Investigating the Effects of Nudges for Facilitating the Use of Trigger Warnings and Content Warnings

Emily C. Altland

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science & Application

Sang W. Lee, Chair
Scott McCrickard
Eugenia Rho

May 6, 2024
Blacksburg, Virginia

Keywords: Social Media, Twitter, X, Trigger Warnings, Content Warnings, Sensitive Content, Nudging Algorithm, OpenAI

Investigating the Effects of Nudges for Facilitating the Use of Trigger Warnings and Content Warnings

Emily C. Altland

(ABSTRACT)

Social media can trigger past traumatic memories in viewers when posters post sensitive content. Strict content moderation and blocking/reporting features do not work when triggers are nuanced and the posts may not violate site guidelines. Viewer-side interventions exist to help filter and hide certain content but these put all the responsibility on the viewer and typically act as 'aftermath interventions'. Trigger and content warnings offer a unique solution giving viewers the agency to scroll past content they may want to avoid. However, there is a lack of education and awareness for posters for how to add a warning and what topics may require one. We conducted this study to determine if poster-side interventions such as a nudge algorithm to add warnings to sensitive posts would increase social media users' knowledge and understanding of how and when to add trigger and content warnings. To investigate the effectiveness of a nudge algorithm, we designed the TWIST (Trigger Warning Includer for Sensitive Topics) app. The TWIST app scans tweet content to determine whether a TW/CW is needed and if so, nudges the social media poster to add one with an example of what it may look like. We then conducted a 4-part mixed methods study with 88 participants. Our key findings from this study include (1) Nudging social media users to add TW/CW educates them on triggering topics and raises their awareness when posting in the future, (2) Social media users can learn how to add a trigger/content warning through using a nudge app, (3) Researchers grew in understanding of how a nudge algorithm like TWIST

can change people's behavior and perceptions, and (4) We provide empirical evidence of the effectiveness of such interventions (even in short-time use).

Investigating the Effects of Nudges for Facilitating the Use of Trigger Warnings and Content Warnings

Emily C. Altland

(GENERAL AUDIENCE ABSTRACT)

Social media can trigger past traumatic memories in viewers when posters post sensitive content. Strict content moderation and blocking/reporting features do not work when triggers are nuanced and the posts may not violate site guidelines. Viewer-side interventions exist to help filter and hide certain content but these put all the responsibility on the viewer and typically act as 'aftermath interventions'. Trigger and content warnings offer a unique solution giving viewers the agency to scroll past content they may want to avoid. However, there is a lack of education and awareness for posters for how to add a warning and what topics may require one. We conducted this study to determine if poster-side interventions such as a nudge algorithm to add warnings to sensitive posts would increase social media users' knowledge and understanding of how and when to add trigger and content warnings. To investigate the effectiveness of a nudge algorithm, we designed the TWIST (Trigger Warning Includer for Sensitive Topics) app then conducted a 4-part mixed methods study with 88 participants. Our findings from this study show that nudging social media users to add TW/CW educates them on triggering topics and raise their awareness when posting in the future. It also shows social media users can learn how to add a trigger/content warning through using a nudge app.

Contents

List of Figures

List of Tables

List of Abbreviations

CW Content Warning

NLP Natural Language Processing

TW Trigger Warning

Trauma: Trauma is the experience and resulting aftermath of an extremely distressing event or series of events [14] like violence, abuse, neglect, loss, disaster, war, and other emotionally harmful experiences [23].

Trigger: A trauma trigger is a psychological stimulus that prompts the involuntary recall of a previous traumatic experience. [3, 6] This can be visual, auditory, or through another sensory stimulus [6]. The stimulus itself need not be frightening or traumatic and maybe only indirectly remind of an earlier traumatic incident such as a scent or a picture of a place [6].

Trauma-Informed Computing is an approach for improving technology while minimizing harm in all phases of technology design, development, and research [14]. The approach uses six key principles of trauma-informed approaches described by SAMHSA [23] — safety, trust, collaboration, peer support, enablement, and intersectionality—to the design, development, deployment, and evaluation of computing systems [14].

NLP is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages.

Chapter 1

Introduction

1.1 Background

People love to use social media, but with the depth of content available, the likelihood of being triggered by something you find online is high for some users. Trauma triggers vary from person to person, but they are a psychological stimulus that prompts the involuntary recall of a previous traumatic experience. [3, 6] This can be visual, auditory, or through another sensory stimulus [6]. The stimulus itself need not be frightening or traumatic and maybe only indirectly reminds them of an earlier traumatic incident, such as a scent or a picture of a place [6]. How you address and cope with these stimuli can vary from person to person, especially when it comes to online spaces like social media platforms. In a study by Andalibi et al. [4], users felt that repeatedly coming across sensitive posts on a troubling topic (to them) was exhausting. Some users choose to even avoid certain platforms altogether if the platform is notorious for not having good systems in place for sensitive viewers [19].

Another approach that authors of social media posts can use is the use of trigger and content warning labels. Trigger warnings are frequently employed in online forums and social media platforms [17, 33]. While trigger warnings have been studied in clinical settings [24, 41], layperson practices in online space have not been less known. In the literature, the words 'trigger warning' and 'content warning' have been used interchangeably across the diverse academic disciplines researching their use. Some but not all studies locate trigger warnings

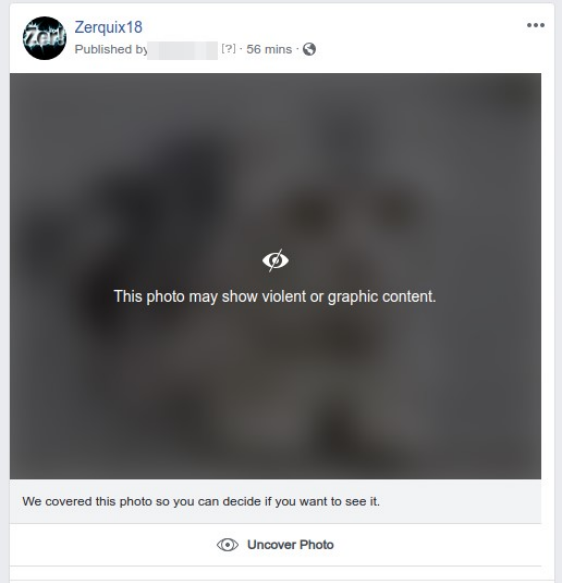
as a particular sub-type of content warnings that are focused specifically on the needs of people with experience of trauma or post-traumatic stress disorder (PTSD) [13].

There have been two contrasting views about the usage of warnings. Advocates claim that warnings allow people to prepare themselves [6, 9], reduce negative reactions toward content [9], and increase individual agency in making informed decisions about engaging with content [13]. In contrast, critics say that warnings may increase negative interpretations [9], with recent empirical evidence from educational sectors suggesting that they may raise anxiety and reinforce the centrality of trauma experience to an individual's identity [13].

Trigger and content warnings may take a variety of forms depending on the platform, content type, and whether they are platform or user-added. Figure 1.1 shows some examples of TW/CW where (a) and (b) are platform-added and (c) and (d) are user-added.

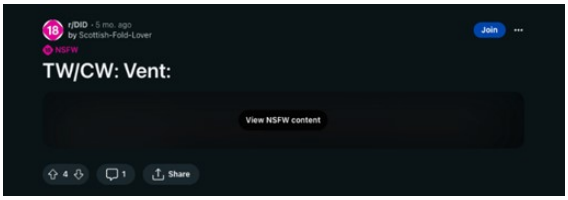
1.2 Motivation

One strategy for handling sensitive content on social media is content moderation [24]. This is when users post freely but either a human moderator or an AI bot will scan the content after posting and determine whether this content goes against the site guidelines and should be removed or hidden from the platform. Some people might argue that blocking, censoring, or removing sensitive content is best since triggering topics should not be on the platform. However, content moderation frequently discriminates against already marginalized groups [21] leading to further censorship of an already oppressed group. Also, triggers are deeply personal and nuanced so they can vary from person to person and even for the same person, season to season. For example, a social media post discussing sensitive content like the #MeToo movement, might be empowering to share for the person who is posting but can be triggering for someone in the audience based on their stage of trauma recovery. There is a

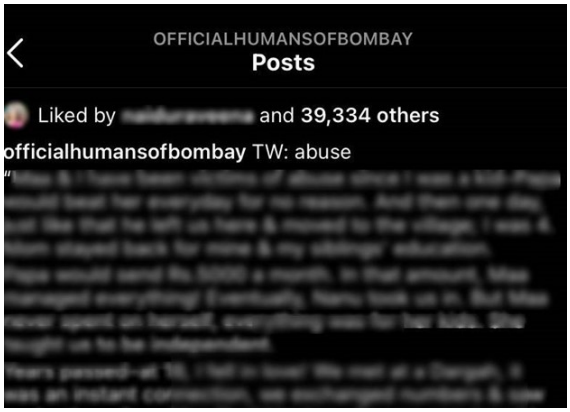


(a) Platform added warning on image post

(b) Platform added warning on video post



(c) User added warning on text post



(d) User added warning in caption of image/video post

Figure 1.1: Examples of TW/CW on Social Media A-D

gray area where the content should not be removed or moderated but some users may want to avoid it because of their personal traumatic experiences. Content and trigger warnings give viewers the agency to make an informed decision to see the content or not see it if they are not in the right mindset [6].

TW/CW can be an effective option to warn viewers proactively about the potentially triggering content to a subset of users so that those who want to avoid such content can choose to read/view it or not. Although the effectiveness of trigger warnings in classrooms and on college campuses has widely been up for debate [6, 8, 10, 11, 27, 39], there has been limited research on their use on social media. *TransTime*, a social media site developed particularly for trans people, effectively enabled its users to selectively view content using *Content Warning Tags* [20]. To hide content that might be troubling or traumatic, the viewers could filter out content based on user-defined tags like Instagram's hashtags. One example of a content warning tag is shown in Figure 1.2.

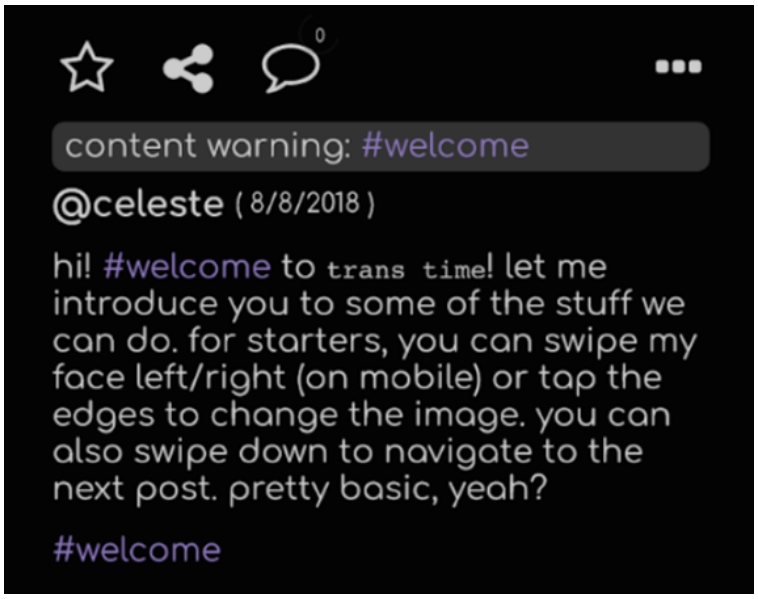


Figure 1.2: Screenshot of TransTime with #welcome content tag

Haimson et al. found that users valued the extra information and care that content warnings

provided enough to be willing to deal with friction and the extra work of labeling content on their posts [20]. Content and trigger warnings are one way to protect users from sensitive content they do not want to see without censoring content that does not violate site guidelines.

Some platforms add content warnings related to common triggers like self-harm, violence, suicidal thoughts, and more. However, most platforms rely on users to create these warnings for their own posts dealing with sensitive content (whether text, pictures, video, or some combination of these). Since most trigger and content warnings (TW/CW) are reliant on the poster to add them, they are inconsistent across users. This is partially due to a lack of knowledge on what topics require them, or what contexts do not. [19] As for the platform-generated warnings, there are inconsistencies across platforms for how they address sensitive content and this typically is discriminatory too. [21] Certain sensitive content gets removed that falls into a gray area concerning social media site policies and community norms even though sometimes they do not violate site policies [21]. There are some interventions that have been designed for the viewer side of social media including *Shinigami Eyes* [2] which shows transphobia in Facebook groups and automated content warnings [40]. Similarly, *DoesTheDogDie.com* [1] is a website that allows people to preview whether certain triggers appear in TV Shows, movies, and videogames using crowdsourcing.

While beneficial, there are limitations of viewer side interventions. All responsibility is on the viewer to hide/block/filter their content. The viewer is limited by the platform's features. Many platforms only allow you to completely block a person and all their content not allowing for partial viewing or connectedness with a person but not all of their content. Another problem is that the viewer may be already exposed to sensitive content when they report/block the post/person. This is what's considered an 'aftermath intervention'. Finally, viewer-side interventions do not teach the posters that the type of content being posted may

be sensitive content. This may lead to posters repeating this posting pattern without a warning in the future.

Overall, little work has been done to investigate whether poster-side interventions may be beneficial though. There is also a disconnect between poster-side interventions and viewer-side interventions for mainstream platforms.

In a prior study with Gupta [19], we conducted an interview study to understand how social media users perceive TW/CW on various platforms as well as inform effective warning design. Throughout the study and conducting axial coding to extract themes from the results, we began to uncover the complexity of topics with TW/CW. We reviewed how warnings are added, who adds them, and what entails the decision to add a warning. This paper also shares the decision-making process participants had to view content even when it contained a warning. We analyzed some factors contributing to TW/CW's (in)effectiveness. At the end of this work, we suggested design implications for how social media platforms may incorporate new and existing features to make these platforms safer for sensitive viewers. Figure 1.3 shows the different codes from axial coding that related to the ideal UI subsection.

This former project motivated the work explained here and helped to design the first iteration of the TWIST App used in this study.

1.3 Knowledge Gap

We conducted this study to determine if poster-side interventions such as a nudge algorithm to add warnings to sensitive posts would increase social media users' knowledge and understanding of how and when to add trigger and content warnings. Many social media users do want to be aware of their audience and how certain topics may be triggering but they may

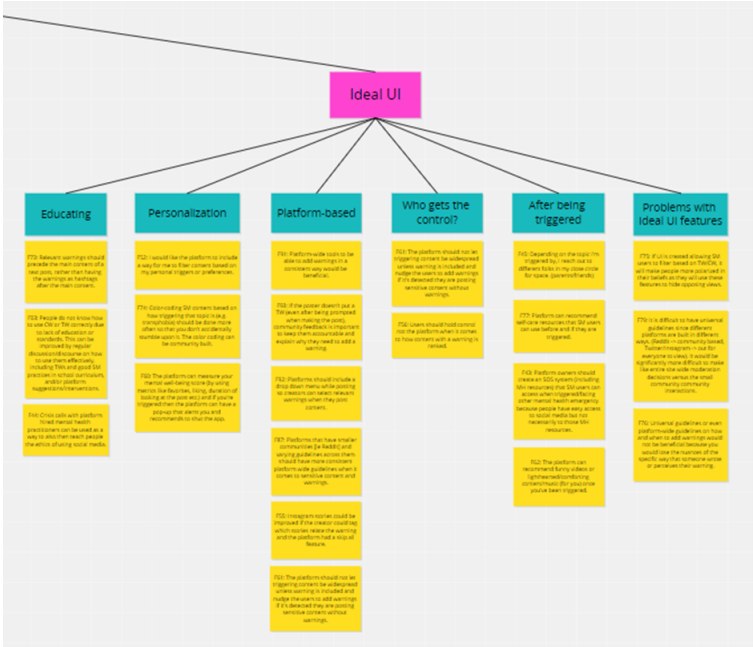


Figure 1.3: Screenshot of Axial Coding From Gupta's Interview Study

not have enough familiarity with how to use trigger and content warnings or what types of topics may require them [19]. This study hopes to bridge that gap by creating a nudge intervention that could be implemented either as an external web extension or ideally built into social media platforms themselves and connected to viewer-side capabilities. This nudge intervention also aims to educate study participants on when trigger warnings are recommended to be added to certain posts. In addition, we also explore if large language models (LLMs) like GPT from OpenAI can be used to predict triggering content and suggest warnings. We will analyze how social media users' understanding and confidence change after using our nudge algorithm and app.

1.4 Research Questions

We aimed to answer the following research questions:

- **RQ1:** How does nudging social media posters influence their knowledge and confidence in determining the inclusion of trigger warnings in their posts?
  - **H1:** Nudging social media users to add TW/CW increases their ability to recognize topics that may be triggering to readers.
  - **H2:** Nudging social media users to add TW/CW increases their perceived understanding TW/CW practice on social media.
- **RQ2:** How does nudging social media users to add TW/CW change their posting behavior in regard to sensitive content on social media?
- **RQ3:** How can LLMs be used to detect TW/CW-related topics and to what extent does it agree with human annotators?

To investigate the effectiveness of a nudge algorithm, we designed the TWIST (Trigger Warning Includer for Sensitive Topics) app. The TWIST app scans tweet content to determine whether a TW/CW is needed and if so, nudges the social media poster to add one with an example of what it may look like. We then conducted a 4-part mixed methods study with 88 participants. The first part (Part A) was a baseline assessment of the participants' knowledge and understanding of trigger and content warnings. Participants were asked to respond to statements about their familiarity with trigger and content warnings with answer choices on a 7-point Likert scale (Strongly disagree to Strongly agree). Next, the participants answered "Does this tweet need a trigger/content warning?" for 12 unique real-life tweets; 8 of them had TW/CW in its original form, which we removed when we presented them in the survey, and 4 of them did not. Every time they responded "yes", an additional question was asked so they could specify how they would add such a warning. They would repeat these steps for 12 tweets. The next part of the study (Part B) was the intervention, where the participants were given a prompt and asked to write a tweet in response. As they attempted

to post their tweet, they would end up interacting with the TWIST app we designed. After completing the 12 simulated post-writing tasks, they were navigated to a survey (Part C) to answer the same questions they did at the beginning (Part A) to see if their knowledge, understanding, and confidence in using trigger/content warnings had changed due to the intervention. Again, the participants were given a tweet and asked whether that tweet needed a trigger/content warning. If they responded "yes", an additional question was asked of them to specify how they would add such a warning. They repeated these steps for 12 more tweets. We reviewed their change in recall and self-efficacy by comparing the pre and post-test results. Finally, the participants were asked to give feedback on the algorithm the TWIST app used as well as answer some exit survey questions (Part D). This was to gain their insight on the prior sections of the study and what they learned, how they would use a tool like this, etc.

