given by the users for *acceptance* and *intent* in Table 4.3 shows higher average value in high prompt questions than low prompt question.

Also, the *futureCommenting* behavior shown in Figure 4.4 shows similar results as study 1. Figure 4.4a, and 4.4b show that, both in low and high prompt question condition follow the same pattern between target and next comments; toxicity score for next comments is lower for majority of user under both conditions. Figure 4.4c compare the percentages of users in each toxicity score range (group as 0.0, 0.2, 0.4, 0.6, 0.8, 1.0). The comparison shows no strong difference between the two condition.

5.3 Significance test explanation

Although the comparisons between the conditions show positive effect for both studies, the significance tests shows no strong differences between the conditions as shown in Table 4.2 and 4.4 for all the DVs, except *reaction* in study 1. I will explain the possible reason behind the results from the findings of qualitative results and limitation of this work.

5.3.1 *Reaction*

For *reaction* DV, study 1 shows that holistic feedback has higher effect than word-specific feedback as I suspected. However, study 2 shows no difference between low and high prompt question. The answer is within the focus of the two studies. In study 1, the focus is on the ML output, and in study 2, the focus is on critical thinking. The word-specific condition in study 1 has component to trigger negative reaction than holistic condition, by throwing the bad words as evidence, which I also found in initial pilot studies. However, there is no difference in the two conditions which can trigger negative reaction, as the focus is on the critical thinking. Thus, the two conditions show similar kind of effect on *reaction* DV for study 2.

5.3.2 *Acceptance*

In both study 1 and 2, significance tests show no strong difference between the conditions. I strongly believe the reason can be explained from the qualitative findings. In both studies, users showed the expression of lacking confidence on ML output regarding detecting contextual and situational toxic comments (details in the subsection 4.3). This indicates users' lack of trust in the ML output, which made the users having no motivation to accept the ML output, whatever the condition is.

5.3.3 *Intent*

Two explanation can be stated as the reason behind not having strong differences in the significance test for *intent*. The first one is the 'reactance' phenomenon, which has shown in political contexts before. When confronted with evidence that a view people hold is false, they tend to become firmer in their beliefs, instead of trying to correct it [33]. Another reason is within the qualitative findings. I found in both studies, the users showed their willingness to stay what they said, and that led them not to change the target comment even though they think the comment is toxic.

5.3.4 *FutureCommenting*

The reason of having no strong difference in the *futureCommenting* DV is connected to the two limitations of this work, explained in the next chapter: in this work, the message sender is a normal user, not moderators of the community, and message is sent one time, not repetitive reminder. This two reasons made the users having no strong motivation to behave in civil way. Another reason I suspect is followed from the negative effect on *acceptance* and *intent*.

Chapter 6. LIMITATION

There are some limitations of this work I like to mention here.

Firstly, I did the studies as a general user of the subreddit, not as the moderator or admin of the platform. That lowers the chance of the response from the user, and the effect on DV, especially in the *acceptance*, *intent*, and *futureCommenting*. The effect of a feedback message is normally higher if the message comes from an admin than a user [21]. My expectation is that if the messages come from an admin, I can see a higher number of responses, and more positive effect on the dependent variables than I found.

Secondly, in this work, I sent a feeedback message once. For better understanding the effect on the dependent variables, specially the *futureCommenting*, a long-term study with iteration of messages is necessary.

Another limitation I like to mention is the limitation of detecting context, or, sarcasm of current ML algorithm. Current ML algorithms are mostly focused on word based detection, context based toxicity detection is yet far away to achieve accurately. As many users also pointed out this restriction, I, as the researcher, also think I need more developed AI/ML tools to help users in identifying the negative content in the comments, which can increase the trust and ultimately, the motivation to behave in civil way form the users.

Chapter 7. CONCLUSION

In this work, I investigated the effect on user's *reaction, acceptance, intent,* and *futureCommenting* behavior after sending feedback messages in online discussion. The result shows promising effect in some aspects, which needs further investigation in this aspect.

Future work includes further iterations on message contents and presentation as well as more long-term studies. Even though the data exhibit promising trends, the dataset is small to draw a meaningful conclusion, long-term study need to be done. Also, the result may vary depending on the platform. I wish to explore how the effect vary depending on the platform environment by doing studies in different platforms.

Bibliography

[1] Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J. and Gilbert, E. *The internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro Scales*, Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), pp.1-25. (2018)

[2] Tiku, Nitasha and Newton, Casey. *Twitter CEO:"We suck at dealing with abuse* Retrieved from https://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the, (February 4, 2015)

[3] Kiesler, S., Kraut, R., Resnick, P., and Kittur, A. *Regulating behavior in online communities. Building Successful Online Communities: Evidence-Based Social Design* MIT Press, Cambridge, MA (2012), 125–178.

[4] Roberts, Sarah T. *Commercial content moderation: Digital laborers' dirty work* (2016).

[5] Roberts, Sarah T. *Behind the screen: The hidden digital labor of commercial content moderation*, Ph.D. Dissertation. University of Illinois at Urbana-Champaign (2014).

[6] Bickert, Monika *Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process*, Retrieved from https://newsroom.fb.com/news/2018/ 04/comprehensive-community-standards/ (Apr. 04, 2018.)

[7] Google. *YouTube Community Guidelines enforcement in Google's Tranparency Report for 2018*, Retrieved from https://transparencyreport.google.com/youtube-policy/removals (2018).

[8] Twitter Public Policy. *Evolving our Twitter Transparency Report: expanded data and insights* Retrieved from https://bit.ly/3gsQIJo (12 December 2018).

