How Affordances and Social Norms Shape the Discussion of Harmful Social Media Challenges on Reddit

Jinkyung Park , Irina Lediaeva , Amy Godfrey , Maria Lopez , Kapil Chalil Madathil , Heidi Zinzow , Pamela Wisniewski

PII: S2772-5014(23)00009-X
DOI: https://doi.org/10.1016/j.hfh.2023.100042
Reference: HFH 100042

To appear in: *Human Factors in Healthcare*

Received date: 11 November 2022
Revised date: 12 April 2023
Accepted date: 24 April 2023

Please cite this article as: Jinkyung Park , Irina Lediaeva , Amy Godfrey , Maria Lopez , Kapil Chalil Madathil , Heidi Zinzow , Pamela Wisniewski , How Affordances and Social Norms Shape the Discussion of Harmful Social Media Challenges on Reddit, *Human Factors in Healthcare* (2023), doi: https://doi.org/10.1016/j.hfh.2023.100042

Original Article

How Affordances and Social Norms Shape the Discussion of Harmful Social Media Challenges on Reddit

Jinkyung Park[a] (Corresponding author)

Department of Computer Science, Vanderbilt University,

2301 Vanderbilt Place, Nashville, United States

jinkyung.park@vanderbilt.edu

Irina Lediaeva[b]

Department of Computer Science, University of Central Florida,

4328 Scorpius St, Orlando, United States

ireneledyaeva@gmail.com

Amy Godfrey[b]

Department of Computer Science, University of Central Florida,

4328 Scorpius St, Orlando, United States

amygodfrey@knights.ucf.edu

Maria Lopez[b]

Department of Computer Science, University of Central Florida,

4328 Scorpius St, Orlando, United States

maria.lopezx96@gmail.com

Kapil Chalil Madathil[c, d]

Department of Industrial and Civil Engineering, Clemson University,

212 Lowry Hall, Clemson, United States

Department of Public Health Sciences, College of Medicine, Medical University of South Carolina,

135 Cannon Street, Charleston, United States

kmadath@clemson.edu

Heidi Zinzow[e]

Department of Psychology, Clemson University,

400-4 College Ave, Clemson, United States

hzinzow@clemson.edu

Pamela Wisniewski[a]

Department of Computer Science, Vanderbilt University,

2301 Vanderbilt Place, Nashville, United States

pamela.wisniewski@vanderbilt.edu

Warning: This paper discusses self-harming behaviors on social media. Therefore, the reader should use their discretion as to whether they want to proceed.

Abstract

Social media challenges are activities performed by an individual or a group and uploaded to social media platforms to achieve a specific goal. We sought to understand how harmful social media challenges are portrayed on Reddit. Therefore, we analyzed 1,552 Reddit posts and 7,092 comments about inherently risky social media challenges (e.g., Cinnamon, Blue Whale, Fire). We found that posts discussed the participation of others (40%), perceptions on the challenges (24%), prevention/raising awareness (18%), seeking information (10%), and one‟s own participation in a challenge (8%). Comments included offensive commentary towards those who participated, tips and tricks on how to perform a challenge, and Reddit moderation of these posts. We uncovered that the affordances and social norms on Reddit contributed to a lack of propagation of harmful challenges on the platform. Our research contributes to an in-depth empirical understanding of how harmful social media challenges are discussed on Reddit and suggests ways to design affordances and reinforce positive social norms to prevent the spread of social media challenges that promote self-harm.

Key Words: Reddit, digital self-harm, social media challenges, affordances, social norms

How Affordances and Social Norms Shape the Discussion of Harmful Social Media Challenges on Reddit

1. Introduction

Social media challenges are activities performed by an individual or a group and uploaded to social media platforms to achieve a specific goal (Emma Hilton, 2017). The challenges vary from the ones that promote social good such as the "Ice Bucket Challenge," which supports finding a cure for Amyotrophic Lateral Sclerosis (ALS) disease (Pressgrove et al., 2018) to ones that put the participant in danger of self-harming behaviors, including setting themselves on fire (Avery et al., 2016), ingesting powdered cinnamon (Grant-Alfieri et al., 2013), and even purportedly attempting suicide (Khasawneh et al., 2021). Self-harming social media challenges are considered online risks since even the low-risk challenges can rapidly spread through social media and encourage others to participate (Falgout et al., 2022; Gomez-Rodrigues et al., 2012; Khasawneh et al., 2019). Particularly, the spread of potentially harmful challenges among youth and young adults on social media is problematic as it can increase the number of vulnerable populations who are engaging in self-harming behaviors (Kırcaburun et al., 2019).

Meanwhile, as a concept that captures the relationship between technology and users, affordance has been applied in social media research as a key concept to examine perceived properties (e.g., features) of social media platforms that enable or constrain users‟ behaviors on the platform (Ronzhyn et al., 2022). Reddit is a pseudo-anonymous social media platform that affords its users to post candidly about sensitive topics without being afraid of social ramifications (Moore & Chuang, 2017). Given the platform‟s openness, Redditors can freely

post provocative topics they may otherwise not feel comfortable discussing in real life (Schouten et al., 2007). Additionally, the Reddit platform is a space where formal rules and implicit social norms are developed within different sub-communities (subreddits) in which users create sub-cultures suited to their specific needs (Morris, 2017). The unique affordances (i.e., features of a social media platform that frame users‟ behavior (Evans et al., 2017)) of the Reddit platform, such as pseudo-anonymity and subreddit communities, have created strong and sometimes unusual social norms that are not observed on other social media platforms (Muller, 2016); and thus, making it an ample platform for researchers to study the novel social phenomenon, such as the emergence of potentially harmful social media challenges (Ksiazek et al., 2015; Robson, 2019).

Literature on human factors and social media domain focused on how human factors interact on social media and the positive and negative impacts of social media usage on users. For instance, there has been substantial work on how individuals collectively discuss health-related topics such as mental illness (De Choudhury & De, 2014), public health (Park & Conway, 2017), and depression-related problems (Tadesse et al., 2019) and the impact of this social interaction on users. A theme among research on human factors and social media is that the positive and negative impacts of social media usage on users vary depending on individual characteristics and context of use. In this work, we study Reddit users from the context of discussing harmful social media challenges. Our research is one of the first empirical studies that examine how the unique affordances of Reddit play a role in the discourse of harmful social media challenges. Understanding the discourse on such harmful challenges on Reddit will inform prevention strategies for promoting the mental health of social media users.

1.1 Harmful Social Media Challenges as Acts of Digital Self-Harm

Digital self-harm is the consumption and production of online content that leads to, supports, or exacerbates non-suicidal yet intentional harm or impairment of an individual‟s physical well-being (Pater & Mynatt, 2017). As the Internet pervaded individuals‟ lives, social media and online communities laid fertile ground for magnifying the issues of digital self-harm through the sharing and/or reinforcing of self-harming behaviors (Pater & Mynatt, 2017). Existing literature on digital self-harm has highlighted the role of social media and online communities in creating and spreading digital self-harm and its adverse impact on public health and safety. Particularly, understanding if and how self-harming behavior propagates through social networks is an emerging area of human factors and healthcare research. Defined as a form of social influence in which the behavior of an individual is influenced indirectly by observing the behavior of others, "behavioral contagion theory" (Polansky et al., 1950) has been widely applied in human factors research to understand decision-making and risk-taking behaviors concerning social conformity and peer influence (Abraham et al., 2022). Another well-known theory is the "social learning theory," emphasizing that human behavior is strongly incentivized by rewarding outcomes such as others‟ acceptance, and such acceptance is expected based on their prior observation of others (Bandura & Walters, 1977). Both theories highlight "social" aspects of individuals' motivations for certain actions, hence, have served as a theoretical lens to study the motivation and the propagation of self-harming behaviors promoted on social media. For instance, online communities that offer understanding and support for self-harming behaviors can contribute to the spreading of self-harming behaviors (Emma Hilton, 2017). This is because discussions and live depictions of self-harm acts can contribute to the normalization and acceptance of self-harm on social media (Dyson et al., 2016). Similarly, public pro-eating

disorder communities are known to motivate their users to continue efforts with anorexia and bulimia (Borzekowski et al., 2010; Norris et al., 2006; Roberts Strife & Rickard, 2011). This normalization of self-harm and exposure to self-harming behavior could further negatively impact other users‟ mental health and create contagion effects (Daine et al., 2013). The studies confirmed that being exposed to self-harming content promotes self-harm behavior and suicidal ideation in vulnerable adolescents (Memon et al., 2018). Even very few self-harming posts on social media sites actively encouraged others to self-harm (Shanahan et al., 2019). As such, the existing research laid the groundwork for studying the motivations and the spread of digital self-harm on social media and called for strategies to alleviate the adverse impact of digital self-harm shared on social media.

Our research provides a unique perspective on digital self-harm by examining harmful social media challenges. Social media challenges involve people encouraging one another to participate by posting images of or recording themselves participating in these activities online (Khasawneh et al., 2021). The challenges vary in their level of risk to individuals; some promote philanthropic causes (e.g., Ice Bucket Challenge), while others pose inherent risks to the individuals who attempt them (Lupariello et al., 2019). For example, the “Cinnamon Challenge” encourages participants to swallow a tablespoon of ground cinnamon in under a minute. It became a viral social media challenge that led to thousands of videos being uploaded to YouTube and other social media platforms (Grant-Alfieri et al., 2013). Although granular cinnamon is not likely to cause long-term damage, swallowing a large quantity can pose a risk of asphyxiation (i.e., choking and possible death) and the burning of the throat or extensive coughing (Grant-Alfieri et al., 2013). The “Blue Whale Challenge” (BWC) is another harmful challenge that is known for its serious nature as it purportedly promotes participants to complete

tasks ranging from cutting themselves to suicide (Lupariello et al., 2019). Although their existence was not clearly observed, the high-risk challenges such as BWC are particularly concerning as they can not only lead participants themselves to serious physical damage (Atherton, 2020; Aver at al., 2016; Grant-Alfieri et al., 2013) but also contribute to self-harm contagion among other users (Khasawneh, et al., 2019).

The existing literature on harmful social media challenges focuses on investigating the prevalence and spread of challenges and their implications for public health and safety. Especially, human factors literature has established that harmful social media challenges can be propagated through behavioral contagion and social modeling (i.e., imitating the behaviors of those we observe) (Abraham et al., 2022). For instance, Roth et al. (2020) analyzed newspaper articles reporting on BWC and found that news reports regarding BWC could unintentionally increase suicide contagion effects and normalize certain self-harming behaviors. Similarly, Khasawneh et al. (2021) investigated how harmful social media challenges such as BWC and Tide Pod Challenges are portrayed on YouTube and Twitter and found prevalent sharing about these behaviors could potentially normalize and contribute to the contagion of self-harming behaviors among youth. At the same time, they showed that many social media posts were intended to educate people and raise awareness about a challenge, indicating that the contagion effect and social modeling can be effective intervention strategies to reduce the creation and spread of harmful social media challenges. The above studies provided insights into how harmful challenges are discussed and spread on social media and potential ways to mitigate the negative impacts of such challenges on the well-being of individuals and society. Building upon the previous work, we examined various inherently dangerous social media challenges (e.g., Cinnamon, Fire, Blue Whale, Tide Pod, etc.) on Reddit (reddit.com) to understand how those

challenges are portrayed on Reddit. Particularly, we addressed how affordances and social norms of Reddit shape discourse on harmful social media challenges and suggest design implications to reduce the prevalence and spread of harmful social media challenges. In the next section, we provide an overview of how affordances are studied in the context of social media and the affordances of Reddit that promote specific social norms and cultures on Reddit.

