Citadel: An Automated Abuse Detection System to Detect And Prevent Abusive Behaviors over Emails

Ishita Haque
1017052031@grad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Rudaiba Adnin
1505032.ra@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Sadia Afroz
1505030.sa@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Faria Huq
1505052.fh@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Sazan Mahbub
smahbub@umd.edu
The University of Maryland
College Park, USA

Sami Azam
sami.azam@cdu.edu.au
Charles Darwin University
Australia

A. B. M. Alim Al Islam
alim_razi@cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

ABSTRACT

Even though emails are identified as a prominent source of exchanging abusive behaviors, very little work has explored abuse over emails. In our accepted paper in NSysS 2021, we explore perceptions of users on types of abuse detection systems for emails, revealing privacy concerns and lack of control in human-moderator-based systems and a noteworthy demand for an automated system. Motivated by the findings, we iteratively develop an *automated* abuse detection system *"Citadel"* for emails in two sequential phases and evaluate in both phases - first over 39 participants through in-person demonstrations, and second over 21 participants through a 3-day field study and over 63 participants through a video demonstration. Evaluation results portray efficacy, efficiency, and user acceptance of *"Citadel"* in detecting and preventing abusive emails.

CCS CONCEPTS

• **Online abuse** → **Email**; *Moderation*; Privacy; • **Human centered computing** → Usability study.

KEYWORDS

Online Abuse, User study, Automation, Abuse detection

ACM Reference Format:
Ishita Haque, Rudaiba Adnin, Sadia Afroz, Faria Huq, Sazan Mahbub, Sami Azam, and A. B. M. Alim Al Islam. 2022. Citadel: An Automated Abuse Detection System to Detect And Prevent Abusive Behaviors over Emails. In *2022 9th International Conference on Networking, Systems and Security (NSysS 2022), December 20–22, 2022, Cox's Bazar, Bangladesh.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3569551.3569555

NSysS 2022, December 20–22, 2022, Cox's Bazar, Bangladesh

ACM ISBN 978-1-4503-9903-6/22/12…$15.00
https://doi.org/10.1145/3569551.3569555

1 INTRODUCTION

Online abuse is becoming more and more prevalent. According to a recent report, about 47% of internet users reported being victims of some form of online harassment or cyberbullying [33]. Such experiences affect the physical and mental well-being of the victims extremely by causing anxiety, depression, low confidence, self-harm, and even suicide contemplation [33]. Therefore, the detection and moderation of online abuse have gained attention in both research and public discourse. Existing approaches can be automated, human-moderated, or incorporate both. Human moderators regulate the abusive contents and verify those reported by the users [15]. Automated systems incorporate machine learning-based approaches to detect abusive content. Some platforms such as Facebook, Twitter, and Youtube incorporate both types of moderation [15], [32]. However, the human moderators are miserable with overwork, and because of their constant exposure to abusive content, they suffer from severe detrimental effects on mental health [13], [30]. Several platforms have been criticized for their inaccurate detection as well [21], [27], as well as a delayed response against such cases.

Despite the growing interest in abuse detection, prior research has focused on abuse in social platforms [6] and mostly ignored the substantial abuse occurring over emails [28]. Moreover, existing approaches for detecting abusive behavior in emails incorporate human moderators [26], which can impose privacy concerns on the users. In delivering important emails, human moderation can further cause frustration with time delay [26]. Even appointing acquaintances such as friends as moderators can be uncomfortable for the victim and can cause secondary trauma for the moderators.

[5]. In our previous research [18], we explore these gaps by performing an interview study and a survey. Findings show a notable demand for automated systems exists (40% of interview participants, 41% of survey participants). Although a hybrid system was preferred overall, users raised several privacy concerns and felt less in control with systems that include human moderation. These findings inspire us to explore an automated abuse detection system for emails. We also uncovered the user expectations of designing such a system.

Considering the user expectations found from our previous study [18], we extend our research in this paper and explore these research questions from the context of Bangladesh.

- **RQ1:** Between human-moderator-based and automated abuse detection systems for emailing platforms, which is preferred by the users? What are the reasons behind their choices?
- **RQ2:** How do users interact with an automated abuse detection system for emailing platforms?
- **RQ3:** What are the advantages and concerns pertinent to an automated abuse detection system for emailing platforms as per the experience of the users?

These questions inspire us to iteratively design and evaluate *"Citadel"*, an automated abuse detection system for emailing platforms. In the first phase, we evaluate the system with 39 participants through in-person demonstrations. Based on the findings, we modify *"Citadel"* and conduct a second phase evaluation with 21 participants through a 3-day field study and with 63 participants through a video demonstration.

1.1 Contributions

We explore the findings obtained from an extensive literature review and our previous research [18] and extend our research by making the following set of contributions in this paper.

- Following an iterative approach, we develop an automated abuse detection system for emails by integrating features to accommodate the concerns of email users.
- Through a rigorous evaluation phase by both in-person demonstrations and field study, we analyze the challenges and limitations of our research. Moreover, we present a comparative analysis of the results from both evaluations.
- Finally, we make design recommendations to further improve the user experience of an automated system for abuse detection in emails.

2 BACKGROUND, RELATED WORK, AND MOTIVATION

Online abusive behaviors can come in forms such as sexual harassment [12], cyber-bullying [9], flaming [33], threat, etc. A 2017 study by Pew Research reports that about 18% of users have been subjected to threats, harassment, or stalking [12]. Another study reports that more than one-third of Americans reported experiencing some type of severe online harassment [16]. Young adults [7], women [33], and those who identify as LGBTQIA+ [19] are more likely to experience online abuse. A study found that 13% of adults in the United States had experienced mental or emotional stress as a result of online harassment, and 7% reported damage to their reputation [12].