1.5 TWIST App's Impact on Trigger Warning/Content Warning Practices

To answer our first research question, "How does nudging social media posters influence their knowledge and confidence in determining the inclusion of trigger warnings in their posts?" we analyzed the growth in performance between parts A and C of the study. Hypothesis 1, "Nudging social media users to add TW/CW increases their ability to recognize topics that may be triggering to readers." was supported. The participants grew in their knowledge of when and how to use TW/CW when posting about sensitive content as evidenced by a 50% increase in the participants' recall in analyzing tweets for sensitive content after our intervention when compared to before. The participants' precision decreased by 10% showing they began to act more cautiously as they classified non-sensitive content as needing a

warning, however, the F1 score rose by 23% signifying the improved classification of sensitive topics outweighed any misclassification of non-sensitive content.

We also reviewed the participants' perceptions of whether they felt they better understood what topics required a warning, how to add a warning, and their likelihood of adding warnings in the future to their own sensitive posts both before and after using the TWIST App. Hypothesis 2, "Nudging social media users to add TW/CW increases their perceived understanding TW/CW practice on social media." was partially supported with significant growth in one question. While only 1/6 of these self-efficacy questions had a significant change, participants did agree more strongly with the statement "I know how to include a trigger/content warning on a social media post that requires it." after the intervention.

To answer our second research question, "How does nudging social media users to add TW/CW change their posting behavior in regards to sensitive content on social media?" we analyzed the results from the participants' use of the TWIST app during Part B of the study. For each tweet writing prompt the participants responded to, we categorized their behavior into 6 different possible cases. The cases categorized the variety of paths participants could follow as they decided whether to add a warning to their post initially, after being prompted due to sensitive content, or not at all. We also looked at whether the participant decided to scan their content or not and if this differed based on whether sensitive content was present or not. We also analyzed how the case distribution varied based on the sensitive topic.

To answer our third research question, "How can LLMs be used to detect TW/CW-related topics and to what extent does it agree with human annotators?" we conducted an inter-rater reliability test to determine the similarity between the LLM and two different human annotators. In all of these comparisons, there is 'substantial agreement' ($0.6 < k < 0.8$) [26], with the LLM comparisons only performing slightly lower than the two human moderators'

comparison.

Lastly, we looked at the user feedback given on the app in Part D of the study which showed mostly positive results when it came to the app's ease of use, perceived effectiveness, integration into workflow capability, opinion on the backend algorithm, likelihood of continued use, and overall satisfaction.

1.6 Contributions of TWIST App For Social Media Content Creators

Our contributions are ...

1. Nudging social media users to add TW/CW educates them on triggering topics and raise their awareness when posting in the future

2. Social media users can also learn how to add a trigger/content warning

3. Grow in understanding of how a nudge algorithm like TWIST can change people's behavior and perceptions

4. Empirical evidence of the effectiveness of such interventions (even in short-time use)

Chapter 2

Literature Review

2.1 Controversy Surrounding Benefits of TW/CW

Trigger warnings are commonly used across various online forums and social media platforms [17, 33]. While clinical studies have examined trigger warnings [24, 41], less is known about their usage in online spaces. In academic literature, the terms "trigger warning" and "content warning" are often used interchangeably by researchers from diverse disciplines. However, some studies suggest that trigger warnings are a specific sub-type of content warnings tailored to individuals with trauma or those experiencing PTSD [13].

There are two contrasting perspectives regarding the effectiveness of warnings. Advocates argue that warnings help individuals prepare themselves, mitigate negative reactions to content [9], and empower individuals to make informed decisions about engaging with it [13]. Conversely, critics posit that warnings might amplify negative interpretations [9], with recent evidence from educational settings suggesting that they could exacerbate anxiety and reinforce trauma-centric identities [13].

2.2 Challenges due to Nuanced Triggers

Social media platforms host a wide variety of sensitive content, spanning topics like politics, religion, race, and gender, as well as material restricted by age, such as nudity and pornography, and graphic imagery depicting violence and gore. What one person finds acceptable may evoke traumatic memories for another due to personal triggers.

Political discourse, for instance, plays a crucial role in civic engagement, yet on social media, issues like homophily, polarization, inequity, misinformation, lack of transparency, and trust pose obstacles to fair civic participation [37]. Balancing the need for open discourse with the risks of misinformation and hate speech presents a challenge: allowing unrestricted free speech can foster the spread of inaccuracies or harmful rhetoric, while content moderation may be perceived as censorship [20, 22]. Topics like transgender individuals' transitions exemplify this dilemma, where discussions involving nudity or discrimination may be sensitive but serve essential purposes in educating, promoting gender affirmation, and nurturing a sense of belonging within the community [20, 22]. Navigating the complexities of social media content demands a delicate balance between promoting free expression and ensuring user safety, underscoring the importance of addressing these issues thoughtfully and inclusively.

Similar ambiguities arise in the realm of trigger warnings and content warnings regarding triggering content on social media. Technology-mediated reflection (TMR) systems, such as Facebook's Memories feature, can unexpectedly evoke strong emotions in users, even when presenting positive memories like wedding photos, which may be tinged with grief due to subsequent events like divorce or loss [28]. Determining what constitutes sensitive content, especially within the gray areas of social media policies and community standards, remains challenging. Content related to transgender transitions, for instance, straddles the line between LGBTQ+ visibility and nudity guidelines, potentially leading to reports or

blocks based on personal identity discrimination [20].

Issues like self-harm, suicide, and past trauma introduce further complexities, with tensions between protecting vulnerable viewers and raising awareness for prevention purposes [6, 34]. Content featuring substance use, for example, may trigger individuals in recovery from addiction [34]. Trigger warnings and content warnings are associated with managing such ambiguous sensitive content, highlighting the difficulties in effective moderation [24].

2.3 Trauma-Informed Computing

Trauma-informed approaches are widely acknowledged as beneficial for all individuals, irrespective of their trauma history [23]. The concept of trauma-informed computing acknowledges that digital technologies can both trigger and exacerbate trauma, aiming to prevent technology-related trauma and retraumatization [14].

Chen et al.'s framework for trauma-informed computing proposes that engineers and designers adopt principles of collaboration and empowerment to integrate individuals' conscious choices into their information streams [14]. However, they caution against the potential drawback of content filtering, noting that *"filter bubbles"* can restrict users to a subset of relevant content, often without their awareness.

In contrast, Randazzo et al. [35] argue that although filter bubbles can lead to polarization, filtering algorithms can assist trauma survivors in challenging societal filters imposed by institutions. They also offer design suggestions, proposing that *"semi-automating trigger warnings and directing word filters can benefit at-risk populations for trauma."*

2.4 What has been done previously to protect social media users from triggering content?

Certain social media platforms offer tools to easily filter out specific users. For instance, Twitter provides a blocklist feature, allowing users to preemptively block accounts to avoid harassment or triggering content [25]. However, these tools only offer the option to either block all content from a user or nothing at all.

To address the complexity of content users may wish to hide from their view, *SquadBox* proposes *friend-sourced moderation* as a potential solution to current moderation challenges [29]. This approach involves selecting friends to help review and block harassing emails, which has proven effective for direct contact but is not scalable for the vast volume of content on social media platforms.

Warnings serve as a means to retain sensitive content that doesn't violate site guidelines while alerting users who may be triggered by such content due to past trauma. For example, *TransTime*, a social media platform for the transgender community, implemented *Content Warning Tags* similar to hashtags, allowing users to filter posts based on these tags [20].

Crowd-sourced methods for labeling sensitive content have shown promise. *Shinigami Eyes*, a browser extension, employs community feedback to color-code social media users and content as either transphobic or trans-friendly [2, 7, 38]. Another extension, *DeText*, automatically generates content warnings and identifies sensitive text related to sexual violence using keyword recognition and sentiment analysis [40]. However, these tools are tailored to specific topics, such as trans issues and sexual violence.

DoesTheDogDie.com offers a comprehensive resource for pre-reviewing movies, TV shows, and video games for various triggers, with over 100 categories covering topics from violence

and abuse to specific plot elements like a dragon dying or the use of shaky cam [1]. While invaluable for traditional media, no equivalent resource currently exists for social media platforms.

2.5 What topics warrant a trigger or content warning?

Since triggers can vary from person to person, how do we determine what needs a trigger or content warning? Charles et al. [12] reviewed content warnings across a variety of disciplines and countries and created the table shown in 2.1 with 14 common topics that content warnings covered.

Since number 13 “flashing lights” and 14 “objects” applied primarily to image or video posts and our study only used text-based posts, we excluded these and focused on the other 12. These 12 sensitive topics were the primary focus of our TWIST app and, therefore, our user study too.

Table 1. NEON content warning typology.

Category (n) and definition	Sub-categories
1. Violence (n = 536)	Violence; War; Weapons; Terrorism; Police brutality; Motiveless killing; Sexual violence; Animal cruelty; Torture; Genocide
Content contains violence	
2. Sex (n = 332)	Nudity; Mild sexual content; Explicit sexual content; Relationship conflict; Reproductive health
Content contains sexual themes, including nudity, sexual content and relationships	
3. Stigma (n = 328)	Racism; Anti-religious (sub-categories: Anti-Semitic; Anti-Christian; Islamophobia); Colonialism; (sub-category: Slavery); Classism; Sexism (sub-categories: Misogyny; Misandry); Transphobia; Gender-identity; Sexuality (sub-category: Homophobia); Anti-disability
Content depicts negative stereotypes about or attitudes towards a specific group, such as racism or sexism	
4. Disturbing Content (n = 236)	Disturbing content with threat; Horror and terror; Disturbing imagery; Medical content; Human bodies and functions
Content contains imagery, sounds, or effects that may frighten, disgust or scare	
5. Language (n = 235)	Sexual language; Adult humour; Swearing; Offensive language
Content contains language which is sexual, crude or offensive	
6. Risky Behaviours (n = 200)	Drug misuse; Alcohol misuse; Tobacco; Gambling
Content depicts risky lifestyle behaviours	
7. Mental Health (n = 108)	Mental health; Eating disorders; Trauma; Self-harm and suicide; Depression; OCD; Panic attacks; Anxiety (sub-categories: Spiders; Snakes; Insects; Needles; Eye contact; Irregular patterns); Hair pulling
Content relates to mental health issues	
8. Death (n = 49)	Death; Accidents; Natural disasters
Content relates to human death or dying	
9. Parental Guidance (n = 47)	Online access; Cyber-bullying; Competitive content; Imitative content; Upsetting content; Non-realistic violence
Content may not be appropriate for children	
10. Crime (n = 38)	
Content depicts or relates to criminal activity	
11. Abuse (n = 37)	Child abuse; Emotional abuse; Physical/sexual abuse; Neglect
Content depicts or relates to abuse	
12. Sociopolitical (n = 27)	Injustice; Political issues; Nazism; Class issues
Content includes social or political issues	
13. Flashing Lights (n = 27)	
Content includes strobe or flashing lighting	
14. Objects (n = 4)	
Content contains specific objects	

https://doi.org/10.1371/journal.pone.0266722.t001

Figure 2.1: Typology of Content Warnings and Trigger Warnings [12] Sensitive Topics

Chapter 3

TWIST App

For this study, we designed an app called Trigger Warning Includer for Sensitive Topics (TWIST for short) to detect sensitive content when someone is posting on social media. The design went through several rounds of iteration and feedback from other graduate students and professors in Computer Science and a variety of design-related majors. This app was originally created as a Chrome extension that connected if the user was on the X (formerly known as Twitter) platform. The planned design for this Chrome extension after iterations over 3 months is shown in Figure 3.1 with a flowchart of how it was planned to work shown in Figure 3.2.

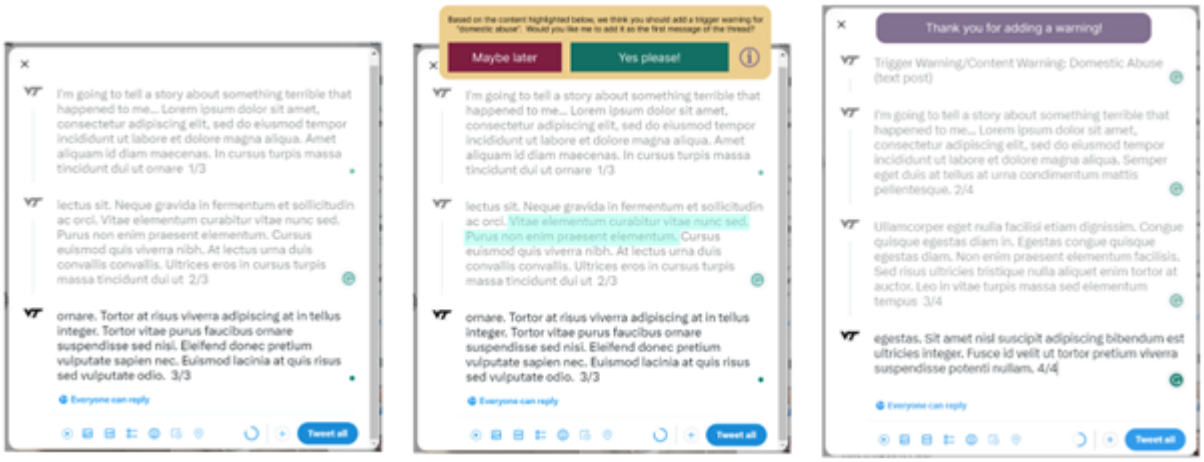


Figure 3.1: TWIST Chrome Extension Design

We decided to move away from the Chrome extension build to a server-hosted site with the TWIST App built in so we could conduct a large-scale quantitative study instead of

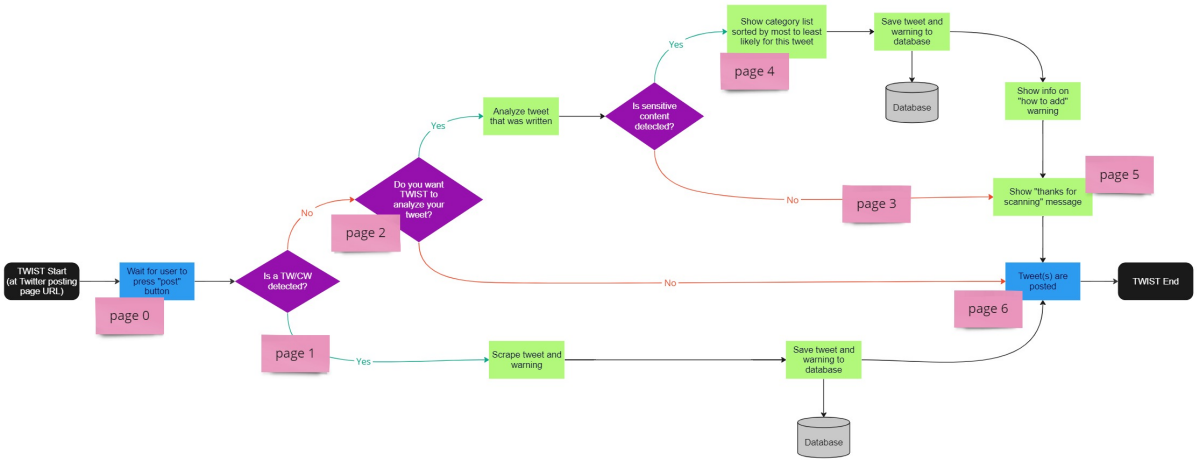


Figure 3.2: TWIST Chrome Extension Flowchart (Old)

a small-scale field study of the extension. This allowed us to focus more on the app and algorithm itself rather than spend time on UI elements of connecting the Chrome extension version seamlessly into the Twitter workflow.

The current version uses an OpenAI GPT 3.5 Turbo to scan the content to see if it would need TW or CW and prompt users to add warnings when sensitive content is detected. Figure 3.3 shows how the TWIST app works.

Users will start on page 0 of the design shown in 3.4. The TWIST App does not scan any content until the user has typed it and pressed the "Post My Tweet" button. When they do, their content will be scanned to route them to their next page.

If the tweet includes a warning already in the post, they will be routed to page 1, which is shown in 3.5. This page thanks the user for adding a warning to their content. The user can then make any changes they wish to the content and press the "Post My Tweet" button again to finish with this tweet-writing prompt.