1.2 Affordances Perspectives

Affordances are a "multi-faceted relational structure" (Faraj & Azab, 2012, p. 54) between technology and the user that frames potential behavioral outcomes in a particular context (Evans et al., 2017). As a concept that captures the relationship between technology and users, affordance has been applied in social media research as a key concept to examine perceived properties (e.g., features) of social media platforms that enable or constrain users" behaviors on the platform (Ronzhyn et al., 2022). Work in this area often uses affordances to focus on the dynamics of social interactions that various social media features afford. Some scholars have used affordance almost synonymously with the features of technology, while others have focused on the social structures that are formed in and through a given technology (Bucher & Helmond, 2018). In the past literature, social media platforms have often been analyzed in terms of having „affordances and constraints" (Ellison & Vitak, 2015). Recently, the concept of social media affordances is also applied to examine the role of algorithms in social media play role in enabling and constraining the use of social media (Ronzhyn et al., 2022). The common theme among social media affordances research is that, unlike technological determinism, they emphasize the role and agency of humans in the use of technology. Affordances approaches allow researchers to incorporate the contextual aspects of technology use, e.g., how social media usage is shaped by the properties of actors and their context (Evans et

al., 2017), hence, making it an appropriate framework for understanding relational interactions between users and social media platforms (Rice et al., 2017). In this study, we applied affordances approaches to explore how the portrayal of harmful social media challenges is shaped by the properties of a social media platform, Reddit.

Reddit is a social discussion website that consists of user-created subcommunities, known as subreddits, where users can post and comment. Each subreddit is distinct, as it pertains to a single topic or theme, and can cover a variety of discussion areas, including sensitive topics such as mental health, self-harm, and suicide. Registered users of Reddit ("Redditors") can contribute to subreddits by creating self-posts (Figure 1) comprised of text, images, videos, or links to other websites and reposting within other subreddits through cross-posting. Redditors can also leave comments on posts (Figure 2), which are hierarchically threaded, where a comment can be in response to the post in general or in reply to another comment.

Fig. 1. Example of a Reddit post about harmful social media challenges

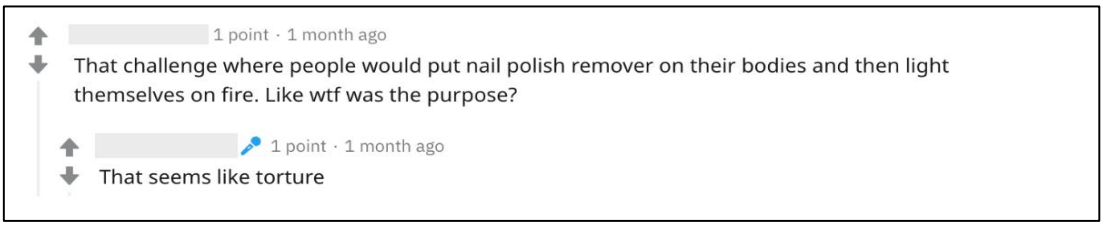

Figure 2: Example of a Reddit comment thread discussing harmful social media challenges

In this study, we focus on how harmful social media challenges are discussed on Reddit to explore the interplay between the unique affordances of the Reddit platform and how the challenges are portrayed and/or propagated. Unlike other popular platforms such as YouTube and Twitter, Reddit provides two unique affordances to its users: identity affordances and social affordances. One of the unique affordances of Reddit is related to *identity affordances*: the opportunities on social media platforms for identity development and portrayal (Moreo & D"Angelo, 2019). Reddit is different from other identity-based platforms such as Facebook or Twitter in that it does not require users to reveal their identity. It does not have a feature to post a profile picture, nor does it have networking features such as Facebook "friends" or Instagram "followers." Instead, users are socially connected within the "subreddit," a distinct community of networked users with shared interests. The low-identity affordance of Reddit allows Redditors to disclose information without repercussions of personal identity (Leavitt, 2015).

Another unique affordance of Reddit is *social affordances* (e.g., the relationship between the technology and the social group that enables or disables certain interactions among members of that group (Bradner et al., 1999)) such as "karma," "upvote," "downvote," "award," "subreddit," "subreddit rules," and "moderation." which help Redditors create the social norm. Social norms are specific rules of the platform that influence what is shared by the user and what the user expects to see being shared (Morris, 2017). "Subreddits" are the ones in which clear social norms can be found, as the norms are set not for the users, but by the users themselves; that is, shared rules and expected online behavior are set by those who participate in the subreddit (Moore & Chuang, 2017). Each subreddit has its own social norms regulated by a list of "subreddit rules" that can be viewed on the subreddit"s community profile. These rules are largely enforced on Reddit through upvotes, downvotes, awards, moderators, and bots examining

the content being shared by users. For instance, posts that do not fit the social norms of the community can be deleted by the community moderator, volunteers who enforce the community-specific rules (Habib et al., 2019). Also, posts and comments can be voted "upvotes" (valued) or "downvotes" (deemed unworthy) by other users and the votes determine where the posts and comments are displayed on the website. That is, posts and comments that do not adhere to the expected norms or the cultural field of a given subreddit are demoted through the down-votes, whereas those that adhere to the rules are promoted by their community through up-votes and are pushed higher to the top of the subreddit page and be seen by a larger audience (Reddit, 2022). Redditors can also reward other Redditors" content by giving awards or they can report a problem with the content (e.g., marking a post or comment as spam, abuse, etc.).

Social norms on Reddit have been widely studied among scholars. For instance, Sharma & De Choudhury (2018) analyzed different mental health subreddit communities to examine how the users conformed to community norms within those subreddits discussing sensitive topics. They found that members of communities that deal with stigmatized and sensitive issues tend to conform to community rules to develop the community as a safe place for candid disclosure. By analyzing 2.8 M comments removed from subreddits, Chandrasekharan et al. (2018) found that some violations of community norms (e.g., racism and homophobia) are universally removed in almost every subreddit by moderators, while other societal issues (e.g., mocking religion, nationality) were not considered norm violations on most of the subreddits. As such, social norms on Reddit play an integral role in guiding acceptable behaviors, and hence in shaping the discourse on certain topics within the subreddits. However, how social norms shape the discussion of harmful social media challenges has been under-explored. In this work, we focus on how harmful social media challenges are portrayed on Reddit and how unique

affordances of Reddit (e.g., pseudo-anonymity and social norms) shape the discourse on harmful social media challenges.

1.3 Problem Statement

The purpose of our study was to investigate how harmful social media challenges are portrayed on Reddit to understand the nature of discourse regarding these challenges and explore how the unique affordances of Reddit contribute to the discussion of the challenges on Reddit. Therefore, we pose the following research questions:

RQ1*: How are harmful social media challenges portrayed through original posts on Reddit?*

RQ2*: Once harmful social media challenge posts are shared on Reddit, how do Redditors engage with (i.e., comment) on these posts?*

Our research contributes an in-depth examination of online users‟ posts and comments related to harmful social media challenges on Reddit and how the affordances of Reddit shape the discourse around harmful social media challenges. By focusing on inherently harmful social media challenges, our research uncovers potential strategies for risk mitigation and prevention through the lens of social media affordances. Previous research in the field of human factors in healthcare studied adolescents‟ and young adults‟ motivations to participate in challenges on social media (i.e., TikTok) (Falgoust, 2022; Khasawneh, 2019; Roth et al., 2022). Our work builds on the prior work by exploring how harmful social media challenges are discussed among Reddit users to provide a basis for the strategies to mitigate risks associated with social media use and to promote the mental health of social media users.

2 Methods

2.1 Data Collection and the Scoping Process

2.1.1 Scoping a Dataset of Harmful Social Media Challenges.

We created a dataset of public posts on Reddit using the BigQuery cloud service (Naidu & Tigani, 2014). We leveraged a publicly available dataset that was uploaded by Redditors from 2006 (shortly after the creation of Reddit in 2005) to 2019. This dataset contained over 3 billion comments and included information about upvotes, downvotes, author, subreddit, flair, and timestamp for each post and comment. First, we explored subreddits and identified specifically addressing harmful social media challenges. These subreddits were navigated directly on the Reddit site with a search of "challenge" and "challenges" to not include/exclude specific challenges or subreddits. We did so because searching for specific subreddits would limit our preliminary search due to the expansive nature of Reddit. We explored several subreddits, including "r/StupidChallenges", "r/GoForGold", "r/Challenges", "r/YouTubeChallenges", and "r/ChallengeAccepted," which addressed harmful social media challenges to identify an initial list of keywords. From those subreddits, we explored the posts, comments, and community rules to understand the norms for discussing harmful social media challenges and identify an initial list of keywords. Next, we used keywords, such as "challenge," AND ("social media" OR "death" OR "fail") to do an initial exploration of the data provided by BiqQuery. We consciously scoped our keyword to target the more risky or harmful social media challenges, rather than the more frivolous (e.g., cheesed challenge) or altruistic challenges (e.g., Ice bucket challenge).

Next, we utilized "Google Trends" to identify the most popular challenges that web users searched for between 2004 to 2019 to finalize the list of keywords. The criteria to select harmful social media challenges were as follows: a) did the challenge pose intentional physical harm?

and b) are the consequences long-term or temporary? (e.g., discomfort vs. suicide). For instance, we eliminated popular challenges that did not intentionally pose harm to their users, even if they posed some level of discomfort (e.g., Ice Bucket Challenge, Chili Pepper). Also, we eliminated challenges that can provide short-term discomfort (e.g., Chili Pepper Challenge) but included those that can lead to long-term injuries (e.g., Cinnamon Challenge). For instance, Chili Pepper Challenge can provide extreme discomfort (high Scoville scale score) to participants, while ingesting a large spoonful of cinnamon (Cinnamon Challenge) is a choking hazard, as cinnamon is hydrophobic, and can lead to lung damage/death. Then, we narrowed our data range to start from 2009 when the harmful social media challenge (i.e., Cinnamon Challenge) first emerged, which left us with data available between January 2009 to May 2019. The final sets of keywords that we used were "challenge," AND ("fire" OR "tide pod" OR "blue whale" OR "cinnamon" OR "boiling water" OR "choking" OR "momo" OR "suicide"). We used standard SQL in BigQuery to create our dataset and the final dataset included a total of 2,255 posts.

2.1.2 Relevancy Coding.

The posts (N = 2,255) were divided between the third and fourth authors for relevancy coding. The first 10% (n = 225) of the posts were coded by both researchers to ensure inter-rater reliability (IRR) (Gwet, 2014). Cohen's κ was run to determine if there was an agreement between the two rater‟s judgment on the relevancy of the post. There was substantial agreement between the two raters‟ judgments (κ = .0.98). Once adequate IRR was confirmed, the remaining 2000 posts were divided among the two coders for relevancy coding. Posts that gave some context related to a social media challenge, including a challenge name and/or an indicator of participation, were deemed relevant. Yet, if the post did not involve a social media challenge, it was removed from our dataset. Of the 2,255 initial posts, 1,552 (~69%) were deemed relevant.