[9] Etim, Bassey, *The Times Sharply Increases Articles Open for Comments, Using Google's Technology*, Retrieved from https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html (June 13, 2017)

[10] Jhaver, S., Appling, D.S., Gilbert, E. and Bruckman, A.*Did You Suspect the Post Would be Removed?" Understanding User Reactions to Content Removals on Reddit. Proceedings of the ACM on human-computer interaction*, 3(CSCW), pp.1-33. (2019)

[11] Jhaver, S., Bruckman, A., and Gilbert, E. *Does transparency in moderation really matter? User behavior after content removal explanations on reddit*, Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-27. (2019)

[12] Myers West, S. *Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms*, New Media and Society (2018).

[13] Jhaver, S., Chan, L., and Bruckman, A. *The view from the other side: The border between controversial speech and harassment on Kotaku in Action*, arXiv preprint arXiv:1712.05851. (2017)

[14] Jhaver, S., Ghoshal, S., Bruckman, A., and Gilbert, E. *Online harassment and content moderation: The case of blocklists* ACM Transactions on Computer-Human Interaction (TOCHI), 25(2), 1-33. (2018)

[15] Blackwell, L., Chen, T., Schoenebeck, S., and Lampe, C. *When online harassment is perceived as justified*, In Twelfth International AAAI Conference on Web and Social Media. (2018, June)

[16] Blackwell, L., Handel, M., Roberts, S. T., Bruckman, A., and Voll, K. *Understanding" Bad Actors" Online*, In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-7). (2018, April)

[17] Maréchal, N. *Gabriella Coleman, hacker, hoaxer, whistleblower, spy: The many faces of anonymous* International Journal of Communication, 9, 5. (2015)

[18] Phillips, W. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*, Mit Press. (2015)

[19] Gillespie, T. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press. (2018)

[20] Kluger, A. N., DeNisi, A. *The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory*, Psychological Bulletin, Vol 119(2), Mar 1996, 254-284. (1996)

[21] Seering, Joseph, Robert Kraut, and Laura Dabbish. *Shaping pro and anti-social behavior on twitch through moderation and example-setting*, Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. (2017).

[22] Jhaver, Shagun, et al. *Human-machine collaboration for content regulation: The case of Reddit Automoderator* , ACM Transactions on Computer-Human Interaction (TOCHI) 26.5 (2019): 1-35.

[23] Jacobs, Julia. *Wikipedia Isn't Officially a Social Network. But the Harassment Can Get Ugly* Retrieved from https://www.nytimes.com/2019/04/08/us/wikipedia-harassment-wikimedia-foundation.html (April 8, 2019)

[24] Krishna, Rachael, *Tumblr Launched An Algorithm To Flag Porn And So Far It's Just Caused Chaos*, Retrieved from https://www.buzzfeednews.com/article/krishrach/tumblr-porn-algorithm-ban (December 4, 2018)

[25] Zhu, H., Kraut, R., and Kittur, A. *Effectiveness of shared leadership in online communities*, In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (pp. 407-416). (2012, February).

[26] Chi, M. T., De Leeuw, N., Chiu, M. H., and LaVancher, C. *Eliciting self-explanations improves understanding*, Cognitive science, 18(3), 439-477. (1994)

[27] Kim, Y. S., Reinecke, K., and Hullman, J. *Explaining the gap: Visualizing one's predictions improves recall and comprehension of data*, In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 1375-1386). (2017, May)

[28] Rezvan, M., Shekarpour, S., Balasuriya, L., Thirunarayan, K., Shalin, V. L., and Sheth, A. *A quality type-aware annotated corpus and lexicon for harassment research*, In Proceedings of the 10th ACM Conference on Web Science (pp. 33-36). (2018, May)

[29] Chi, M. *Self-explaining expository texts: The dual processes of generating inferences and repairing mental models*, Advances in instructional psychology 5 : 161-238. (2000)

[30] Facebook Artificial Intelligence *AI advances to better detect hate speech*, Retrieved from https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech (May 12, 2020)

[31] Kizilcec, René F. *How much information? Effects of transparency on trust in an algorithmic interface*, Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. (2016)

[32] Eslami, M., Vaccaro, K., Lee, M. K., Elazari Bar On, A., Gilbert, E., and Karahalios, K. *User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms*, In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-14). (2019, May)

[33] Nyhan, Brendan, and Jason Reifler. *When corrections fail: The persistence of political misperceptions*, Political Behavior 32, no. 2: 303-330. (2010)

[34] Choi, D., Han, J., Chung, T., Ahn, Y. Y., Chun, B. G., and Kwon, T. T. *Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors* In Proceedings of the 2015 acm on conference on online social networks (pp. 233-243). (2015, November).

Acknowledgment

I want to thank my advisors and colleagues, without whom I could not have done this work. First and foremost, my advisor, Juho Kim, who kept me on track and provided a great amount of invaluable feedback and ideas. All my dear colleagues at KIXLAB for providing me with feedback at every stage and the good environment that enabled me to conduct this research.

Tae-Hyeon An, who helped me in the initial prototyping, and formulating the research framing. I am gratefully indebted to him for his valuable comments on this thesis.

Finally, I must express my profound gratitude to my parents and to my spouse, for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Curriculum Vitae

Name : Jibon Naher

Date of Birth : March 07, 1992

Birthplace : Tangail, Bangladesh

Address :

Educations

2010.05.22 – 2015.09.20 B.Sc. in CSE (4 years)

Publications

1. **Naher, J.**, An, T., and Kim, J., *Improving Users' Algorithmic Understandability and Trust in Content Moderation*, CSCW Workshop. (2019, November)
2. Kaniz, S. T., **Naher, J.**, and Hashem, T. *Authentication of k nearest neighbor queries in the presence of obstacles*, In 2017 4th International Conference on Networking, Systems and Security (NSysS) (pp. 1-9). IEEE. (2017, December)