If the tweet is long enough but no warning is already present, they will go to page 2, shown

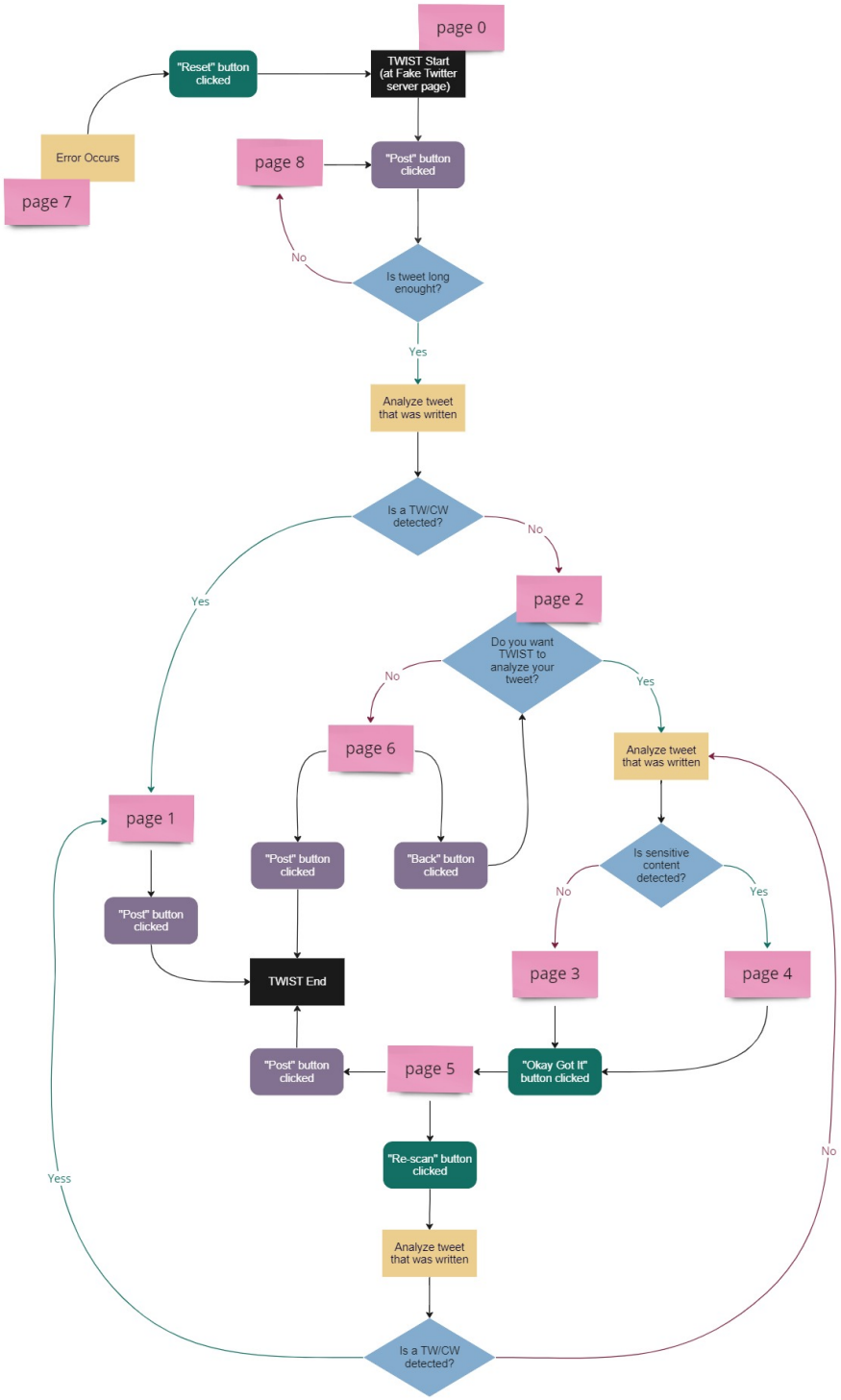


Figure 3.3: Flowchart of how TWIST app works

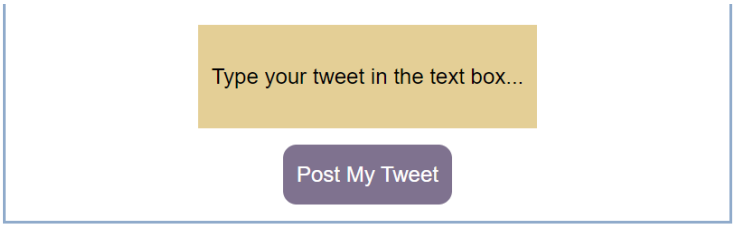


Figure 3.4: TWIST App Page 0 - Start Page

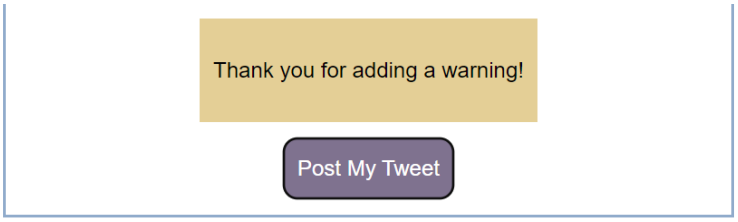


Figure 3.5: TWIST App Page 1 - Warning Already Added Page

in 3.7. This page asks the user if they would like their content to be scanned to see if it needs a trigger or content warning. At this point, the user could click "No" to skip ahead and be able to finish posting their tweet. The ability for the user to decline at any point, giving them the autonomy to select no and continue posting as normal was especially important in our original design as a Chrome extension. However, if they click "Yes", their tweet will be scanned to see if sensitive content is detected.

Whether the participant chooses to scan their content or not, the backend program will scan to save the response from GPT. The code used in the server for the prompting of OpenAI is shown in Figure 3.6. Lines 42 to 45 are the lines that we used to call OpenAI API and the variable text contains the system prompt we used.

GPT 3.5 Turbo was used with the system role, and the prompt given (content: text) was the following: "Here is a list of 12 sensitive topics with a short definition and some sub-categories each:" then a text version of the table from [12] which lists 12 sensitive topics (for text-based posts) then "Based on this sensitive topic list, does the following tweet contain any of those topics?" then the tweet text from the user input box then some additional instructions for

```
// Endpoint for handling OpenAI requests
app.post("/openai", jsonParser, async (req, res) => {
    try {
        // Get text from the request body
        const text = req.body.text;

        const completion = await openai.chat.completions.create({
            messages: [{ role: "system", content: text }],
            model: "gpt-3.5-turbo",
        });

        // Send the OpenAI response back to the frontend
        res.json({ response: completion.choices[0].message.content });
    } catch (error) {
        console.error(error);
        res.status(500).json({ error: "Internal Server Error" });
    }
});

```

Figure 3.6: Screenshot of Server Side Code for the Open AI API Call

how the response should be formatted. The full GPT prompt can be found in the appendix (see Appendix A).

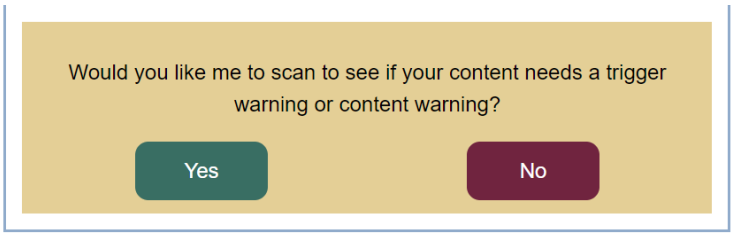


Figure 3.7: TWIST App Page 2 - No Warning Detected Page

If no sensitive content is detected, then the app will tell them it has not detected sensitive content while still alerting the user to what categories they may post about in the future that are considered potentially sensitive content. This is shown in 3.8

To determine the amount users agreed with the OpenAI algorithm, we added a radio button to both pages 3 (see 3.8) and 4 (see 3.11) of the app. The user could select whether they agreed with the scan results (see 3.9) or disagreed with the scan results (see 3.10).

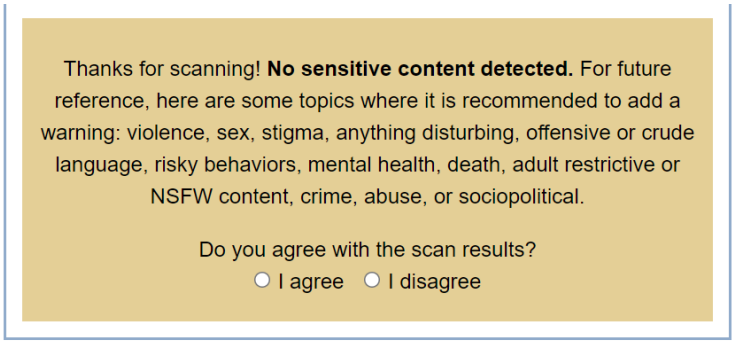


Figure 3.8: TWIST App Page 3 - No Sensitive Content Detected Page

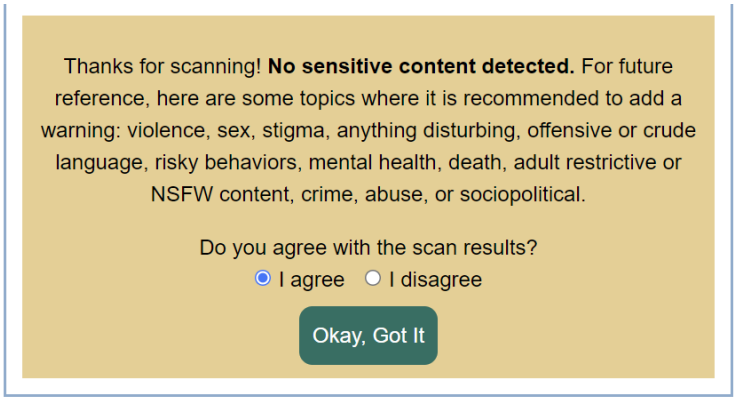


Figure 3.9: TWIST App Page 3 - Agree Radio Button Selected

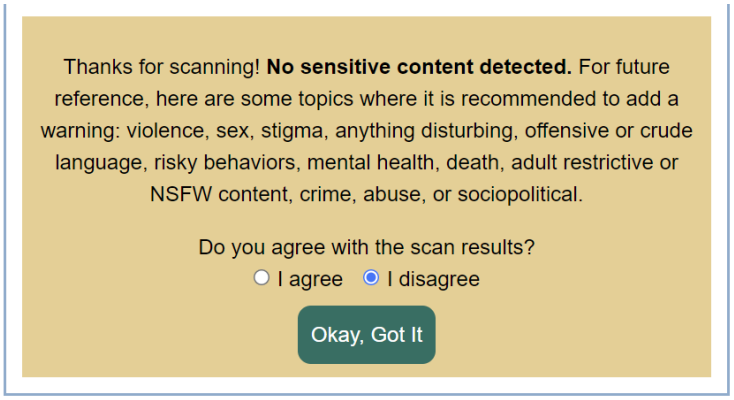


Figure 3.10: TWIST App Page 3 - Disagree Radio Button Selected

If sensitive content is detected, the app shows the user the top 5 most likely categories for their post's content. The app also recommends how the user might include a trigger or content warning. You can see an example of this in 3.11.

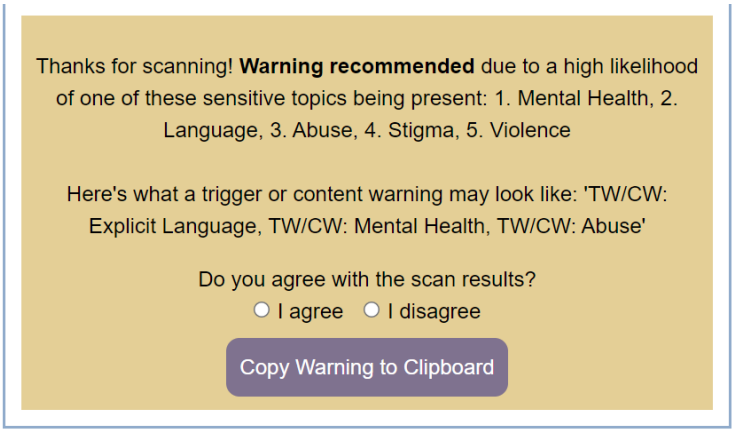


Figure 3.11: TWIST App Page 4 - Sensitive Content Detected Page

Similar to the "No Sensitive Content Detected" page, when the algorithm detected sensitive content, we asked the user to agree or disagree with this analysis. An example of the user selecting "I agree" and the "Post" button returning is shown in 3.12.

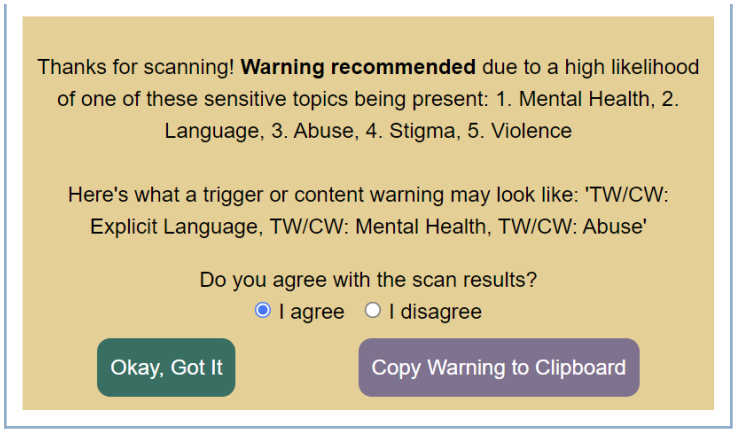


Figure 3.12: TWIST App Page 4 - Agree Radio Button Selected

If the user engaged with the TWIST app allowing it to scan, independent of whether sensitive content was detected or not, the user will then be routed to a "Thank You" page. This

"Thank You" page is shown in 3.13 and 3.14. 3.13 shows what the thank you page looks like after no sensitive content was detected. This one does not include a back button. 3.14 shows what the thank you page looks like when sensitive content is detected. This one does include a back button so the user can re-gain access to the trigger and content warning example, as well as the "Copy to Clipboard" button to save the warning to add to their post.

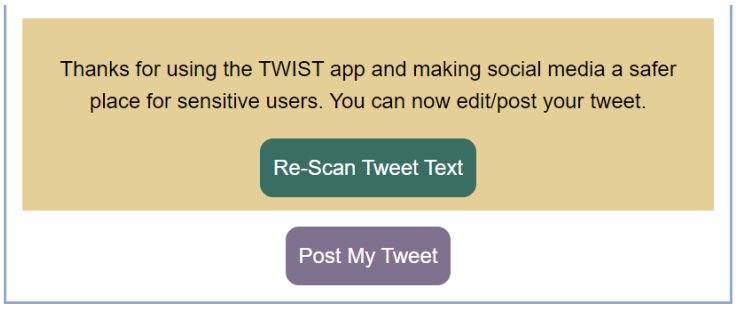


Figure 3.13: TWIST App Page 5 - Thanks Page After No Sensitive Content Detected

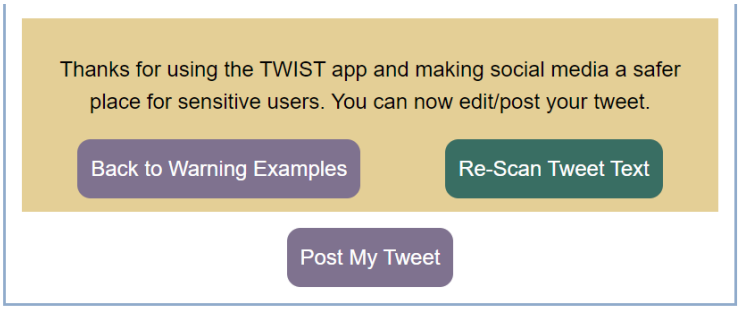


Figure 3.14: TWIST App Page 5 - Thanks Page After Warning Recommended

If the user decides to skip ahead and not scan their content, they are navigated to page 6, shown in 3.15. They can still make any changes and choose to scan it again or post their content with or without further editing. Just like the "Thank You" page (shown in 3.13 and 3.14, clicking post a final time will actually post the content, or in our study, it will just move on to the next tweet-writing prompt. 3.15 also includes a "Back" button if the user changes their mind and decides to scan their content.

There are a few places in the app where errors are possible. Some of these are where the

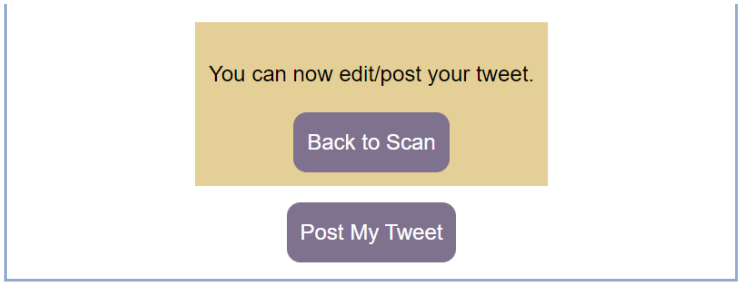


Figure 3.15: TWIST App Page 6 - You Can Edit/Post Again Page

tweet content is sent to OpenAI and a response is retrieved, and whenever a database save is done. While we tried to reduce any possibility of the users getting to the error page shown in 3.16, we intentionally included the error page in our design with the "Reset" button to keep their tweet text, while re-setting the rest of the TWIST app from the beginning.

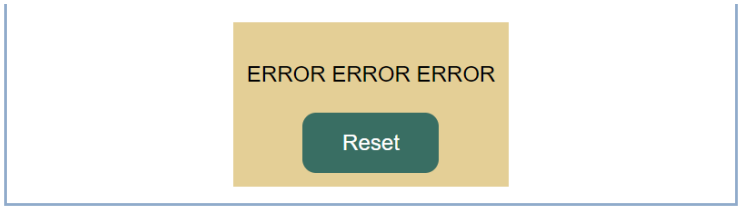


Figure 3.16: TWIST App Page 7 - Error Page

Page 8, shown in 3.17, was added later specifically for the user study. We wanted to ensure that the posts were long enough to get a more accurate OpenAI response, so we required that posts had to be at least 6 words long.

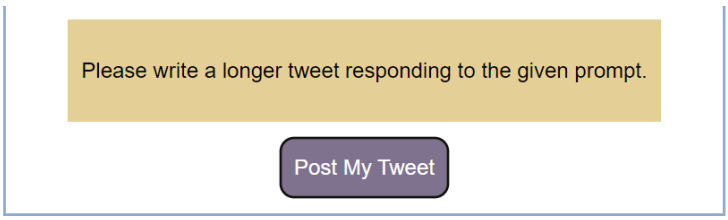


Figure 3.17: TWIST App Page 8 - Write More Page

At any point, no matter what page of the TWIST app the user is on, they can change their tweet content.