Conflicts were resolved by forming a consensus among the two coders. Based on the 1,552 relevant posts, we then used BiqQuery to retrieve the associated comments (N = 11,941) for the relevant posts.

2.2 Data Analysis Approach

2.2.1 Qualitative Thematic Analysis on Original Posts (RQ1).

To answer RQ1, we conducted a qualitative thematic analysis (Mayring, 2000) on the original posts discussing harmful social media challenges (N = 1,552). First, we familiarized ourselves with the data to generate initial codes (e.g., stranger, known participants, propagation, social commentary). Once initial codes were generated, we group the conceptually related codes together (e.g., stranger and known participants) and assigned a higher-level label, "theme" (e.g., others‟ participation). Each code remained as a "subtheme" under the relevant theme. Then, the third and fourth authors coded 10% (n = 153) of the posts to ensure consistency between their codes. Cohen‟s κ was run to determine if there was an agreement between the two rater‟s judgments on the use of the developed themes and subthemes. The agreement between the two raters‟ judgments ranged from substantial agreement (κ = .0.72) to complete agreement (κ = 1.00), depending on themes. The two coders met with the second and last authors to form a consensus on conflicting themes and subthemes and finalize the codebook. Next, the posts were divided between the two coders to finish the coding process. During this process, a couple of new subthemes emerged. In these cases, the second, third, fourth, and last authors met to discuss these new subthemes and determine they needed to be added and revisited the coded data to recode for the emergent subthemes. Through the qualitative content analysis on original posts, we generated five themes including "Others‟ Participation" (posts related to other users‟ participation in harmful challenges), "Perspectives on the Challenges" (posts expressing

Redditors‟ opinion on the harmful challenge) “Prevention/Raising Awareness” (posts trying to prevent and/or raise awareness regarding harmful challenges), “Information Seeking” (posts seeking information related to harmful challenges), and “Self-participation” (posts sharing Redditors‟ own experience of participating in harmful challenges), and subthemes under each theme. The final codebook is presented in **Table 1**. Posts were coded mutually exclusively based on the subtheme in which they fit the best. During the qualitative content analysis, we identified eight distinct harmful social media challenges that were mentioned in original posts. The eight challenges included:

1. Cinnamon Challenge (*n* = 829): Participants attempt to swallow a spoonful of ground cinnamon in under one minute without the use of water or fluids (Grant-Alfieri et al., 2013).
2. Fire Challenge (*n* = 189): Participants cover their own bodies with a flammable liquid and then set the liquid on fire (Avery et al., 2016).
3. Tide Pod Challenge (*n* = 186): Participants place a Tide detergent pod in their mouth, chewing, and often swallowing (Robson, 2019).
4. Momo Challenge (*n* = 150): Participants are encouraged by a character named „Momo‟‟to complete dangerous tasks including self-harm, violent attacks, and suicide (Weekend Edition Saturday, 2019).
5. Blue Whale Challenge (*n* = 124): Participants are purportedly given 50 tasks over 50 days involving self-harm (e.g., cutting) and completing the last day of suicide (Khasawneh et al., 2021).
6. Suicide Challenge (*n* = 132): A reference to both the Momo and Blue Whale Challenges (Roth et al., 2020).

7. Boiling Water Challenge (n = 70): Participants pour boiling water on another person, either willingly or by surprise, or participants may attempt to drink boiling water (Gray, 2019).

8. Choking Challenge (n = 9): Participants either strangle themselves or have another person attempt to choke them to deprive their brains of oxygen (Daine et al., 2013).

We note that while most of the challenges involved self-harm, in some cases (e.g., the Boiling Water and Choking Challenges), the challenges involved harm to others. Since Reddit is pseudo-anonymous, we have no personally identifiable information such as the demographics of the users who made the posts and comments we analyzed. Additionally, the usernames were not included in our data scope**.**

Table 1: Final Codebook for Reddit Posts (N = 1,552)

Theme		
Subtheme	**Description**	**Example Post**
Others' Participation (40%, n = 619)		
Stranger's Participation (39%, n = 238)	The participant was not known by who shared it.	*"Guy does the boiling water challenge*" -r/videos
Unclear Relationship (38%, n = 235)	The user did not make it clear if the participant was known or not.	*"cinnamon challenge fail!"* -r/videos
Known Participant (12%, n = 77)	The participant was stated to be personally known by the user.	"*My wife attempting the cinnamon challenge."* -r/pics
Public Figure (11%, n = 66)	The participant was someone with a large online following (i.e., actor, professional athlete).	*" PewDiePie does the tide pod challenge with an intense twist"* -r/PewdiepieSubmissions
Perspectives on the Challenge (24%, n = 381)		
Propagation (31%, n = 118)	Posts suggesting a new challenge idea based off of existing challenge(s).	*"You've heard of the boiling water challenge, and now I present to you the liquid metal challenge"* -r/memes
Social Commentary (30%, n = 113)	Posts discussing challenges as social phenomena and/or implications of attempting challenges at a social level.	*"First they challenge kids to commit suicide and now they crash the market, we gotta stop this whales!"* -r/CryptoCurrency
Critique of Challenges (22%, n = 82)	Posts negatively criticize and/or redicule specific challenges and the participants.	"*Someone I know tried to kill himself doing the cinnamon challenge. I send him this saying it is more effective"* -r/ImGoingToHellForThis
Challenge Portrayal (17%, n = 65)	Posts sharing challenge name(s), generally with attached media and short description.	"*The real momo challenge"* -r/funny
Prevention/Raising Awareness (18%, n = 273)		
Media Sources (93%, n = 255)	Posts linking external media sources discussing harms of a challenge to raise awareness.	*"@washingtonpost: An 11-year old was doused with boiling water at a sleepover. Her mother blames an online challenge. https://t.co/JqLoYt8PB0"* -r/newsbotbot
Cautionary Messages	Posts including warnings through	"*Don't do the Cinnamon Challenge! It's*

(7%, *n* = 18)	testimonies and personal experiences.	*deadly! It's pointlesspointless..."* -r/skeptic
Information Seeking (10%, *n* = 154)		
Challenge Experience (57%, n = 87)	Posts seeking information about other users" participation.	*"Hey guys any idea about that momo challenge? Any links or vids of proof"* -r/creepypasta
Understanding Challenge (23%, n = 36)	Posts wanting to know the rules or origins of a challenge.	*"Why are people doing stupid internet dares like the tide pod challenge or the bird box challenge?"* -r/AskReddit
Challenge Trends (20%, n = 31)	Posts discussing past challenges or wondering about next challenge trends.	*"When will we see the boiling water challenge?"* -r/AskReddit
Self-Participation (8%, *n* = 125)		
Unknown Outcome (43%, *n* = 54)	The user did not specify whether their challenge attempt was successful or not.	*"So i tried the cinnamon challenge earlier today"* -r/fffffffuuuuuuuuuuuu
Successful Attempt (28%, *n* = 35)	The user specifically stated they successfully participated in a challenge.	*"i beat the cinnamon challenge...is it a big deal?"* -r/reddit.com
Solicitation (15%, (*n* = 19)	The user was asking for likes and/or views of their attempts or encourage participation.	*"My goal is to get my cinnamon challenge video to 30,000 views by the end of the day! Help?!?"* -r/videos
Failed Attempt (14%, *n* = 17)	The user participated in a challenge but failed or ended badly.	*"my cinnamon challenge fail"* -r/funny

2.2.2 Topic Modeling on Comments (RQ2).

To address RQ2, we first subdivided the comments (*N* = 11,941) based on the five qualitative themes generated from the original posts (**Table 1**). We did this because we anticipated that different topics would surface depending on how the challenges were being portrayed in the original posts and hence, we wanted to capture nuanced differences in the comments across different themes of original posts. The volume of comments was appropriate for topic modeling (Wallach, 2006) across all themes except for the "self-participation" theme; therefore, we performed content analysis to code for the major topics discussed in this category.

To perform topic modeling, we compared and contrasted two topic modeling approaches - a conventional latent Dirichlet allocation (LDA) topic model (Qiang et al., 2020) and a biterm topic model (BTM) (Yan et al., 2013) - to move forward with the model that works best for our

short-sized text data**.** We chose the two approaches because they both are widely applied for discovering topics from short-sized text data without requiring any prior annotations of the documents. The difference between the two is that with LDA, each document is modeled as a mixture of various topics, and each topic is characterized by a probability distribution over all the words (Qiang et al., 2020), while BTM models the word co-occurrence pattern (i.e., biterms), instead of documents, and uses the aggregated patterns in the whole corpus for the topic learning (Yan et al., 2013). Both models generate the topic words (set of terms to generate topic) and coherence score (how often the topic words for each topic appear together in a document, closer to zero is better) (Mimno et al., 2011), depending on the number of topics researchers set for the models (*K*) (see Table 4 in Appedix for detials).

Due to its largest size, the comments in the "Others" Participation" theme (44%, *n* = 5,271) were used to run both topic modeling approaches on the data. We compared the coherence scores (Yan et al., 2013) and the quality of the topic words generated by the two approaches. Overall, BTM gave us a better average coherence score and more semantically meaningful topic words as compared to the LDA model, hence, we moved forward with BTM for further analysis.

We trained the model using the BTM package in Python 2.7 (Yan, 2020). Using Jupyter Notebooks in Google Colaboratory (Google Colaboratory, n.d.), we processed the comments through the normalization steps by 1) removing newline characters and hyperlinks, 2) tokenizing – dividing a text input into tokens such as words or phrases (Albalawi et al., 2020), 3) removing single character words and non-Latin characters, 4) lemmatizing – returning the base of dictionary form of a word to enhance the model"s accuracy (Albalawi et al., 2020), 5) removing stopwords that did not add any semantic value when making the topical inferences, 6) removing

words with a term frequency less than 10, 7) filtering out comments with less than 2 and more than 30 tokens (to exclude significantly small and significantly large comments), 8) removing duplicates. After the normalization steps were finished, we listed the top 20 most frequent words found in the comments (N = 7,092). If the words were not specific enough and meaningful to the analysis, we added those words to the pre-configured stopword list and went through the normalization steps again. We proceeded to analyze the comments for each theme by performing a biterm topic model on them. We configured the biterm model‟s hyperparameters according to the previous literature (Yan et al., 2013), where $\alpha = 50/K$ (K is the number of topics) and β = 0.01, and ran the model for 1,000 iterations. We first began with two topics (K=2) and increased the topic number by increments of one. We measured the coherence score for each number of topics to identify the number of topics that would provide succinct cohesion for the comments under a particular theme (**Table 2)**.

Table 2: Average coherence score on the K number of topics for each theme. A larger coherence score means the topics are more coherent

Theme	K = 2	K = 3	K = 4	K = 5	Final number of chosen topics
Others‟ Participation	-106.57826	-87.040472	**-82.887618**	-99.163049	4
Perspectives on Social Media Challenges	**-70.065144**	-73.452742	-92.819477	-90.094707	3
Information Seeking	-115.22124	-118.52841	-114.91572	**-111.9328**	3
Prevention/Raising Awareness	-63.294054	**-57.431352**	-75.184303	-73.893802	3

Since BTM generates different sets of topic words depending on the number of topics (K), it is up to the researchers‟ discretion to choose the sets of topic words to name topics discussed.