Emails have become an important form of personal and professional communication [22]. In 2019, the number of global email users amounted to 3.9 billion and is set to grow to 4.48 billion users in 2024 [8]. However, emails have become a significant medium of online abuse. A study shows that 16% of participants recall their most recent experience of online harassment in a personal email account [11]. Another study shows that 13% of teenagers said that someone had sent them a threatening or aggressive email, instant message, or text message [25]. Although much research has been done on spam, very few have explored abuse over emails. Very few organizations exists that offer help through such incidents [**?**], [31]. Moreover, internet service providers are more concerned with avoiding spam and unwanted pop-ups, rather than protecting against online harassment [29].

2.1 Related Work

Existing moderation approaches can be broadly categorized as human-moderated and automated. Human-moderated systems can be further divided into centralized and distributed systems [6]. The centralized approach uses paid or unpaid moderators or externally contracted companies by the platform to moderate according to the platform policies [6]. In the distributed approach, users down-vote and report the undesirable content [23]. Reddit, Stack Overflow and Yik Yak use distributed moderation [23]. Automated approaches use machine learning-based models to detect abusive content [6]. A combination of automation and human moderators is used to triage content before moderators review them [6].

Among recent research on abuse over emails, a tool named SquadBox [26] keeps a list of trusted friends, volunteers, or paid moderators. It gives the trusted contacts the authority to read and delete the incoming messages before they can reach the users based on the users' priorities. However, abusive content moderation can potentially become an overload and create risks of secondary trauma for the friends [4]. Therefore, the need for an automated abusive behavior detection system persists. Although features like blocking and reporting allow users to defend themselves to some extent [14], [35], these features do not allow users to prevent abusive behavior beforehand, and the victim has already been impacted.

CrossMod is an automated abusive behavior detection system for Reddit, which assists human moderators in lessening their workload [6]. Reddit depends on human moderators to regulate abusive content. BoC is a cross-platform automated abuse detection system that is built to allow communities to deal with abusive behaviors [7]. Tune is an experimental Chrome extension that detects abusive comments and allows customization of the level of toxicity people want to see in comments across the internet [2].

Numerous platforms depend on human moderators; however, the perception of abuse to a human moderator may differ significantly from the user. Often gap in the perception of abuse between moderator and victim exists because of social and language barriers. Users report privacy concerns with human moderators and prefer that sensitive information such as financial information should not be viewed by moderators, even if they are friends of theirs [26]. For time-sensitive platforms such as emailing platforms, such systems

can create time delays. Due to the low number of moderators and ever-increasing content, each content receives attention for merely a few seconds [17]. Moreover, moderators working on numerous platforms face numerous mental health problems such as self-harm, depression, anxiety, etc. [3].

2.2 Motivation

From prior literature, we uncover the necessity of an abuse detection system for emails. We also find the shortcomings of human-moderator-based systems. To explore these gaps further, in our recent study [18], we conducted an interview study and survey. Our results show that 40% of the participants preferred an automated system, 53% of the participants preferred a combined system of automatic and human moderation, and 7% preferred human moderation. We perform a Chi-squared test that confirms a significant difference between the level of privacy concern in cases using human moderator-based and automated systems. We also find that users want more control over moderation and detection. In the online survey, we explore our research questions further on a wider scale to gain results concerning the themes we developed. Our results show that 51% of the participants preferred a combined system of automatic and human moderation, 41% preferred an automated system, and 8% preferred a human moderator-based system. Similar to our interview results, we find that users raised serious privacy concerns with human moderation.

Insights gathered from our literature study and result from our previous study motivated us to design and develop *"Citadel"*, an automated online abusive behavior detection system for emailing platforms. 93% of participants in the interview study and 89% of participants in the survey study mentioned Gmail as the most used emailing platform. Therefore, we choose to develop *"Citadel"* as a Chrome extension.

3 METHODOLOGY

From the insights obtained from our prior research [18], we design and implement *"Citadel"*, an abuse detection system for the emailing platform Gmail. Later, we evaluate the system by demonstrating the features and workflow to the participants and collecting feedback from them which leads us to modify the system by integrating the findings from their feedback. Finally, we evaluate the system through a 3-days field study as well as through a survey where a video demonstration is shown to the participants. Figure 1 shows the flowchart of the methodology.

3.1 Research Context

Our research context includes email users in Bangladesh. All participants have high computer literacy and use emails on a regular basis. Participants frequently use Gmail as an emailing platform. The target participants are people of all ages. We maintained diversity in age and gender during recruitment. The timeline of evaluation of the final design was between May 2020 to July 2020. Prior research shows that victims of online abuse in South-East Asia are uncomfortable sharing their experiences with abuse and mostly share with their family and friends [33]. Therefore, to gain valuable insights into user interactions with "Citadel," we recruited several of our participants through personal networks.

3.2 Research Ethics

The study and data collection were approved by the Ethics Committee, a part of the Integrity Strategy and Innovation of the institution of the authors. To guarantee the ethical conduction of our research, we ensured the confidentiality and anonymity of the participants in our study. Before recruitment, the participants were notified about the purpose of the study, the data collection process, and the affiliations of the researchers. They were also notified that none of their emails would be stored by us. Before short interviews of chosen participants from the field study, we sought verbal consent from each participant for audio recording and using the provided information for research purposes. Interview participants were also given the freedom to choose between a phone call and one-to-one Zoom meetings. We stored collected data in a private Google Drive accessible to the authors only.

4 SYSTEM DESIGN AND IMPLEMENTATION

"Citadel" is a Chrome extension that can be added to a Gmail account with the user's permission. At each stage of our system design process, we consider the user's needs gathered from the previous study [18].

4.1 Feature Identification

The top three expected features reported by the participants from our previous study were to view abusive emails in a notification window, contact friends after detection, and contact relatives after detection. We name these features as *"Show in Notification Window"* and *"Contact Trusted People"*.

4.2 Implementation

The system comprises two parts - *Backend engine* and *Frontend interface.*

4.2.1 ***Backend Engine.*** The backend engine is built with a deep neural network model to detect abusive emails, with a database to save user details and trusted contact details. With respect to the privacy concern about storing their personal emails by automated systems [18], our system does not store any personal emails of the users.

Dataset Collection

87.1% of participants from the survey and 93.3% of participants from the interviews from our previous study [18] who opted for automated or a combined system of automated and human moderation say they are not willing to allow any third party to store their emails. With correspondence to this, we find no trainable dataset about abusive behavior in emails. For this reason, we use the dataset of competition on toxic comment detection on Kaggle for training our models of abusive text detection [1]. There are six types of indicated toxicity in the original dataset which are 'toxic,' 'severe toxic,' 'obscene,' 'threat,' 'insult,' and 'identity hate.' This dataset has 159,571 data points.

Model Building

We build a deep learning-based architecture that comprises a convolutional module [24], Long short-term memory (LSTM) layers [20], and fully connected feed-forward layers. The convolutional module is inspired by the inception module introduced in [34].

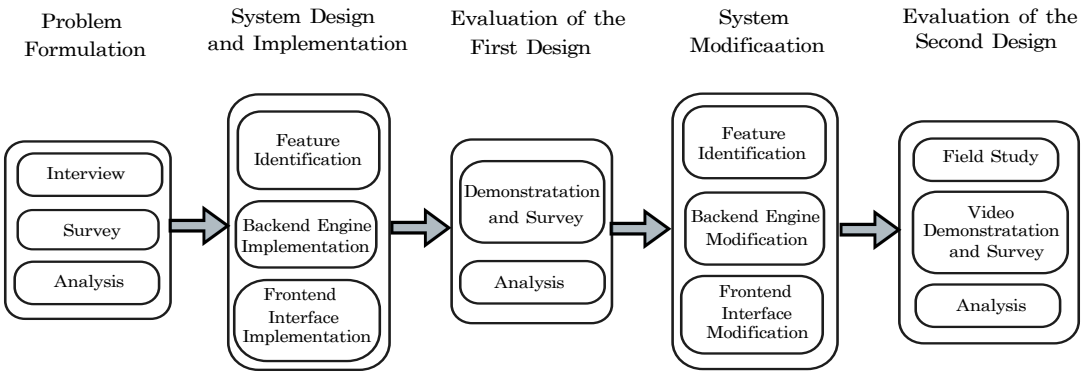

Figure 1: Flow chart of methodology

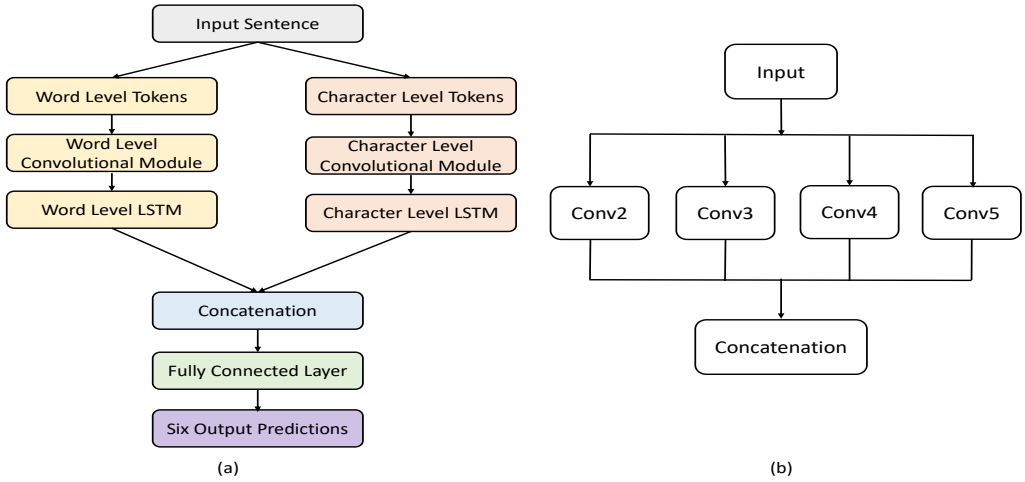

Figure 2: (a) Architecture of our complete model, (b) architecture of our convolutional module (ConvX represents a one-dimensional convolutional layer with filter size X)

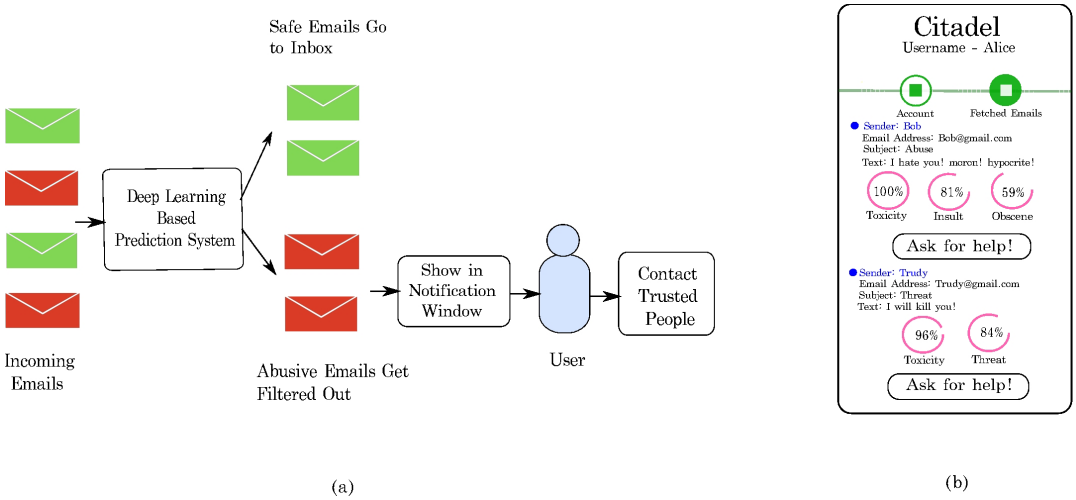

Figure 3: (a) Diagram of the system flow, and (b) notification window of the first design of the abuse detection system

It is useful for extracting relationships among input sentence tokens, and LSTM is useful for capturing long-range dependencies among these tokens. We use both character-level and word-level convolution and the LSTM module so that the word-level branch

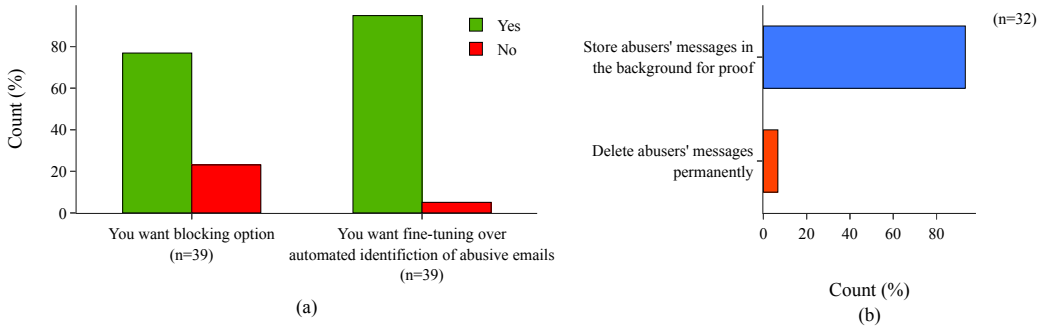

Figure 4: An analysis of the results obtained from the evaluation of the first design. (a) shows an analysis of users' desire for two features, and (b) shows an analysis of users' preferred actions associated with blocking an abuser.

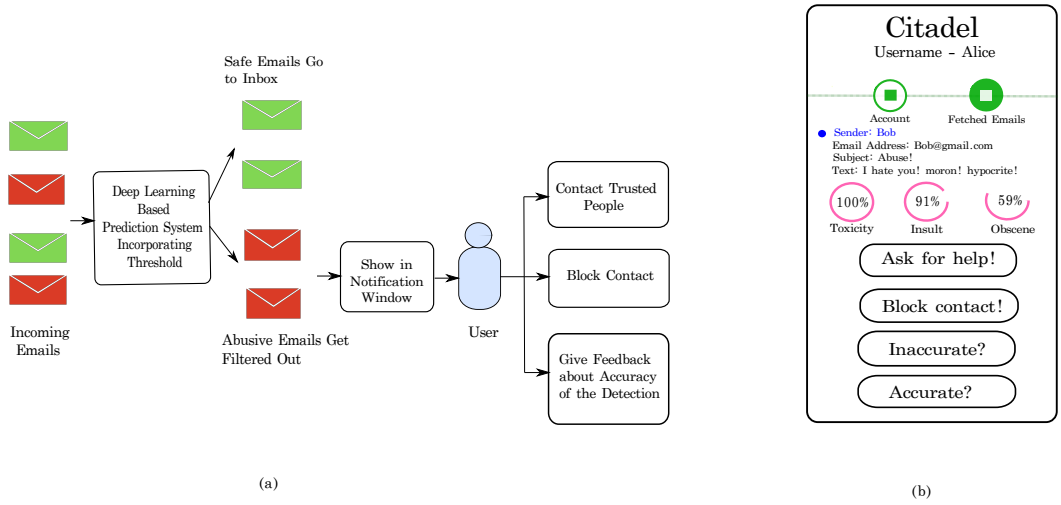

Figure 5: (a) Diagram of the system flow, and (b) notification window of the system of the second design

can capture the context of the sentence very effectively, while the character-level branch helps the model achieve better accuracy in case of intentional or unintentional spelling mistakes [36]. The model achieves 98.4% accuracy on the test dataset that we created by randomly choosing 9,571 data points from a total of 159,571 data points, while the rest 150,000 data points were used for training the model (this research spans over the last two years; hence the slightly lower accuracy compared to current detection algorithms). The full architecture of this deep neural network is shown in Figure 2.

4.2.2 ***Frontend Interface.*** When an abusive email is sent to a user, the system detects the email as abusive and shows it in the notification window along with the severity percentage of the toxicity of the email. Figure 3 (a) shows the system flow. There is an option, *"Ask for help!"* in the notification window users can seek help from their added trusted contacts as we see in Figure 3 (b). When the users use this option, an email with abusive content is sent to the added trusted contacts of the users. Users can add and remove trusted contacts and can edit their profiles in the profile management window.

5 EVALUATION OF THE FIRST DESIGN

We demonstrate our automated abuse detection system, *"Citadel"* to 39 participants (22 males and 17 females). They range in age from 20 years to 40 years. The demonstrations are 20 minutes to 30 minutes long. After the demonstration, participants are asked to fill up a questionnaire and give their feedback. As we see from Figure 4, 75.6% of the participants express their desire for a blocking option. The type of blocking option most participants (about 73.8%) want is where the abusive emails will be stored in the background as a means of proof. Around 89.74% of participants are interested in having auto-tuning incorporated into this detection system through which they can specify the system about their known acquaintance.

Users express their interest in using the system and optimism about the usability of our system. They also feel that there should be a strategy to understand the difference between a known acquaintance and a stranger.

6 SYSTEM MODIFICATION

After the evaluation of the first design, we identify new features and modify our system to integrate those.

6.1 Feature Identification

As observed from the evaluation of the first design, we discover that people often do not consider a certain content abusive when it comes to a close acquaintance, while the same content is abusive when it is sent by a stranger. According to one participant in our evaluation of the first design, *"While chatting on social media, my friends and I often use words that we do not consider to be offensive, and I am always okay with this. But I am never okay with any stranger using the same words while texting me."* Existing systems integrate white-list and blacklists [26], but creating blacklists is not feasible because the abuse has already taken place before the abuser is added to the blacklist. Moreover, whitelists and blacklists create a burden on the user to regularly update it.

These findings inspire us of a new feature called *"Threshold per sender"*, where users can specify the system if the detection of abusive emails is agreeable to them or not.

Another feature that we identify strongly with the participants is *"Block Contact"*. When an abusive email reaches a user, the notification window shows it to the user. The blocking option enables the users to directly block the sender all future emails sent by the abuser go to the trash box.

6.2 Implementation

To incorporate the new features, we modify the backend engine and frontend interface of our system.

6.2.1 ***Backend engine.*** If the user chooses that the prediction is not right, the system will automatically increase the threshold of that specific sender, and it will not be sent to the notification window of the user. If the user considers the prediction to be right, the system will do just the opposite but will not decrease the value to less than a specific constant because, according to our design, that is the lowest possible threshold for any type of abusive email. Some modifications in the system's database are done to add specified senders' thresholds. The system shows the abusive email in the notification window of the user if the detection's toxicity scores exceed the senders' threshold; otherwise, it does not. Some changes in the database of the system are done to keep block lists of the users as well.

6.2.2 ***Frontend interface.*** Users have the options *"Accurate?"* and *"Inaccurate?"* in the notification window. The system accordingly sets the threshold of detection of abusive emails for that specific sender. The notification window has a *"Block Contact!"* option for the users as we can see from Figure 5 (b). Users can see their block list in a window of the system as well. Users can remove and manually add blocked contacts as well. The emails sent by blocked contacts go to the trash box, and users do not see any emails from those blocked contacts in their inboxes.

7 EVALUATION OF THE SECOND DESIGN

We evaluate our modified system after integrating the features of the second design with the participants in two steps. We conduct a 3-day field study with 21 participants as well as a survey showing a video demonstration of our system *"Citadel"* with 63 participants.

7.1 Field Study

Due to the sensitive nature of online abusive behaviors and the vulnerable position of the recipients, we are cautious about conducting a field study. We conduct a 3-day field study with 21 participants aged between 15 to 50 years, 10 males and 11 females. They are recruited by non-probabilistic sampling and snowball sampling. The participants are required to use Gmail.

At first, we give the participants detailed documentation on how to install and use the system. We ask the participants to add the extension to their Chrome browsers and use it for 3 days with their personal email accounts. We observe the usage pattern through the server-side logs. After the field study, we provide a google form with both open-ended and closed-ended questions to the participants to share their experiences of those 3 days. The questionnaire consists of seven categorized groups of 5-point Likert scale questions to find out users' desire for an abuse detection system, desire for an *automated* abuse detection system, features they approve, ease of use, ease of learning, and satisfaction after using the system.

We randomly picked 5 participants for an additional short interview of 5-10 minutes over Zoom meetings. We ask a few additional in-depth questions about their experiences with errors in our system and their overall impression of our system.

7.2 Video Demonstration and Survey

Due to the current situation with COVID-19, we are unable to conduct usability testing in controlled settings. We conduct an online survey showing the participants a video demonstration of the workflow of the system. 63 participants aged between 15 to 50 years (32 males and 31 females) are selected by non-probabilistic sampling and snowball sampling.

The participants are asked to watch the video and then fill up a questionnaire containing both open and closed-ended questions. It consists of seven categorized groups of 5-point Likert scale questions, and the questionnaire for the video demonstration survey consists of six categorized groups of 5-point Likert scale questions to find out users' desire for an abuse detection system, desire for an *automated* abuse detection system, features they approve, ease of use, ease of learning, and satisfaction, after watching the system's workflow in case of the video demonstration survey.

8 FINDINGS FROM THE EVALUATION OF SECOND DESIGN

We present our findings from the field study and video demonstration and provide a comparative analysis.

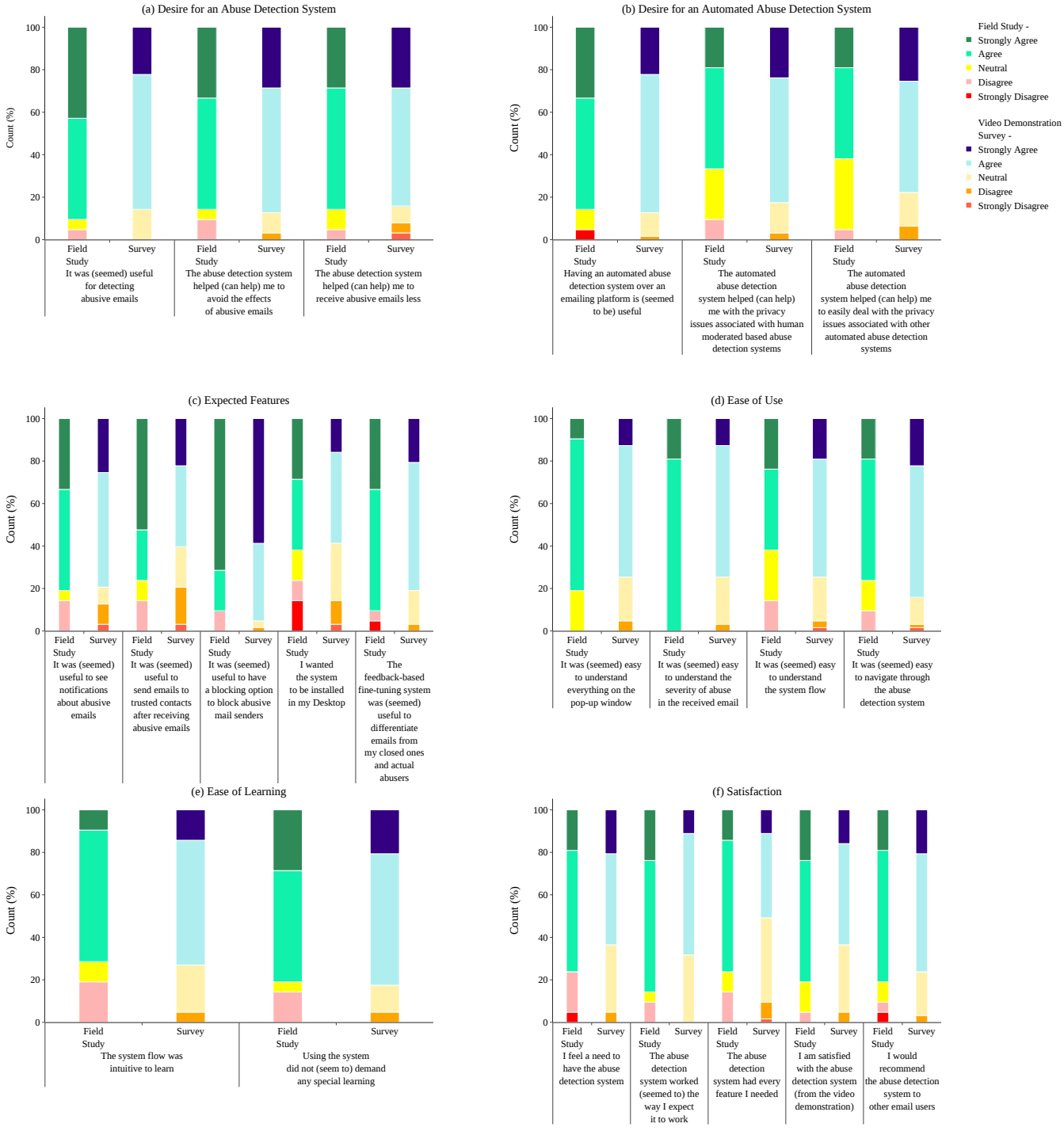

Figure 6: Participants' (a) desire for an abuse detection system, (b) desire for an automated abuse detection system, (c) expected features, (d) ease of use, (e) ease of learning, (f) satisfaction with the system in the field study (n=21) and the video demonstration survey (n=63) in the evaluation of the second design (Response scale: 1=Strongly Disagree, 5=Strongly Agree)

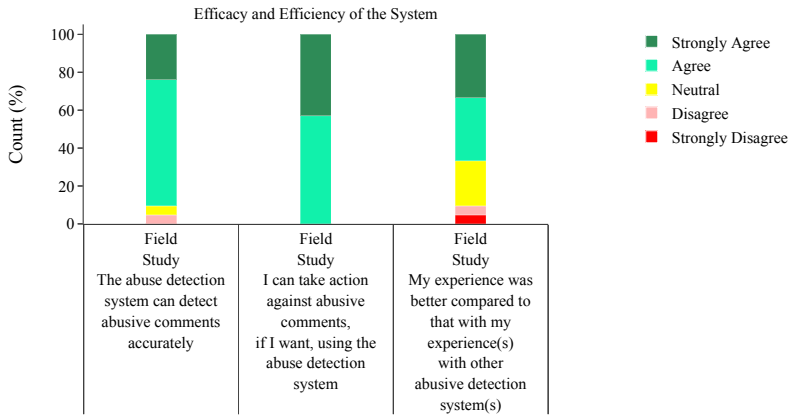

Figure 7: Participants' perceived efficacy and efficiency of the system in the field study (n=21) in the evaluation of the second design (Response scale: 1=Strongly Disagree, 5=Strongly Agree)

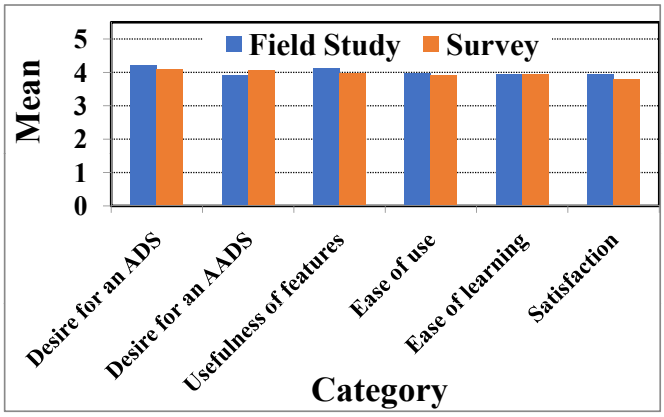

Figure 8: Comparative analysis over averages of user ratings in different categories used to rate user experiences in field study and survey through in-person demonstrations ("ADS" refers to "Abuse Detection System" and "AADS" refers to "Automated Abuse Detection System")

8.1 Findings from Field Study

Among all, 11 participants use *"Fetched Emails"* section more than 7 times for checking the abusive emails, which indicates their curiosity about this feature and urge to have something like this: *"I am totally satisfied. It detects so fast and shows the toxicity level. I did not face any bugs or glitches at all. Overall it was very easy to use and efficient as well"*.

"Threshold per Sender", the feature implemented for fine-tuning over automated identification of abusive emails. After the first evaluation, it is liked by most of our participants in the field study. 7 participants say they like the *"Block Contact"* feature the most out of the 21 participants. Other liked features include *"Contact Trusted People"*, the percentage of the toxicity of abusive emails shown in the notification window.

From the short interview conducted after the online survey of the field study, most participants were impressed with the efficiency in the detection of abusive emails. However, one participant says, "I didn't see the system detect any non-abusive email as abusive. But sometimes, I found that it was not able to detect contextual abuse. I mean, if the abuse wasn't very straightforward, it would fail sometimes." All of the 5 participants interviewed expressed their preference for showing toxicity ratings for abusive emails. One participant also suggested that toxicity rating may be helpful in the "Ask for help" feature or asking for legal help if necessary.

Figure 6 and Figure 7 show the mean values of the six categorized groups of questions in the field study and survey. In the category efficacy and efficiency of the automated abuse detection system, the mean is *4.14* which means most participants in the field study perceive the system as efficient.

8.2 Findings from Video Demonstration and Survey

Analyzing the video demonstrated survey results, we find that among all the features, the *"Block Contact!"* feature seems most likable to the participants. They express their desire for a similar type system in social media platforms as well: *"If can be extended to popular social media platforms, that will be really useful"*. Nearly all the participants show interest in using this system.

Analyzing results from both studies, we find that the field study result surpasses the survey by video demonstration result, as shown in Figure 8. Although the survey is done in a broader sector (63 participants), lack of interaction with the system makes them give less legit feedback compared to field study participants since they are not only getting familiarized with the system but also have hands-on experience. In the categories desire for an abuse detection system, desire for an automated abuse detection system, and expected features get a more positive vibe from all the participants of the field study and the video demonstration survey.

9 DISCUSSION

While conducting the whole study, we come up with the following research findings.

9.1 User's Appreciation for an Abuse Detection System for Emails

From our user studies, we confirm that email is a significant source of online abuse. This is reflected through the responses of the users in the in-person demonstration, field study, and video demonstration-based survey. Our detection system was very well received among the users, who acknowledged its need even for teenagers and children.

9.2 User's Appreciation to Have More Control

From our previous study, we find that users want to have control of abuse detection and moderation in emailing platforms. Rather than having complete filtering of abusive emails by another human, the users prefer to have control of their own in the abuse detection and prevention system. The control can be achieved by enabling users to block abusive emails and letting them ask for help in case of receiving abusive emails. Besides, the users can also be given control by ensuring that abusive emails are brought to their attention rather than removing the abusive emails without their oversight. Our evaluation studies confirm that these types of user control are desirable to the users.

9.3 Design Implementation Guided by User Preferences

We learn that designing an efficient automated abuse detection system is not possible without incorporating feedback from the users. Therefore, following an iterative approach, we gathered users' requirements for an automated abuse detection system. Guided by these preferences, we introduced two features in our study - 1) introducing user control by enabling a threshold for filtering abusive content, and 2) enabling blocking contacts of abusers. We further evaluate the desirability and usefulness of these features through rigorous user studies.

9.4 User's Appreciation for Abuse Detection Systems in Other Platforms

Our results show that users have faced abuse on other online platforms as well, and therefore, are also very eager to use abuse detection systems on those platforms. Inspired by these results, we look forward to expanding the capabilities of *Citadel* to other platforms as well in the future.

10 LIMITATIONS

Our system currently works only on English literature. The system may not work as perfectly as it does now for other languages. Since all our participants are from South-East Asia origin, our study is limited in ensuring the diversity of the participants. Most of the participants are mostly from researchers' primary and secondary networks since they were recruited through Snowball sampling. Therefore, our work is not free from participation bias and selection bias. Thus, arguments and opinions driven by the interviews and surveys may not represent the collective view of the people of the whole nation. As our system is not deployed globally yet, we do our field study using our local server, and the participants face slow responses while using the system.

11 SCOPE OF FURTHER RESEARCH

In the future, we plan to expand the scale of our research by incorporating more participants in our user studies, including participants from other countries, and exploring the differences in their perceptions. From our user studies, we come across numerous expected features from an automated abuse detection system. One explorable research directions include developing a feature where users can add specific words that they find abusive in a customizable list. We want to explore showing only the toxicity ratings to the users instead of showing the whole content and allowing users to have the option to view more or delete directly. We plan to develop a dashboard showing statistics of abusive and non-abusive emails. Inspired by the results of our user study, we look forward to exploring abuse on other online platforms as well. One possible solution may be to develop a single API for accessing messaging platforms. Such an API will provide support across multiple platforms and also help developers of small-scale platforms. In the future, we want to explore the feasibility and risks associated with creating a single API to detect abuse.

12 CONCLUSION

Abuse detection over emails is little explored in the literature. Moreover, automated abuse detection over emails is yet to be explored in the literature to the best of our knowledge. In this work, we explore automated abuse detection in emailing platforms. Incorporating feedback, desires, and expectations obtained from the users' perspectives as per our previous study, we iteratively design and develop *"Citadel"*, a new automated abuse detection system for emails. We evaluate our system in two phases of our design process - first, through in-person demonstration and second, through a 3-day field study as well as through a survey conducted based on

a video demonstration. Results confirm the efficiency and efficacy of our system as well as its user acceptability in real usage.

REFERENCES

[1] 2017. Toxic Comment Classification Challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

[2] 2019. Tune (experimental). https://chrome.google.com/webstore/detail/tune-experimental/gdfknffdmmjakmlikbpdngpcpbbfhbnp?hl=en/.

[3] Daniel Etcovitch Andrew Arsht. 2018. The Human Cost of Online Content Moderation. https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation/.

[4] Jennifer Beckett. 2018. We need to talk about the mental health of content moderators. https://theconversation.com/we-need-to-talk-about-the-mental-health-of-content-moderators-103830.

[5] Michael L. Bourke and Sarah W. Craun. 2014. Secondary Traumatic Stress Among Internet Crimes Against Children Task Force Personnel: Impact, Risk Factors, and Coping Strategies. *Sexual Abuse* 26, 6 (2014), 586–609. https://doi.org/10.1177/1079063213509411 arXiv:https://doi.org/10.1177/1079063213509411 PMID: 24259539.

[6] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[7] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3175–3187.

[8] J. Clement. 2020. Number of e-mail users worldwide 2017-2024. https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/#statisticContainer.

[9] M. Dadvar and Franciska M.G. de Jong. 2012. Cyberbullying Detection; A Step Toward a Safer Internet Yard. In *Proceedings of the 21st International World Wide Web Conference, WWW 2012 - PhD-Symposium*. Association for Computing Machinery (ACM), United States, 121–125. https://doi.org/10.1145/2187980.2187995 null ; Conference date: 16-04-2012 Through 20-04-2012.

[10]]DonotPay DonotPay. [n. d.]. A Guide on Reporting Email Abuse. https://donotpay.com/learn/reporting-email-abuse/.

[11] Maeve Duggan. 2014. Online Harassment. https://www.pewresearch.org/internet/2007/06/27/cyberbullying/.

[12] Maeve Duggan. 2017. Online Harassment. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/.

[13] Jeanne Whalen Elizabeth Dwoskin and Regine Cabato. 2019. Content moderators at YouTube, Facebook and Twitter see the worst of the web — and suffer silently. https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/.

[14] Facebook. 2020. Facebook Report something. https://www.facebook.com/help/263149623790594/.

[15] FaceBook. 2020. Facebook Terms and Policies. https://www.facebook.com/communitystandards/.

[16] ALYSSA FOOTE. 2019. 1 in 3 Americans Suffered Severe Online Harassment in 2018. https://www.wired.com/story/severe-online-harassment-2018-adl-survey/.

[17] Tarleton Gillespie. 2018. *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. 1–288 pages.

[18] Ishita Haque, Rudaiba Adnin, Sadia Afroz, Faria Huq, Sazan Mahbub, and A. B. M. Alim Al Islam. 2021. *"A Tale on Abuse and Its Detection over Online Platforms, Especially over Emails": From the Context of Bangladesh*. Association for Computing Machinery, New York, NY, USA, 19–28. https://doi.org/10.1145/3491371.3491374

[19] Kayla Hendricks, Pitso Tsibolane, and Jean-Paul Van Belle. 2020. Cyber-Harassment Victimization Among South African LGBTQIA+ Youth. In *Conference on e-Business, e-Services and e-Society*. Springer, 135–146.

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[21] Rachael Krishna. 2018. Tumblr Launched An Algorithm To Flag Porn And So Far It's Just Caused Chaos. https://www.buzzfeednews.com/article/krishrach/tumblr-porn-algorithm-ban/.

[22] Joanne Kuzma. 2013. Empirical study of cyber harassment among social networks. *International Journal of Technology and Human Interaction (IJTHI)* 9, 2 (2013), 53–65.

[23] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[25] AMANDA LENHART. 2007. Cyberbullying. https://www.pewresearch.org/internet/2007/06/27/cyberbullying/.

[26] Kaitlin Mahar, Amy X Zhang, and David Karger. 2018. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[27] Casey Newton Nitasha Tiku. 2015. Twitter CEO: 'We suck at dealing with abuse. https://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the.

[28] Karen L Paullet, Daniel R Rota, and Thomas T Swan. 2009. Cyberstalking: An exploratory study of students at a mid-Atlantic university. *Issues in Information Systems* 10, 2 (2009), 640–649.

[29] Michael Pittaro. 2007. Cyber stalking: An Analysis of Online Harassment and Intimidation. 1 (01 2007). https://doi.org/10.5281/zenodo.18794

[30] Benjamin Plakett. 2018. Unpaid and abused: Moderators speak out against Reddi. https://www.engadget.com/2018-08-31-reddit-moderators-speak-out.html?guccounter=1/.

[31] New Zealand Police. 2020. Someone has been sending me offensive emails or threatening / harassing me over the internet. What should I do? https://www.police.govt.nz/faq/someone-has-been-sending-me-offensive-emails-or-threatening--harassing-me-over-the-internet.-what-should-i-do.

[32] Twitter Public Policy. 2018. Evolving our Twitter Transparency Report: expanded data and insights. https://blog.twitter.com/official/en_us/topics/company/2018/evolving-our-twitter-transparency-report.html.

[33] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. 2019. T̈hey Don't Leave Us Alone Anywhere We GoG̈ender and Digital Abuse in South Asia. In *proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper With Convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Inc Twitter. 2020. Report abusive behavio. https://help.twitter.com/en/safety-and-security/report-abusive-behavior.

[36] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*. 649–657.