Chapter 4

Study Design

To answer our research questions, we conducted a 4-part study that used a mix of online tasks and surveys. It was conducted completely online from recruitment to consent to participation. The flow of the study is shown in 4.1.

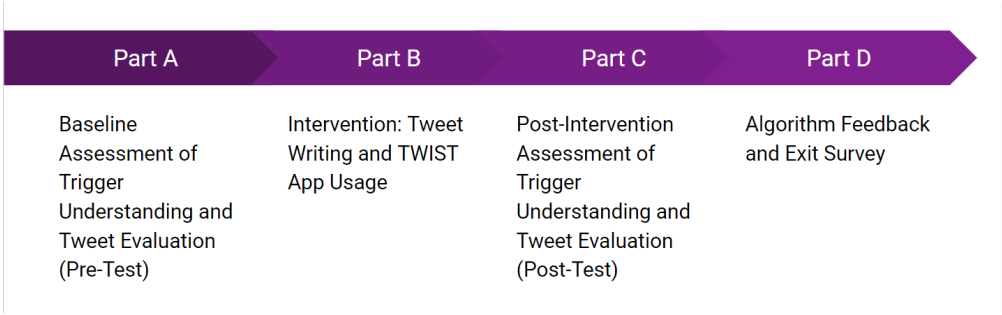


Figure 4.1: Study Flowchart

The first part (Part A) was a baseline assessment of their knowledge and understanding of trigger warnings and content warnings. Participants were asked to respond to statements about their familiarity with trigger and content warnings with answer choices on a 7-point Likert scale (Strongly disagree to Strongly agree). Next, the participants answered "Does this tweet need a trigger/content warning?" for 12 unique real-life tweets. Every time they responded "yes", an additional question was asked so they could specify how they would add such a warning. They would repeat these steps for 12 tweets. The next part of the study (Part B) was the intervention, where the participants were given a prompt and asked to write a tweet in response. As they attempted to post their tweet, they would interact

with the TWIST app we designed. After completing the 12 simulated post-writing tasks, they were navigated to a survey to answer the same questions they did at the beginning to see if their knowledge, understanding, and confidence in using trigger/content warnings had changed due to the intervention (Part C). Again, the participants were given a tweet and asked whether that tweet needed a trigger/content warning or not, similar to Part A. If they responded "yes", an additional question was asked of them to specify how they would add such a warning. They repeated these steps for 12 more tweets. We reviewed their change in performance and self-efficacy by comparing the pre and post-tests. Finally, the participants were asked to give feedback on the algorithm the TWIST app used as well as answer some exit survey questions (Part D). This was to gain their insight on the prior sections of the study and what they learned, how they would use a tool like this, etc.

4.1 Tweets Datasets Collection

For the pre and post-test performance in identifying when a tweet needed a content or a trigger warning, we created a dataset of real tweets. The Tweets datasets were created by finding actual tweets that had used a trigger warning or content warning on a text-based post. We tried to find a variety of tweets that contained warnings similar to the various top 12 triggers from the [12] paper. By finding tweets that already had a warning included, we trusted the authors' judgment that the topic being discussed could be triggering to other users. As we collected these 24 tweets, we assigned them each a category from the list of 12. While there was definite overlap in some tweets across multiple categories, the intention was to make sure each of the 3 datasets covered at least 8/12 sensitive topics. In addition to the 24 tweets that included sensitive content and warnings, we found 12 tweets that did not include warnings but still evoked an emotional response like anger, bittersweet, happiness,

love, optimism, and sadness. These acted as decoys in our dataset where the participant will likely agree these topics do not need a warning.

Figure 4.2 shows a screenshot of one original tweet with a TW/CW due to sensitive content and one tweet that did not include a TW/CW. All the selected tweets can be found in the appendix (see Appendix D).

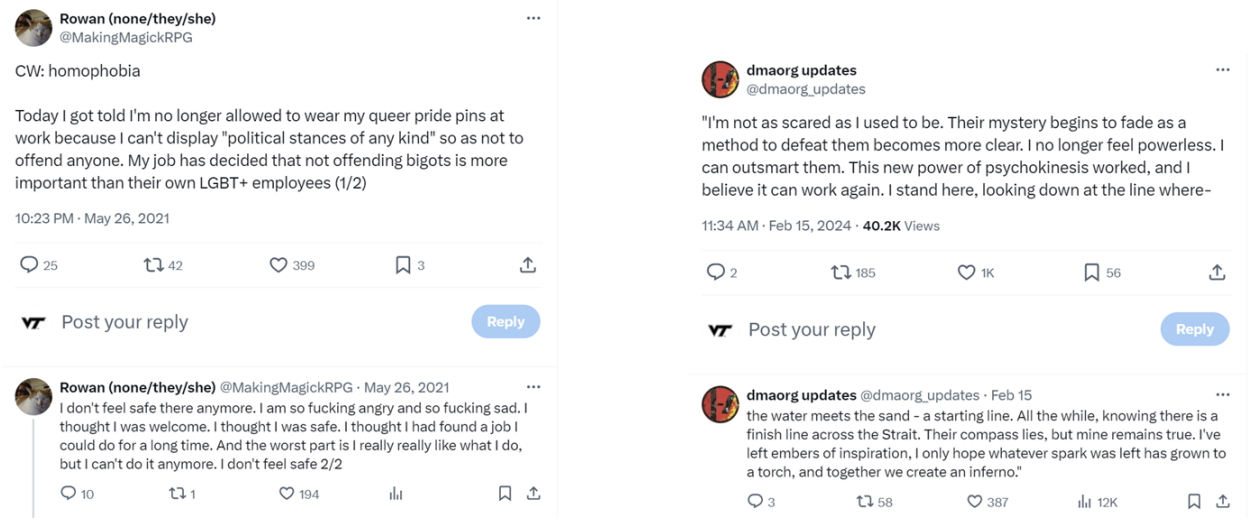


Figure 4.2: Two original tweets we used in our Tweets Dataset (a) one with a content warning and (b) one without a TW/CW or sensitive content

4.2 Scenario Prompts Creation

After we had collected the 36 real tweets to be used in our three tweets datasets used in parts A and C of the study, we needed to create the prompts for part B of the study where the participant would be asked to respond to a simulated post-writing task given a prompt. About 50% of the prompts were created by taking the tweets from the tweets dataset and creating a prompt that would evoke a similar tweet. For example, one of the tweets was "tw sv you would think that with an operating budget of $140,515,333, the london police would be able to have proper training about how to speak about sex assault and violence

against women without resorting to regressive talking points, and to hire a moderator for the live chat". We created a prompt based on this that was "You recently heard news about police officers using outdated and offensive language to talk publicly about sexual assault and sexual violence. Write a post that shows your outrage." The remaining prompts were created by trying to fill in any gaps so we could include as many topics as possible from [12]. The full prompt text for all 36 prompts can be found in the appendix (see Appendix E).

4.3 Methods

Since we had three parts of the study that included sensitive topics, and we wanted to be sure each of the 12 topics (from [12]) was covered twice for each participant, 8/12 topics would need to be covered in each of the three main parts (total 24 tweets with sensitive topics). Each participant would be randomly assigned the 3 datasets in a random order for Parts A, B, and C. Table 4.1 shows an example of how the participants could be assigned the three datasets across the three main parts of the study.

	Part A: Baseline Assessment	**Part B: Tweet Writing**	**Part C: Post Assessment**
Participant 1	Tweets DS1	Prompts DS2	Tweets DS3
Participant 2	Tweets DS1	Prompts DS3	Tweets DS2
Participant 3	Tweets DS2	Prompts DS1	Tweets DS3
Participant 4	Tweets DS2	Prompts DS3	Tweets DS1
Participant 5	Tweets DS3	Prompts DS1	Tweets DS2
Participant 6	Tweets DS3	Prompts DS2	Tweets DS1
Participant 7	Tweets DS1	Prompts DS2	Tweets DS3
Participant 8	Tweets DS1	Prompts DS3	Tweets DS2
...	...	...	...

Table 4.1: Participants Dataset Randomization

Table 4.2 shows the distribution of the 12 topics across the 3 datasets. 8 of the sensitive topics appeared in each dataset. Each dataset also has 4 decoys. The reason we included decoys is

so that participants would not assume every topic needed a trigger or content warning. We recognize that some topics can evoke emotions without necessarily being triggering, and we wanted to convey this to our study participants too.

Sensitive Topic	Dataset 1	Dataset 2	Dataset 3
1: Violence	X	X	
2: Sex		X	X
3: Stigma	X		X
4: Disturbing Content	X	X	
5: Language		X	X
6: Risky Behaviours	X		X
7: Mental Health	X	X	
8: Death		X	X
9: Parental Guidance	X		X
10: Crime	X	X	
11: Abuse		X	X
12: Sociopolitical	X		X
Decoy	X	X	X
Decoy	X	X	X
Decoy	X	X	X
Decoy	X	X	X

Table 4.2: Sensitive Topics and Decoys Distribution in Datasets of Study

4.3.1 Part A: Baseline Assessment of Trigger Understanding and Tweet Evaluation (Pre-Test)

The goal of this part of the study was to gain a baseline assessment of the participants' understanding of trigger and content warnings. Participants were asked to respond to 6 statements about their familiarity with trigger and content warnings. Each statement had answer choices on a 7-point Likert scale: Strongly disagree, Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Agree, Strongly agree. Figure 4.3 shows a screenshot of the part of the study where the participants answered questions about their

experience with trigger and content warnings.

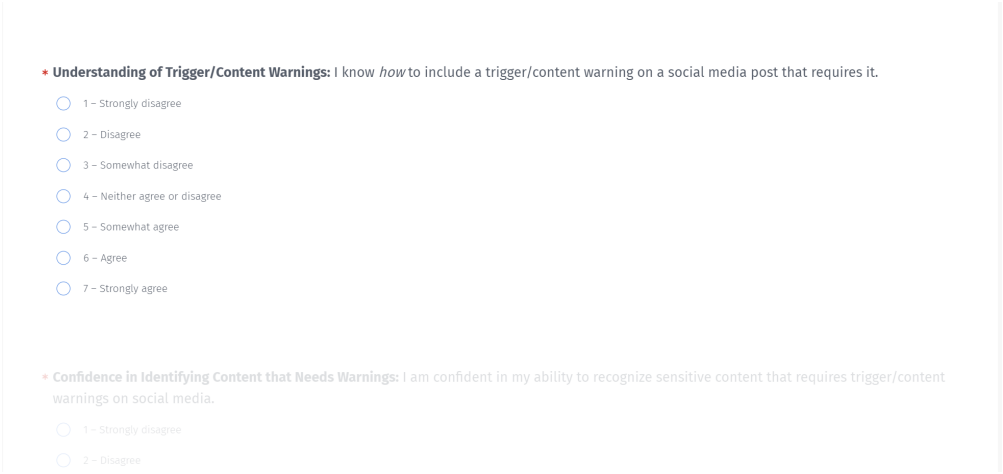


Figure 4.3: Study Part A: Baseline Assessment of Trigger/Content Warning Understanding

The statements the participants responded to were the following:

- **Awareness of Trigger/Content Warnings:** I am familiar with the concept of trigger/content warnings on social media platforms.
- **Understanding of Trigger/Content Warnings:** I understand *what topics* should include a trigger/content warning on a social media post.
- **Confidence in Identifying Content that Needs Warnings:** I am confident in my ability to recognize sensitive content that requires trigger/content warnings on social media.
- **Understanding of Trigger/Content Warnings:** I know *how* to include a trigger/-content warning on a social media post that requires it.
- **Perceived Effectiveness:** I believe that trigger/content warnings are effective in preventing people with traumatic experiences on social media from being exposed to sensitive content.

- **Willingness To Use:** I will begin or continue using trigger/content warnings when posting sensitive content on social media.

Next, the participants were given a tweet and asked whether that tweet needed a trigger/-content warning or not. If they responded "yes", an additional question was asked of them to specify how they would add such a warning. An example of the part of the study where the participant answered that tweets should include a trigger/content warning is shown in Figure 4.4. They would repeat these steps for 12 tweets and then begin the next part of the study.

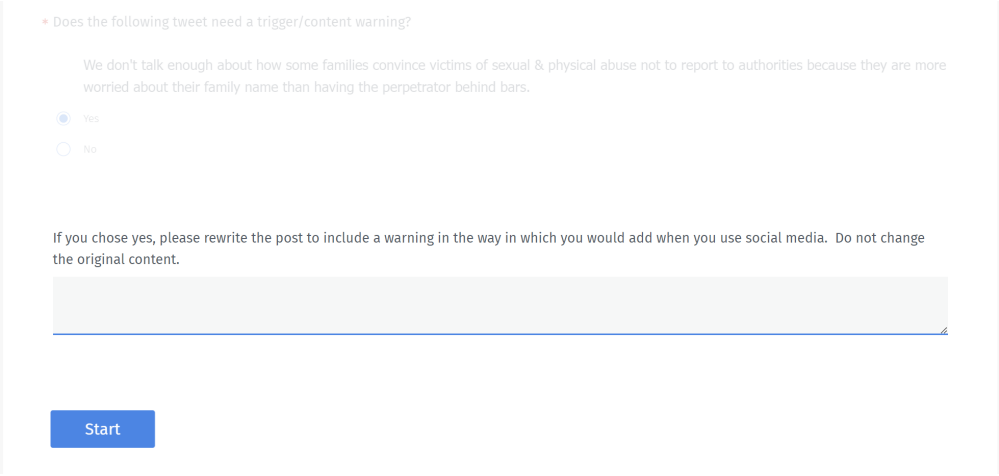


Figure 4.4: Study Part A: Tweet Evaluation (Pre-Test)

4.3.2 Part B: Intervention: Tweet Writing and TWIST App Usage

Participants were next navigated to use our TWIST web application and complete a series of 12 simulated post writing tasks. They started this part of the study by getting instructions which is shown in Figure 4.5.

In each task, participants were given context and then asked to write a post following that prompt. One example of the blank prompt page is shown in Figure 4.6, and a started

Figure 4.5: Study Part B: Tweet Writing and TWIST App Usage Start Page

response is shown in Figure 4.7. For each prompt writing task, the TWIST app would operate as explained in 3. The TWIST app always intervened, but the participant could decide to skip ahead to be able to post their tweet despite any sensitive content that was present.

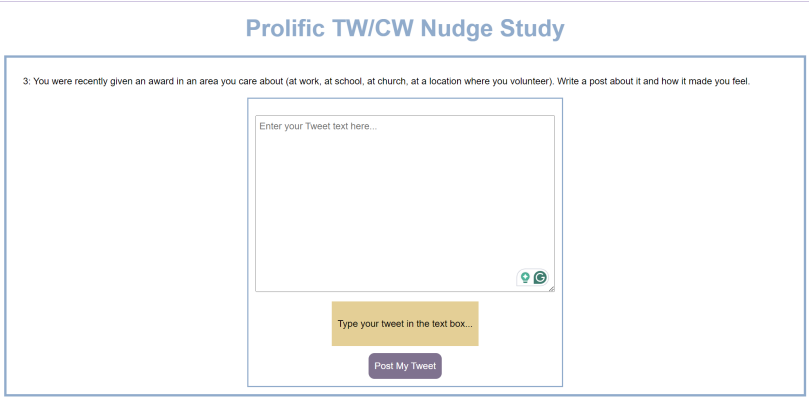


Figure 4.6: Study Part B Prompt Page

After the participant finished responding to the post-writing task, they would be navigated to a break page which is shown in 4.8. This included a progress update on how many of the prompts the participant had completed so far. This also gave the participants a break between prompts during the study.

By clicking the "Continue" button, the next prompt will appear for the participant to respond

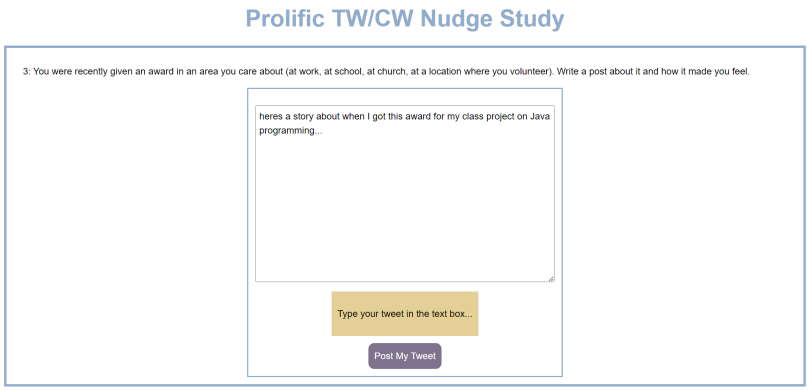


Figure 4.7: Study Part B Tweet Page

Figure 4.8: Study Part B Break Page

to. The creation of prompts is explained further in 4.2, and the full prompt text for all prompts across the 3 datasets can be found in the appendix (see Appendix E). Each of the 3 datasets contained 12 prompts with 8 expected to cover sensitive topics and 4 non-sensitive topics. Even though we anticipated 4 of the responses not containing sensitive content, participants could respond to the prompts in unexpected ways, which included sensitive content when we didn't expect it and vice versa. After the participant finishes their 12 prompt-writing tasks, they will be routed to the next part of the study.

4.3.3 Part C: Post-Intervention Assessment of Trigger Understanding and Tweet Evaluation (Post-Test)

This section was the same as Part A, however, the participants were assigned to a different dataset of questions when analyzing whether the tweets needed warnings this time. The change in performance and self-efficacy was measured by comparing the pre and post tests. As a reminder, a screenshot of this part of the study where the participants answered questions about their experience with trigger and content warnings is shown in Figure 4.3. An example of the part of the study where the participant answered tweets should include a trigger/content warning is shown in Figure 4.4.