Therefore, we reviewed all sets of topic words generated by BTM (K=2, 3, 4, 5) and went back to our comment data to see which sets of words can be used to best describe the major topics discussed among the comments under the five themes. Then, we used semantic validation to compare the results of the BTM with expert reasoning and confirm that comment groupings into topics made semantic sense (Baumer et al., 2017; Chang et al., 2009). For instance, for the "others' participation" category, the coherence scores for K = 3 and K = 4 were close and expert reasoning found that topics made more semantic sense when K = 4, thus, K = 4 was selected. The final number of topics chosen by the expert is shown in **Table 2**.

Next, for each category, we sorted comments by their concentration on a topic and analyzed the top 50 representative responses (i.e., the comments that more likely belonged to a topic) based on the comments' probability scores generated by BTM. We reviewed the top words generated by BTM and read through the 50 representative comments under all four themes of original posts to deductively name the topics. The five topics generated by topic modeling were "Offensive Commentary," "Tips and Tricks for Specific Challenge," "Suicide Challenges," "Unexpected Tangents," and "Reddit Moderation." We recognize that conceptually, the term "topics" are similar to the term "subthemes" that we used to answer RQ1. However, we chose to use the term "topics" here to address RQ2 to distinguish between the harmful challenges discussed among the original posts (RQ1) and among the comments (RQ2). In addition, given a small number of comments in the Self-Participation theme, rather than applying a computational model, we conducted a content analysis of the comments in this theme using the same five topics generated by topic modeling. The five topics we generated from topic modeling and the proportions of topics discussed are shown in **Table 3**.

Each cell in the table represents a percentage of the topic discussed among the comments under each theme (totaling 100% for each column). The percentage values were generated by BTM except for the self-participation theme. For each topic under the theme, we sorted the top 50 representative comments (i.e., the comments that more likely belonged to a topic) to understand how the topics are discussed in the comments. For the self-participation theme, we report the results of a qualitative content analysis using the five topics generated by topic modeling.

Table 3: Proportions of topics discussed in the comments in each post theme

Themes / Topics	Others' Participation (*n* = 3350)	Perspectives on Challenge (*n* = 1283)	Information Seeking (*n* = 1353)	Prevention/ Raising Awareness (*n* = 1046)	Self-Participation (*n* = 62)
Offensive Commentary	74.9%	55.2%	-	84.3%	-
Tips and Tricks for Specific Challenges	15.6%	31.9%	36.7%	-	10.2%
Suicide Challenges	-	-	58.8%	-	-
Unexpected Tangents	4.4%	-	-	9.7%	89.8%
Reddit Moderation	5.1%	12.9%	4.5%	6%	-
Total	100%	100%	100%	100%	100%

3 Results

3.1 Original Posts about Harmful Social Media Challenges (RQ1)

In this subsection, we present the main themes that emerged from the original posts and the subthemes identified within each theme, along with some example quotes from the dataset.

To illustrate the themes in detail, we included example quotes with the names of the subreddit from which the comments were taken. The usernames of the example quotes are not included.

3.1.1 Posts About Other People Participating in the Challenges.

The most frequent theme among the original posts was "Others' Participation" in harmful social media challenges (40%, *n* = 619). In this theme, we observed that the challenges were attempted by the posters' friends, celebrities, or strangers (not by the original posters themselves). The original posts on this theme mainly discussed others' participation accompanied by commentaries, often *critical*, about the persons who are attempting the challenges.

> *"Complete opposite of ice bucket challenge, stupid girl dumps boiling water on unsuspecting brother as a prank"* -r/videos

Furthermore, the posters often included links to or attached different media types (e.g., YouTube videos) in their posts to share the harmful social media challenges. While Redditors were mostly sharing media along with their criticisms, we also aimed to identify the connection between participants and posters. The most mentioned subtheme was "stranger's participation" (39%, *n* =238) in which the posters did not know the participants of the challenges. In this subtheme, the posters describe a stranger who is attempting a challenge as "*a man*" or "*this kid,*" making it clear that the poster did not know the participants yet chose to share their attempt. Many of the posts describing strangers' participation were critical in tone. For example, Redditors often used derogatory words to describe the person performing the challenge to express their disagreement with the participation:

> *"Dumbass decides to do the „boiling water challenge"."*-r/DarwinAwards

The next most frequent subtheme was "Unclear Relationship" (38%, *n* = 235) where the persons who attempted the challenge were not known by the original posters. The posts under this subtheme were vague about the relationship between the posters and the participants, stating only that a challenge had been completed or not with a media attachment (e.g., photo or video). These posts often included the commentary that the challenge "fails," which added the critical subtext to the post, likely because the Reddit posters had no personal connection to the persons getting hurt by failing the challenges. The following post describes a failed attempt at the Cinnamon Challenge, but the attached YouTube link to the recording of the failed challenge had been removed by the moderator:

"Girl fails the cinnamon challenge" -r/videos

Meanwhile, 12% (*n* = 77) of the posts in this theme were about the challenges attempted by the "Known Participant" of the original posters such as friends, family members, or significant others. The challenges participated by the known parties of the Redditors were primarily expressed in a positive or humorous tone; for instance, laughing (e.g., "*LOL*") about how the attempt had not been successful. The posts under this subtheme were mostly about the challenges with low-level risks and hence, Redditors consider their know person's participation as a form of physical comedy:

"My friend Karl doing the cinnamon challenge, and getting a bloody nose." -r/videos

The least mentioned subtheme was "Public Figure" (11%, *n* = 66) where Redditors posted public figures who are attempting harmful challenges, most often the Cinnamon Challenge. For example, 12 posts were sharing the same video of Maisie Williams, an actress, performing the Cinnamon Challenge. Similar to the posts under the "Known Participant" theme, the tone of the posts in this subtheme was less critical and more lighthearted in nature:

"Maisie Williams does the cinnamon challenge…oh sweet summer child" - r/gameofthrones

In summary, we observed that Redditors most often posted about someone else‟s participation in one of the harmful social media challenges. Redditors reacted critically toward the participants when there was no personal relationship established between them and the participants. On the other hand, Redditors tend to become less critical as their connections to the participants became personal (e.g., family and friends).

3.1.2 Perspectives on Harmful Challenges.

The "Perspectives on Harmful Challenges" theme (24%, *n* = 381) consisted of posts sharing Redditors‟ opinions on the potentially harmful challenges. Media attachments were present in half of the posts under this theme, many of which were images or memes making fun of the challenges or videos of fake challenges that are low in physical risks. At the same time, some interactions helped create new challenges. The posts under the "Propagation" subtheme (31%, *n* = 118) facilitated the harmful social media challenge discourse by suggesting new challenges, both intentionally and unintentionally. Posts in this subtheme often manifested humorous attitudes toward serious challenges. For instance, one Redditor discussed a variant of the Cinnamon Challenge with cocaine which could cause serious health risks to participants.

"Does the cinnamon challenge WITH COCAINE" -r/WTF

The posts under the "Social Commentary" subtheme (30%, *n* = 113) presented an understanding of the implications of harmful challenges at a social level. For example, there were posts discussing the media coverage of the cinnamon challenge through the lens of the media ecosystem:

"You know it's a slow news day when they are focused on how dangerous the cinnamon challenge is -r/news

The posts under the "Critique of Challenge" subtheme (22%, *n* = 82) were negative in tone as they manifested disagreement with the harmful social media challenge. This subtheme consisted of posts criticizing participants by calling them "*stupid*" or "*idiot.*" Similar aggressive words were used to describe the challenges themselves and to express how surprised the Redditors were by the challenges:

"Please tell me this 'fire challenge' started on 4chan or somewhere to troll stupid people." -r/videos

In the "Challenge Portrayal" subtheme (17%, *n* = 65), we observed posts with memes and humorous images that reflected the current popular culture, or video parodies of named challenges. Most of the posts in this subtheme had media attachments with short texts in the title. The posts under this theme rarely had commentary about the challenge itself, rather, they simply presented the name of a challenge and embedded a link to a meme:

"The cinnamon challenge." -r/ffffffuuuuuuuuuuuu

In summary, we observed that the pseudo-anonymity of the Reddit communities allowed users to freely express their thoughts. Regardless of intention, some Redditors propagated harmful social media challenges by suggesting new challenges. At the same time, behind a shield of humor, many Redditors criticized the ideas of the harmful challenge, and those who participated in those challenges often used offensive words to emphasize the criticisms.

3.1.3 Prevention/Raising Awareness of Potential Dangers of Harmful Social Media Challenges.

18% (n = 273) of the original posts belonged to the "Prevention/Raising Awareness" theme in which warnings against the potential dangers of harmful social media challenges dominated. In the posts under this theme, sensationalized words that could appeal to one‟s emotion, such as "*dangerous*," "*deadly,*" "*warn,*" and "*kill*" were frequently used. Furthermore, referencing authority figures ("*MDs,*" "*Fire Marshall(s),*" or the "*American Poison Control*") to their claim was a commonality among those posts.

"Charlotte woman arrested after teen burned in Facebook fire challenge" -r/news

In the majority of the posts, Redditors attached "Media Sources" (93%, n = 255) outside of Reddit (e.g., news articles or videos from other social media platforms) to support their warnings against the challenges. The following is an example of a post referring to the consequences of the fire challenge with a video attached that shows the second-degree burns a participant acquired after attempting the challenge, along with the bandages and medication needed to heal his wounds:

"The white guy, who posted the fire challenge, shows an extended aftermath." -r/videos

Media attachments were critical to the prevention/raising awareness theme as they added credibility to the post. We also observed some posts that share "Cautionary Messages" (7%, n = 18), more personal perspectives on the challenges and advice toward the challenge participants. In the following post, a Redditor expresses their personal opinion about the alleged leader and intentions behind the Momo challenge:

"Guy exposes the people who were behind, and were profiting off of, the exploitation of children through the momo challenge." -r/youtube

Original posts discussing more severe consequences were frequently observed in this theme. The cautionary messages related to high-risk challenges such as “suicide,” “fire,” or “Blue Whale” appeared more than twice as frequently as those related to “Tide Pod” and “Cinnamon” did. The titles of the posts related to high-risk challenges resembled the news headline without humorous tones.

“Pradesh police increase efforts to stop the spread of „Blue Whale" challenge, an online challenge ending in the player's suicide.” - r/worldnews

As seen from the above, Redditors took the high-risk challenges seriously and tried to warn fellow Redditors against the negative consequences of participating in harmful challenges. The posts under the “Prevention/Raising awareness” theme surfaced with the idea that the Reddit platform can potentially be used to increase risk awareness and prevent participation in harmful social media challenges.

3.1.4 Seeking Information about Harmful Challenges.

The “Information Seeking” theme (10%, *n* = 154) included posts inquiring about what the challenges are about, as well as, those seeking to learn about others‟ experiences in participating in harmful challenges. The posts under this theme can be characterized by active interactions among Redditors as they navigated the information regarding harmful social media challenges.

“What are the exact 50 challenges in the "blue whale challenge"?” - r/morbidquestions

Most of the posts did not have media attachments as Redditors asked questions about other Redditors‟ personal experiences of challenge participation. Some Redditors were seeking to understand the rules and motivations behind the challenges and participation, and others were

recollecting past challenges. The posts under the "Challenge Experience" subtheme (57%, *n* = 87) asked questions to participants regarding their experience with the challenge such as how the challenge went. Some Redditors asked for links for the high-risk challenges such as Momo and Blue Whale Challenge to consider participating in them despite their dangerous nature.

