



Research Article

Cyberhate aware social media: A software engineering perspective

Malik ALMALIKI^{1,*}

¹College of Computer Science and Engineering, Taibah University, Yanbu, 966144, Saudi Arabia

ARTICLE INFO

Article history

Received: 04 November 2024

Revised: 08 January 2025

Accepted: 28 April 2025

Keywords:

Cyberhate; Persuasive
Technology; Social Media;
Software Engineering

ABSTRACT

Incidents of cyberhate directed at individuals or groups based on perceived identities, such as ethnicity, religion, or nationality, are escalating rapidly across social media platforms. This trend adversely affects both the quality of online content and the experiences of users. Fortunately, researchers are showing a growing interest in combating cyberhate by developing algorithms capable of detecting such incidents. However, these studies have largely overlooked the promotion of healthier user behavior to actively challenge cyberhate. This paper advocates for the utilization of persuasive technology to encourage users to adopt prosocial roles in confronting and mitigating the spread of cyberhate. An empirical study was conducted to explore users' perceptions of adopting various persuasive techniques on social media aimed at motivating them to confront cyberhate. The results revealed variations in users' preferences and indicated that users' cultural backgrounds play a role in shaping their opinions regarding the adoption of specific persuasive strategies for combating cyberhate. This emphasizes the need for systematic and innovative approaches to incorporate persuasive technology into the design of social media platforms to effectively counter the spread of cyberhate. Building upon the results, the author devised a conceptual framework that can offer guidance to software engineers in designing persuasive cyberhate-aware social media platforms. The study innovates by empirically evaluating persuasive technologies against cyberhate, quantifying their effectiveness and offering actionable recommendations for stakeholders.

Cite this article as: Almaliki M. Cyberhate aware social media: A software engineering perspective. Sigma J Eng Nat Sci 2026;44(3):2195–2206.

INTRODUCTION

Social networking sites such as Facebook, X (formerly known as Twitter), and WhatsApp are widely used platforms for creating and sharing content. They offer a convenient and efficient way for users to quickly transmit information, making them a valuable source of information [1-4]. However, these platforms can also be utilized as

channels for distributing harmful and offensive material, including cyberhate, misinformation, and cyberbullying. The presence of such content on social media can have detrimental effects on the quality of the shared content and the experiences of users [5]. Additionally, it can potentially lead to criminal activities or violence against others.

Cyberhate, also known as online hate speech, refers to the expression of hate or incitement to hatred against

*Corresponding author.

*E-mail address: mrmalki@taibahu.edu.sa

This paper was recommended for publication in revised form by
Editor-in-Chief Ahmet Selim Dalkilic



individuals or groups based on their perceived characteristics, such as ethnicity, religion, nationality, or sexual orientation [6-8].

It has become rampant and rapidly proliferates across social media networks. While there is currently no universally agreed-upon international definition or legal framework for cyberhate, it has garnered significant attention from researchers, technology companies, social media platforms, and programmers over the past decade as they strive to address this problem.

According to a recent report by the anti-defamation league (ADL), 41% of Americans have experienced cyberhate and harassment [9]. Figure 1 illustrates the percentage of cyberhate incidents reported by respondents based on their perceived identities. Similarly, the European Union Kids Online Survey conducted between 2017 and 2019 revealed an increase in reports of cyberbullying and cyberhate [10]. In response to this issue, the European Commission and Information Technology (IT) companies introduced the Code of Conduct 1 in 2016. This code included commitments from companies like Facebook, Twitter, YouTube, and Microsoft to combat the spread of hate speech online in Europe.

As a result, many IT companies have pledged their efforts to tackle cyberhate. Moreover, certain countries have implemented national laws, including criminal sanctions, to prevent and combat cyberhate [11]. Despite these efforts, the problem of cyberhate continues to escalate [12]. Users of social media platforms often employ acronyms and common phrases that may not appear hateful on the surface but are used to express implicit hatred. This poses a significant challenge in addressing the issue effectively.

Recently, a number of studies conducted in the fields of computer science and information systems have delved into cyberhate spread, aiming to develop machine learning models capable of detecting instances of cyberhate on social media platforms [13-15]. Nevertheless, concerns have been raised regarding the applicability of current models [16].

Although these models exhibit impressive performance on their respective test datasets, their effectiveness noticeably declines when applied to a different set of hate speech data [17]. This lack of generalizability undermines the practical utility of these cyberhate detection models [18].

Several additional studies have focused on investigating the dynamics and flow patterns of cyberhate on social media platforms [19,20], while a limited number of studies have put forth approaches to prevent the occurrence and spread of cyberhate in these environments [11]. Furthermore, some research was conducted to gain insight into how individuals receive and propagate cyberhate on social media, along with exploring their attitudes and behaviors in relation to this issue [20,21]. Nevertheless, none of these studies specifically focus on the promotion of healthier behavior among social media users as a mean to counteract the dissemination of cyberhate [22,23].

Fortunately, software can serve as a medium for implementing strategies aimed at encouraging behavioral changes among users, thereby persuading bystanders to actively participate in combating cyberhate on social media platforms. According to [24,25], uninvolved bystanders, who are often the largest group of witnesses in incidents involving hateful speech, play a crucial role because their non-involvement could be interpreted as agreement with such statements. Therefore, it is vital to encourage bystander intervention in addressing online hate speech, as it is believed to help mitigate the negative consequences faced by targeted groups [26]. This is particularly true for counter arguing in publicly visible user comments [27].

Counter-speech, described as a collective response to extremism or hateful content, is widely recognized as a crucial form of intervention in combating hate speech [27]. Though it may not always succeed in persuading or silencing those promoting hate, it has the potential to shape discourse norms and demonstrate to uninvolved witnesses that hate speech does not reflect the majority opinion, thereby encouraging them to intervene as well. Additionally, it can

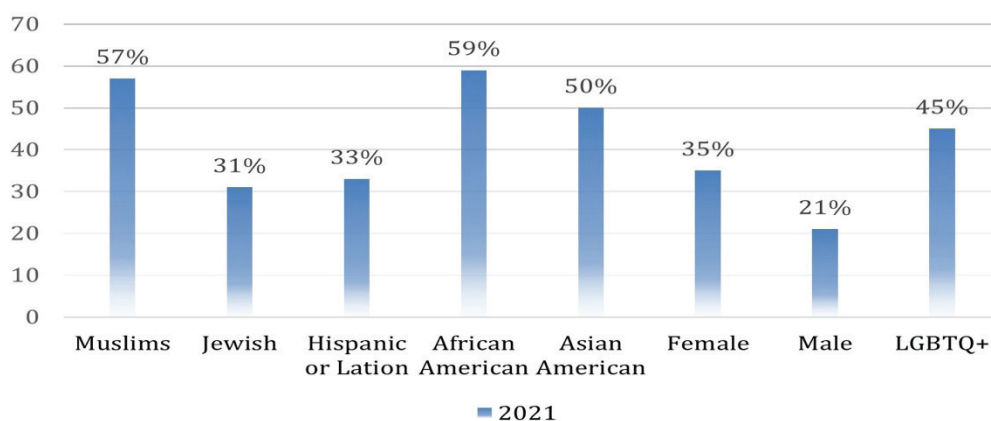


Figure 1. Reported cyberhate and harassment instances based on respondents' identities [9] [Open access].

lead to the exchange of diverse perspectives on a broader scale, which is believed to be highly effective in combating hate speech [28].

The bystander intervention model (BIM), introduced by [29], was developed to provide an explanation for why bystanders often fail to intervene in emergency situations, including instances of hate speech. According to the BIM, uninvolved bystanders need to go through three essential steps in order to intervene effectively. Firstly, they must assess the situation as being threatening and recognize it as an emergency for the individuals affected. Secondly, they need to perceive a sense of personal responsibility for taking action and intervening. Finally, they must make the conscious decision to intervene and follow through with appropriate actions. That is, if any of these steps are not adequately fulfilled, bystanders are less likely to intervene.

In order to encourage bystanders to take action against cyberhate, it is valuable to develop social media platforms that are more effective in persuading and motivating bystander intervention. Within the digital realm, information systems can be purposefully designed to influence behavioral changes [30]. These systems are often referred to as persuasive systems [31] or persuasive technology (PT) [32]. Effective application of PT is growing in various domains such as health, education and wellness [33–36]. Additionally, research has demonstrated that incorporating persuasive techniques into social media platforms can result in increased behavior change and higher levels of engagement [37,38]. Therefore, applying such techniques in motivating bystanders to speak up against cyberhate can contribute to reducing the prevalence and impact of this issue. This can ultimately reduce the detrimental impact of cyberhate on users and enhance the overall user experience on social media platforms.

However, how can software engineers employ persuasive techniques to motivate bystanders to counter-speech cyberhate on social media platforms? What are users' opinions, dislikes and likes regarding the use of such techniques? Do cultural backgrounds of users impact the design and selection of persuasive techniques? There is a research gap in understanding how to effectively utilize persuasive techniques to encourage bystander intervention and mitigate the dissemination of cyberhate on social media platforms, all while maintaining a positive user experience and promoting higher quality online content. This emphasizes the importance of studying and exploring, from the perspective of users, how persuasive techniques can be systematically incorporated in the design of social media to address cyberhate.

In this study, the author empirically investigates the feasibility of integrating persuasive techniques into the design process of social media platforms with the goal of empowering bystanders to challenge the spread of cyberhate. The research explores users' diverse needs, cultural backgrounds effect, and preferences regarding the application of several persuasive techniques in combating cyberhate. Based on

the findings, a conceptual framework is developed to guide software engineers in designing persuasive social media platforms that are cyberhate aware. Ultimately, this framework aims to enhance users' behavior towards cyberhate on social media platforms, thereby improving their overall online experience.

The novelty of the study lies in its empirical evaluation of persuasive technologies designed to combat cyberhate, an area previously underexplored. It goes beyond existing literature by quantifying the effectiveness of these interventions, analyzing user engagement across various demographics and cultural backgrounds, and providing actionable recommendations for stakeholders, thereby offering a conceptual framework for designing cyberhate-aware social media platforms.

MATERIALS AND METHODS

This paper presents an empirical investigation into the potential integration of persuasive techniques within the design process of social media platforms, with the goal of motivating bystanders to actively combat the propagation of cyberhate. The study utilizes BIM model as a foundation for selecting six persuasive techniques that can be aligned with the three steps of the BIM model, thereby encouraging bystanders to challenge cyberhate. The selection process takes into consideration the practicality and feasibility of incorporating these techniques into social media design without imposing significant costs. Detailed descriptions of the techniques employed in this study, adapted from [37], are provided in Table 1. Subsequently, the author designed multiple interfaces that mimic the current user interface of X (formerly known as Twitter) to demonstrate potential integration of each technique within the context of this research. Lastly, a survey is conducted to evaluate users' perceptions regarding the impact of each integrated persuasive technique on their willingness to confront cyberhate.

The designed interfaces included a fabricated news post regarding refugees' unemployment and a cyberhate comment as a response to the post. The interfaces were presented to participants as a scenario and the participants were explicitly informed that the news post and comments were not real, but rather designed to evaluate their perceptions of integrated persuasive techniques. To enhance the realism of the scenario and avoid participants perceiving the accounts as anonymous, the name and photo of the news account and commenters' accounts were intentionally left unaltered. The following is a breakdown of each interface used and how it exemplifies the principles of persuasive design:

Private Reply (Reduction)

A private mode of communication such as direct messaging (DM) or private messaging (PM) can be preferred by some bystanders as a way of reminding the aggressor of netiquette rules in a more private manner without

Table 1. BIM steps mapping to adopted persuasive techniques and the corresponding implementations

Matching	Technique	Description	Our Implementation
Steps of BIM			
Step 3	Reduction	A technique of breaking down complex behaviors into manageable and easy tasks.	Private Reply (PR): This feature simplifies private messaging for bystanders by eliminating the need to copy the post link and paste it in the private message section. Bystanders can directly comment privately on the post by clicking the “reply privately” button.
Step 1	Suggestion	A technique that provides appropriate recommendations	Facts Popups (FP): pop up messages that give facts about cyberhate negative impact to encourage bystanders to take prosocial actions against it.
Step 3	Self-monitoring	A technique that facilitates self-monitoring of one’s status or advancement.	Comment Analyzer (CA): It is an indicator that allows users to visualize how their comments may be perceived by someone reading them. As users compose their comments, the indicator on the scale dynamically adjusts based on word choices, writing style, and punctuation, distinguishing between tones such as aggressive and friendly.
Steps 2-3	Recognition	A technique that offers public recognition for adhering to specific behavior.	ProSocial Badge (PB): It is a digital recognition symbol awarded to users who occasionally take a stand against cyberhate.
Steps 1-2	Normative influence	A technique that demonstrates social norms by illustrating typical behavior and approved conduct.	Social Expectation (SE): Pop-up reminder messages of what others would expect from bystanders when they are victims of cyberhate.
Steps 2-3	Competition	A technique to incentives users to embrace a desired attitude or behavior by harnessing humans’ inherent tendency for competition.	First-Second-Third place Trophies (FT): Digital trophies are given to bystanders based on their order of intervention to stop cyberhate.

causing unnecessary defensiveness [39,40]. While Twitter does offer a private communication channel, the concept of private replying is suggested as a mean to streamline the complex process. Engaging in private messaging to combat cyberhate requires more effort by bystanders, whereas private replying enables bystanders to challenge cyberhate posts effortlessly as they browse through their feeds.

Facts Popups (Suggestion)

By presenting to bystanders some facts about the negative impact of cyberhate, we mean to assist bystanders who find it difficult to start challenging cyberhate. Presenting facts about the negative impact of cyberhate can potentially raise awareness and increase users’ understanding of the issue. By highlighting the harmful consequences of cyberhate, individuals may be more inclined to take prosocial actions against it (e.g., supporting victims) [41,42].

Comment Analyzer (Self-monitoring)

Self-monitoring is recognized as an effective persuasive strategy [43], as it aids individuals in gaining a deeper understanding of their behavioral patterns [44] and facilitates self-regulation [45]. In this context, the implementation of a comment analyzer can play a valuable role by assisting bystanders in enhancing their self-awareness

regarding the tone and impression of their comments to challenge cyberhate.

ProSocial Badge (Recognition)

Alongside offering public recognition, digital badges can serve as motivating incentives for bystanders who actively address instances of cyberhate. Social networking sites such as Foursquare or StackOverflow have successfully employed badges as a mechanism to stimulate users and encourage them to enhance their level of participation [46].

Social Expectation (Normative Influence)

Normative social influence can effectively be harnessed in persuasive technologies to modify existing behaviors [47]. Social norms pertain to individuals’ beliefs about the attitudes and actions of others within a social group [48]. When individuals perceive a behavior to be widely accepted by others, they are more likely to engage in that behavior [49]. In light of this, we proposed the use of pop-up reminder messages that convey information about injunctive norms, which reflect the approval of the majority of people.

First-Second-Third place Trophies (Competition)

The utilization of competition as a persuasive strategy aims to encourage individuals to adopt a specific behavior by benefiting from their inherent inclination to compete

[39]. In the context of this study, bystanders compete against each other to act quickly and win a digital trophy based on the order of their involvement to challenge cyberhate.

Survey Design

The questionnaire developed for this study comprised 26 questions of various types, including single-choice, multiple-choice, and Likert scale questions. Initially, the questionnaire was tested on a small sample of five respondents who met the sampling criteria established for this study. The feedback and responses obtained from these respondents were then utilized to enhance and refine the questionnaire before it was distributed to a larger sample of respondents. The participants were recruited to participate in the study through ClickWorkers (www.clickworkers.com), an online platform specializing in the recruitment of research participants. As an incentive, a payment of one Euro was offered to each participant upon completing the survey. To ensure the participants' familiarity with the subject matter, a summary of the study's topic was presented at the beginning of the questionnaire itself. The respondents were also provided with a briefing on the expectations from their participation and how their contributions would be utilized. A copy of the questionnaire submitted to the participants can be found at <https://shorturl.at/gipDT>.

Sampling

A convenience sampling method was utilized to select 113 participants who were 18 years of age or older and represented various cultural backgrounds. The participants' characteristics were analyzed and summarized using

cross tabulation and frequency counts. A comparison was made among the respondents based on variables such as age, gender, nationality, and educational level. The results of this analysis indicate a high level of diversity among the participants, which enhances the potential generalizability of the findings of this study, aligning with our objective of promoting a culturally adaptable approach to combating cyberhate. Table 2 presents an overview of the respondents' characteristics.

Analysis

Initially, the questionnaire responses underwent a process of preparation and refinement to ensure suitability for analysis. Any responses that were unrelated or inconsistent were eliminated, specifically, 21 forms that were incomplete or randomly filled out were excluded. Subsequently, a descriptive analysis was conducted to represent the collected responses [50]. Additionally, a chi-square test of independence was performed to examine the relation between cultural backgrounds of the participants and their perceptions regarding certain persuasive techniques in the context of this study.

RESULTS AND DISCUSSION

Users' Reaction to Cyberhate

Before investigating the integration of persuasive techniques into the design of social media platforms to encourage bystanders to take a stand against cyberhate, the study examined how users perceive and react to instances of cyberhate

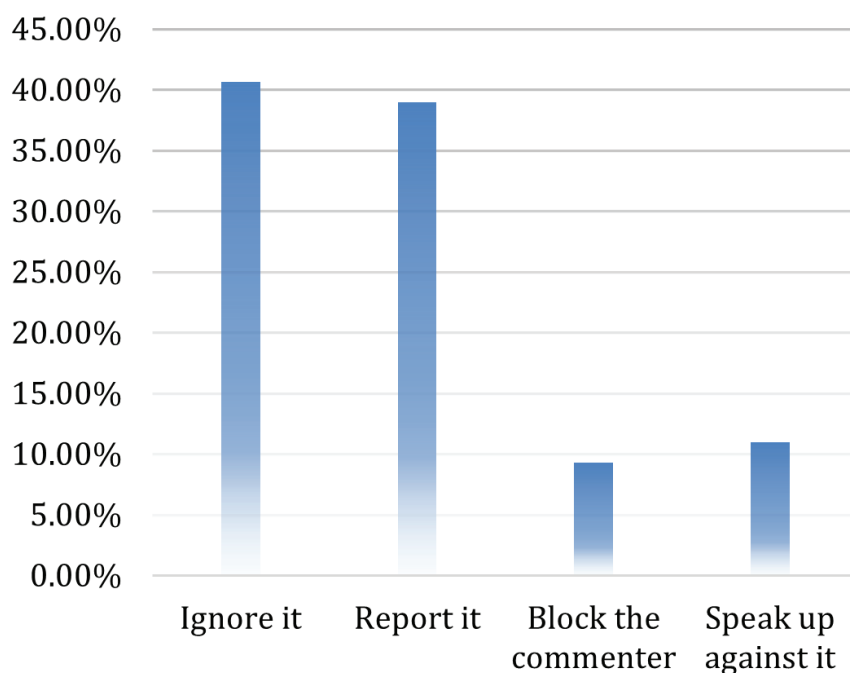


Figure 2. Users' reactions to cyberhate.

in the first place. Participants were explicitly asked about their typical response when they encounter cyberhate in the form of written or visual posts or comments. Approximately 41% of the participants stated that they simply ignore it, while around 38% said they would report such content (Fig. 2). Moreover, only about 11% of the participants mentioned that they would actively voice their opposition to such hateful content. These findings indicate the necessity for innovative and appealing approaches to motivate social media users to play an active role in combating the spread of cyberhate, such as the utilization of persuasive technology. Ultimately, this can foster a healthier online environment where users actively contribute to enhancing the quality of shared and generated online content.

Persuasive Technology to Combat Cyberhate

This subsection investigates users' perception on the use and integration of the previously mentioned PTs (Table 3) through designed interfaces that mimic the current user interface of X (formerly known as Twitter).

Private reply (PR)

The participants were asked about their opinions and attitudes regarding the use of the proposed "Reply Privately" feature as a mean to address cyberhate on social media platforms. RP allows users to comment privately on posts that contain cyberhate instances by clicking a "private reply" button which eliminates the need to use messages. Overall, the responses indicate that around 48% of the respondents perceive the "Reply Privately" feature as a valuable tool for combating online hate speech. They believe it can increase their willingness to address cyberhate, provide an easier means to challenge it, prevent public embarrassment, protect against public attacks, and potentially foster acceptance of different viewpoints (Table 3). These findings align with the objective of this study that implementing such a feature could contribute to creating a more comfortable and effective environment for bystanders to challenge cyberhate on social media platforms.

Facts popups (FP)

The participants were asked about their opinions regarding the use of a "Facts Pop-ups" feature to motivate them to intervene against cyberhate on social media platforms. The popups present to users facts about cyberhate's negative impact to encourage bystanders to take prosocial actions against it. In general, the responses suggest that around 54% of the respondents perceive the FP feature as effective in raising awareness about the negative impact of cyberhate, motivating them to challenge it, and providing the justification needed to intervene and take action (Table 4). This finding supports the study's objective of encouraging bystanders to take actions against cyberhate.

Comment analyzer (CA)

The survey investigated the participants' perceptions regarding the use of a "Comment Analyzer" feature that

helps them speak up against cyberhate on social media platforms works as an indicator that allows users to visualize how their comments to cyberhate may be perceived by someone reading them (e.g., aggressive, friendly, etc). Generally, the results suggest that around 61% of the respondents view the CA feature beneficial in various ways. They perceive it as a tool that helps them reconsider their comments, utilize a more acceptable language, address cyberhate effectively, and potentially make their replies more convincing to others (Table 5). These findings suggest that implementing such a feature could contribute to fostering more responsible and effective communication when challenging cyberhate on social media platforms. This indeed aligns with the study's objective of promoting healthier online interactions.

ProSocial badge (PB)

The participants were asked about their perception on the use of a "ProSocial badge" feature to motivate them to challenge cyberhate on social media platforms. PB is a digital recognition symbol awarded to users who occasionally take a stand against cyberhate. Overall, the results suggest that around 50% of the respondents perceive the PB feature as valuable in several ways. They see it as a way to receive recognition for their prosocial behavior, a continuous motivator to challenge cyberhate, and an encouragement (when seen on other users' profiles) to actively challenge cyberhate in order to earn such a badge (Table 6). PB feature has the potential to create a positive impact by recognizing and motivating users who challenge cyberhate, fostering a sense of community, and promoting a culture of respect and inclusion on social media platforms which highly associates with the objective of this study.

Social expectation (SE)

The participants were asked about their opinions on the use of a "messages popups" feature to motivate them to speak up against cyberhate. The pop-up messages serve as reminder messages of what others would expect from bystanders (Social Expectation (SE)) when they are victims of cyberhate. SE feature aims to leverage persuasive techniques through normative influence to foster a stronger sense of responsibility, motivation, and justification among bystanders when it comes to addressing cyberhate. The survey results suggest that around 50% of the respondents view SE feature as valuable in several ways (Table 7). They perceive these pop-up messages as reminders of their moral responsibility, motivators to challenge cyberhate, and sources of justification for intervening and addressing cyberhate. By incorporating such a persuasive technique into the user experience, social media platforms can encourage a proactive and engaged community that actively challenges cyberhate, promotes inclusivity, and contributes to a more positive online environment which supports the objective of this study.

Table 2. Respondents' characteristics

		Age Groups				Gender			Cultural Background			
		18-25	26-34	35-54	55 or above	Total	Male	Female	Total	European	Middle eastern	Total
level of education	High school	7	1	7	1	16	12	4	16	9	7	16
	Associate degree	0	2	2	2	6	3	3	6	6	0	6
	Bachelor's degree	14	28	21	3	66	40	26	66	20	46	66
	Master's degree	1	8	9	2	20	13	7	20	10	10	20
	Doctorate degree	0	1	1	0	2	0	2	2	1	1	2
	Others	0	1	2	0	3	0	3	3	1	2	3
	Total	22	41	42	8	113	68	45	113	47	66	113

Table 3. Participants' responses to the adoption of PR feature

Statement	Disagree	Neutral	Agree
I prefer using this feature to challenge hate speech rather than replying publicly	31%	31%	38%
This feature makes me more willing to challenge hate speech	28%	26%	46%
This feature helps me to challenge hate speech easily	24%	26%	50%
This feature helps me not to embarrass the attacker publicly	27%	21%	52%
This feature helps me not to be attacked by the attacker or his supporters publicly	18%	21%	61%
This feature could make the attacker accept others' point of views	34%	26%	40%

First-second-third place trophies (FT)

The survey explored the participants' opinions regarding the use of digital trophies (FT) feature as

motivator to challenge cyberhate on social media platforms. FTs involve awarding bystanders with virtual medals or badges in recognition of their quick and proactive responses against cyberhate based on their order of intervention (i.e., the first, second and third place). The survey results suggest that around 48% of the respondents perceive the FT feature as valuable in several ways (Table 8). They view digital trophies as a means of receiving recognition for their quick reactions against cyberhate, as a motivation to continue acting swiftly, and as a stimulus for their competitive nature to combat cyberhate promptly. By providing tangible rewards in the form of virtual trophies, this feature seeks to create a positive environment where users are inspired to respond quickly, actively contribute to countering cyberhate, and earn recognition for their efforts. This supports the overall goal of this study to encourage bystanders to a timely intervention against cyberhate.

Cultural Impact on Persuasive Technology Adoption for Cyberhate

Upon conducting a deeper analysis of the gathered responses, it became apparent that users' perceptions and opinions regarding certain proposed PTs were influenced by their cultural backgrounds (i.e., Middle Eastern and European). To investigate the connection between users' perceptions and opinions on some of the proposed PTs and their cultural backgrounds, A chi square test of independence was performed. The relation between these variables was significant with values less than 0.05 as seen in Table 9.

Generally, Middle Eastern users appear to value virtual digital badges, such as FT and PB, more than Europeans do. They are also more likely to be influenced by these features in order to confront cyberhate. Similarly, this trend can be observed with the SE feature, where Middle Easterners are more inclined than Europeans to be motivated by this feature to combat cyberhate. The message pop-ups function as reminders of the expectations society has for bystanders (Social Expectation- SE) when they encounter instances of cyberhate. The variations in users' perceptions and opinions about the adoption of specific

Table 4. Participants' responses to the adoption of FP feature

Statement	Disagree	Neutral	Agree
Such messages remind me of the negative impact of online hate speech on people	21%	16%	63%
Such messages motivate me to challenge online hate speech	21%	27%	52%
Such messages give me the justification I need to intervene and challenge online hate speech	23%	28%	49%

Table 5. Participants' responses to the adoption of CA feature

Statement	Disagree	Neutral	Agree
The comment analyzer helps me to reconsider my comment before posting it.	21%	19%	60%
The comment analyzer helps me to use a more acceptable and respectful language to challenge hate speech.	14%	21%	65%
The comment analyzer makes it easier for me to challenge hate speech in an acceptable language and tone.	19%	17%	64%
The comment analyzer could make my reply look more convincing to others and the attacker.	20%	27%	53%

Table 6. Participants' responses to the adoption of PB feature

Statement	Disagree	Neutral	Agree
This ProSocial digital badge gives me the recognition I deserve for being a prosocial user	23%	26%	51%
Having such a digital badge on my profile makes me always motivated to challenge cyberhate	25%	22%	53%
Seeing ProSocial digital badges on other users' profiles encourages me to challenge cyberhate to earn the same digital badge	28%	25%	47%

Table 7. Participants' responses to the adoption of SE feature

Statement	Disagree	Neutral	Agree
Such messages remind me of my moral responsibly to challenge hate speech	22%	28%	50%
Such messages make me more willing to challenge online hate speech	27%	24%	49%
Such messages give me the justification I need to intervene and challenge online hate speech	24%	26%	50%

persuasive techniques to address cyberhate, based on their cultural backgrounds, emphasize the need for software engineers to consider these differences when designing cyberhate-aware social media platforms. Furthermore, this emphasizes the need for further research and investigation to gain a better understanding of how cultural backgrounds influence the adoption of persuasive technology, with the goal of encouraging bystanders to speak out against cyberhate on social media platforms. Taking cultural nuances into account can ensure that social media platforms are more inclusive and tailored to the needs and preferences of diverse user groups.

Cyberhate-Aware Social Media

Figure 3 presents an initial application-independent conceptual framework for the design of a persuasive cyberhate-aware social media. It provides a concise

summary of the findings discussed in Section 3. The framework acknowledges that the proposed persuasive techniques have the potential to motivate users, specifically bystanders, to actively address the issue of cyberhate on social media. However, it also recognizes that users have diverse needs and preferences related to these persuasive techniques, which necessitates careful consideration by software engineers in order to cater to these variations effectively.

Furthermore, users' adoption and acceptance of certain proposed persuasive techniques for combating cyberhate are influenced by their cultural backgrounds. This emphasizes the need for software engineers to consider these differences when designing cyberhate-aware social media platforms. By catering to these diverse preferences, it becomes possible to effectively utilize persuasive

technology to motivate bystanders to actively address instances of cyberhate and encourage them to speak out against it.

As a result, the overall quality of content shared on social media, user engagement, satisfaction, experience, and critical digital literacy are expected to improve. Future work on the framework will involve further exploration of potential classifications and natural groupings of user behaviors regarding the adoption of persuasive technologies (PTs) to combat cyberhate, as well as strategies for catering to each user group. Additionally, a software engineering method will be proposed to guide software developers in implementing the framework effectively. Furthermore, considering that users' preferences and needs may evolve over time, it is crucial for the proposed framework to adapt to these behavioral changes in order to better meet users' evolving preferences and needs. This highlights the importance of incorporating runtime adaptation capabilities into the framework in future research endeavors.

In addition, the findings of this study present significant implications for various stakeholders, including social media platforms, educators, and policymakers. Social media companies are encouraged to implement features such as Facts Popups and Comment Analyzers to promote healthier online environments, while educational institutions can leverage these insights to enhance digital citizenship and raise awareness of cyberhate. Policymakers can utilize the study's findings to inform regulatory frameworks aimed at mitigating online hate,

highlighting the necessity for collaboration between technology firms and governmental bodies. Furthermore, it delineates avenues for future research, such as longitudinal studies to evaluate the sustained effectiveness of these interventions and investigations into cultural factors that may influence user responses. Ultimately, stakeholders are urged to apply these findings to facilitate meaningful advancements in the fight against cyberhate, fostering a collaborative effort to enhance the online experience for diverse user populations.

Validity Threats

Despite the author's careful adherence to the principles of conducting quantitative studies, there are three main potential threats to the validity of this study:

- While the adopted methodology successfully identified and described users' perceptions and preferences regarding the use of persuasive technology (PT) to combat cyberhate, it is possible that certain significant aspects that could impact their behaviors in this context were not fully captured.
- A common concern when using questionnaires is whether respondents interpreted and understood the questions as intended. To address this, a pilot study was conducted with five participants who met the study's inclusion criteria. Based on their feedback, some questions were reviewed and improved to ensure a shared understanding among all respondents.
- The sample size of the study, consisting of 113 respondents, can be considered medium-sized. A larger sample

Table 8. Participants' responses to the adoption of FT feature

Statement	Disagree	Neutral	Agree
Such digital trophies give me the recognition I deserve for my quick reactions against cyberhate.	27%	25%	48%
Winning such digital trophies make me always motivated to act fast against online hate speech and win a digital medal.	30%	20%	50%
Seeing such digital trophies on other users' replies motivates my competitive nature to act fast against cyberhate.	30%	24%	46%

Table 9. Values of Chi square test on the relation between cultural backgrounds and PTs

Statement	Chi-Square statistic (χ^2)	p-value
Seeing such digital trophies on other users' replies motivates my competitive nature to act fast against cyberhate.	15.6831	0.003
Having ProSocial badge on my profile makes me always motivated to challenge cyberhate	14.8194	0.005
Such digital trophies give me the recognition I deserve for my quick reactions against cyberhate.	11.7619	0.019
Seeing ProSocial digital badges on other users' profiles encourages me to challenge cyberhate to earn the same digital badge	17.303	0.001
This ProSocial badge gives me the recognition I deserve for being a prosocial user	11.0985	0.025
Message pop ups (SE) makes me more willing to challenge online cyberhate	11.4431	0.022

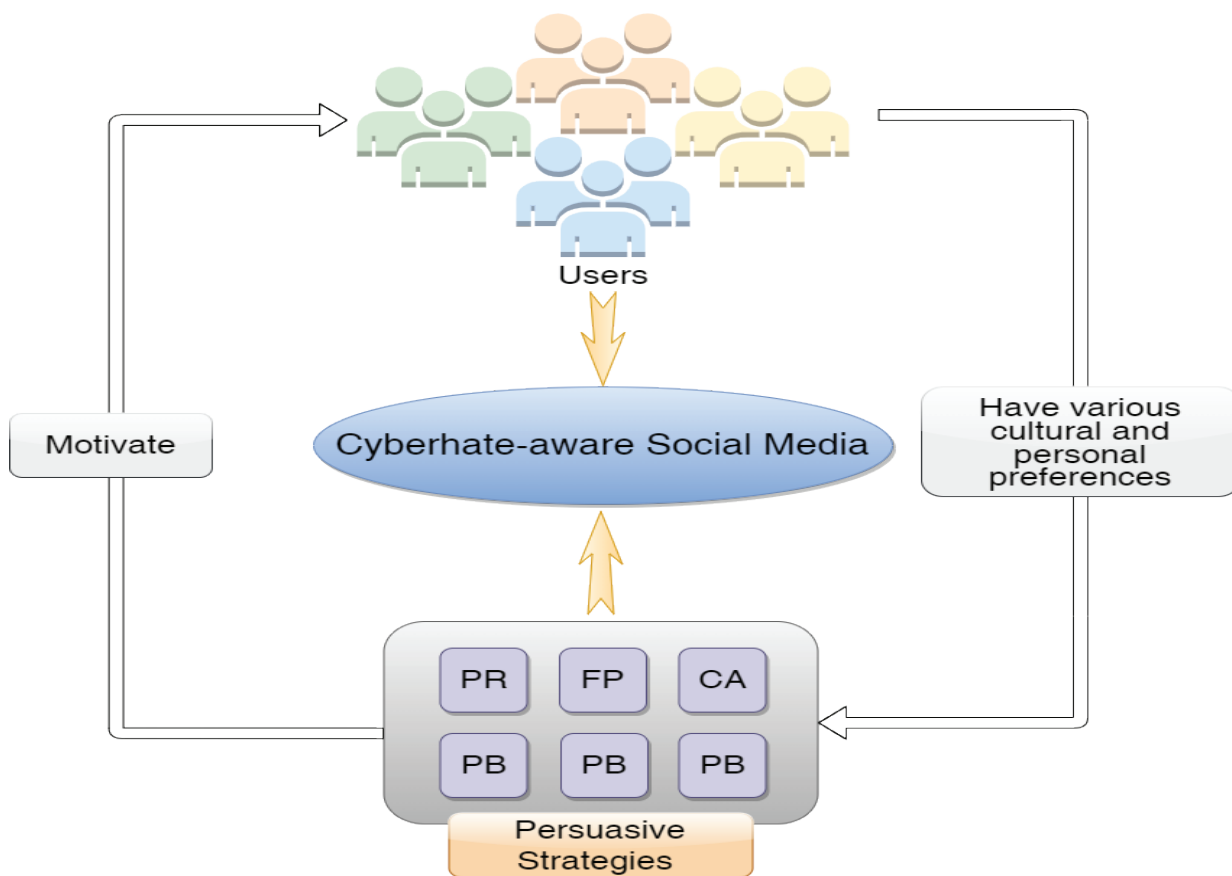


Figure 3. A conceptual framework for the design of a persuasive cyberhate-aware social media.

would enable the findings to be more generalizable to larger population groups. Further investigation of the study's results on a larger population will be conducted in future research.

- The study utilized a convenience sampling method, which may not fully represent the broader population. To mitigate this issue, the author recruited participants from diverse cultural backgrounds and demographics to enhance the generalizability of the findings.
- Participants may have provided socially desirable responses rather than their true opinions, particularly regarding sensitive topics like cyberhate. To minimize this threat, the author ensured anonymity in responses and emphasized that there were no right or wrong answers, encouraging honesty.

CONCLUSION

In this paper, the author carried out and reported on a quantitative study focusing on users' perceptions and opinions regarding the use of PT to encourage bystanders of cyberhate to speak out against it on social media platforms. The study innovates by empirically evaluating persuasive technologies against cyberhate, quantifying their effectiveness and offering actionable recommendations

for stakeholders. The study's findings revealed that users' opinions and needs within the context of this study exhibit significant variations and are affected by multiple factors. Additionally, the findings indicated that users' cultural backgrounds play a role in shaping their perceptions regarding the adoption of specific PTs for combating cyberhate. These results highlight the necessity for new systematic approaches and methods to integrate PT into the design of social media platforms, with the aim of motivating bystanders to actively address cyberhate. Such approaches and methods should consider and accommodate users' diverse preferences and needs, particularly in relation to the aspects that affect their perceptions within the study's context. The availability of such systematic approaches has the potential to significantly enhance users' online experiences, improve the quality of content generated online, and ultimately contribute to the development of social media platforms that are more conscious and proactive in addressing cyberhate. In future research and based on the proposed conceptual framework, the author intends to develop a software engineering method for the design of a persuasive cyberhate-aware social media. This method aims to enable software engineers to implement the framework at a large scale, considering its technical, social and cultural challenges. Additionally, further investigation of the research

topic with a larger, more diverse and multicultural sample of participants will be conducted to validate the findings, thus providing a clearer context for the study's conclusions and a better generalizability of its results. Furthermore, an exploration of the adoption of a larger set of PTs, gamification and other behavioral intervention methods will be experimented in future research to enhance the proposed framework across different social media contexts.

DATA AVAILABILITY STATEMENT

The Data is available on: <https://shorturl.at/gipDT>

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

- [1] Müller K, Schwarz C. Fanning the flames of hate: Social media and hate crime. *J Eur Econ Assoc* 2021;19:2131–2167. [\[CrossRef\]](#)
- [2] Almars AM, Almaliki M, Noor TH, Alwateer MM, Atlam E. HANN: Hybrid attention neural network for detecting COVID-19 related rumors. *IEEE Access* 2022;10:12334–12344. [\[CrossRef\]](#)
- [3] Rana MRR, Ali T, Nawaz A. Exploring multilingual reviews for aspect-based sentiment analysis using Lexicon and BERT. *Sigma J Eng Nat Sci* 2024;42:1469–1479. [\[CrossRef\]](#)
- [4] Almaliki M, Almars AM, Gad I, Atlam ES. ABMM: Arabic BERT-mini model for hate-speech detection on social media. *Electronics* 2023;12:1048. [\[CrossRef\]](#)
- [5] Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y. Abusive language detection in online user content. In: *Proceedings of the 25th International Conference on World Wide Web*. Geneva: International World Wide Web Conferences Steering Committee; 2016. p. 145–153. [\[CrossRef\]](#)
- [6] Wachs S, Wright MF. Associations between bystanders and perpetrators of online hate: The moderating role of toxic online disinhibition. *Int J Environ Res Public Health* 2018;15:2030. [\[CrossRef\]](#)
- [7] Blaya C. Cyberhaine: Les jeunes et la violence sur Internet. Paris: Nouveau Monde; 2019. [\[CrossRef\]](#)
- [8] Machackova H, Blaya C, Bedrosova M, Smahel D, Staksrud E. Children's experiences with cyberhate. London: EU Kids Online, The London School of Economics and Political Science; 2020.
- [9] Anti-Defamation League. Online hate and harassment: The American experience 2021. New York: Center for Technology and Society; 2021.
- [10] Ponte C, Batista S. EU kids online Portugal: Usos, competências, riscos e mediações da internet reportados por crianças e jovens (9-17 anos). Lisbon: EU Kids Online e NOVA FCSH; 2019.
- [11] Costa S, Mendes da Silva B, Tavares M. Video games and gamification against online hate speech? In: *10th International Conference on Digital and Interactive Arts*. New York: ACM; 2021. p. 1–7. [\[CrossRef\]](#)
- [12] Citron DK, Norton H. Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston Univ Law Rev* 2011;91:1435–1484.
- [13] Cao R, Lee RKW, Hoang TA. DeepHate: Hate speech detection via multi-faceted text representations. In: *12th ACM Conference on Web Science*. New York: ACM; 2020. p. 11–20. [\[CrossRef\]](#)
- [14] Nugroho K, Noersasongko E, Fanani AZ, Basuki RS. Improving random forest method to detect hate-speech and offensive word. In: *2019 International Conference on Information and Communications Technology (ICOIACT)*. Piscataway: IEEE; 2019. p. 514–518. [\[CrossRef\]](#)
- [15] Zhou Y, Yang Y, Liu H, Liu X, Savage N. Deep learning based fusion approach for hate speech detection. *IEEE Access* 2020;8:128923–128929. [\[CrossRef\]](#)
- [16] Swamy SD, Jamatia A, Gambäck B. Studying generalizability across abusive language detection datasets. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Stroudsburg: Association for Computational Linguistics; 2019. p. 940–950. [\[CrossRef\]](#)
- [17] Arango A, Perez J, Poblete B. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Inform Syst* 2022;105:101584. [\[CrossRef\]](#)
- [18] Yin W, Zubiaga A. Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Comput Sci* 2021;7:e598. [\[CrossRef\]](#)
- [19] Das M, Mathew B, Saha P, Goyal P, Mukherjee A. Hate speech in online social media. *ACM SIGWEB Newsl* 2020;2020:1–8. [\[CrossRef\]](#)
- [20] Mathew B, Dutt R, Goyal P, Mukherjee A. Spread of hate speech in online social media. In: *Proceedings of the 10th ACM Conference on Web Science*. New York: ACM; 2019. p. 173–182. [\[CrossRef\]](#)
- [21] Bedrosova M, Machackova H, Šerek J, Smahel D, Blaya C. The relation between the cyberhate and cyberbullying experiences of adolescents in the Czech Republic, Poland, and Slovakia. *Comput Human Behav* 2022;126:107013. [\[CrossRef\]](#)