4.3.4 Part D: Algorithm Feedback and Exit Survey

The goal of this part of the study was to conduct an exit survey where participants could respond to questions about whether they would use a tool like this and what improvements could be made to make the TWIST app more likely to be used. Participants were asked to respond to 7 statements with answer choices on a 7-point Likert scale: Strongly disagree, Disagree, Somewhat disagree, Neither agree or disagree, Somewhat agree, Agree, Strongly

agree. Figure 4.9 shows a screenshot of the part of the study where the participant answered the exit survey questions.

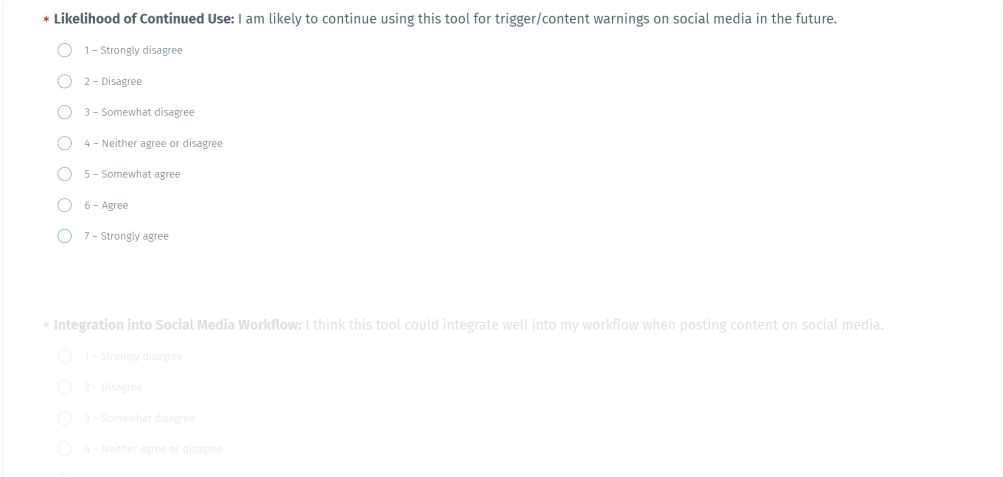


Figure 4.9: Study Part D: Algorithm Feedback and Exit Survey

These questions have response options on a 7-point Likert scale from 'Strongly Agree' to 'Strongly Disagree':

- **Familiarity with LLMs/OpenAI/ChatGPT:** I am familiar with Large Language Models, such as ChatGPT or OpenAI.
- **Ease of Use:** I found it easy to use the tool to generate trigger/content warnings for social media posts.
- **Perceived Effectiveness:** I believe the tool effectively identifies content that may require trigger/content warnings.
- **Integration into Social Media Workflow:** I think this tool could integrate well into my workflow when posting content on social media.
- **Opinion on Backend Algorithm:** I feel the algorithm used to generate warnings was effective in showing me how and when to add a warning.

- **Likelihood of Continued Use:** I am likely to continue using this tool for trigger/-content warnings on social media in the future.

- **Overall Satisfaction:** I am satisfied with the tool provided to assist with trigger/-content warnings on social media.

4.4 Recruitment

For the user study, we recruited participants from Prolific, an online survey platform. We originally intended to recruit 100 participants, but 12 did not complete all the required sections. We were able to analyze the results from 88 participants.

4.5 Participants and Eligibility

We used five eligibility criteria to recruit our participants:

1. they were over the age of 18

2. they are fluent in English

3. they currently reside in the United States

4. they used Twitter more than 3 times over the past 12 months

5. they were willing to participate in a study that involved potentially triggering content in text (No images included)

Individual-Level Variables	User Count (N)	Percentage (%)	Mean	Std. Dev
Age	88		36.22	9.87
18-24	8	9.1%		
25-34	33	37.5%		
35-44	29	33.0%		
45-54	11	12.5%		
55+	5	5.7%		
Unknown	2	2.3%		
Sex				
Male	49	55.7%		
Female	37	42.0%		
Unknown	2	2.3%		
Race				
White	54	61.4%		
Black	14	15.9%		
Asian	6	6.8%		
Mixed	8	9.1%		
Other	4	4.6%		
Unknown	2	2.3%		
Tweeting frequency (in the past 12 months)				
4-20 times	40	45.5%		
20-100 times	25	28.4%		
More than 100 times	21	23.9%		
Unknown	2	2.3%		

Table 4.3: Demographics Breakdown for the User Study

4.5.1 Participant Demographics

In Table 4.3, you can see the demographic breakdown of our 88 participants. (Note: Two participants did not complete demographic information.)

4.6 Analysis

To analyze our results, we utilized both the quantitative and qualitative data we collected. We compared both the pre and post-intervention data to see if the performance of the participants reviewing tweets for sensitive content improved or declined. While we hypothesized

their recall would improve, we were unsure if their perception of their own performance would increase or decrease. To find out, we compared the responses to the 6 self-efficacy questions they answered in Part A to their same responses in Part C.

From the use of the TWIST app section (Part B of the study), we looked at whether participants selected to scan their content or not, whether they started the prompt with a warning even before intervention or if they chose to add one after the use of the app, whether they agreed or disagreed with the OpenAI response on whether sensitive content was detected, whether they added a warning or re-phrased their content based on the OpenAI response in the TWIST app, and what their warning looked like (if they chose to add one).

In Part D of the study, we analyzed the usability, effectiveness, workflow integration capability, backend algorithm strengths, and overall satisfaction of the TWIST app. These results helped us to learn where the TWIST app should be improved or changed if it is to be adapted into a Chrome extension or even further as part of a social media platform.

Chapter 5

Results

To answer our first research question, "How does nudging social media posters influence their knowledge and confidence in determining the inclusion of trigger warnings in their posts?", we analyzed the change in performance between the pre and post intervention review of tweets and responding to whether each post needed a warning or not. I calculated precision, recall, and F1 scores. We wanted to see if people just became more cautious overall, leading to a higher false positive rate after the intervention, or if participants truly understood when to add and when not to add a warning. All 3 of these values had a significant change after the intervention, with precision decreasing while recall and F1 score increased.

To answer our second research question, "How does nudging social media users to add TW/CW change their perceptions on TW/CW related topics?", we reviewed the participants' perceptions of whether they felt they better understood what topics required a warning, how to add a warning, and their likelihood of adding warnings in the future to their own sensitive posts both before and after using the TWIST App. While only 1/6 of these self-efficacy questions had a significant change, participants did agree more strongly with the statement "I know how to include a trigger/content warning on a social media post that requires it." after the intervention. To answer our third research question, "How can LLMs be used to detect TW/CW-related topics and to what extent does it agree with human annotators?", we looked at both the TWIST App usage data as well as the user feedback given on the app in Part D of the study.

5.1 [RQ1] Participants' Performance Change Pre vs Post Intervention

5.1.1 [H1] Awareness Change Pre vs Post Intervention

Part of the user study assessed the effectiveness of an intervention aimed at educating users on trigger and content warnings, measured through pre-intervention and post-intervention evaluations. I calculated precision, recall, and F1 scores to see if people just became more cautious overall, leading to a higher false positive rate after the intervention, or if participants truly understood when to add and when not to add a warning. Precision slightly decreased after the intervention, while recall and F1 score both increased. The decline in precision was due to users acting more cautiously saying tweets that didn't need a warning did need one. However, the precision and F1 scores explained above show that while false positives did rise, the growth in true positives outweighed this. All results were statistically significant and these values are shown in Figure 5.1.

Precision measures the accuracy of positive predictions. In this context, a precision of 90.5% indicates that nearly all of the tweets that they chose that they thought it needed TW/CW the ones were sensitive topics.

Recall quantifies the ability of the participants to identify all relevant instances (e.g., correct responses) within the dataset. A recall of 47.7% suggests that approximately 47.7% of sensitive topics were correctly identified by the participants.

The F1 score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. An F1 score of 58.8% in Part A indicates a reasonable balance between precision and recall for this segment of the study.

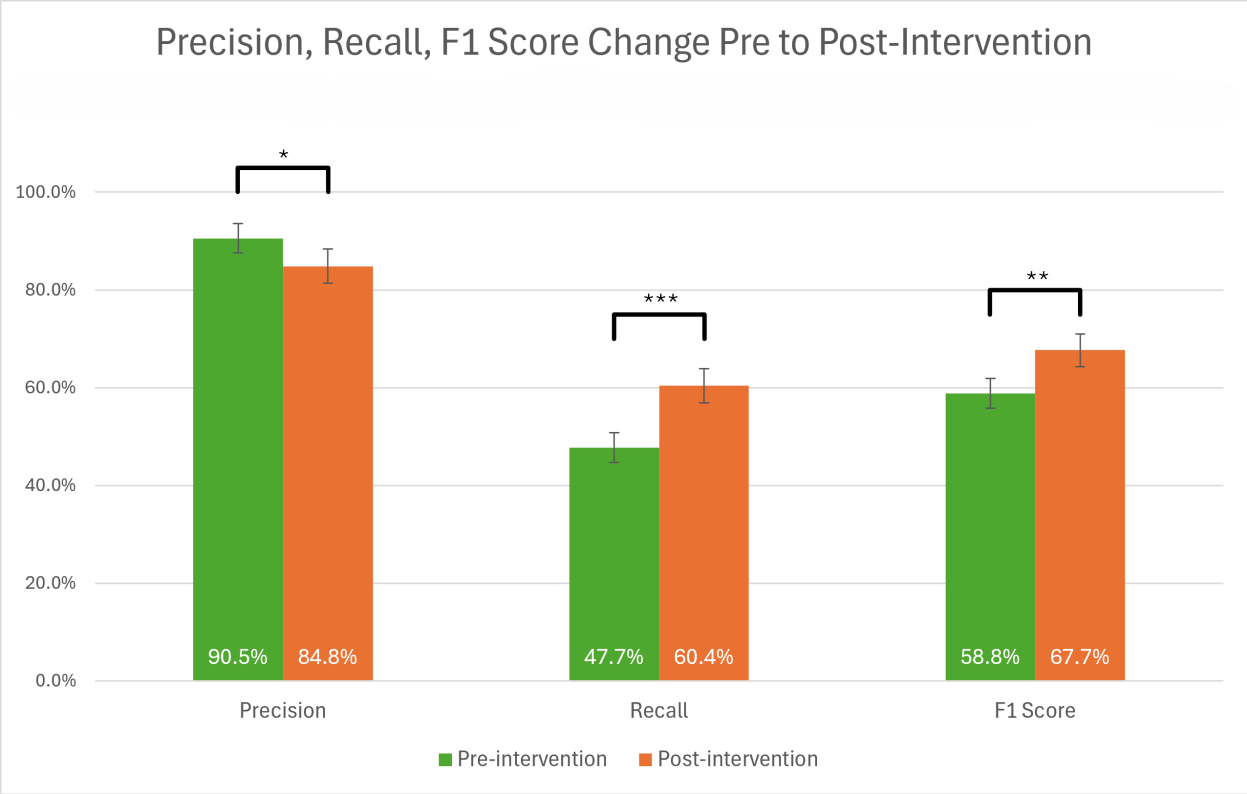


Figure 5.1: Change in Precision, Recall, and F1 Score Pre vs Post-Intervention

We also calculated the post-intervention precision, recall, and F1 score. Similar to Part A, the precision of 84.8% in Part C suggests a high level of accuracy in the positive predictions made by the participants. However, this decrease from Part A signifies that participants acted more cautiously leading to a rise in false positives.

The recall of 60.4% in Part C signifies an improvement over Part A, indicating that a greater proportion of relevant instances were successfully identified by the participants in this segment.

With an F1 score of 67.7% in Part C, there is a notable enhancement in the balance between precision and recall compared to Part A, highlighting an overall improvement in the participants' performance after using the TWIST App for the tweet writing part.

We chose to look specifically at the 12 sensitive topics to see if any of the sensitive topics had a greater change from the pre-intervention recall to post-intervention recall. Table 5.1 shows the 12 topics in the same order as presented by [12] where "1: Violence" was the most common sensitive topic in their results and "12: Sociopolitical" was the least common of their text-based sensitive topics. Since every participant didn't encounter every topic in both the pre-test and post-test, we could not run a paired test to determine statistical significance for the recall change.

Interestingly, our results on which sensitive topics were the most familiar (and therefore highest recall in Part A) do not align with the ordering Charles et al. [12] had. 4: Disturbing Content had the highest recall in the initial part of the study, with 10: Crime following in second place. The most commonly missed topics were 12: Sociopolitical and 9: Parental Guidance.

As we reviewed growth from Part A to Part C, 12: Sociopolitical and 3: Stigma had the highest growth in understanding that these topics require trigger/content warnings.

	Sensitive Topic	Pre-Intervention Recall	Post-Intervention Recall	Change from Pre to Post Intervention
1	Violence	45.61%	49.15%	7.76%
2	Sex	25.86%	46.77%	**80.86%**
3	Stigma	27.87%	56.36%	**102.25%**
4	Disturbing Content	91.23%	86.44%	-5.25%
5	Language	56.90%	64.52%	13.39%
6	Risky Behaviours	40.98%	63.64%	**55.27%**
7	Mental Health	71.93%	61.02%	-15.17%
8	Death	58.62%	72.58%	23.81%
9	Parental Guidance	24.59%	43.64%	**77.45%**
10	Crime	78.95%	79.66%	0.90%
11	Abuse	39.66%	62.90%	**58.63%**
12	Sociopolitical	16.39%	34.55%	**110.73%**

Table 5.1: Sensitive Content Recall Change Before and After Using TWIST App

5.1.2 [H2] Self Efficacy Change Pre vs Post Intervention

While the change in determining whether tweets needed a trigger or content warning was one measured factor, we also chose to look at how participants' understanding of trigger and content warnings changed. Participants were asked to respond to 6 statements about their familiarity with trigger and content warnings both before and after their use of the TWIST App:

- **Awareness of TW/CW:** I am familiar with the concept of trigger/content warnings on social media platforms.
- **Understanding of TW/CW (Topic):** I understand *what topics* should include a trigger/content warning on a social media post.
- **Confidence in Identifying Content that Needs TW/CW:** I am confident in my ability to recognize sensitive content that requires trigger/content warnings on social media.
- **Understanding of TW/CW (Method):** I know *how* to include a trigger/content warning on a social media post that requires it.

- **Perceived Effectiveness:** I believe that trigger/content warnings are effective in preventing people with traumatic experiences on social media from being exposed to sensitive content.
- **Willingness To Use:** I will begin or continue using trigger/content warnings when posting sensitive content on social media.

Each statement had answer choices on a 7-point Likert scale: Strongly disagree, Disagree, Somewhat disagree, Neither agree or disagree, Somewhat agree, Agree, Strongly agree.

Figure 5.2 shows the degree to which the participants agreed with each of these statements before the intervention as well as the degree to which the participants agreed with each of these statements after the intervention.

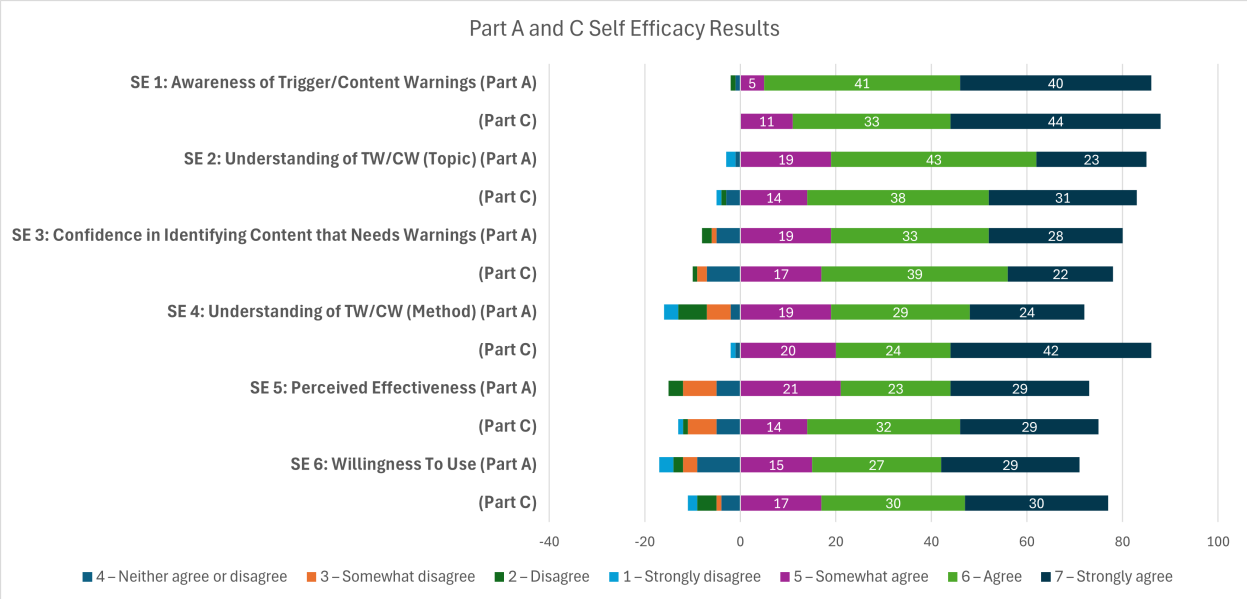


Figure 5.2: Self Efficacy Results Pre and Post-Intervention (p-value for SE4 < 0.001)

As you can see in Figure 5.2, our participants already had some basic familiarity with trigger and content warnings. They felt they understood when to add a warning, how to add a warning, and what topics may require a warning. Surprisingly to us, they also said

they felt TW/CW are effective in preventing people with traumatic experiences on social media from being exposed to sensitive content and they were likely to add warnings to their own content when required. In Part C, when we asked the same self-efficacy questions, there were some small shifts where people switched from somewhat agree to agree or agree to strongly agree. The calculated growth from the pre-intervention to post-intervention perceptions is shown in Table 5.2. The only change with recognized statistical significance was the statement "Understanding of Trigger/Content Warnings: I know how to include a trigger/content warning on a social media post that requires it." The mean grew from 5.398 (SD = 1.644) to 6.170 (SD = 1.008), which is a 36.4% growth ($p < 0.001$).

	Part A Mean	Part A Standard Deviation	Part C Mean	Part C Standard Deviation	Growth
SE 1: Awareness of TW/CW	6.330	0.798	6.375	0.700	2.7%
SE 2: Understanding of TW/CW (Topic)	5.909	1.046	6.023	1.061	3.6%
SE 3: Confidence in Identifying Content that Needs Warnings	5.864	1.106	5.784	1.055	0.6%
SE 4: Understanding of TW/CW (Method)	**5.398**	**1.644**	**6.170**	**1.008**	**36.4%**
SE 5: Perceived Effectiveness	5.602	1.386	5.750	1.324	5.5%
SE 6: Willingness To Use	5.591	1.513	5.727	1.436	9.9%

Table 5.2: Change in User Perceptions of Trigger and Content Warnings Before and After Using the TWIST App

5.2 TWIST App Usage Results

To classify the 1056 different paths through responding to the prompts in Part B and using the accompanying TWIST App, we labeled 6 cases where each prompt could lead the participant to take a different path through their usage of the app. We defined these 6 cases as follows:

- Case 1: Added TW/CW initially
- Case 2: Scanned, but TW/CW was not needed
- Case 3: Added TW/CW as suggested by GPT

- Case 4: Did not scan, and TW/CW not needed
- Case 5: TW/CW was suggested, but no TW/CW was added
- Case 6: Did not scan even though TW/CW was suggested

The flowchart in Figure 5.3 shows the 6 cases, and these are ranked where Case 1 is the best case scenario and Case 6 is the worst case scenario.

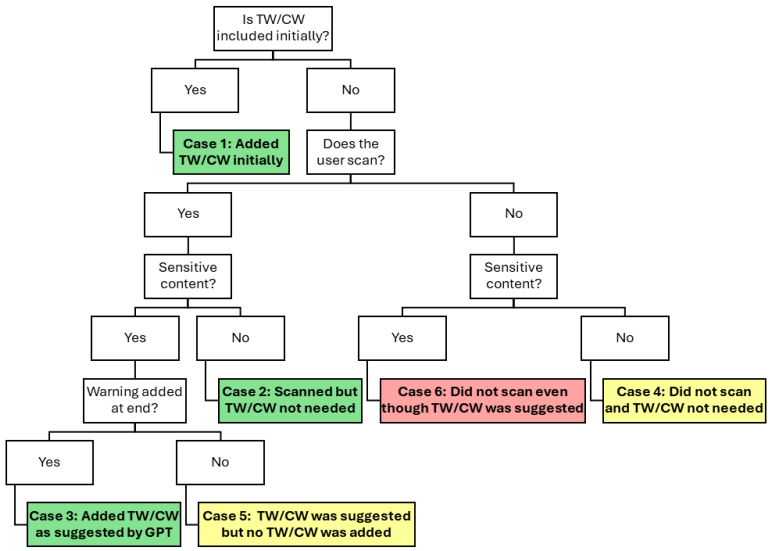


Figure 5.3: Flowchart of the TWIST App Cases Based on 1056 Responses

We considered that behaviors presented in Cases 1, 2, and 3 are desirable since they scanned their content to detect sensitive content or added a warning initially/when suggested.

Cases 4 and 5 could be less desirable, colored in yellow, since the participants chose not to add a warning when one was recommended or they chose not to scan, respectively. However, more information would need to be considered for these two cases. With Case 4, the participant could have intentionally chosen not to scan because they knew their content would not need a warning. It's also possible the participant decided not to scan without any thought into whether that content was triggering and got lucky that sensitive content was not present.

As for Case 5, the participant could have rephrased their content so a warning was no longer needed, or the AI could have incorrectly determined that a warning was needed; therefore, these are not necessarily bad actors.

Case 6 was when a participant wrote a tweet containing sensitive content and chose not to scan it, but sensitive content was present. This is why we classified this group as bad actors, colored in red.

Next, I reviewed which cases were most common depending on the sensitive topic. This is shown in Figure 5.4. We aligned the graph between Cases 4 and 5 to illustrate a distinction: bad actors appear on the left side of the center line, while good actors are mostly on the right side. As previously mentioned, both Cases 4 and 5 may include individuals exhibiting both positive and negative behaviors. However, we observed that Case 5 participants typically engage without due regard for sensitive topics in their posts, whereas Case 4 participants generally consider the AI's recommendation but opt not to add a warning for reasons other than simply not wanting to do so.

5.3 How Did LLM Perform Compared to Human Annotators (Inter-Rater Reliability)?

Since the AI reviewing the tweet content can make mistakes, we wanted to determine inter-rater reliability (IRR) between the LLM and human reviewers of the tweets. We had two human annotators review 939 responses to the prompts in Part B of the study that were reviewed by the LLM (939 = 1056 - 91 where a warning was already added - 26 error cases) and then calculated the inter-rater reliability for different comparisons. The IRR for the two human annotators was 0.7213. The IRR for the LLM and human annotator 1 was 0.6477.

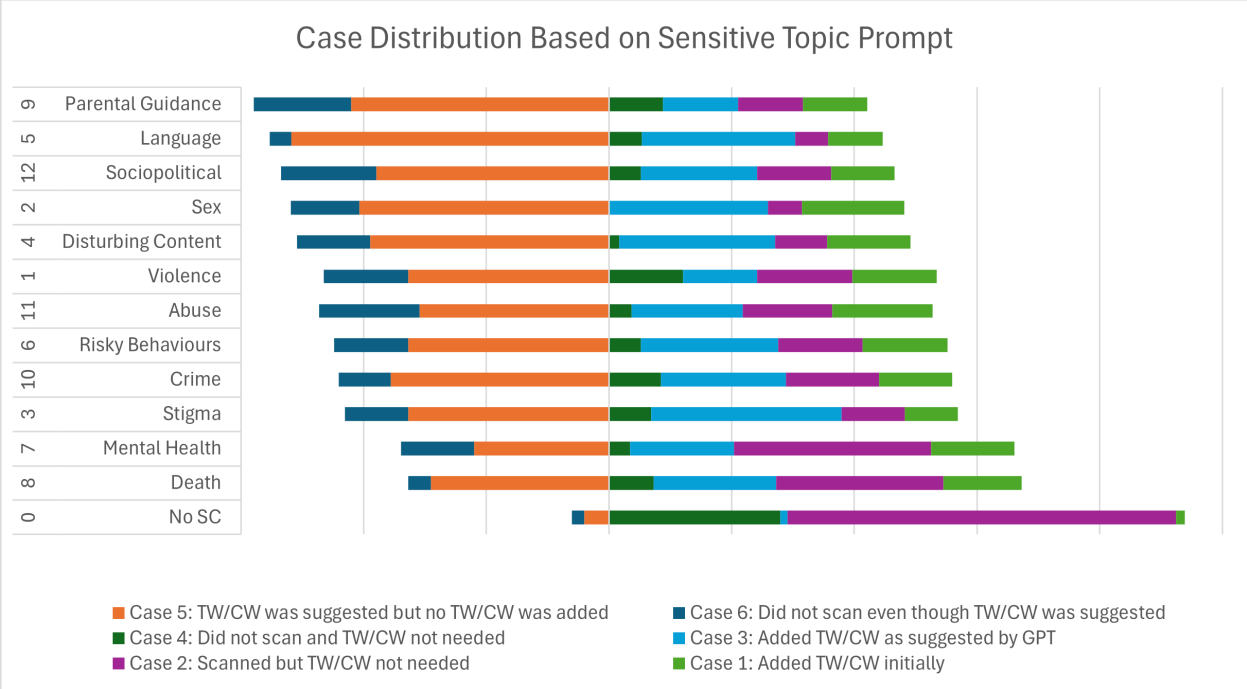


Figure 5.4: Case Distribution Based on Sensitive Topic Prompt (Number on left is from Charles et al. topic ranked list [13])

The IRR for the LLM and human annotator 2 was 0.6798. In all of these comparisons, there is 'substantial agreement' ($0.6 <$ kappa value (k) < 0.8) [26], with the LLM comparisons only performing slightly lower than the two human moderators' comparison. The LLM recommended a warning for 53% of cases while the human annotators, H1 and H2, recommended a warning 65% of the time and 59% of the time respectively. The LLM was actually more conservative than we expected in its decision to recommend a warning due to sensitive content. It would be interesting to learn more about how this varied based on the topic and if it acted more cautiously when certain topics were involved as compared to others. These IRR scores along with the ratio of TRUE/FALSE for each reviewer agrees with prior work showing that triggering topics are nuanced, and it is difficult to say definitively whether a certain post needs a warning [19].

5.4 Algorithm Feedback Results

The goal of Part D of the study was to conduct an exit survey where participants could respond to questions about whether they would use a tool like this and what improvements could be made to make the TWIST app more likely to be used. Participants were asked to respond to 7 statements with answer choices on a 7-point Likert scale.

- **Familiarity with LLMs/OpenAI/ChatGPT:** I am familiar with Large Language Models, such as ChatGPT or OpenAI.
- **Ease of Use:** I found it easy to use the tool to generate trigger/content warnings for social media posts.
- **Perceived Effectiveness:** I believe the tool effectively identifies content that may require trigger/content warnings.
- **Integration into Social Media Workflow:** I think this tool could integrate well into my workflow when posting content on social media.
- **Opinion on Backend Algorithm:** I feel the algorithm used to generate warnings was effective in showing me how and when to add a warning.
- **Likelihood of Continued Use:** I am likely to continue using this tool for trigger/-content warnings on social media in the future.
- **Overall Satisfaction:** I am satisfied with the tool provided to assist with trigger/-content warnings on social media.

Figure 5.5 shows the responses to the multiple-choice questions.

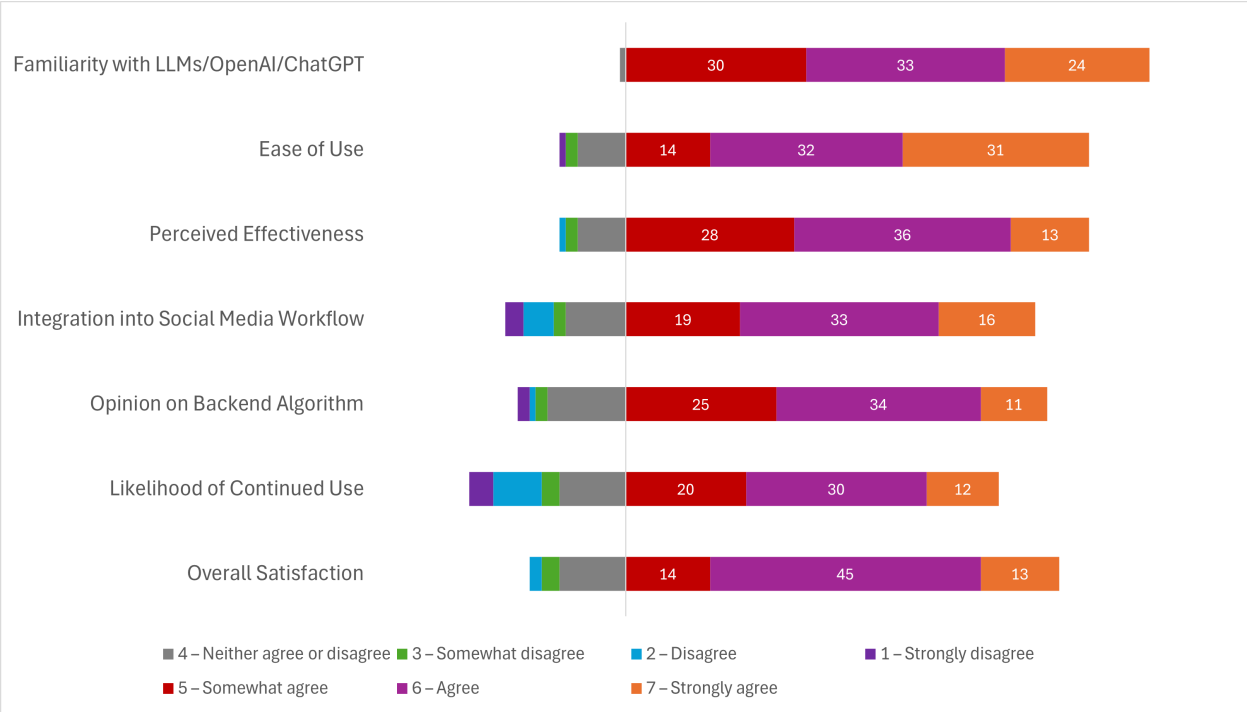


Figure 5.5: Algorithm Feedback Results

Chapter 6

Discussion

6.1 Comparative Analysis and Implications of the TWIST App's Impact on Trigger Warning/Content Warning (TW/CW) Awareness

Our study shows that the TWIST App effectively increases users' objective understanding of TW/CW and the topics that necessitate them, as evidenced by the pre and post-intervention surveys. This finding aligns with existing literature that underscores the role of digital interventions in promoting awareness on sensitive topics [5, 14]. However, an intriguing observation from our study is the lack of significant change in users' subjective perception of their awareness. This phenomenon could be attributed to the Dunning-Kruger Effect [15], where individuals with limited knowledge overestimate their understanding of a concept. As our participants gained more knowledge about TW/CW through the TWIST App, they might have concurrently realized the depth and complexity of the subject matter, leading to a perceived plateau in their understanding [19]. This unexpected outcome provides valuable insights into the cognitive processes involved in learning about sensitive topics and highlights the need for designing interventions that address both objective and subjective aspects of awareness. Our findings contribute to the broader understanding of the topic by demonstrating the potential of digital interventions like the TWIST App in raising awareness about

TW/CW. However, they also underscore the complexity of changing perceptions, extending the existing results [19] by revealing the gap between objective knowledge and subjective perception. Given these findings, there is a clear call for further research to explore strategies that can effectively bridge this gap. Future projects could focus on developing features within the TWIST App that not only educate users but also help them accurately gauge their understanding. This may include a feedback measure as they start adding warnings without being nudged that gives the user metrics on how they are progressing. The observed discrepancy between objective knowledge and subjective perception also connects to previous literature, which suggests that individuals often struggle to accurately assess their understanding of complex topics [16, 32]. This provides some explanation for our work and emphasizes the need for interventions that address this cognitive bias.

6.2 Variations in Awareness Levels Based on Triggering Topic

Our study reveals that awareness levels vary significantly across different triggering topics. Certain topics, such as Disturbing Content (91.23%) and Crime (78.95%), were easily recognized by users as requiring a TW/CW, even without the aid of the TWIST App. This finding is consistent with existing literature that identifies these topics as commonly associated with trigger warnings [18, 31]. Conversely, we identified topics such as Sociopolitical (16.39% in Part A) and Parental Guidance (24.59% in Part A) that had low initial awareness levels. Despite improvements following the intervention, awareness levels for these topics remained relatively low (under 50% in Part C). This suggests that these topics may be less intuitive or familiar to users, highlighting an area for further research and intervention development.

When comparing our results to the study by Charles et al. [12], which provided the 12 sensitive topics for our study, we observed a difference in the ordering of topics. The order in our study was not chronological (1-12), but rather based on the level of awareness. For Part A, the order was: 4, 10, 7, 8, 5, 1, 6, 11, 3, 2, 9, 12, and for Part C, it was: 4, 10, 8, 5, 6, 11, 7, 3, 1, 2, 9, 12. Notably, the categories of Mental Health (7) and Violence (1) changed their positions in the sorted order, indicating that the growth in awareness was not as prominent for these two categories as it was for the other ten. This discrepancy between our findings and the frequency-based order in Charles et al. [12] suggests that the difficulty in recognizing the need for a TW/CW does not necessarily correlate with the frequency of exposure to each topic. This insight extends the existing literature by highlighting the complex interplay between exposure, recognition, and awareness in the context of TW/CW [18, 19].

6.3 Understanding the Reason Behind Omitting Trigger Warnings/Content Warnings (TW/CW)

We posit several hypotheses to explain why users might choose not to add warnings even when the TWIST App identifies sensitive content.

One hypothesis is that users acknowledge the sensitivity of the content but deem the dissemination of the information as crucial for raising awareness on the topic. This is consistent with the findings of Gupta et al. [19], which highlight instances where sensitive content, such as suicide risk resources, is shared without warnings to ensure maximum reach and visibility. Another example is the sharing of information on political or civil unrest, which, despite its potential to trigger some viewers, is considered necessary due to attempts by

mainstream media to suppress such content. These scenarios underscore the complex ethical considerations involved in the use of TW/CW, where the perceived benefits may outweigh the potential risks.

Another group of users, referred to as the 'rephrasing' group in our study, may choose to modify their language to reduce the triggering potential of the content, thereby eliminating the need for a warning. This raises interesting questions for future research, such as whether rephrasing effectively removes triggering content or merely masks it.

Lastly, some users may disagree with the Language Learning Model's (LLM) suggestion to add a warning, perceiving the content as mild and not warranting a warning. This discrepancy could stem from inaccuracies in the AI model or differences in personal beliefs about what constitutes triggering content. This situation presents a complex challenge in determining the 'correct' approach, given that the LLM aims to educate users, but its training is ultimately influenced by human input.

These findings extend the existing literature by highlighting the nuanced factors influencing the use of TW/CW and underscore the need for further research to better understand these dynamics and inform the development of more effective and user-centric digital interventions.

6.4 Leveraging AI for Social Good: A Dual Approach

The question arises, 'Is a Language Learning Model (LLM) an effective tool for managing sensitive topics?' In addressing this, we identify two distinct strategies for employing AI in the context of sensitive content review.

The first strategy, termed the 'detection approach', is primarily utilized for content moderation or viewer-side interventions [31]. This approach focuses on identifying and filtering

sensitive content to protect viewers from potential triggers.