"Can anyone give me a genuine link that includes the momo challenge?" - r/NoStupidQuestions

The posts inquiring about the challenge experiences and seeking for information to attempt the challenge gave us insights into how potential participants may be drawn or exposed to harmful challenges on Reddit. In contrast, posts under the "Understanding Challenge" (23%, *n* = 36) questioned the reasoning behind participation or the purpose/origins of a challenge of posts. Some Redditors doubted the authenticity of the high-risk challenges such as the Momo challenge and BWC aided by a general lack of understanding of the purpose of the challenges.

"What's the story with the tide pod challenge, As in, Is it a meme that got out of control, or was it actually intended to be attempted?" -r/AskReddit

Under the "Challenge Trend" subtheme (20%, *n* = 31), Redditors also shared posts that recollected old social media challenges or drew comparisons to older challenges to develop new challenges. Posts within this subtheme reminded past social media challenges and at times wondered how the trend in challenges will evolve.

"Older people of reddit, what was your "tide pod challenge" as a kid?" -r/AskReddit

"What trendy Internet challenge would you create like the tide pod challenge?" - r/AskReddit

In summary, Redditors sought information about harmful challenges such as what they were, how they were started, and how to get involved. Oftentimes, Redditors expressed a lack of understanding of harmful challenges, thus, facilitating the discussion of challenges by posting about them on Reddit.

3.1.5 Sharing Self-Made Content and Personal Experiences of Harmful Challenges.

The "Self-Participation" theme (8%, *n* = 125) was the least frequently discussed theme among the original posts. The posts under this theme provided a firsthand perspective of harmful social media challenges such as the experiences with the challenges and descriptions of how the Redditors either failed or completed the challenge. In the following post, a Redditor expresses the reaction after completing the Cinnamon Challenge with a link to moving images:

"*MRW I complete the cinnamon challenge*" -r/MRW

By attaching media files, the original posters expressed how their experiences of the challenges went. For instance, the videos revealed the posters‟ attempts and the images often included memes that showed regret or made fun of their participation. Those who failed challenges expressed their failures often in a light tone, joking about themselves participating in the challenge. In some cases, Redditors were making fun of their younger selves and their prior choices to participate in the challenges retrospectively:

> *"Hey /r/teenagers, have a laugh at 12 year old me attempting to do the cinnamon challenge. I can't even watch the whole thing."* -r/teenagers

Overall, 14% (*n* = 17) of the posts stated that they (Redditors) failed the challenge, while 28% (*n* = 35) claimed that they successfully completed a challenge and wanted to talk about their experiences. For 43% (*n* = 54) of the posts, Redditors did not specify if their attempt was successful or not. The posts in this theme can be characterized by a light and self-reflective tone

compared to the "Others" Participation" theme as we observed less of a harmful and derogatory tone in the posts:

"Our cinnamon challenge went very wrong - but not the way I expected [Video]"

-r/teenagers

While the motivation behind attempting a challenge was usually unclear, Redditors sometimes mentioned the influence of social influence on their decision to attempt the challenge.

"Decided to jump on the bandwagon and try the tide pod challenge..." -r/HotVids

Meanwhile, 15% (n = 19) of the posts in this theme were asking for views, likes, and comments for attempting challenges. The posters of this type of post were seeking encouragement from other Redditors to attempt one of the potentially harmful challenges or trying to promote content involving their participation in the challenges. In the following example, a Redditor is seeking attention (upvotes) from other Redditors by announcing their willingness to participate in harmful challenges:

"If this post gets over 500 upvotes I will try to drink whole gallon of milk, eat a ghost pepper and take the cinnamon challenge back to back." -r/AskReddit

"My goal is to get my cinnamon challenge video to 30,000 views by the end of the day! Help?!?" -r/youtube

In summary, sharing experiences of participating in challenges and self-reflection on challenges from the first-person perspectives was representative of the "Self-Participation" theme. The posts under this theme showed that few Redditors were engaged in harmful challenges (mostly low-risk challenges such as Cinnamon Challenge) or were willing to participate in the challenge for attention.

3.2 Comments Shared in Response to Posts about Harmful Challenges (RQ2)

In this subsection, we present the five topics that were discussed in comments across five different themes in the original posts. To illustrate the topics in detail, we included the example comments with the names of the subreddit from which the comments were taken.

3.2.1. Offensive Commentary.

The "Offensive Commentary" topic was present across multiple themes, including **Others' Participation** (74.9%, *n* = 2510), **Perspectives on Challenge** (55%, *n* = 708), and **Prevention/Raising Awareness** (84%, *n* = 881) themes. In the comments discussing this topic, Redditors mostly considered participants of harmful social media challenges as "*stupid*" or "*dumb,*" often referencing "*Darwin*" and "*natural selection*," to imply that their stupidity would lead to their demise. For example, under the "Others' Participation" theme, Redditors usually commented about participants of the Fire Challenge using the phrase "natural selection," as if they are happy to see the participants getting hurt by the challenge:

> *"Nope, don't feel bad for these people whatsoever. Glad that natural selection is finally putting some pressure on the idiots."* -r/videos

Even in the comments under the "Prevention/Raising Awareness" theme, where the original posters tried to promote prevention and awareness of harmful social media challenges, Redditors called participants of harmful challenges "*idiots*" and "*stupid.*" Instead of engaging in constructive discourse on promoting risk prevention and awareness, some Redditors expressed that the participants deserved negative consequences and that they wanted "*stupid"* enough participants to go ahead and hurt themselves:

> *"If you are stupid enough to do this you actually do deserve to get injured…"* -r/nottheonion

In another example, a funny video of a man doing the Cinnamon Challenge generated a considerable number of offensive commentaries such as how the man looked "pathetic" or simply tried to get more likes and attention by hurting himself and performing the Cinnamon Challenge. Even though the video was fun to watch for many Redditors, some expressed mixed feelings toward the man who put himself at inherent risk to get more attention:

> *"Who the fuck is that guy? Oh my God, I hate him! He looks like he's 30 and dicking around like a retarded kid. Goddamn it, now I have to take a drink. Jesus fuck!"* - r/cringe

In extreme cases, Redditors not only ridiculed but also expressed hate speech toward the participants of the challenges. For example, in the comments under the "Others' Participation" theme, Redditors were laughing at a participant of the Cinnamon Challenge, calling him a "*fat*" and "*midget*" and using racially charged messages to make fun of participants:

> *"A black midget... What's the point of being black when you live inside a mountain."* -r/videos

Overall, in most offensive commentaries, the Redditors ridiculed and criticized the participants of harmful social media challenges in an aggressive and sarcastic tone. Even in the comment thread whose purpose was to promote risk prevention and raise awareness, Redditors still shamed people who attempted those challenges, rather than sharing constructive opinions.

3.2.2. Tips and Tricks for Specific Challenges.

Providing "Tips and Tricks for Specific Challenges" was another prevalent topic across **Others' Participation** (16%, *n* = 536), **Perspectives on Challenge** (32%, *n* = 411), **Information Seeking** (19%, *n* = 257), and **Self-Participation** (10%, *n* = 6) themes. In these comments, Redditors usually provided recommendations on how to perform a specific challenge

or described the strategies that could help perform the challenge successfully (i.e., without getting hurt badly). Some Redditors supported others‟ comments by adding scientific knowledge to explain how the challenge works. Likewise, Redditors often referred to their experience or the experiences of others who attempted that challenge to demonstrate how the challenge should be done properly:

> *"Not that difficult if you know what you're doing. Just take a spoonful of cinnamon and prepare to sit and do absolutely nothing but wait for your mouth to slowly work up moisture. Just make sure not to get any in your throat (don't breathe through your mouth, try to swallow any, or talk) and you should be fine if you wait it out."* - r/AskReddit

Interestingly, we found that new challenges were propagated through comments, where Redditors provided recommendations on how to adjust the harmful challenges to make them easier to accomplish without severe consequences or provided more explanations on how the challenge works and affects the human body. For example, in the comments under the "Perspectives on Challenges" theme, we identified that a new challenge called Hot Chili Pepper Challenge was suggested and how pepper can change one‟s body temperature and, thus, cause the shivering of the participant. In the following example, a Redditor mentioned multiple harmful challenges and how completing one challenge can help the participant to manage the other challenges:

> *"If he does the gallon of milk last he might have a shot. That's assuming he's one of the very small percentage of people who can physically pull it off. If he does them all concurrently the only real problem would still be the physical volume of the milk. The*

milk would most certainly make the ghost pepper and cinnamon challenge manageable." -r/AskReddit

Overall, we found that Redditors helped others navigate harmful social media challenges. Sometimes, Redditors help others ease the consequences of harmful social media challenges by providing strategies based on their personal experience or the experiences of others. In doing so, some Redditors propagated ideas for new challenges.

3.2.3. Suicide Challenges.

In the **Information Seeking** theme (59%, *n* = 798), Redditors discussed topics related to "Suicide Challenges" such as Momo and Blue Whale Challenges. For the suicide challenges, Redditors mostly asked questions about the suicide challenges out of curiosity (e.g., what the challenge is, when and where it started, etc.). Some Redditors expressed the negative emotions that they felt while watching the videos related to suicide challenges.

"That's (video regarding Momo Challenge) terrifying Jesus" -r/NoStupidQuestions

Given the senseless nature of suicide challenges, many Redditors claimed that the challenges are hoaxes, not real phenomena and that the sources of information related to suicidal challenges are unreliable.

*"Apparently it's just a hoax, and the few numbers that *work* are fake too..."* -r/AskReddit

"I've seen a million warning posts, but not a single legitimate post citing a child that was injured or showing a video with Momo embedded in a video.." -r/creepypasta

Redditors sometimes initiated discussions around the suicide challenge at a social level. For instance, some Redditors asked if there are larger problems underlying the phenomena of suicide challenges:

"Are Indian teenagers really committing suicide because of the Blue whale challenge? Nope. It's just kulcha blaming blue whale. They are committing suicide due to depression financial issue unable to find a job and study pressure." -r/india

"I swear this stuff is just mapping the information networks of various social groups for the purposes of opinion manipulation." -r/worldnews

Unlike other topics, we found that comments discussing suicide challenges emerged exclusively under the "Information Seeking" theme. Although the suicide challenges (Momo Challenge and BWC) were frequently discussed in the original posts under other themes (e.g., Prevention/Raising Awareness theme), those posts did not motivate active discussions among Redditors in the comments. That is, Redditors tend not to actively engage with posts discussing high-risk challenges in their comments.

3.2.4. Unexpected Tangents.

Some of the comments were not necessarily about the challenges themselves. We labeled these "off-topics" that did not fit cohesively with one another as "Unexpected Tangents." For instance, in the **Others' Participation** theme (4%, n = 134), Redditors talked about a video of a boy doing the Cinnamon Challenge and a father blaming his son for participating in the challenge and calling him "*eejit*" or "*idiot*." This video has been shared on the Reddit platform for many years and sparked a lot of discussion among Redditors, where they argued about the differences between the word "*eejit*" in Irish and "*idiot*" in English. Overall, most of the Redditors agreed with the father calling his son "*idiot,*" in a humorous ironic tone:

"Angry is just the default Irish response to someone doing something stupid. It's the anger of love.." -r/videos

In the **Prevention/Raising Awareness** theme (10%, *n* = 105), many Redditors discussed a police officer (or "*cop*") who pressured a clerk to swallow a tablespoon of cinnamon and eat 10 crackers in less than a minute. Although some Redditers condemned the officer"s behavior by saying how unlawful this was, the majority of the Redditors discussed something other than the challenge itself:

> *"These cops actually seem funny. Also, I love how the article makes it seem like ADHD is some kind of mental retardation."* -r/CFB

In the **Self-Participation** theme (90%, *n* = 56), Redditors discussed in detail the challenges in which the original posters claimed to participate. For example, when one Redditor posted that they decided to attempt Boiling Water Challenge, Redditors reacted to the post by discussing off-topics such as what solar can do or how to survive a heat wave in a humorous tone.