- [22] Mathew B, Illendula A, Saha P, Sarkar S, Goyal P, Mukherjee A. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction* 2020;4:1–24. [\[CrossRef\]](#)
- [23] Almaliki M. Cyberhate dissemination: A systematic literature map. *IEEE Access* 2023;11:63250–63264. [\[CrossRef\]](#)
- [24] Leonhard L, Rueß C, Obermaier M, Reinemann C. Perceiving threat and feeling responsible: How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *SCM Stud Commun Media* 2018;7:555–579. [\[CrossRef\]](#)
- [25] Schieb C, Preuss M. Governing hate speech by means of counterspeech on Facebook. In: 66th ICA Annual Conference. Fukuoka: International Communication Association; 2016. p. 1–23.
- [26] Preuß M, Tetzlaff F, Zick A. "Publizieren wird zur Mutprobe": Studie zur Wahrnehmung von und Erfahrung mit Angriffen unter Journalist*innen. Bielefeld: Mediendienst Integration & Institut für interdisziplinäre Konflikt- und Gewaltforschung; 2017.
- [27] Bartlett J, Krasodonski-Jones A. Counter-speech: Examining content that challenges extremism online. London: Demos; 2015.
- [28] Delgado R, Stefancic J. Hate speech in cyberspace. *Wake Forest Law Rev* 2014;49:319–342.
- [29] Latané B, Darley JM. The unresponsive bystander: Why doesn't he help? New York: Appleton-Century-Crofts; 1970.
- [30] Oinas-Kukkonen H. A foundation for the study of behavior change support systems. *Person Ubiquit Comput* 2013;17:1223–1235. [\[CrossRef\]](#)
- [31] Oinas-Kukkonen H, Harjumaa M. Towards deeper understanding of persuasion in software and information systems. In: *First International Conference on Advances in Computer-Human Interaction*. Piscataway: IEEE; 2008. p. 200–205. [\[CrossRef\]](#)
- [32] Fogg BJ. Creating persuasive technologies: An eight-step design process. In: *Proceedings of the 4th International Conference on Persuasive Technology*. New York: ACM; 2009. p. 1–6. [\[CrossRef\]](#)
- [33] Alhasani M, Mulchandani D, Oyeboode O, Orji R. A systematic review of persuasive strategies in stress management apps. In: *BCSS@PERSUASIVE 2020*. Aachen: CEUR-WS; 2020. p. 48–59.
- [34] Orji R, Moffatt K. Persuasive technology for health and wellness: State-of-the-art and emerging trends. *Health Inform J* 2018;24:66–91. [\[CrossRef\]](#)
- [35] Widyasari YDL, Nugroho LE, Permasari AE. Persuasive technology for enhanced learning behavior in higher education. *Int J Educ Technol Higher Educ* 2019;16:15. [\[CrossRef\]](#)
- [36] Win KT, Mullan J, Howard SK, Oinas-Kukkonen H. Persuasive systems design features in promoting medication management for consumers. In: *Proceedings of the 50th Hawaii International Conference on System Sciences*. Honolulu: HICSS; 2017. p. 1–7. [\[CrossRef\]](#)
- [37] Elaheebocus SMRA, Weal M, Morrison L, Yardley L. Peer-based social media features in behavior change interventions: Systematic review. *J Med Internet Res* 2018;20:e8342. [\[CrossRef\]](#)
- [38] Wiafe I, Koranteng FN, Owusu E, Ekpezu AO, Gyamfi SA. Persuasive social features that promote knowledge sharing among tertiary students on social networking sites: An empirical study. *J Comput Assist Learn* 2020;36:636–645. [\[CrossRef\]](#)
- [39] Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: Key issues, process model, and system features. *Communications of the Assoc Inform Syst* 2009;24:485–500. [\[CrossRef\]](#)
- [40] Carnegie D. How to win friends and influence people. New York: Simon & Schuster; 1936.
- [41] Miškolci J, Kováčová L, Rigová E. Countering hate speech on Facebook: The case of the Roma minority in Slovakia. *Soc Sci Comput Rev* 2020;38:128–146. [\[CrossRef\]](#)
- [42] Osama Y. Countering online hate speech: A qualitative study of digital literacy training programs. *Egypt J Media Res* 2023;2023:189–240.
- [43] Oyeboode O, Alqahtani F, Orji R. Exploring for possible effect of persuasive strategy implementation choices: Towards tailoring persuasive technologies. In: *International Conference on Persuasive Technology*. Cham: Springer; 2022. p. 145–163. [\[CrossRef\]](#)
- [44] Halttu K, Oinas-Kukkonen H. Need for cognition among users of self-monitoring systems for physical activity: Survey study. *JMIR Format Res* 2021;5:e23968. [\[CrossRef\]](#)
- [45] Bandura A. Social cognitive theory of self-regulation. *Organiz Behav Hum Decis Process* 1991;50:248–287. [\[CrossRef\]](#)
- [46] Anderson A, Huttenlocher D, Kleinberg J, Leskovec J. Steering user behavior with badges. In: *Proceedings of the 22nd International Conference on World Wide Web*. New York: ACM; 2013. p. 95–106. [\[CrossRef\]](#)
- [47] Stibe A, Oinas-Kukkonen H, Lehto T. Exploring social influence on customer engagement: A pilot study on the effects of social learning, social comparison, and normative influence. In: *2013 46th Hawaii International Conference on System Sciences*. Piscataway: IEEE; 2013. p. 2735–2744. [\[CrossRef\]](#)
- [48] Lapinski MK, Rimal RN. An explication of social norms. *Commun Theory* 2005;15:127–147. [\[CrossRef\]](#)
- [49] Perkins HW. The emergence and evolution of the social norms approach to substance abuse prevention. In: Perkins HW, editor. *The social norms approach to preventing school and college age substance abuse: A handbook for educators, counselors, and clinicians*. San Francisco: Jossey-Bass; 2003. p. 3–17.
- [50] Williams A. How to write and analyse a questionnaire. *J Orthodont* 2003;30:245–252. [\[CrossRef\]](#)