The second strategy, the 'empowering approach', is proactive and targeted at content creators [30]. This approach has significant educational benefits as it provides social media posters with insights into triggering topics. Importantly, the knowledge gained through this approach extends beyond social media contexts, contributing to a broader understanding of sensitivity in communication. This proactive use of AI can complement existing viewer-side interventions, fostering safer online spaces for all users [5, 30]. An example of this is the *TransTime* application, which offers a tailored solution for the trans-community.

However, these strategies raise complex questions about the balance between content moderation and freedom of expression, the accuracy of AI in detecting sensitive content, and the ethical considerations involved in managing online content. These issues underscore the need for ongoing research and dialogue in this field, as well as the development of AI models that are both effective and respectful of users' diverse experiences and perspectives. Our findings contribute to the broader discourse on the use of AI for social good, highlighting the potential of LLMs in promoting sensitivity and inclusivity in online spaces. They also call for further research to refine these approaches and explore their implications in different contexts.

6.5 Self Reflection

Throughout this project, I learned even quantitative studies can be emotionally exhausting when it comes to triggering material like this study. While none of the 12 topics we focused on were personally triggering to me, reading them over-and-over was draining as I built the surveys, checked the data was being recorded correctly, analyzing results, etc. This made me appreciative of the online spaces where there are still human moderators filtering sensitive

content. I also believe more needs to be done to support the mental health of HCI researchers when conducting studies with triggering topics. Other researchers have looked into how the mental health of researchers studying sensitive topics can be affected as well [36].

6.6 Limitations and Future Work

One limitation of this work is that we were not able to analyze the rephrasing of tweets. Rephrasing the tweet content could make it no longer require a warning, but we only analyzed the detection of sensitive content on the first pass through, not after rephrasing could have occurred. We did collect the data on this, so we could look at this in a future work, but so far, we have not manually reviewed all the tweets.

Another limitation is that LLMs are sometimes wrong. While our inter-rater reliability with human annotators showed the LLM did agree most of the time, we are unsure what type of impact this may have had on the participants in their decision-making process. We need to look closer at the cases LLM was incorrect, but how can we do this objectively (human reviewing LLM reviewing human)?

Another limitation is that the selected tweets and our designed prompts could have impacted the results. Certain categories are more clearly defined than others, like Violence can be clearly found to either be included in a tweet or not. A more challenging topic was Parental Guidance which could be why this was a low-performing group for recall and high disagreement as shown by many of the participants not adding a warning to their final post when this was flagged. If this study were to be repeated, a norming study could be conducted to ensure all the prompts and tweets have relatively similar severity so as not to impact the other results.

This project focused on a poster-side intervention through the use of a nudge algorithm. However, poster-side interventions need to be combined with viewer side interventions for best results as neither should operate in isolation.

Another future work would be to take the existing TWIST App, make any improvements, then convert this to a Chrome extension for field testing. Hopefully, this type of nudge algorithm could then be built into social media platforms themselves and connected with viewer-side interventions like selective filtering.

Chapter 7

Conclusions

Our key findings from this study include (1) Nudging social media users to add TW/CW educates them on triggering topics and raise their awareness when posting in the future, (2) Social media users can learn how to add a trigger/content warning through using a nudge app, (3) Researchers grew in understanding of how a nudge algorithm like TWIST can change people's behavior and perceptions, and (4) We provide empirical evidence of the effectiveness of such interventions (even in short-time use). We also feel that combining this poster-side intervention with viewer-side interventions could lead to a more effective result overall.

Chapter 8

Summary

This study shows that a nudge algorithm implemented either as an external web extension or ideally built into social media platforms themselves and connected to viewer-side capabilities could make social media safer for sensitive users. We conducted this study to determine if poster-side interventions such as a nudge algorithm to add warnings to sensitive posts would increase social media users' knowledge and understanding of how and when to add trigger and content warnings. To investigate the effectiveness of a nudge algorithm, we designed the TWIST (Trigger Warning Includer for Sensitive Topics) app. The TWIST app scans tweet content to determine whether a TW/CW is needed and if so, nudges the social media poster to add one with an example of what it may look like. We then conducted a 4-part mixed methods study with 88 participants. Hypothesis 1, "Nudging social media users to add TW/CW increases their ability to recognize topics that may be triggering to readers." was supported. The participants grew in their knowledge of when and how to use TW/CW when posting about sensitive content as evidenced by a 50% increase in the participants' recall in analyzing tweets for sensitive content after our intervention when compared to before. The participants' precision decreased by 10% showing they began to act more cautiously as they classified non-sensitive content as needing a warning, however, the F1 score rose by 23% signifying the improved classification of sensitive topics outweighed any misclassification of non-sensitive content. We also reviewed the participants' perceptions of whether they felt they better understood what topics required a warning, how to add a warning, and their

likelihood of adding warnings in the future to their own sensitive posts both before and after using the TWIST App. Hypothesis 2, "Nudging social media users to add TW/CW increases their perceived understanding TW/CW practice on social media." was partially supported with significant growth in one question. While only 1/6 of these self-efficacy questions had a significant change, participants did agree more strongly with the statement "I know how to include a trigger/content warning on a social media post that requires it." after the intervention. Our key findings from this study include (1) Nudging social media users to add TW/CW educates them on triggering topics and raise their awareness when posting in the future, (2) Social media users can learn how to add a trigger/content warning through using a nudge app, (3) Researchers grew in understanding of how a nudge algorithm like TWIST can change people's behavior and perceptions, and (4) We provide empirical evidence of the effectiveness of such interventions (even in short-time use).

Bibliography

[1] Doesthedogdie.com trigger warning database for movies, tv, books and more. https://www.doesthedogdie.com/.

[2] Shinigami Eyes. https://github.com/shinigami-eyes/shinigami-eyes.

[3] Vietnam veterans association of australia. URL http://www.vvaa.org.au/experience.htm. Accessed Oct. 28, 2021.

[4] Nazanin Andalibi and Andrea Forte. Responding to sensitive disclosures on social media: A decision-making framework. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(6):1–29, 2018.

[5] Karla Badillo-Urquiola. A social ecological approach towards empowering foster youth to be safer online. *Electronic Theses and Dissertations, 2020-*, January 2022. URL https://stars.library.ucf.edu/etd2020/1459.

[6] Keely Ball. Let's talk about trigger warnings | keely ball | tedxwarwick, April 2021. URL https://www.youtube.com/watch?v=Zcn1RMZCVI4.

[7] Jack Bowker and Jacques Ophoff. Reducing exposure to hateful speech online. In *Science and Information Conference*, pages 630–645. Springer, 2022.

[8] Guy A. Boysen. Evidence-based answers to questions about trigger warnings for clinically based distress: A review for teachers. *Scholarship of Teaching and Learning in Psychology*, 3(2):163, 2017.

[9] Victoria ME Bridgland, Deanne M Green, Jacinta M Oulton, and Melanie KT Takarangi. Expecting the worst: Investigating the effects of trigger warnings on re-

actions to ambiguously themed photos. *Journal of Experimental Psychology: Applied*, 25(4):602, 2019.

[10] Katie Byron. From infantilizing to world making: safe spaces and trigger warnings on campus. *Family Relations*, 66(1):116–125, 2017.

[11] Angela M. Carter. Teaching with trauma: Trigger warnings, feminism, and disability pedagogy. *Disability Studies Quarterly*, 35(2):10, 2015.

[12] Ashleigh Charles, Laurie Hare-Duke, Hannah Nudds, Donna Franklin, Joy Llewellyn-Beardsley, Stefan Rennick-Egglestone, Onni Gust, Fiona Ng, Elizabeth Evans, Emily Knox, Ellen Townsend, Caroline Yeo, and Mike Slade. Typology of content warnings and trigger warnings: Systematic review. *PLOS ONE*, 17(5):1–14, 05 2022. doi: 10.1371/journal.pone.0266722. URL https://doi.org/10.1371/journal.pone.0266722.

[13] Ashleigh Charles, Laurie Hare-Duke, Hannah Nudds, Donna Franklin, Joy Llewellyn-Beardsley, Stefan Rennick-Egglestone, Onni Gust, Fiona Ng, Elizabeth Evans, Emily Knox, et al. Typology of content warnings and trigger warnings: Systematic review. *PloS one*, 17(5):e0266722, 2022.

[14] Janet X Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. Trauma-informed computing: Towards safer technology experiences for all. In *CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2022.

[15] David Dunning. Chapter five - the dunning–kruger effect: On being ignorant of one's own ignorance. volume 44 of *Advances in Experimental Social Psychology*, pages 247–296. Academic Press, 2011. doi: https://doi.org/10.1016/B978-0-12-385522-0.00005-6. URL https://www.sciencedirect.com/science/article/pii/B9780123855220000056.

[16] Todd E. Feinberg and Jon Mallatt. Subjectivity "demystified": Neurobiology, evolution, and the explanatory gap. *Frontiers in Psychology*, 10, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.01686. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.01686.

[17] Evan George and Angela Hovey. Deciphering the trigger warning debate: a qualitative analysis of online comments. *Teaching in Higher Education*, 25(7):825–841, 2020.

[18] Parush Gera, Nadia Thomas, and Tempestt Neal. Hesitation while posting: A cross-sectional survey of sensitive topics and opinion sharing on social media. In *International Conference on Social Media and Society*, SMSociety'20, page 134–140, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450376884. doi: 10.1145/3400806.3400822. URL https://doi.org/10.1145/3400806.3400822.

[19] Muskan Gupta. Understanding social media users' perceptions of trigger and content warnings. Master's thesis, Virginia Polytechnic Institute and State University, 2023.

[20] Oliver L Haimson, Justin Buss, Zu Weinger, Denny L Starks, Dykee Gorrell, and Briar Sweetbriar Baron. Trans time: Safety, privacy, and content warnings on a transgender-specific social media site. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27, 2020.

[21] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35, 2021.

[22] Blake W Hawkins and Oliver Haimson. Building an online community of care: Tumblr use by transgender individuals. In *Proceedings of the 4th Conference on Gender & IT*, pages 75–77, 2018.

[23] Larke N Huang, Rebecca Flatow, Tenly Biggs, Sara Afayee, Kelley Smith, Thomas Clark, and Mary Blake. Samhsa's concept of truama and guidance for a trauma-informed approach. 2014.

[24] Orla Hyland. *A qualitative investigation of people's attitudes towards the use of trigger warnings on social media*. PhD thesis, Dublin, National College of Ireland, 2023.

[25] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33, 2018.

[26] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529310.

[27] Eleanor Amaranth Lockhart. Why trigger warnings are beneficial, perhaps even necessary. *First Amendment Studies*, 50(2):59–69, 2016.

[28] Caitlin Lustig, Artie Konrad, and Jed R Brubaker. Designing for the bittersweet: Improving sensitive experiences with recommender systems. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

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

[30] Nora Mcdonald, Afsaneh Razi, Karla Badillo-Urquiola, John S. Seberger, Denise Agosto, and Pamela J. Wisniewski. Ai through the eyes of gen z: Setting a research agenda for emerging technologies that empower our future generation. In *Companion Publication*

of the 2023 Conference on Computer Supported Cooperative Work and Social Computing, CSCW '23 Companion, page 518–521, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701290. doi: 10.1145/3584931.3611281. URL https://doi.org/10.1145/3584931.3611281.

[31] Meriem Mejhed Mkhinini, Aboubacar Sidiki Sidibe, Khaoula Benali, Nouha Bentaarit, and Aymen Khelifi. Image and signal processing to detect violent content in social media videos. In *Proceedings of the 2023 15th International Conference on Machine Learning and Computing*, ICMLC '23, page 309–315, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398411. doi: 10.1145/3587716.3587767. URL https://doi.org/10.1145/3587716.3587767.

[32] Dwayne H. Mulder. Objectivity | Internet Encyclopedia of Philosophy. URL https://iep.utm.edu/objectiv/.

[33] Irene V Pasquetto, Briony Swire-Thompson, Michelle A Amazeen, Fabrício Benevenuto, Nadia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio, et al. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review*, 2020.

[34] Chanda Phelan, Jeremy Heyer, Rachel Pfafman, Connie Kerrigan, Golfo K Tzilos Wernette, Lynn Dombrowski, Andrew D Miller, and Jessica Pater. The work of digital social re-entry in substance use disorder recovery. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–33, 2022.

[35] Casey Randazzo and Tawifq Ammari. "if someone downvoted my posts—that'd be the end of the world'': Designing safer online spaces for trauma survivors. 2023.

[36] Afsaneh Razi, John S. Seberger, Ashwaq Alsoubai, Nurun Naher, Munmun De Choudhury, and Pamela J. Wisniewski. Toward trauma-informed research practices with youth

in hci: Caring for participants and research assistants when studying sensitive topics. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1), apr 2024. doi: 10.1145/3637411. URL https://doi.org/10.1145/3637411.

[37] Scott P Robertson. Social media and civic engagement: History, theory, and practice. *Synthesis Lectures on Human-Centered Informatics*, 11(2):i–123, 2018.

[38] Henrik Skaug Sætra and Jo Ese. Shinigami eyes and social media labelling as a technology for self-care.

[39] Mevagh Sanson, Deryn Strange, and Maryanne Garry. Trigger warnings are trivially helpful at reducing negative affect, intrusive thoughts, and avoidance. *Clinical Psychological Science*, 2019.

[40] Manuka Stratta, Julia Park, and Cooper deNicola. Automated content warnings for sensitive posts. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

[41] Nina Vlodder et al. Social representations of harm: Trigger warning practices on facebook watch. 2023.

Appendices

Appendix A

OpenAI Full Prompt

Here is a list of 12 sensitive topics with a short definition and some sub-categories each:

1. **Topic 1: Violence**
    - Definition: Content contains violence
    - Subcategories:
        (a) Violence
        (b) War
        (c) Weapons
        (d) Terrorism
        (e) Police brutality
        (f) Motiveless killing
        (g) Sexual violence
        (h) Animal cruelty
        (i) Torture
        (j) Genocide
2. **Topic 2: Sex**
    - Definition: Content contains sexual themes, including nudity, sexual content and relationships

- Subcategories:
    (a) Nudity
    (b) Mild sexual content
    (c) Explicit sexual content
    (d) Relationship conflict
    (e) Reproductive health

3. **Topic 3: Stigma**
    - Definition: Content depicts negative stereotypes about or attitudes towards a specific group, such as racism or sexism
    - Subcategories:
        (a) Racism
        (b) Anti-religious
            – Anti-Semitic
            – Anti-Christian
            – Islamophobia
        (c) Colonialism
            – Slavery
        (d) Classism
        (e) Sexism
            – Misogyny
            – Misandry
        (f) Transphobia
        (g) Gender identity

(h) Sexuality

- Homophobia

(i) Anti-disability

4. **Topic 4: Disturbing Content**

- Definition: Content contains imagery, sounds, or effects that may frighten, disgust or scare
- Subcategories:

(a) Disturbing content with threat

(b) Horror and terror

(c) Disturbing imagery

(d) Medical content

(e) Human bodies and functions

5. **Topic 5: Language**

- Definition: Content contains language which is sexual, crude or offensive
- Subcategories:

(a) Sexual language

(b) Adult humour

(c) Swearing

(d) Offensive language

6. **Topic 6: Risky Behaviours**

- Definition: Content depicts risky lifestyle behaviours
- Subcategories:

(a) Drug misuse

(b) Alcohol misuse

(c) Tobacco

(d) Gambling

7. **Topic 7: Mental Health**

- Definition: Content relates to mental health issues
- Subcategories:

(a) Mental health

(b) Eating disorders

(c) Trauma

(d) Self-harm and suicide

(e) Depression

(f) OCD

(g) Panic attacks

(h) Anxiety

- Spiders
- Snakes
- Insects
- Needles
- Eye contact
- Irregular patterns

(i) Hair pulling

8. **Topic 8: Death**

- Definition: Content relates to human death or dying
- Subcategories:
  (a) Death
  (b) Accidents
  (c) Natural disasters

9. **Topic 9: Parental Guidance**
   - Definition: Content may not be appropriate for children
   - Subcategories:
     (a) Online access
     (b) Cyber-bullying
     (c) Competitive content
     (d) Imitative content
     (e) Upsetting content
     (f) Non-realistic violence

10. **Topic 10: Crime**
    - Definition: Content depicts or relates to criminal activity

11. **Topic 11: Abuse**
    - Definition: Content depicts or relates to abuse
    - Subcategories:
      (a) Child abuse
      (b) Emotional abuse

(c) Physical/sexual abuse

(d) Neglect

12. **Topic 12: Sociopolitical**

- Definition: Content includes social or political issues
- Subcategories:

  (a) Injustice

  (b) Political issues

  (c) Nazism

  (d) Class issues

Based on this sensitive topic list, does the following tweet contain any of those topics?

< tweet text inserted here >

Please answer with only a 'no' or if the answer is 'yes', respond with a 'yes' and a ranking of the top 5 topics the tweet exhibits with 1 as the most likely and 5 as the fifth likely. The formatting of the ranking should look like this: '1. Violence, 2. Death, 3. Sociopolitical, 4. Crime, 5. Stigma'

After this ranking, include 2-3 sample trigger or content warnings to complete this sentence and fill in the blank with ONLY 1-5 words: 'Here's what a trigger or content warning may look like: 'TW/CW: <blank>"

Appendix B

Screening Questionnaire

Participants must be:

- 18 or older
- Fluent in English
- Reside in the US
- Have posted on social media more than 3 times in the past 12 months
- Willing to participate in a study that involves potentially triggering content in text (No images included). If you have ANY triggering topics (e.g., sexual abuse) from your own traumatic experience that makes you uncomfortable, you are NOT eligible in participating in the study.

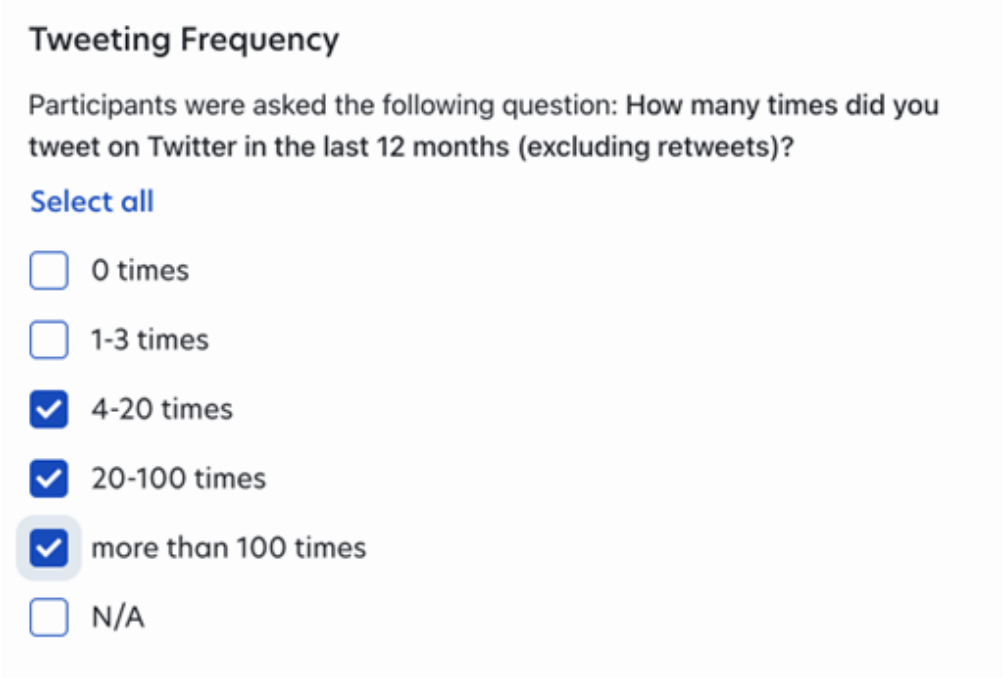


Figure B.1: Screen capture of Prolific Tweeting Frequency Criteria

Appendix C

Self Efficacy Questions

- **Awareness of TW/CW:** I am familiar with the concept of trigger/content warnings on social media platforms.
- **Understanding of TW/CW:** I understand *what topics* should include a trigger/-content warning on a social media post.
- **Confidence in Identifying Content that Needs Warnings:** I am confident in my ability to recognize sensitive content that requires trigger/content warnings on social media.
- **Understanding of TW/CW:** I know *how* to include a trigger/content warning on a social media post that requires it.
- **Perceived Effectiveness:** I believe that trigger/content warnings are effective in preventing people with traumatic experiences on social media from being exposed to sensitive content.
- **Willingness To Use:** I will begin or continue using trigger/content warnings when posting sensitive content on social media.

Appendix D

Parts A and C Tweets Datasets

Each question in Parts A and C begins with "Does the following tweet need a trigger/content warning?" and then the tweet text follows. For each dataset, the answer to the first 8 tweets was "Yes, this needs a warning" while the last 4 did not need a warning. Note: the newline characters had to be removed for the study but were replaced by a space character. Emojis from the original tweet tried to be retained in the survey platform, but are not included in the tweets as shown in the appendix.

D.1 Tweets Dataset 1

- with the super bowl tonight , i want to remind you all that domestic violence rates can be high due to a team losing. if you plan to watch the game tonight, please stay safe. here's the number for the national domestic hotline: 1-800-799-7233
- 14 years ago today I was walking to my teacher training placement in Rochdale and I was stabbed in the back by a stranger in a random attack. The knife went through my liver and lung and i was lucky to survive. It was a life threatening but life changing
- Today's that one day of the year where not only do bigots feel extra bold, but every news media and political analyst shares a half baked opinion on terrorism that gets real triggering real fast, so be kind to and check in on your brown and Muslim friends.

- Today I got told I'm no longer allowed to wear my queer pride pins at work because I can't display "political stances of any kind" so as not to offend anyone. My job has decided that not offending bigots is more important than their own LGBT+ employees I don't feel safe there anymore. I am so fucking angry and so fucking sad. I thought I was welcome. I thought I was safe. I thought I had found a job I could do for a long time. And the worst part is I really really like what I do, but I can't do it anymore. I don't feel safe

- i once saw the most disturbing video of a palestinian boy w his face, his private part and his legs missing/burnt off, his skin is white as snow, i don't understand how anyone could be so cruel to support israel, fuck yall

- i'm gonna go on a bit of a rant here, so bear with me. do you know what m3th does to the face? i don't understand how people can even possibly believe that Britney is on it and use it to drag her. i'm not posting pics, but you can easily google what a m3th addict looks like. m3th addicts are also unhealthily thin, have several burns and sores, rotting or no teeth, violent behavior, heart problems, and have reduced coordination- all of which Britney clearly DOES NOT have. I mean the girl can spin endlessly in her living room, you know how much coordination that takes?! she's doing a lot better than most of us would under 13 years of continuous trafficking and abuse, bffr. there's no doubt she has mental health issues. i think we all do. but calling her a m3th addict is ridiculous and it needs to stop. even fucking KEVIN FEDERLINE denied that claim. there are people who seriously need help or have lost someone from it. it's not a joke. idgaf whether you like Britney or not, it's wrong.

- What's the best help for someone having suicidal thoughts? What type of counseling would help? How can I help them? P.S. Not asking for me, it's for someone I know.

- I can't believe Stevie was bullied. Have you people no fucking shame? It's one thing to disagree on something and another to send death threats. Shame on you!
- Camden's own and my guy @CoachSElliott is leaving Georgia State to become the new tight ends coach at South Carolina, according to reports today. Elliott was previously on staff and USC from 2010-2016. HUGE get for the #Gamecocks.
- It's raining cats and dogs, so instead of Walkies we went to the pet shop. There was lots of lovely smells to sniff. Mum stocked up on doggy sausages. Cheers Mum, they should last til the end of the week #IsaMary
- "Keep love in your heart. A life without it is like a sunless garden when the flowers are dead." Oscar Wilde #CulturesWays #Quote
- Sad at the thought of how some people who once played a huge part in building your dreams are no longer part of when it's finally happening.

D.2 Tweets Dataset 2

- A year ago today there was a shooting at my school, I was in the building next door in a bathroom with my 75 classmates for several hours terrified. And now today I'm going to my school placements and appointments and it just feels so strange to exist today.
- Just watched a true crime story on tiktok about a transgender girl being murdered by two people, and some man in the comments went like "girl". i feel sick on all levels possible, how can you even comment something like that
- We don't talk enough about how some families convince victims of sexual & physical

abuse not to report to authorities because they are more worried about their family name than having the perpetrator behind bars.

- you would think that with an operating budget of $140,515,333, the london police would be able to have proper training about how to speak about sex assault and violence against women without resorting to regressive talking points, and to hire a moderator for the live chat

- stop posting the video of that woman being r*ped during her c-section. no one wants to see that and i highly doubt her or her family wants to see that all over social media or even relive that. it's beyond disturbing.

- You're retarded

- this time last year, i was in hospital after deciding to try and take my own life. to be honest, i actually forgot about it until a memory came up on my phone. this was probably one of the worst moments in my life. i talk about this subject, not to get sympathy but to educate and to show people that they are far from alone. trust me, you're not alone. if this thought ever crosses your mind, please don't hesitate to message me or even call me. a lot of people called me selfish, and told me that it was all in my head. and that will stick with me forever, i wasn't being selfish. i was stuck, and i was sad. just like i still am, i feel like i'm stuck in a time loop, and it just keeps repeating itself and keeps getting worse. that's why this time of year is hard. and that's why i will be taking time off social media if i feel like i need to. you're not alone. don't forget that.

- "Death Due to Accidental Air Conditioner Compressor Explosion" Case 2: This is a case of a 27-year-old mechanic was fixing compressor of a nonworking AC when there was an explosion. The blast was such of high intensity that a substantial part of his

skull was blown off and the entire brain matter was spilling out from the cranial cavity. He died on the spot. The AC was a split type, and the compressor unit was fitted to the outside wall of a jewellery shop. Autopsy findings: Scalp hairs were burnt and singed. The cranial bone was absent in an area of 17 x 15 cm in left frontal and both parietal regions. The brain matter was not present. The left eye ball was completely lacerated. Multiple lacerated wounds and multiple abrasions were present on the right ear and front of right hand, respectively. Chest hairs were burnt and singed. A bruise was present on the upper back of the trunk in the midline. Both hands were stained with black greasy material and dust. Internal organs were pale. Cause of death was craniocerebral damage due to explosion. Reference: Behera C, Bodwal J, Sikary AK, Chauhan MS, Bijarnia M. Deaths due to accidental air conditioner compressor explosion: a case series. Journal of forensic sciences. 2017 Jan;62(1):254-7

- Anger eh , it can make and unmake you . Change your sleeping position to a prison so fast . Don't act on your anger , it's hard but when you think about the consequences of acting on your anger, your body go calm down fast . Just walk away !

- Love is the foundation to all things. Learning and relearning how to move energy, in order to give and receive love - without fear, anger, or resentment - is necessary for a life of abundance. This is the state of divine compassion.

- I'm not as scared as I used to be. Their mystery begins to fade as a method to defeat them becomes more clear. I no longer feel powerless. I can outsmart them. This new power of psychokinesis worked, and I believe it can work again. I stand here, looking down at the line where the water meets the sand - a starting line. All the while, knowing there is a finish line across the Strait. Their compass lies, but mine remains true. I've left embers of inspiration, I only hope whatever spark was left has grown to a torch, and together we create an inferno.

- no one talks about how hard it is to open up to someone about being sad for no reason. how hard it is to explain to your friends & family that you have a heavy feeling in your chest for no reason

D.3 Tweets Dataset 3

- People say they forgot their childhood bc of how bad it was but I am the only one who remembers it vividly, too vividly? who still hears the screams and feels the pain and even smells and sort of rooms will bring me back there?

- Now we have tapes of the president trying to extort and manipulate the governor of Georgia. We have congressmen and senators openly committing sedition. If you still think he is some Christian and deserving of president, there is the door

- Hey just a reminder that folks here with uteruses in the U.S. are still very much suffering from the overturning of Roe v. Wade Instead of treating me in the midst of what turned out to be the passing a decidual cast, I was interrogated as to whether or not I had induced a miscarriage, while an officer was strolling up and down the hall. I was suffering from a condition they had documented in my medical history, and their first priority was making sure I hadn't broken the law.

- I stumbled onto an Instagram thread of abled mothers discussing why they will never let their kids identify as disabled because it implies that they are "undesirable" or "less than" and when I tell you that line of thinking just makes me SO. VIOLENT.

- Depp said he would "fuck her burnt corpse" in a group chat but yes an op-ed, with facts of abuse Amber faced btw, is what's dangerous

- it's not anyone's place to speculate why a player enters the assistance program, but making addiction/drug use jokes is disgusting as hell recovery isn't easy, and it takes a lot to get help for serious issues like that don't be a fucking asshole
- A year ago today a girl I went to high school with, Brianna Ghey, was murdered in cold blood by two monsters. And although her murderers have now been found guilty and charged with their crimes, no amount of prison time will undo what they did and bring her back Brianna, I am so sorry for they did to you that day, you didn't deserve that. And I'm sorry that people still refuse to understand and accept who you are. May you rest in peace and never go through that pain again <3
- HEY ! did u know that causing ppl anxiety attacks , repeated cyber bullying and creating drama unnecessarily isn't cool ! woah crazy right ????? so stop.
- Steelers players have the right to be angry. The NFL took their bye away and now they took their mini bye too. I'd be angry if I was a Steelers player too. The Niners played 2 weeks ago when they had a lot of positive cases on Thursday. So why not let the game be on for tmrw?
- This year taught me how important it is to show up for the ones that show up for you and to never take the people that's truly in your corner for granted…. love on yo real 1s it's hard to come across them kind
- Have your gaming preferences changed as you've gotten older? I'm more into Rpg's now than I've ever been! And thanks to Animal Crossing/Stardew Valley, I've starting to really enjoy farming sims and the like #Gaming #GamersUnite #AnimalCrossingNewHorizons
- "Opportunities are like sunrises. If you wait too long, you miss them. Seize every chance to create wealth and manifest your dreams into reality."

Appendix E

Parts B Prompts

For each dataset, the first 8 prompts are intended to have the participant write a post that included sensitive content, while the last 4 were intended to not include sensitive content but still have an emotional aspect to them.

E.1 Prompts Dataset 1

- Suppose you were seriously injured in a car accident. Share your experience and describe how you felt during/after the incident.
- Share with your followers a time when you felt judged or marginalized because of your sexual orientation, gender identity, race, ethnicity, religious beliefs, disability status, etc. If you cannot think of one, write about a time when someone close to you (friend or family member) experienced something like this and write about that.
- It is the middle of a war that has made international news. You just saw a graphic photo of an injured boy wounded by one of these attacks. Write a post showing your outrage that people are taking this country's side in the conflict. (Feel free to select the war and country you can best relate to.)
- Suppose you have a gambling addiction that has affected multiple areas of your life, including your relationships, your work, etc. Share about your experiences and its

impact on your life.

- Write a post discussing the challenges of seeking help for mental health problems and the barriers to accessing treatment for the public who may be misinformed.

- You are a middle school teacher and there have been recent reports of cyberbullying in your school district. Write a post about the impact of cyberbullying on children's mental health and self-esteem.

- It has been several years since you were the victim of a random act of violence on the street. Write a post about how this has been a transformative process for you.

- It is the anniversary of the 9/11 terrorist attacks, and you want to be empathetic to those who lost their lives that day, but you also know people can be racist under the guise of patriotism. Write a post to show your care for your brown and Muslim friends.

- It's the beginning of the year and many of your friends are sharing their New Year's Resolutions online. Share about a personal goal you're currently working towards and the steps you're taking to achieve it.

- Write a post about your favorite pet or an animal that's special to you.

- You've just returned from traveling to a place you've always wanted to visit. Write a post about a memorable experience you had there.

- You've had a tough year, but you want to share your appreciation for those who have stood by you as you stand by them. Write a post about this.

E.2 Prompts Dataset 2

- It is the one-year anniversary of a school shooting that occurred at your school. Write a post about how you're feeling today.

- Suppose you are a parent and you found some sex educational material from your children's school that frames rape victims as being accountable for their actions, write a post that criticizes the material.

- You have seen a horrific and sexually graphic video continue to be reposted on social media. Write a post to show your frustration and beg people to stop reposting it.

- Write a post using offensive language.

- It has been one year since you were in the hospital for trying to take your own life. Write a post encouraging others that they are not alone if they feel the same way.

- Share your experiences with grief and coping with the loss of a loved one.

- You just watched a true crime story about a transgender girl being murdered and in the comments people are joking and taking this story lightly. Write a post showing your anger.

- Share about a situation where you intervened to help a close friend who was experiencing abuse or mistreatment in their relationship.

- Suppose you were just on a jury and participated in a case where someone let their anger lead them to terrible actions and they are now in prison. Write a post that warns your followers to regulate their anger because it can have consequences.

- You were recently given an award in an area you care about (at work, at school, at church, at a location where you volunteer). Write a post about it and how it made you feel.
- Write about a favorite comfort food or meal that always brings you joy.
- You recently heard one of your favorite sports coaches is leaving your alma mater to take a new position at another university. Write a post showing your support.

E.3 Prompts Dataset 3

- You recently heard news about police officers using outdated and offensive language to talk publicly about sexual assault and sexual violence. Write a post that shows your outrage.
- You have been working for months at a business that allows you to proudly wear your LGBTQ+ pins on your work uniform. You have been really enjoying this job, but today, you were told that you can no longer wear your pins as they are considered "political". Write a post sharing your disappointment with this new policy.
- Write a post using offensive language.
- There has been recent news about a famous football player entering a rehab facility for his struggles with addiction and people are making jokes about it and him. Write a post to call out these people.
- One year ago, a girl you went to high school with was murdered for being transgender. Write a post both informing your followers of what happened while also sharing your sympathy.

- You are a parent of four children ranging in age from high school to elementary school. Share a post about how you've felt concerned about your children's exposure to violent or upsetting content online.
- Suppose you have witnessed how family members can discourage victims from reporting predatory behaviors that they have experienced due to the family's reputation. Write a post that can raise awareness of the issue.
- It is an election year and the candidate people appear to publicly endorse has been proven to be committing crimes during his time as a public official. Write a post both informing people of what's going on while showing your outrage people would vote for this candidate. (Feel free to choose a politician that you can relate to.)
- You have recently broken up with your significant other of 3 years who always supported your dreams. Write a post about how you're feeling.
- Share about a hobby or activity you've enjoyed lately and why it brings you happiness.
- Write about a song that holds special significance to you and why it resonates with you.
- Today, you took your dogs for a walk and to some of their favorite places. Write a post about your day.

Appendix F

Parts D Exit Survey Questions

These questions have response options on a 7 point Likert scale from 'Strongly Agree' to 'Strongly Disagree':

- **Familiarity with LLMs/OpenAI/ChatGPT:** I am familiar with Large Language Models, such as ChatGPT or OpenAI.
- **Ease of Use:** I found it easy to use the tool to generate trigger/content warnings for social media posts.
- **Perceived Effectiveness:** I believe the tool effectively identifies content that may require trigger/content warnings.
- **Integration into Social Media Workflow:** I think this tool could integrate well into my workflow when posting content on social media.
- **Opinion on Backend Algorithm:** I feel the algorithm used to generate warnings was effective in showing me how and when to add a warning.
- **Likelihood of Continued Use:** I am likely to continue using this tool for trigger/-content warnings on social media in the future.
- **Overall Satisfaction:** I am satisfied with the tool provided to assist with trigger/-content warnings on social media.

These questions have open-ended responses:

- **Hesitations or Concerns:** Do you have any hesitations or concerns about using a tool like this in a real-world setting? If yes, please specify.
- **Suggestions for Improvement:** Are there any improvements you would suggest for the tool or its implementation?
- **Overall Comments:** Do you have any additional comments or feedback you would like to share about your experience with the tool?