> "*Beaver events are funny in the hot, I think they should have to leave from a different side than they enter!*" -r/RimWorld

Overall, we observed some Redditors engaged with posts about harmful social media challenges by shifting the discussion topics to "off-topics" that are not closely related to the challenges. When discussing those tangential topics, Redditors tend to express their opinions in a humorous or ironic tone.

3.2.5. Reddit Moderation.

One of the topics that emerged consistently across multiple post themes was "Reddit Moderation" posted by human moderators or Reddit bots. The moderation comments were observed in all post themes except for the "**Self-Participation"** theme (likely due to its small number of comments). The moderation comments were intended for both the original post and

the comments. For instance, some moderation comments were to inform or warn the Redditors about the removal of the original posts due to a violation of the subreddits‟ rules:

> *"We loved your submission, *The new tide pod challenge*, but it has been removed because it doesn't quite abide by our rules, which are located in the sidebar."* -r/wholesomememes

Particularly, personally identifiable information was strongly protected by community rules so that Redditors can freely share their opinion on sensitive topics without worrying about brigading (e.g., a coordinated attack by a group of users) and harassment.

> *"Please censor all information that can be used to identify a person. This includes, but is not limited to: first and last names, usernames (including your own), the symbol next to your username identifying it as your own, your own stupid comments tagging the sub, profile pictures where a person's face is visible, subreddit, group, and online community names, titles of specific posts, and other information like locations (city, state, etc.), addresses and license plates. This is to prevent brigading and harassment, so we take this rule very seriously"* -r/iamverybadass

Oftentimes, the comments were removed because the commenting Redditors‟ accounts did not have enough karma or account age required by the subreddit rules. The minimum requirements to submit comments varied depending on the subreddits and they acted as a preventive mechanism against spam messages posted by new Redditors.

> *"This post has been removed due to Rule 9. This is a forum for reputable investors. Your account must be older than 7 days and have at least 50 comment karma to post."* -r/MemeEconomy

Finally, we found some moderation comments to warn Redditors against posting jokes or off-topic comments. The following example comment explicitly describes that comments that do not conform to the rules of "serious replied only" set by the original poster will be removed.:

> *"**Attention! ** Please keep in mind that the OP (original poster) of this thread has chosen to mark this post with the **[Serious] replies only** tag, therefore any replies that are jokes, puns, off-topic, or are otherwise non-contributory will be removed. If you see others posting comments that violate this tag, please report them to the mods!"* -r/AskReddit

From the above example, we could see that not only community rules that were collectively set by a group of Redditors, but also rules and standards set by an individual Redditor can shape the norms and expected behaviors on Reddit. Overall, different community rules and expectations were set by Redditors within the subreddits and actively enforced by moderating mechanisms. That is, community rules and moderation policies helped set social norms and expected behaviors on subreddits.

4 Discussion

In this study, we explored how harmful social media challenges are discussed on Reddit. Below, we discuss the implications of our major findings through the lens of affordances and social norms of Reddit. Then we suggest design implications to prevent the spread of social media challenges that promote self-harm.

4.1 Reddit: Where Self-Harming Posts Do Not Prevail, but Offensive Comments Do

Our results confirmed that rather than posting their own participants in harmful challenges, most Redditors shared videos or images of others participating in harmful challenges posted on other platforms (i.e., YouTube and Twitter). One explanation for low instances of self-participation posts could be the low-identity affordance of Reddit (i.e., pseudo-anonymity). First, when we looked at the Reddit moderation comments, it was clear that the pseudo-anonymity policy was strictly enforced by Reddit moderation. Since sharing self-participation posts can potentially reveal one‟s identity and violate the platform-wise rules, Redditors may have refrained from sharing videos or images of self-participation.

Another explanation for low instances of self-participation could be the lack of social pressures to participate in harmful social media challenges on Reddit. According to previous literature, social pressure (perceived peer influence on participants‟ social sharing behaviors) may impact individuals‟ behavior (Maheux et al., 2020). Especially, on social media platforms with high-identity affordances (e.g., Facebook), social pressure to share viral content on social media positively affects individuals‟ willingness to participate in viral communications (e.g., viral social media challenges) (Abraham et al., 2022; Borges-Tiago et al., 2019; Roth et al., 2021). Our results indicate that pseudo-anonymity, the low-identity affordance on Reddit, could lower social pressure on Redditors to participate in harmful challenges and share self-participation content with their networks. Taken together, the low-identity affordances of Reddit could contribute to Redditors discussing others‟ participation rather than sharing their own participation.

Meanwhile, our results showed that offensive expressions prevail among comments to mainly highlight Redditor‟s criticism toward the participants of harmful social media challenges. Even the comments under the posts whose nature is to prevent risks and raise awareness did not

take the discussion seriously, but rather continue to ridicule the participants of harmful challenges. According to deindividuation theory, when others cannot identify and no one can evaluate them, individuals are less concerned about social evaluations, hence, lose inner restraints and allow uninhibited behavior to be released (Festinger et al., 1963). On the one hand, the pseudo-anonymity of Reddit allowed its users to share their opinion without being afraid of social judgment or pressure. On the other hand, the low identity afforded by Reddit motivated Redditors to lose their inner restraints and react aggressively toward the participants of harmful social media challenges. This way, the low identity affordance of Reddit could have impacted the prevalence of offensive comments discussing harmful social media challenges.

Given that most offensive commentaries were intended to highlight the negative perceptions toward the idea or participants of harmful social media challenges (rather than to attack other Redditors discussing the same topic), we acknowledge that some level of offensive comments may help set social norms against harmful social media challenges. At the same time, we worry that if inflammatory posts and comments (e.g., mocking other people or abusive verbal attacks) are not properly moderated, this can also contribute to creating environments where the use of inflammatory expressions becomes a norm and culture. In addition, such inflammatory posts and comments may push others in an opposite direction than constructive discussion to demote harmful social media challenges. More importantly, we recognize that the reactions to offensive commentaries can be much more complex than we can foresee. Hence, future work should address moderation mechanisms to reduce offensive commentary to help create healthy norms and cultures to support vulnerable populations rather than ridicule and mock them with abusive language.

4.2 Reddit: The Place Where Harmful Social Media Challenges Are Not Valued

Our results confirmed that the majority of the original posts about harmful social media challenges were critical in tone and that many comments contained negative reactions to the idea of participating in those challenges. That is, there was a dominating theme among original posts and comments that harmful social media challenges are not welcomed. According to social norms theory, social media users are likely to observe the dominating themes among prior posters in their social networks and use those themes to develop social norms and to determine "appropriate" behavior (Festinger, 1954; Khasawneh et al., 2019). In addition, individuals learn what is acceptable and what is not by observing others and behave in ways that they can expect rewarding consequences (Bandura & Walters, 1977). That means, when Redditors observed original posts criticizing harmful social media challenges, they may perceive that harmful social media challenges are not appreciated in their communities, hence, being critical of the challenges is appropriate behavior. Similarly, when Redditors witnessed comments against the idea of participating in harmful social media challenges, fellow Redditors could realize that criticizing the challenges is an expected norm on Reddit. In this way, Redditors can create strong social norms and cultures around the discussion of harmful social media challenges in ways that such challenges are not valued in their communities.

Furthermore, our results showed that the unique social affordances of Reddit (i.e., upvotes and downvotes) can contribute to reinforcing the social norms and cultures surrounding the discussion of harmful social media challenges on Reddit. For instance, Redditors were motivated to post comments criticizing the participants of the challenge to get more "upvotes" from fellow Redditors so that their comments can be promoted or to get more "Karma" or "Awards." At the same time, they were demotivated to post comments encouraging participation in harmful challenges because they did not want their comments to be downvoted and demoted.

Unlike Reddit, most social media platforms (e.g., Facebook, Instagram, Twitter) do not have social features such as upvotes/downvotes. Although YouTube has similar features (i.e., likes/dislikes), the number of likes/dislikes does not affect whether the videos are promoted/demoted. Therefore, in most social media platforms, it would be difficult to observe social norms similar to what we observed on Reddit. As such, the unique social affordance of Reddit such as upvotes/downvotes could have played important roles in shaping the discourse of harmful social media challenges on Reddit.

Additionally, our results confirmed that suicide challenges (e.g., Momo or BWC) or new challenge ideas were not frequently discussed among Redditors. This could be due to another unique social affordance of Reddit, "subreddit rules" and "moderation." In our dataset, some comments were removed by moderators due to a violation of community rules or rules set by the original posters (e.g., "serious replies only"). This indicates that we did not observe high-risk content or content suggesting new challenge ideas prevail because they were already removed by the moderation mechanisms implemented on Reddit. Previous research highlighted that subreddit rules and the degree of moderation play a strong role in setting social norms around what online behavior is expected within the subreddits (Chandrasekharan et al., 2018). Taken together, our results indicate that on Reddit, subreddit rules discouraging posting high-risk self-harming content, coupled with strict moderation mechanisms can indeed create strong *social norms* and environments where suggesting and/or spreading self-harming content is not accepted. That is, social modeling of positive norms such as promoting prosocial challenges and raising awareness toward risky challenges can be an effective intervention strategy to reduce the creation and spread of harmful social media challenges. Future research can examine how effective social modeling strategy is in reducing the creation and spread of risk challenges on social media and

the motivations of online users to promote awareness and/or participate in harmful social media challenges. It can also compare how different subreddit rules and moderation policies impact the portrayal of harmful social media challenges in different subreddits.

4.3 Implications and Applications

The accessibility of suicide-related information on social media can negatively influence the mental health of vulnerable populations to attempt suicide (Gould et al., 2003; Nock, 2008). At the same time, it can provide preventive information and resources for those who seek help (Robert et al., 2015). To minimize the spread of harmful social media challenges and nourish healthy conversations in online communities, we advise the following strategies.

First, social media platforms can design features to provide guidelines for creating safe content. For example, the Reddit platform recently incorporated a banner (including helpful links to external websites) on its main page regarding COVID-19 acknowledging Redditors to stay safe and informed. The same approach can be implemented in the communities discussing social media challenges (e.g., „r/Challenges", „r/StupidChallenges", etc.). Other social media platforms can also implement advisory warnings of potentially harmful social media challenges to help raise awareness. Social media platforms can also consider designing social features that could incentives users for reporting others" posts or comments sharing/suggesting harmful social media challenges. For instance, recently the Reddit platform added a feature to flag posts and comments as self-harming behavior and teamed up with "Crisis Text Line" to support users who might be suicidal or hurt themselves (Mitroff, 2020). They can also consider rewarding users (e.g., giving social currency) for flagging self-harming content or reporting those who indicated participation in self-harming behaviors to motivate users to actively pay attention to and support those who

are at risk. Other social media platforms can add similar features to support their users who could potentially be at risk of self-harm.

As seen from our study, community rules and guidelines help set social norms and expected behaviors within the community. Therefore, we urge online communities to incorporate appropriate rules and guidelines to help users set positive social norms for raising awareness and prevention of self-harming behaviors. For instance, online communities can consider creating rules where sharing self-harming content must be devalued (e.g., downvote, dislike) or even posting self-harming content is banned to actively demotivate users to share such content. This way, social media platforms can clearly signal to their users that risky content is not acceptable in their communities, hence, contributing to creating social norms and culture toward a safe online environment. We also acknowledge the necessity of setting community rules to reduce offensive comments toward participants of harmful social media challenges. The problem of offensive comments can grow in scale because public posts (such as Reddit posts) may be seen by many. More importantly, viewing such comments may have a serious impact on the most vulnerable populations who seek help. Therefore, social media platforms should set clear community rules against posting offensive commentary toward participants of harmful challenges to promote positive social norms around prevention and awareness, rather than abusive norms toward the participants.

Since incorporating rules and guidelines itself does not warrant users to follow them (Khasawneh et al., 2019), there should be appropriate moderation measures in place. As moderating all self-harming and abusive content is labor-intensive, automated algorithms can help detect and moderate such content in scale. For instance, machine learning algorithms can be used to alert community moderators about self-harming or abusive content so that they can

proactively ask the users to remove or edit such content before it is posted. By reducing the density and volume of risky content, social media platforms can contribute to reducing the behavioral contagion effect which has motivated individuals to attempt risky behavior because others are doing it (Polansky et al., 1950). In addition, this will help social media platforms to reduce the chances of self-harming content and offensive comments being publicly available and adversely impacting the mental health of vulnerable populations.

At the same time, our results also confirmed that social modeling of positive norms can be an effective intervention strategy to reduce the creation and spread of harmful social media challenges. Given that positive challenges were more likely to encourage others to participate than negative challenges (Abraham et al., 2022), social media platforms can proactively create and spread positive online content that promotes positive challenges or raises awareness about risky challenges on their platforms. Finally, online communities can also consider adding social features that allow users to moderate self-harming and offensive content themselves. This will also help users to be actively involved in creating strong social norms and cultures where self-harming challenges and verbal attacks are not welcomed. Creating and enforcing positive community rules, along with proper moderation measures will help ensure positive social norms are set for risk prevention and raising awareness toward the harmful social media challenges and eventually promote the mental health of social media users.

4.4 Limitations and Future Work

Due to Reddit's policy of user anonymity, we were unable to ascertain the demographic information of the Redditors included in our dataset. Therefore, we cannot make generalizable statements regarding our results to specific populations. Future research can examine how harmful challenges are discussed on other social media platforms with various populations.

Second, our dataset was scoped based on keywords identified through an exploratory and iterative process. Therefore, other keywords may have yielded other challenges, themes, and topics beyond those that were identified in our results. For example, “social media” AND “challenge” could have given us results with more neutral tones, while “challenge” AND “death” might have given us more results with a critical or warning style. Future work should investigate a wider variety of harmful social media challenges, especially those that are emerging.

In addition, the dataset that we obtained through BigQuery did not contain posts and comments that had been removed by Reddit moderation, and, thus, more risky content (e.g., inflammatory comments) may have already been deleted prior to our analysis. Also, we focused on the analysis of digital trace data to examine the discourse on harmful social media challenges on Reddit. Hence, we were not able to identify the underlying emotions and/or motivations of the Redditors who posted or commented regarding harmful social media challenges. Future work can utilize other empirical methods, such as interviews, surveys, or focus groups to explore the motivations or emotions of users discussing harmful social media challenges.

Finally, while there was a dominating theme among original posts and comments that harmful social media challenges are not welcomed, we found some posts and comments where Redditors were providing tips about how to participate and suggesting new challenges. This means that for some Redditors, the online discussion could potentially be a place where harmful challenges are encouraged or originate. Along with analysis of user posts and comments, future work can explore network features of user posts and comments (e.g., upvote/downvote) to determine which types of posts were encouraged or propagated, and how the propagation could have been subjected to social influence.

5. Impact Statement

This study was one of the first to examine the discourse on harmful social media challenges on Reddit from the perspective of affordances. Through qualitative content analysis and topic modeling, we found that the overall tone of the discourse on harmful social media challenges was critical and that self-harming content did not prevail on Reddit. We explained that low-identity affordance of Reddit contributed to low instances of self-harming content. Also, we argued that the social affordances of Reddit helped create social norms and cultures where sharing/suggesting harmful social media challenges is not welcomed. We highlight that designing affordances to set positive social norms and culture is important to creating healthy online communities that promote risk prevention and raise awareness toward harmful social media challenges. Our work provides the basis for developing interventions to mitigate the risks associated with social media use and promote the well-being of social media users.

Acknowledgment

This research was supported by National Science Foundation grants #1832904 and #1844881. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsor.

References

Abraham, J., Roth, R., Zinzow, H., Madathil, K. C., & Wisniewski, P. (2022). Applying Behavioral Contagion Theory to Examining Young Adults‟ Participation in Viral Social Media Challenges. *Transactions on Social Computing*, *5*(1-4), 1-34.

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence, 3*, 42.

Atherton, R. R. (2021). The „Nutmeg Challenge": a dangerous social media trend. *Archives of disease in childhood*, *106*(5), 517-518.

Avery, A. H., Rae, L., Summitt, J. B., & Kahn, S. A. (2016). The fire challenge: a case report and analysis of self-inflicted flame injury posted on social media. *Journal of Burn Care & Research*, *37*(2), e161-e165. DOI:https://doi.org/10.1097/BCR.0000000000000324

Bandura, A., & Walters, R. H. (1977). *Social learning theory* (Vol. 1). Prentice Hall: Englewood cliffs.

Baumer, E. P., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?. *Journal of the Association for Information Science and Technology*, *68*(6), 1397-1410. DOI:https://doi.org/10.1002/asi.23786

Borges-Tiago, M. T., Tiago, F., & Cosme, C. (2019). Exploring users' motivations to participate in viral communication on social media. *Journal of Business Research*, *101*, 574-582. DOI:https://doi.org/10.1016/j.jbusres.2018.11.011

Borzekowski, D. L., Schenk, S., Wilson, J. L., & Peebles, R. (2010). e-Ana and e-Mia: A content analysis of pro–eating disorder web sites. *American journal of public health*, *100*(8), 1526-1534. DOI:https://doi.org/10.2105/AJPH.2009.172700

Bradner, E., Kellogg, W. A., & Erickson, T. (1999). The adoption and use of „Babble": A field study of chat in the workplace. In *ECSCW''99*(pp. 139-158). Springer, Dordrecht.

Bucher, T., & Helmond, A. (2018). The affordances of social media platforms. *The SAGE handbook of social media*, *1*, 233-254.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate

speech. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1-22. DOI:https://doi.org/10.1145/3134666

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., ... & Gilbert, E. (2018). The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1-25. DOI:https://doi.org/10.1145/3274301

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, *22*.

De Choudhury, M., & De, S. (2014, May). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

Daine, K., Hawton, K., Singaravelu, V., Stewart, A., Simkin, S., & Montgomery, P. (2013). The power of the web: a systematic review of studies of the influence of the internet on self-harm and suicide in young people. *PloS one*, *8*(10), e77555. DOI:https://doi.org/10.1371/journal.pone.0077555

Dyson, M. P., Hartling, L., Shulhan, J., Chisholm, A., Milne, A., Sundar, P., ... & Newton, A. S. (2016). A systematic review of social media use to discuss and view deliberate self-harm acts. *PloS one*, *11*(5), e0155813. DOI:https://doi.org/10.1371/journal.pone.0155813

Ellison, N. B., & Vitak, J. (2015). Social network site affordances and their relationship to social capital processes. *The handbook of the psychology of communication technology*, 203-227.

Emma Hilton, C. (2017). Unveiling self- harm behaviour: what can social media site Twitter tell us about self- harm? A qualitative exploration. *Journal of clinical nursing*, *26*(11-12), 1690-1704. DOI:https://doi.org/10.1111/jocn.13575

Evans, S. K., Pearce, K. E., Vitak, J., & Treem, J. W. (2017). Explicating affordances: A conceptual framework for understanding affordances in communication research. *Journal of Computer-Mediated Communication*, *22*(1), 35-52. DOI:https://doi.org/10.1111/jcc4.12180

Falgoust, G., Winterlind, E., Moon, P., Parker, A., Zinzow, H., & Madathil, K. C. (2022). Applying the uses and gratifications theory to identify motivational factors behind young adult's participation in viral social media challenges on TikTok. *Human Factors in Healthcare*, *2*, 100014.

Faraj, S., & Azad, B. (2012). The materiality of technology: An affordance perspective. *Materiality and organizing: Social interaction in a technological world*, *237*, 258.

Fazlioglu, M. (2013). Forget me not: the clash of the right to be forgotten and freedom of expression on the Internet. *International Data Privacy Law*, *3*(3), 149-157. DOI:https://doi.org/10.1093/idpl/ipt010

Festinger, L. (1954). A theory of social comparison processes. *Human relations*, *7*(2), 117-140.

Festinger, L., Pepitone, A., & Newcomb, T. M. (1963). Some Consequences of De-individuation in a Group. In N. J. Smelser & W. T. Smelser (Eds.), *Personality and social systems* (pp. 125–135). John Wiley & Sons, Inc.. https://doi.org/10.1037/11302-012

Flaskerud, J. H., & Winslow, B. J. (1998). Conceptualizing vulnerable populations health-related research. *Nursing research*, *47*(2), 69-78.

Gabriel, F. (2014). Sexting, selfies and self-harm: Young people, social media and the performance of self-development. *Media International Australia*, *151*(1), 104-112. DOI:https://doi.org/10.1177/1329878X1415100114

Gibson, A. (2019). Free speech and safe spaces: How moderation policies shape online discussion spaces. *Social Media+ Society*, *5*(1), 2056305119832588. DOI:https://doi.org/10.1177/2056305119832588

Gomez-Rodriguez, M., Leskovec, J., & Krause, A. (2012). Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *5*(4), 1-37.

Google Colaboratory. (n.d.). *Welcome to Colab!* Google Colaboratory. https://colab.research.google.com/notebooks/intro.ipynb#scrollTo=5fCEDCU_qrC0

Gould, M., Jamieson, P., & Romer, D. (2003). Media contagion and suicide among the young. *American Behavioral Scientist*, *46*(9), 1269-1284.

Grant-Alfieri, A., Schaechter, J., & Lipshultz, S. E. (2013). Ingesting and aspirating dry cinnamon by children and adolescents: the "cinnamon challenge". *Pediatrics*, *131*(5), 833. DOI:https://doi.org/10.1542/peds.2012-3418

Gray, M. (2019). The boiling water challenge is sending people to the hospital. *CNN Newsource Sales, Inc.* Retrieved from https://www.cnn.com/2019/02/07/us/burns-from-boiling-water-challenge/index.html

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Habib, H., Musa, M. B., Zaffar, F., & Nithyanand, R. (2019). To act or react: Investigating proactive strategies for online community moderation. *arXiv preprint arXiv:1906.11932*.

Haralabopoulos, G., Anagnostopoulos, I., & Zeadally, S. (2015). Lifespan and propagation of information in On-line Social Networks: A case study based on Reddit. *Journal of network and computer applications*, *56*, 88-100. DOI:https://doi.org/10.1016/j.jnca.2015.06.006

Khasawneh, A., Chalil Madathil, K., Dixon, E., Wisniewski, P., Zinzow, H., & Roth, R. (2019, November). An investigation on the portrayal of Blue Whale Challenge on YouTube and Twitter. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, No. 1, pp. 887-888). Sage CA: Los Angeles, CA: SAGE Publications. DOI:https://doi.org/10.1177/1071181319631179

Khasawneh, A., Madathil, K. C., Dixon, E., Wiśniewski, P., Zinzow, H., & Roth, R. (2020). Examining the self-harm and suicide contagion effects of the Blue Whale Challenge on YouTube and Twitter: qualitative study. *JMIR mental health*, *7*(6), e15973. DOI:https://doi.org/10.2196/15973

Khasawneh, A., Madathil, K. C., Zinzow, H., Wisniewski, P., Ponathil, A., Rogers, H., ... & Narasimhan, M. (2021). An investigation of the portrayal of social media challenges on YouTube and Twitter. *ACM Transactions on Social Computing*, *4*(1), 1-23. DOI:https://doi.org/10.1145/3444961

Kırcaburun, K., Kokkinos, C. M., Demetrovics, Z., Király, O., Griffiths, M. D., & Çolak, T. S. (2019). Problematic online behaviors among adolescents and emerging adults: Associations between cyberbullying perpetration, problematic social media use, and psychosocial factors. *International Journal of Mental Health and Addiction*, *17*(4), 891-908. DOI:https://doi.org/10.1007/s11469-018-9894-8

Ksiazek, T.B., Peer, L., and Zivic, A. (2015). Discussing the News: Civility and hostility in user comments. *Digit. Journal. 3*(6), 850-870. DOI:https://doi.org/10.1080/21670811.2014.972079

Leavitt, A. (2015, February). " This is a Throwaway Account" Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 317-327). DOI:https://doi.org/10.1145/2675133.2675175

Lupariello, F., Curti, S. M., Coppo, E., Racalbuto, S. S., & Di Vella, G. (2019). Self- harm risk among adolescents and the phenomenon of the "Blue Whale Challenge": case series and review of the literature. *Journal of forensic sciences*, *64*(2), 638-642. DOI:https://doi.org/10.1111/1556-4029.13880

Maheux, A. J., Evans, R., Widman, L., Nesi, J., Prinstein, M. J., & Choukas-Bradley, S. (2020). Popular peer norms and adolescent sexting behavior. *Journal of adolescence*, *78*, 62-66. DOI:https://doi.org/10.1016/j.adolescence.2019.12.002.

Mayring, P. (2000). Qualitative Content Analysis. *Forum Qual. Sozialforschung Forum Qual. Soc. Res.* 1, 2 (June 2000). DOI:https://doi.org/10.17169/fqs-1.2.1089

Memon, A. M., Sharma, S. G., Mohite, S. S., & Jain, S. (2018). The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature. *Indian journal of psychiatry*, *60*(4), 384. DOI:https://doi.org/10.4103/psychiatry.IndianJPsychiatry_414_17

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272).

Mitroff, S. (2020). Reddit now lets you report users who might be suicidal. *CNET*. https://www.cnet.com/health/reddit-now-lets-you-report-users-that-you-worry-might-self-harm/

Moore, C., & Chuang, L. (2017, January). Redditors revealed: Motivational factors of the Reddit community. In *Proceedings of the 50th Hawaii International Conference on System Sciences*. DOI:https://doi.org/10.24251/HICSS.2017.279

Moreno, M. A., & D'Angelo, J. (2019). Social media intervention design: applying an affordances framework. *Journal of Medical Internet Research, 21*(3), e11014. DOI:https://doi.org/10.2196/11014

Morgan., D. L. (1996). *Focus Groups as Qualitative Research*. SAGE Publications.

Morris, J. (2017). *Reddit as a Cultural Field – Norms and Affordances in Social Media Platforms.* MDIA104 Social and Interactive Media Weekly Blogs. https://mdia104socialandinteractivemediaweeklyblogs.wordpress.com/2017/09/29/reddit-as-a-cultural-field-norms-and-affordances-in-social-media-platforms/

Mueller. C. (2016). Positive Feedback Loops: Sarcasm and the Pseudo-Argument in Reddit Communities. *Appl. Linguist. Teach. Engl. Speak. Lang,* 14. DOI:https://doi.org/10.7916/D8SX7R41

Naidu, S., & Tigani, J. (2014). *Google BigQuery Analytics*. John Wiley & Sons.

Nambiar, P. (2021). *WHY DID INSTAGRAM REMOVE LIKES? CHANGE EXPLAINED!* HITC. https://www.hitc.com/en-gb/2021/03/03/why-did-instagram-remove-likes-change-explained/

Nock, M. K. (2008). Actions speak louder than words: An elaborated theoretical model of the social functions of self-injury and other harmful behaviors. *Applied and preventive psychology*, *12*(4), 159-168.

Norris, M. L., Boydell, K. M., Pinhas, L., & Katzman, D. K. (2006). Ana and the Internet: A review of pro- anorexia websites. *International Journal of Eating Disorders*, *39*(6), 443-447. DOI:https://doi.org/10.1002/eat.20305

Park, A., & Conway, M. (2017). Tracking health related discussions on Reddit for public health applications. In *AMIA annual symposium proceedings* (Vol. 2017, p. 1362). American Medical Informatics Association.

Pater, J., & Mynatt, E. (2017, February). Defining digital self-harm. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1501-1513). DOI:https://doi.org/10.1145/2998181.2998224

Polansky, N., Lippitt, R., & Redl, F. (1950). An investigation of behavioral contagion in groups. *Human Relations*, *3*(4), 319-348.

Pressgrove, G., McKeever, B. W., & Jang, S. M. (2018). What is Contagious? Exploring why content goes viral on Twitter: A case study of the ALS Ice Bucket Challenge. *International Journal of Nonprofit and Voluntary Sector Marketing*, *23*(1), e1586. DOI:https://doi.org/10.1002/nvsm.1586

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*, *34*(3), 1427-1445.

Reddit. (2022). *How does voting work on Reddit.* Reddit. https://www.reddithelp.com/hc/en-us/articles/7419626610708-How-does-voting-work-on-Reddit-

Rice, R. E., Evans, S. K., Pearce, K. E., Sivunen, A., Vitak, J., & Treem, J. W. (2017). Organizational media affordances: Operationalization and associations with media use. *Journal of Communication*, *67*(1), 106-130. DOI:https://doi.org/10.1111/jcom.12273

Robert, A., Suelves, J. M., Armayones, M., & Ashley, S. (2015). Internet use and suicidal behaviors: internet as a threat or opportunity?. *Telemedicine and e-Health, 21*(4), 306-311.

Roberts Strife, S., & Rickard, K. (2011). The conceptualization of anorexia: The pro-ana perspective. *Affilia*, *26*(2), 213-217. DOI:https://doi.org/10.1177/0886109911405592

Robson, K. (2019). Dangerous Detergent: Dealing With the Tide Pod Challenge. In *SAGE Business Cases*. SAGE Publications: SAGE Business Cases Originals. DOI:https://doi.org/10.4135/9781526476739

Ronzhyn, A., Cardenal, A. S., & Batlle Rubio, A. (2022). Defining affordances in social media research: A literature review. *New Media & Society*, 14614448221135187.

Roth, R., Abraham, J., Zinzow, H., Wisniewski, P., Khasawneh, A., & Chalil Madathil, K. (2020). Evaluating News Media Reports on the'Blue Whale Challenge'for Adherence to Suicide Prevention Safe Messaging Guidelines. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1-27.

Roth, R., Ajithkumar, P., Natarajan, G., Achuthan, K., Moon, P., Zinzow, H., & Madathil, K. C. (2021). A study of adolescents‟ and young adults‟ TikTok challenge participation in South India. *Human Factors in Healthcare*, *1*, 100005.

Schouten, A. P., Valkenburg, P. M., & Peter, J. (2007). Precursors and underlying processes of adolescents' online self-disclosure: Developing and testing an “Internet-attribute-

perception" model. *Media Psychology*, *10*(2), 292-315. DOI:https://doi.org/10.1080/15213260701375686

Shanahan, N., Brennan, C., & House, A. (2019). Self-harm and social media: thematic analysis of images posted on three social media sites. *BMJ open*, *9*(2), e027006. DOI:https://doi.org/10.1136/bmjopen-2018-027006

Sharma, E., & De Choudhury, M. (2018, April). Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-13). DOI:https://doi.org/10.1145/3173574.3174215

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *IEEE Access*, *7*, 44883-44893. DOI:https://doi.org/10.1109/ACCESS.2019.2909180

Tosun, L. P. (2012). Motives for Facebook use and expressing "true self" on the Internet. *Computers in human behavior*, *28*(4), 1510-1517. DOI:https://doi.org/10.1016/j.chb.2012.03.018

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015, April). Recognizing depression from twitter activity. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3187-3196). DOI:https://doi.org/10.1145/2702123.2702280

Wallach, H. M. (2006, June). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning* (pp. 977-984). DOI:https://doi.org/10.1145/1143844.1143967

Weekend Edition Saturday. (2019). The Latest Internet Hoax: "Momo Challenge." *Weekend Edition Saturday*. https://www.npr.org/2019/03/02/699663319/the-latest-internet-hoax-momo-challenge

Yan, X. (2020). *Code of Biterm Topic Model.* xiaohuiyan/BTM. https://github.com/xiaohuiyan/BTM

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1445-1456). DOI:https://doi.org/10.1145/2488388.2488514

Appendix.

A. Method: Topic Modeling

Table 4 shows the average coherence scores and topic words generated by LDA and BTM.

Table4: Average coherence score on the top 20 words in topics generated by LDA and BTM.

Topics / Model	Topic 1	Topic 2	Topic 3	Topic 4
LDA (avg. coherence: -136.04)	cinnamon, go, fire, eat, never, take, die, try, mouth, love, year, happen, find, use, maybe, look, hold, keep, old, stop	people, kid, post, right, get, say, guy, meme, stupid, cause, fucking, video, tell, teacher, probably, show, wrong, person, bit, least	lung, water, cinnamon, give, lol, bot, get, sure, thank, try, first, question, cough, guess, breathe, talk, link, make, remember, eat	people, video, think, comment, make, sound, man, day, watch, back, feel, life, kill, thing, word, mean, real, fuck, ever, idiot

BTM (avg. coherence: -82.89)	video, people, guy, cinnamon, watch, try, kid, stupid, tell, friend, seem, fucking, man, eat, year, never, show, pretty, comment, youtube	cinnamon, lung, try, mouth, eat, water, swallow, cough, hold, cause, breathe, nose, air, pepper, saliva, inhale, throat, people, back, fire	post, bot, comment, subreddit, remove, message, moderator, meme, link, contact, video, question, rule, thank, automatically, perform, vote, action, follow, compose	word, irish, idiot, fire, eejit, people, use, top, dad, swear, joke, sell, part, person, accent, sound, first, different, wrong, call

Declaration of Competing Interest

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: