ABSTRACT

Title of dissertation: DESIGNING CYBERBULLYING PREVENTION AND MITIGATION TOOLS

Zahra Ashktorab, Doctor of Philosophy, 2017

Dissertation directed by: Professor Jennifer Golbeck
College of Information Studies
Professor Jessica Vitak
College of Information Studies

While cyberbullying is prevalent among adolescents, attempts by researchers to evaluate mechanisms for it's prevenion and mitigation have been largely non-existent. In this dissertation, I argue that the complex nature of cyberbullying, made more challenging by the affordances of diverse social media, cannot be solved through strictly algorithmic approaches. Instead, I employ multidisciplinary methods to evaluate data generated by teens on social media and work with teens to develop and test potential cyberbullying mitigation solutions. I further argue that solutions focused on improving users' well-being after being targeted online offer designers a valuable tool in fighting back against the harm caused by cyberbullying. Based on the interdisciplinary studies conducted in this dissertation, I offer design recommendations for cyberbullying prevention and mitigation tools. I address the mitigation of adolescent cyberbullying through a multi-methodological approach: 1) data-centric exploratory study of discourse occuring alongside cyberbullying 2) an experimental design of reactions to positive messages in response to cyberbul-

lying 3) human-centered participatory design to design cyberbullying mitigation prototypes and 4) a longitudinal study evaluating the effectiveness of cyberbullying mitigation tools. I offer design recommendations for building and administering cyberbullying mitigation tools. This dissertation begins with a data-centric study to understand *why* users are motivated to post and interact through ASKfm, a social media platform that affords cyberbullying and how anonymity and the site's other affordances affect these interactions. I discuss the unique affordances specific to semi-anonymous Q&A social media platforms and how such affordances enable users to engage in self-disclosure and gaining social support on sensitive topics. I then present two studies to first determine if users will be receptive to anonymous positive messages responding to bullying messages, then to administer positive messages or *Cyberbully Reversal Pings* to Ask.fm users who have received bullying messages. I then use a human-centered approach methodology to co-design cyberbullying prototypes with teens. I use the design recommendations derived from the participatory design study to test the impact of a cyberbullying mitigation system. I address technological mechanisms to mitigate sadness and decline in well-being caused by negative online experiences and cyberbullying. I administer cyberbullying mitigation through technology-mediated memory; in other words, I use positive posts and images participants have previously shared on social media to remind them of existing social support in users social networks. The studies in this dissertation comprise of a mixed methods approach to understand social media platforms on which cyberbullying occurs, work collaboratively with users to design mitigation platforms and ultimately evaluate a cyberbullying mitigation platform with real users. These

aforementioned studies result in design recommendations for building cyberbullying mitigation tools and design recommendations for designing a study to evaluate a cyberbullying mitigation tool.

DESIGNING CYBERBULLYING PREVENTION AND MITIGATION TOOLS

by

Zahra Ashktorab

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Jennifer Golbeck, Chair/Advisor
Professor Jessica Vitak, Chair/Advisor
Professor William Kules
Professor Katie Shilton
Professor Kathryn Wentzel

Dedication

I dedicate this thesis to my loving mother and father, Dr. Farideh Chitsaz and Dr. Hassan Ashktorab.

Acknowledgements

I owe my gratitude to all the people who have made this thesis possible.

First and foremost I'd like to thank my advisors, Dr. Jessica Vitak and Dr. Jennifer Golbeck for their continued support and feedback throughout this process. Without Dr. Vitak's extraordinary theoretical ideas and guidance, and Dr. Golbeck's computational expertise and direction, this thesis would have been a distant dream. Id like to thank my committee members, Dr. William Kules, Dr. Katie Shilton, and Dr. Kathryn Wentzel for their valuable feedback.

I would also like to thank my colleagues for providing much appreciated assistance throughout the various studies in this dissertation. Id like to thank Leyla Norooz for being "researcher number 2" during coding of qualitative data during the last phase of analysis of this dissertation and providing her valuable expertise of qualitative data analysis.

I owe my deepest thanks to my family - my mother and father who have always stood by me and guided me throughout this journey, and have always encouraged me to be the best at whatever I love to do. Words cannot express the gratitude I owe them. Id like to thank my little brother, Yusuf, for making me laugh, bringing me so much joy, and always encouraging me to accomplish my dreams and goals, no matter how difficult. Thanks to my sister, Parnak and my brother-in-law Hesam for your continued support throughout these years. Samaneh, thank you for always being there for me.

Table of Contents

List of Figures

List of Abbreviations

LDA	Lated Dirichlet allocation
CTMM	Curated Technology Mediated Memory
TMM	Technology Mediated Memory

Chapter 1: Introduction

Cyberbullying is an umbrella term that captures instances of bullying, harassment, and intimidation through online social media platforms. With the growing popularity of social media and other forms of computer-mediated communication technologies, incidences of cyberbullying have significantly increased [179]. At least 42% of teens in the United States have experienced cyberbullying [156]. Victims of cyberbullying experience emotional problems like anxiety and depression [123, 141]. Teens who are bullied have a higher risk of suicide, which is currently the third leading cause of death among young people [123, 141, 142].

The advent of social media platforms like YouTube, Facebook, Instagram, and ASKfm provide bullies with a larger and more widely visible platform through which they can harass their victims regardless of temporal or spatial constraints. The affordances of these platforms, such as the high visibility and persistence of posted content make it more difficult for victims to seek refuge from their tormentors [40]. The anonymity/pseudonymity that many sites offer further enables cyberbullying because bullies can hide their identity from their victims. For example, on Twitter, users create a "handle" that need not be connected to their real identity. Likewise, the high rate of abusive comments on question-asking sites ASKfm and spring.me

are attributed (in part) to the ability for users to post anonymously, i.e., without any identifying information showing [145]. Even in identified spaces such as Facebook or messaging services (texts, instant message), we are seeing an increase in the frequency and severity of cyberbullying messages [245]. The negative effects of cyberbullying, which include depression, anxiety and even suicide [2, 46, 112, 132, 219, 224, 242] highlight the critical need for interventions to protect adolescents from the negative emotional effects that such harassing activities cause.

The modes of cyberbullying (flaming, harassment, denigration, impersonation, outing and trickery, exclusion, and cyberstalking) are enacted based on the social and technical affordances of a given platform [149, 195]. Platforms that contain private messaging features enable outing and trickery by allowing a bully to take a private conversation/personal photo and sharing it with a wider audience. Among adolescents, there have been numerous examples of private content, such intimate photos shared between a couple, later becoming widely viewable on social media or through a texting chain, which may have significant social and emotional consequences for the involved parties [176]. Addressing the cyberbullying affordances of some social media platforms becomes even more complicated when considering supposedly ephemeral communication platforms like Snapchat, which purport to delete a message after a specified number of seconds. However, there are several workarounds to capturing shared images and sharing them with a wider audience than the sender intended [283].

Previous research has largely focused on the automatic detection of cyberbullying [80, 145, 279] and has considered embedding such automatic detection systems

within platforms to prevent or mitigate cyberbullying and evaluate how users respond to such systems. Furthermore, there have been no studies that have attempted to evaluate cyberbullying mitigation tools with users. The studies in this dissertation comprise of a mixed methods approach to understand social media platforms where cyberbullying occurs, work with users to design mitigating platforms and ultimately implement and evaluate one such cyberbullying mitigation platform.

While Facebook has held the majority of the public and researchers' interest over the last decade, adolescents are increasingly flocking to other platforms, including Instagram and Snapchat. Teens are seeking privacy from their superiors on social networking platforms on which their parents are not active [87]. Cyberbullying is prevalent among adolescents but attempts to mitigate it so far have been largely lacking or ineffective. In this dissertation, I argue that the complex nature of cyberbullying made more challenging by the affordances of social media, cannot be solved through strictly algorithmic approaches. Instead, interdisciplinary methods should be employed to evaluate data generated by teens on social media and work with teens to develop and test potential solutions. I further argue that solutions focused on improving users' well-being after being targeted online offer designers a valuable tool in fighting back against the harm caused by cyberbullying.

1.1 Dissertation Overview

This dissertation presents a mixed methods approach to build and evaluate cyberbullying mitigation tools. I first begin with an exploratory study of under-

standing the motivations behind the use of ASKfm, a platform infamous for cyberbullying leading to the suicide of youth [156]. The results of this study give insight into why ASKfm users continue to use ASKfm despite it's propensity for cyberbullying. I discuss the unique affordances specific to semi-anonymous Q&A social media platforms and how such affordances enable users to engage in self-disclosure and gaining social support on sensitive topics. In the next chapter, I use the results presented in Chapter 3 as motivation to administer positive messages or *Cyberbully Reversal Pings* to Ask.fm users who have received bullying messages. I discuss three iterations of *Cyberbully Reversal Ping* administration and conclude with a set of design heuristics for cyberbullying mitigation through this mode, as well as ethical considerations for researchers studying similar populations. This study represents one of the first to empirically evaluate the feasibility of a cyberbullying mitigation solution "in the wild" and provides a useful case study for social media developers addressing this critical issue. Chapter 5 describes a Participatory Design study to design cyberbullying mitigation tools with teens. This study presents a human-centered approach to designing cyberbullying mitigation tools for various types of cyberbullying.

The study described in the final chapter addresses cyberbullying and online harassment among young adults and investigates the design and effectiveness of technological mechanisms to mitigate sadness and decline in well-being caused by negative online experiences and cyberbullying. I administer cyberbullying mitigation through curated technology-mediated memory; in other words, I use positive posts and images participants have previously shared on social media to remind

participants of existing social support in users social networks. In a final survey, I measure well-being and negative online experiences to allow for pre-test/post-test comparisons. Based on my results, I provide design recommendations for designing cyberbullying mitigation tools, and recommendations for designing a study to test the effectiveness of a cyberbullying mitigation tool. The overview of the four studies described in this dissertation and the order in which they were implemented can be seen in Figure 1.1.

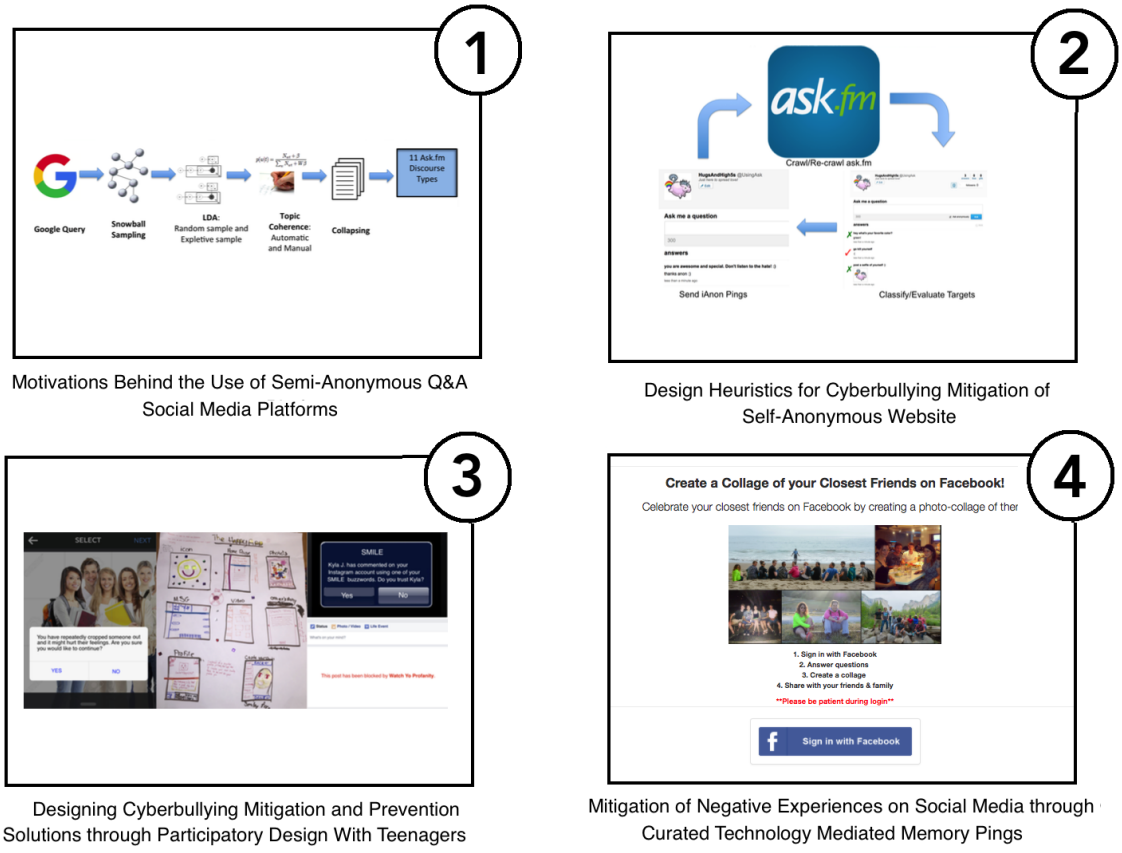


Figure 1.1: Overview and ordering of four studies described in this dissertation.

1.2 Contributions

In this dissertation I offer design recommendations for building and evaluating the effectiveness of a cyberbullying mitigation system. While the primary contribution of the studies presented in this dissertation are design recommendations for cyberbullying mitigation systems, there are other secondary contributions of this study 1) the introduction of a longitudinal data collection method, 2) exploration of methods of automatic detection of cyberbullying, and 3) design heuristics for cyberbullying mitigation tools. The data collection method that is proposed in this study is a contribution to practice and explores the most effective way to engage respondents over a period of time to respond about their social media usage. The feedback from the evaluations in these studies can help form design heuristics for cyberbullying mitigation tools for future designers of cyberbullying mitigation tools.

This dissertation is the first of its kind to merge a diverse array of methods (LDA, Naive Bayes classification, experimental design, participatory design with teens, and experience sampling) to address cyberbullying mitigation and prevention. Furthermore, the participatory design sessions (described more in depth in Chapter 5) include teens in the design process for cyberbullying and I collaborate *with* them to introduce new themes for designing cyberbullying mitigation tools. Furthermore, I evaluate a cyberbullying mitigation tool through experience sampling (described in depth in Chapter 6). This study is the first of its kind to evaluate the effectiveness of a cyberbullying mitigation tool whose design and execution have been informed by both human-centered studies and data-centered studies.

Chapter 2: Literature Review

2.1 Chapter Summary

This chapter provides an overview of literature related to social media use by adolescents, the current state of cyberbullying, and existing cyberbullying mitigation tools. To address the nuanced methods and approaches in the coming chapters, I provide a more specific and granular review of literature for each study in the coming chapters in this dissertation (Chapters 3-7). The literature reviewed in this section is meant to provide background on the use of social media by adolescents, review the current state of cyberbullying and online harassment as well as reviewing existing literature intervention ans support and how it relates to cyberbullying. This chapter's goal to help the reader understand the motivations for my research questions and overarching goals of my dissertation.

2.2 Social Media Use by Adolescents

Adolescents are among the earliest adopters and prolific users of social media sites. Nearly 80% of all U.S. teens own a cell phone, while nearly half own a smartphone and 58% have downloaded apps to their phone [158]. Fully 81% of

online teens use social media [166]. While Facebook remains the most popular site among this age group [166], many other sites have seen significant growth among adolescents in recent years, including Twitter, Snapchat, Yik Yak, ASKfm, and Formspring. Adolescents primary motivations for using social media include social connection [23,199]. As danah boyd [42] notes, teens today view their mobile phones and social media as crucial to keeping in touch, coordinating, and maintaining relationships with their friend network. Other motivations include feeling part of a larger community [263], engaging in creative activities and identity work [23,88] and entertainment purposes [202].

Research is mixed regarding potential outcomes of social media use by adolescents. Several studies have highlighted both positive and negative correlations between specific behaviors (e.g., frequency of use, engagement in social compensation) and adolescents self-esteem [23, 262, 263]. These spaces also enable posting of mean, unflattering, and bullying content. In a large study of adolescents by Patchin and Hinduja [124], they found that 20% of youth in reporting school districts reported being victims of cyberbullying, and 20% reported engaging in cyberbullying at some point in their lives. In a separate study, the Pew Internet Project [158] found that nearly all teens had witnessed someone being cruel online, while 15% said they had been the recipient of mean comments in the last year.

Boyd introduces four affordances that distinguish "network publics" (i.e., social media platforms) from traditional public spaces and highlight the differences between bullying and cyberbullying: persistence, searchability, replicability, and invisible audiences [40]. While communication in unmediated publics is ephemeral,

communication recorded on social networks has a degree of permanence, and a potentially large invisible audience looking on. Hurtful communication in networked publics can affect the recipient as long as it remains visible. Comments and profiles in networked publics are more searchable than they are in unmediated publics, where it is difficult to locate a particular person at any particular time. While it is difficult to replicate a insidious comment that was uttered from one person to another in an unmediated public, the simple act of copy/pasting makes this task extremely easy to replicate in networked publics. Understanding the affordances of social networks and the surrounding discourse and context of insidious communications to be able to efficiently address cyberbullying allow us to efficiently address cyberbullying because they help us more clearly detect it and potentially leverage the social networks themselves to counter malicious behavior.

2.3 The State of Cyberbullying

Bullying has been an area of research in the social sciences long before cyberbullying was even a possibility. Understanding the key reasons behind malicious behavior from bullies is important in this area of research. To understand and study cyberbullying, one must understand the different types of cyberbullying. Understanding the different types of bullying is vital to differentiating which social networking platforms allow which specific types of cyberbullying based on the interactions extant on their platforms. Certain types of cyberbullying may manifest on certain social networking platforms. For example, any social networking plat-

form that allows public discussion and comments can yield flaming, harassment, cyberstalking and denigration. Any social networking platform that allows private messaging can allow flaming, harassment, and cyberstalking. Furthermore, social networking platforms that allow posting photos can lead to outing and trickery. In fact, the suicide by Amanda Todd was purported to be caused after suggestive photographs of her of her were posted on Facebook without her permission [112]. Any social networking platform that allows the formation of exclusive groups, (i.e group pages on Facebook) enables the exclusion of particular individuals and thus qualifies as a type of cyberbullying if done intentionally. Impersonation is possible with any social networking platform that enables logging and having ones own personal account. Thus, different modes of interaction (groups, private messages, public messages) that are available through the various social networking platforms enable different types of cyberbullying.

Research in the social sciences about bullying and cyberbullying informs us about the reasons behind these kinds of behaviors. Beran et al. reveal through surveys that kids who experience cyberbullying are also more likely to participate in cyberbullying [30]. The authors also draw a link between cyberbullying and school bullying through the social rank theory: the theory that one social group will be aggressive in order to dominate their peers so that a hierarchy is established and they can get access to prestige, power and access to resources [93, 207]. Beran et al. develop the claim by Espelage and Swearer, that children cannot be dichotomously classified as bullies or victims [30, 93]. The results of this paper show that children who have been cyberbullied are most likely to cyberbully others. This finding is

particularly relevant to the automatic detection of online bullying. If one wants to label roles (bully, victim, bystander) within a social network, it is important to understand that bully and victim roles are not mutually exclusive [30].

As noted above, researchers have identified a number of affordances of social media platforms that differentiate them from more traditional interaction spaces [40]. Extending this research from networked publics more generally to focus specifically on cyberbullying, they highlight how instances of cyberbullying differ from more traditional conceptualizations of bullying, such as bullying that occurs at school. For example, the high visibility and persistence of mediated interactions allows people to revisit and share content long after it has originally been posted; in the case of cyberbullying, this gives online harassment a longer lifespan than in more ephemeral, face-to-face interaction [121]. Likewise, searchability and replicability enabled through search features on Youtube and other platforms, create new audiences for cyberbullying acts and may encourage particularly negative events to spread virally [3, 66, 209]. When combined with the public nature of these communications and the presence of "invisible audiences" (i.e., individuals outside of the intended audience who can view the content) and disinhibition [252], these spaces can become breeding groups for negative interactions.

While there is no consensus about the exact definition of cyberbullying [177], many researchers agree that cyberbullying inherits the criteria for defining bullying: intentionality, repetition, and imbalance of power [198]. Researchers claim that defining cyberbullying has two additional circumstances to consider: anonymity and public interaction versus private interaction [244]. Understanding the key rea-

sons behind malicious behavior from bullies is important in this area of research. The negative outcomes associated with adolescent cyberbullying – including suicide in some cases– highlight the critical need for researchers to develop tools to alleviate the depression and anxiety that victims commonly experience. Most research on cyberbullying has focused either on the social science theories that explain the underlying causes of cyberbullying or the automatic detection of cyberbullying posts [30,145,279]. With few exceptions [79], very little research has occurred in the realm of interface design and mitigation of depression and anxiety specifically for cyberbullying, and none to date have examined cyberbullying mitigation interfaces "in the wild". Findings from this study provide important implications for future research combining automatic detection of cyberbullying with human-generated responses, including ethical considerations for research in this field.

2.4 Bullying and Intervention

School bullying is a pervasive longstanding phenomenon [218]. Schools address bullying by using the curriculum to teach about bullying, and working directly with students who are involved in bullying situations [246]. Results from a survey of 15686 students in grades 6 through 10 in US in public and private schools who took the World Health Organizations Health Behavior in school-aged children revealed that 29.9% were involved in frequent or moderate bullying [187]. A study found that direct verbal aggression is the most common form of bullying which occurs at the same frequency with both sexes. Indirect forms of bullying are more common

with girls, while more direct forms of bullying are more common with boys. Among indirect forms of bullying frequent among girls are: taking of personal belongs, rumors, teasing, reject and name-calling [218]. A 25-year study (1980-2004) looked at intervention mechanisms by schools in Europe and the United States, and found that interventions produced clinically important positive affects for about one third of the 15,387 participants in the study. The authors conclude school bullying interventions are likely to influence self-perceptions and self-esteem positively than actually preventing bullying behaviors [178].

2.5 Existing Cyberbullying Mitigation Tools

Existing literature reflects that there have not been many formal studies focusing on the design of cyberbullying mitigation tools in the realm of research. However, many independent developers have created tools to help promote well-being of social media users. For example, the application"you're valued" searches Twitter for tweets that say "nobody likes me" and then tweets the user with a response of "I like you", "You're valued" or "You matter" [272]. Another application, Honestly, looks to combat cyberbullying by asking friends of a particular user questions about a person like "Can I sing well?" In an attempt to boost the self-esteem of the user, only the positive responses are shared with the user [241]. While the intent of all of these applications is to mitigate low self-esteem and low confidence (one of the effects of cyberbullying), none have included children and adolescents in the design process to gauge the potential impact of such interventions. Until this point, there

has been very little work on the design of technological solutions for cyberbullying. For instance, there has been work on cyberbullying that focuses on community involvement and parental responsibility to address the problem (i.e., education) [74]. Dinakar et al. [79] introduce "reflexive interface" prototypes as a means to prevent cyberbullying across a limited range of subjects: appearance, intelligence, racial and ethnic slurs, social acceptance, and rejection. The reflective interface encourages positive digital behavioral norms and consists of the following interactions in order to deter malicious behavior: notifications, action delays, displaying hidden consequences, system-suggested flagging, and interactive education. The reflective interfaces to mitigate cyberbullying did not involve youth in the design process or been evaluated.

In sum, research to date has focused largely on categorizing types of cyberbullying, its potential harms and, to a lesser extent, detection and mitigation of cyberbullying. In the next chapters, I present results from four mixed methods studies to expand upon this research by developing and testing a cyberbullying tool designed in partnership with high school students.

Chapter 3: Motivations Behind the Use of Semi-Anonymous Q&A Social Media Platforms

3.1 Chapter Summary

ASKfm is a social media platform popular among teens and young adults where users can interact anonymously or semi-anonymously. In this chapter, I identify the modes of disclosure and interaction that occur on the site, and evaluate why users are motivated to post and interact on the site, despite its reputation for facilitating cyberbullying[1]. By understanding motivations behind the use of anonymous platforms infamous for cyberbullying, design recommendations for improving such platforms will be better informed since the users will be better understood. Through topic modeling–supplemented with manual annotation–of a large dataset of ASKfm posts, I identify and classify the rich variety of discourse posted on ASKfm, including both positive and negative forms, providing insights into the why individuals continue to engage with the site. These findings are complemented a survey of young adult (ages 18-20) ASKfm users, which provides additional insights into users' motivations and interaction patterns. I discuss how the affordances specific to platforms like ASKfm,

[1]This study was a collaborative project done with Jennifer Golbeck, Eben Haber and Jessica Vitak. It was presented at the 2017 ACM Web Science Conference.

including anonymity and visibility, might enable users to respond to cyberbullying in novel ways, engage in positive forms of self-disclosure, and gain social support on sensitive topics. I conclude with design recommendations that would highlight the positive interactions on the website and help diminish the repercussions of the negative interactions.

3.2 Introduction

Recent years have seen a rise in the use of social media platforms that afford anonymous communication such as ASKfm and Formspring [36,285] and mobile applications that allow anonymous sharing like YikYak and Kik [137]. While anonymous online communication has existed for decades (e.g., Usenet, anonymous chat rooms) [117,220], platforms like ASKfm are novel because they allow users to anonymously communicate with known recipients (i.e., semi-anonymous communication). Because anonymity has been shown to lower people's inhibitions [252], it is not surprising that these platforms have been used for cyberbullying [128,129,161]. While Formspring shut down in 2015, ASKfm (which is based on Formspring's interaction model) remains quite popular among young users, suggesting that the anonymity that likely leads to problematic interactions may also enable positive outcomes for users. In this chapter, I examine ASKfm to better understand how users interact, as well as the impact of semi-anonymity on those interactions.We make design recommendations for semi-anonymous platforms that foster the positive interactions that lead to social support and self-disclosures. I make recommendations that would

help to mitigate and diminish the repurcussions of the negative interactions on the platform.

ASKfm facilitates a variety of interactions between users with different degrees of anonymity: it allows users to follow others anonymously and to send anonymous or pseudonymous questions to specific known recipients in exchanges visible to all of the recipients' followers. ASKfm users can also non-anonymously indicate approval of an exchange (i.e., "liking" it), and give virtual "gifts." While ASKfm describes its central interactions in terms of questions (the profile post prompt is "Ask me a Question"), users' interactions are much more diverse and represent a variety of types of discourse. ASKfm should not be confused with Q&A sites that allow the posting of information-seeking questions to a broad community, soliciting answers from any member (e.g., Yahoo! Answers, Google Answers, Stack Overflow) [6, 73] or with more general platforms like Facebook and Twitter, where interactions can include non-anonymous questions directed to an individual or broadcast to a wide audience. On ASKfm, questions are directed to a particular individual, and are often posed anonymously.

Cyberbullying is prevalent among adolescents, but attempts to mitigate it so far have been largely lacking or ineffective. In this dissertation, I argue that the complex nature of cyberbullying, made more challenging by the affordances of diverse social media cannot be solved through strictly algorithmic approaches.

To ultimately make design recommendations to improve social media platforms, we must understand the behaviors and interactions pervasive on these platforms. In the later chapters, I take a human-centered approach to the probing and

exploration of the social media environments by asking users directly about their experiences and including them in the design process. However, there is value and insight to be gained from examining the first-hand data that is created by users when they interact on social media platforms. Automatic methods of analysis of users' exchanges and interaction can give insight into their behaviors and the specific characteristics of social media platforms that afford these behaviors. In this chapter, I take a data-centered approach to understand the prevalent behaviors on ASKfm. I evaluate the following research questions:

RQ1 What kinds of interactions occur alongside cyberbullying discourse on ASKfm?

RQ2 How does anonymity on ASKfm shape users' disclosure and interaction behaviors on the platform?

RQ3 What kinds of design changes can make ASKfm a safer space?

I address these research questions through two studies. First, I analyze user data collected from ASKfm–the publicly available data for each user, including basic profile information as well as exchanges with other users within an individual's network–and conduct topic modeling on these exchanges, followed by coding of data to discover the diverse kinds of interactions and discourse that occur on ASKfm. I identify 11 types of discourse posted on ASKfm, including many positive modes that provide examples of why individuals continue to engage with the site. I extend these computational findings through a survey of young ASKfm users to further understand users' motivations for disclosure and interaction, especially when they have negative experiences on the site.

This study contributes new insights into the benefits and drawbacks of online anonymity, as well as how adolescents and young adults navigate an online space that can be fraught with negativity and harm. The findings from these studies will in later chapters support my decision to focus on cyberbullying mitigation Facebook , a platform on which people are identified. Based on my findings, I push the discussion of anonymous interactions [137] beyond the standard focus on negative and bullying messages to consider the range of positive and negative outcomes associated with site use. While I acknowledge the importance of minimizing user risks, my study highlights how and why these sites are useful to young people, i.e., by providing an outlet for interactions that may be perceived as stigmatizing in less anonymous environments.

3.3 Related Work

Current research on semi-anonymous websites has largely focused on automatic detection of cyberbullying [127, 145, 216] and specific exploration of cyberbullying practices [36,128,183]. By broadening the focus to consider all potential motivations for disclosure on these platforms, I can better explain users' motivations for engagement and continued use. In order to make recommendations on how to improve the platform, I must understand the context in which the cyberbullying occurs. Below, I provide an overview of the ASKfm platform and discuss the state of existing research on anonymous and semi-anonymous social media platforms.

3.3.1 ASKfm: Description of Platform

ASKfm claims 37 million unique global visitors per month [196]. ASKfm's interaction model comprises of asking questions and reacting to those questions. On ASKfm's official Linkedin page, it says "At ASKfm, my premise is simple: I believe questions and answers are the building blocks of conversation, self-expression and deeper understanding." When navigating to a user's profile, a box prompts you to "Ask @User" a question, as seen in figure 3.1. The question can also be asked anonymously (by checking the "Ask anonymously" box) or non-anonymously. A user's profile displays the questions that they have chosen to answer. These questions can be be liked and shared on other social media platforms (Facebook, Twitter). When a question is sent to a user, a user views it on their "Questions" page which is only visible by the recipient. The question is only published to the user profile if the recipient chooses to answer the question. The "Friend" feed displays all questions answered by individuals that a ASKfm user follows.

In recent years, ASKfm has received significant media coverage related to cyberbullying occurring on the site; for example, a Google Search of "ASKfm" reveals the headlines "10 Frightening facts about ASKfm all parents should know" and "ASKfm: A Guide for Parents and Teachers - Webwise". These headlines are a reflection of the reputation garnered by suicide incidents reported over the years that were thought to be the direct result of cyberbullying on the website [47]. While this reputation persists, not all users are engaging in cyberbullying. Other types of interaction exist on the website, and this work aims to explore these types of

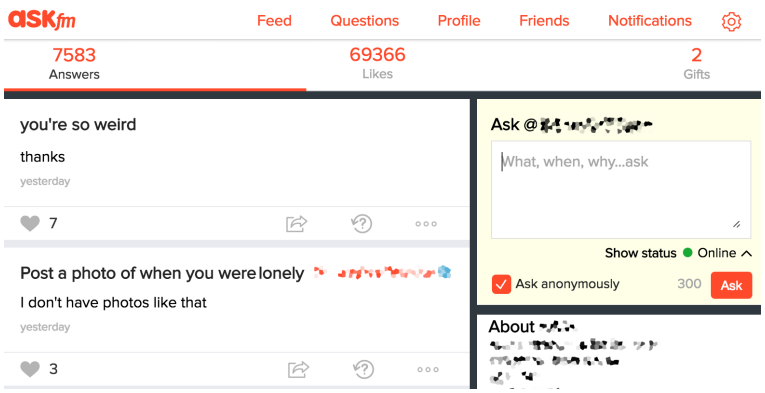


Figure 3.1: Example of a ASKfm user profile. From the top, the first question is asked anonymously. The second question is posted by an identified person (their name has been filtered next to the question). All those who liked the question-answer pair are visible if heart under the question-answer pair is clicked. The box that prompts a user to ask a question is on the upper right hand corner. *Content has been changed and filtered to both reflect the reality of the content of these posts and to protect the identity of people involved in this post.*

interactions.

3.3.2 Discourse on Semi-Anonymous Social Media

Previous studies on semi-anonymous Q&A websites have focused on detecting and understanding the nature of cyberbullying behaviors [128,129,145]. Research on semi-anonymous websites such as Formspring and ASKfm has primarily explored negative interactions, with a specific focus on cyberbullying because of the links between cyberbullying and teen suicide [112, 161]. For example, Kontostathis et al. [145] used data from Formspring to automatically detect cyberbullying content on the site. Hosseinmardi et al. [128] explored ASKfm by examining the occurrence

of "negative" words and interactions on ASKfm and found that individuals with negative content on their profiles are less active and the least sociable. They also found that as positive words increase on a user's profile, the more active and engaging that user will be. Moore et al. [183] evaluated cyberbullying and anonymity on Formspring and identified aggression in both online attacks and defense posts (i.e., posts that defend the victim); they further noted that anonymity correlated positively with attacks and negatively with defense posts.

3.3.3 Anonymity, Disinhibition, and Online Behavior

Anonymity has been associated with anti-social and disinhibited behavior [152, 252]. Suler first introduced "dissociative anonymity" by noting that when anonymity results in online disinhibition, people act out more frequently or intensely than they would in person [252]. Anonymity enables a process of disinhibition where users can separate their actions from their actual identity, making it easy for them to act out [133]. Yet disinhibition is not necessarily bad; Christopherson [57] gives the examples of "Catharsis" and "Autonomy" as positive outcomes of anonymity, where people can experiment with new behaviors and express themselves unhindered without fear of social consequences.

Kang et. al. contend that ephemerality is an intrinsic part of anonymous communication platforms [137]. For example, the median life of a post on 4chan is 3.9 minutes [33]. However, semi-anonymous social web applications like ASKfm do not embrace the same ephemerality as other fully anonymous social web applications since posts are recorded on user profiles. Literature on anonymous media

applications further reveals that anonymity empowers individuals to disclose personal information [137]. Yun identifies three types of anonymity: "Self Anonymity", "Other Anonymity", and "Discursive Anonymity". Self Anonymity is when a user is anonymous to others; Other Anonymity is when other people are anonymous to a user and Discursive Anonymity is when language use and writing style are not personally identifiable pieces of information about a particular user [285]. All three types of anonymity identified by Suler are present on semi-anonymous social media platforms: users can opt to be anonymous to others; others can opt to be anonymous to a particular user; and someone's language use and writing style may further anonymize them if it is not personally identifiable. Through an analysis of personal journal blogs, Hollenbaugh et al. demonstrate that those who share photos of themselves tend to participate more in self-disclosure, revealing more information about themselves. However, they found that younger participants and women participate more in self-disclosure, revealing more information and discussing a breadth of topics when their names are anonymous (discursive anonymity). This study makes a distinction between discursive anonymity and visual anonymity, suggesting that users believe visual cues (such as photos) to be less identifying than discursive cues (like real names) [126].

3.3.4 Positive Outcomes of Anonymous Disclosures

Research has identified a number of positive outcomes associated with anonymous disclosures. Self-presentation is done through self-disclosure [233], revealing personal information about yourself which is compatible with the image a person is

trying to project about themselves and is an important step for the development of close relationships [138]. Kang et. al. [137] observed that a high degree self-disclosure (sharing of private personal information) occurs in anonymous mobile communications like YikYak because users felt comfortable sharing private information about themselves without the risk of being judged by their network of friends. They found comfort in the anonymity and thus were able to disclose information about themselves. Likewise, numerous researchers have identified benefits to pseudonymous health forums, especially in cases of stigmatized or rare diseases, where individuals may find it difficult to find people to talk to in their offline settings [70, 253, 261]. More recent work has highlighted positive uses of the social media platform Reddit for highly sensitive topics like discussions of sexual abuse [11]; in addition to pseudonymity, the site offers additional features to further separate a poster from their permanent identity (e.g., temporary accounts; see [154]). Individuals who have experienced any form of past trauma are more likely to use Web-based services when they can do so anonymously [136]. Schoenebeck contends that websites like You Be Mom, a online social outlet for mothers that allows anonymous communication provides a safe forum for moms to "trespass social norms and expectations" [236].

In summary, the existing literature on anonymous mediated interactions provides a conflicting picture. On one hand, sites that facilitate anonymous interactions may encourage cyberbullying and other negative behaviors. On the other, there is significant potential for positive outcomes, especially in the form of social support, to be generated from semi-anonymous disclosures. While research has already established the benefits of anonymous platforms to narrowly focused communities like

cancer forums, I will now explore the motivations for disclosure on more general question-asking sites. In the following sections, I describe findings from two studies, including topic modeling of data from more than 40,000 ASKfm users and survey data from 243 young adults who actively used the site.

3.4 Study I: Discourse Discovery on ASKfm

In my first study, my goal was to discover the the kinds of interactions and discourse that occur alongside cyberbullying discourse on ASKfm (**RQ1**). The primary mode of interaction on ASKfm involves one user sending a question or message to another user. By default, these messages are anonymous, but with an affirmative action (unchecking the box immediately under the post) the initiator can make their identity visible. The recipient can then reply to the post publicly or privately. On ASKfm, users have public profiles, yet the act of "following" or friending another user occurs anonymously – i.e., a user is aware of the number of followers but does not know who is following them [19]. A user can infer who is following them by the exchanges they receive on their profile like questions and likes on questions. Each person's feed consists of exchanges which belong to individuals whom that person follows. Each person can express approval for other people's exchanges using a "like" mechanism, and each user has a page in which their "best" exchanges are viewable, i.e., those exchanges that received the greatest number of "likes" from user's network. The "Like" mechanism is exchanged non-anonymously. This enables a mix of anonymous and non-anonymous interactions; for example, I

observed anonymous posts soliciting "likes" from the broader anonymous network. This interaction is better described through Figure 3.1, which shows an anonymous question being asked. Those who like the question however, are identified.

3.4.1 Data Collection

Studies have demonstrated that websites that allow anonymous question-asking experience greater cyberbullying [127,145,216]. I therefore began the data collection process by searching ASKfm using common terms that are *unambiguously* associated with cyberbullying. My first step was to query ASKfm through Google for variations of the terms "go kill yourself" and "go die" [112, 161]. I used this as the starting point for my data collection for two reasons: (1) to capture discourse that occurred alongside cyberbullying and these terms are unambiguously malicious and (2) to explore how individuals who have instances of such behavior on their profiles or within their network used ASKfm to engage in other types of interactions. I pulled data for subsequent analysis by crawling users who interacted with the original Google search result users (through likes and questions) and then crawling their interactions iteratively (snowball sampling) with a python crawler). The data was stored in a SQL database for analysis.

I acknowledge that my sampling choice is limited by the fact that the "like core" my sample are users who used a variation of terms "go kill yourself", so the data explored in this study are likely biased toward more negative forms of interaction. However, my results show that despite the source of my sample, my snowballing was able to capture a wide variety of positive discourse types in addition

to negative discourse types. Furthermore, my sampling choice was influenced by media attention directed towards ASKfm centered around the many individuals who have taken their lives because of comments instructing them to using terms like: "drink bleach", "go die", and "every1 will be happy if u died" [9, 36, 47, 248]. I used these terms to capture the most extreme cases of cyberbullying and the discourse that occurs alongside these negatively valenced interactions. I break down the percentage of discourse types later in this chapter and demonstrate the large presence of other discourse types despite the search query starting point.

I used Google search to find these posts, with the initial search yielding 19 public profiles. Before collecting user information, I contacted ASKfm and notified them of my study and data collection process.At the time of data collection (October 4, 2013), ASKfm's Terms of Service had no restrictions on users scraping or otherwise collecting user information. The interactions on ASKfm have remained the same between 2013 and the current version with the only changes occurring in color and aesthetic. AskFM does not have an Application Programming Interface (API). I expanded my search in a snowball fashion, collecting user information from all public profiles of those users who had liked exchanges on the "best" pages of the 19 original profiles. I repeated this method of data collection until I yielded over 8 million exchanges from over 40,000 users. I collected user profile information includes username, user biography, user headline, the 100 most recent exchanges with time posted, author information if applicable, and the respective answer. Additionally, I collected the 25 "best" exchanges for each user, which also includes respective time posted, answer, author information (if question was non-anonymous), and likers of

that particular exchange.

3.4.2 Topic Modeling to Discover Discourse Types

Once I collected the very large corpus of posts, my next task was to discover the different types of discourse that occurred between the users. I didn't find prior examples of similar analysis of social media text, so I decided to approach the data using Latent Dirichlet Allocation topic modeling (LDA). Previous researchers have found topic modeling to be an efficient way to automatically discover topics, organize, and categorize large amounts of text [37] [213] [250] [287]. I detail how I refine the LDA process, first by using Minmo et al.'s topic coherence algorithm followed by human annotation of topics to ensure only topics that were coherent were included in the study.

While LDA is widely used for discovering topics and analyzing text, I acknowledge it's limitations. One limitation is that a user pre-defines the number of topics K. The size of K can lead to nuanced topics that overlap semantically or more general topics. Another limitation of topic modeling is that each topic is generated in the form of the most common keywords found across documents in the topic - interpretation of the meaning left to the user [20]. I addressed both limitations by 1) repeating the analysis with a wide variety of values for K, 2) using an automated measure to select those topics with the highest topic coherence values for each value of K, and 3) using multiple human annotators to validate topic coherence, label each topic, and collapse overlapping topics. The flow from data collection to the resulting topics is summarized in Figure 3.2.

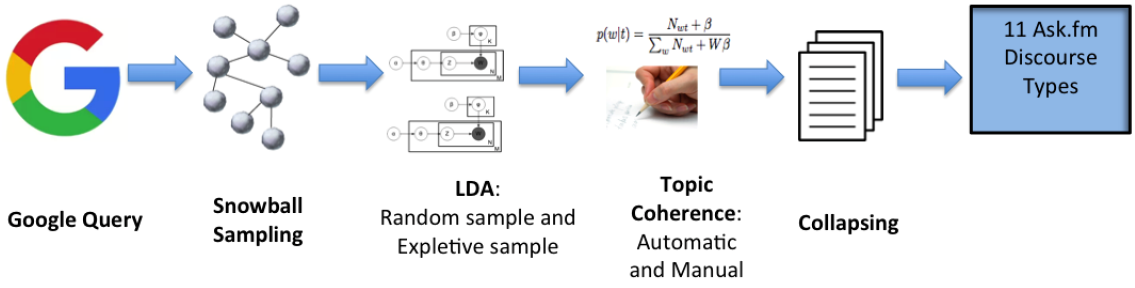


Figure 3.2: Methodology Pipeline: Query, LDA, Topic Coherence, and Collapsing

For this analysis, I ran LDA on two samples of data. The full dataset was not analyzed because of processing limitations and I felt a random sample would be representative of the different types of discourse. My first sample of data consisted of 300,000 random documents from my data set. For LDA purposes, I define documents as ASKfm exchanges comprising of a question and answer combination. My second sample of was filtered for a list of expletives to contain approximately 80,000 documents. I used an expletive sample because I wanted to ensure that I captured any hint of cyberbullying in my dataset. While cyberbullying discourse may not necessarily include expletives, previous studies have found that the existence of expletives in documents are indicative of the presence of cyberbullying [145].

LDA requires, as input, a specific number K of topics to search for, and not all topics found will be coherent. Since I did not know, a priori, the number of topics covered on ASKfm, I ran LDA asking for $K = 10$, then again with $K = 20$, then with $K = 30$, all the way to $K = 100$ topics. This produced a total of 550 topics. My next step was to automatically calculate topic coherence for each topic. Minmo et al. [181] demonstrate that calculating topic coherence is an efficient

way of evaluating a topic model. They define topic coherence is the measure of co-occurrence of the top words within a topic in a document. Highly coherent topics were found within each LDA run. I wanted the data to inform my categories, so from each LDA run, I selected the most coherent 20% of topics as my initial categories for qualitative coding, resulting in 110 coherent topics.

We then performed a manual evaluation for each coherent topic. My evaluators were graduate students in Computer Science who had conducted research on social networking platforms. A total of 12 annotators were recruited to annotate the coherence for each topic and rate whether a particular topic was coherent. Each topic was evaluated by three annotators. For each topic model, I presented the words in the topic and I computed the most probable documents that would fall in that topic model group. I asked evaluators whether they thought that the documents were coherent and belonged in the same category. I also asked evaluators to "in their own words" to label the topic group. These labels were later used to collapse similar topic models.

Cohen's Kappa inter-rater reliability score was calculated for each pair of annotators, and the total average was 0.8125 [174]. From the topic models, I selected those categories deemed coherent by all evaluators (scored higher than 3 by all three raters on a Likert score), resulting in a total of 53 coherent topics.

I then analyzed the 53 user-validated coherent categories and collapsed the categories based on similarities in the labels given by annotators. For example a topic model resulting from $K = 30$ (top words: love beautiful perfect xxx amazing gorgeous aw babe girl xx thankyou sweet lovely stay he) was judged very simi-

lar to a topic model resulting from $K = 10$ (top words:love lt xxx haha xx omg thoughts hahaha amazing funny pretty aw nice cute hahah bby aha omfg). The annotators labelled both with: "complimenting a friend", "positive sentiment" and "compliments". Given the similarity in topic labels, I combined these topics into the emergent "Compliments and Positivity Discourse" category. I combined overlapping topics manually based on similarities in labels generated by human annotators. The categories represent the most coherent discourse types from the LDA sample.

Combining the topic model groups based on overlapping labels resulted in 11 distinct categories (detailed below). I acknowledge that these categories do not cover *all* types of interactions, but these represent the most common modes of interaction from my sample.

1. **Bullying/Inflammatory/Insulting Discourse:** malicious messages aimed to threaten or insult the recipient and may include in/direct threats and expletives. Responses may reciprocate inflammatory remarks.

   Question *oi mate when you go back to school am gona f****** stab you and im gona beat the shit out of you and im gona put you in hospital*

   Answer *i never new that mate come find me you c****

2. **Compliments and Positivity Discourse:** kindness and compliments; characterized by compliments directed at the recipient.

   Question *you are so beautiful you are the nicest girl ever you have the coolest personality*

   Answer *awe c: it did make me smile .thats super nice of you to take your time and make me feel good about myself cx and umm...can i know who you*

are ?

3. **Defense of Bullied Victim Discourse:** message to cyberbullying victim in defense of previously receiving negative comment. Posters often tell victim to disregard inflammatory remarks and are sometimes include a compliment to mitigate harm.

   Question *dont listen to people who send you hate just remember that i love you and haters are gonna hate on how pretty you are its probably some fat little c*** behind a computer screen who cant say it to your face*

   Answer *phhahahhahahaha that made me laugh :)*

4. **Like Solicitation and Rating Discourse:** asks that whoever "likes" the discourse will receive some sort of interaction on the website through "rating", "compliments", or reciprocated "likes". In the example below, each liker is promised a certain amount of reciprocated "likes" on their profile.

   Question *Likers get 5 likes and 5 questions?*

   Answer *like if you want this x*

5. **Listing All people you follow:** asks a user to list everyone they follow on the site (via @username). This discourse type reveals "hidden" information as the site structure prevents users from seeing their followers list unless they receive a "like" interaction or are tagged in a discourse type like this one.

   Question *list of people you follow*

   Answer *@[redacted] @[redacted] @[redacted]*

6. **Picture/Video Request:** asks for a picture/video of the recipient; sometimes coupled with a conditional that states if a user receives more than a certain

amount of likes, they are deemed "pretty."

Question *selfie?*

Answer *[redacted].jpg*

7. **Preference Questions:** asks about a user's preference in movies, music, pets, jewelry, etc. Answers associated with these questions tend to be straightforward.

   Question *do you prefer gold or silver jewelry?*

   Answer *haha gold but i only have silver jewelry*

8. **Self-Harm Discourse:** questions inquiring about someone's opinion on self harm, whether they participate in it, and how they engage in self-harm (e.g., cutting, starvation).

   Question *what is your opinion on self harming? x*

   Answer *i think its a horrible thing. for someone or something make someone feel like they should use their skin as paper. some people right now feel unwanted or ugly or fat or like they just dont belong with the world because theyre being bullied*

9. **Sexual Content Discourse:** exchanges that are sexual in nature. These questions often ask for sexual favors or preferences in sexual exchanges.

   Question *big boobs, small butt OR small boobs, big butt?*

   Answer *small boobs, big butt*

10. **Things that Annoy you/you Hate:** questions about users' dislikes. Answers to these question vary from hatred of things like "spiders" or hatred of things that "guys do."

Question :*something you hate?*

Answer *actually hate everything. i hate guys who say they hate this girl and then they text every f****** day. i hate people who always try to start shit*

11. Thoughts and Opinions: asks a user's attitudes toward the question asker or mutual acquaintances; responses are expected to be honest appraisals.

Question *opinions on [redacted] [redacted]?*

Answer *haseena:shes soooo funnny we always get the giggles ive know.her for agesss she always knows how to mke mea laugh we hv the weirdest memories! and yea she just amazing and i tell her*

Table 3.1: Naive Bayes Classification Results

Discourse Type	Precision	Recall	F-Measure
Complimenting/Positivity	0.987	0.865	0.922
Bullying/Inflammatory	0.798	0.753	0.775
Picture or Video Request	0.888	0.978	0.93
Preference Question	0.989	1	0.994
Like Solicitation and Rating	0.89	0.91	0.9
Thoughts/Opinions	0.971	0.382	0.548
Defense Discourse	1	0.933	0.965
Sexual Content	0.883	0.933	0.907
List of Followers	0.953	0.91	0.931
Things that you Hate	0.944	0.955	0.95
Self Harm	0.888	0.807	0.845
None of the Above	0.53	0.865	0.658
Weighted Avg.	0.886	0.858	0.856

3.4.2.1 Classifying Discourse Types

LDA topic models predict the probability that a given document belongs to a topic. To permit automatic categorization of exchanges on ASKfm, I built a Naive Bayes classifier to assign documents to the above 11 categories. The features for the

classifier were the key words identified by each topic model. For my training set, I manually annotated a sample of 1100 documents according to my derived discourse types. The collapsing of topics for each discourse type generated key words for my feature engineering process. Since a document in each of my topic models consisted of a question-answer combination (the way discourse types are presented on ASKfm and other self-anonymous social websites), my features checked for the existence of the keywords in the question-answer combination. The performance of this classifier can be seen in Table 3.1. The performance of my classifer is very encouraging, suggesting that is it possible to perform reasonably accurate automatic detection of different discourse categories. The performance is not suprising since the LDA topics are based on keyword frequency and a predictive model based on the same keywords should be accurate. However, the performance of my model indicates that the topics represent real distinct discourse categories on ASKfm.

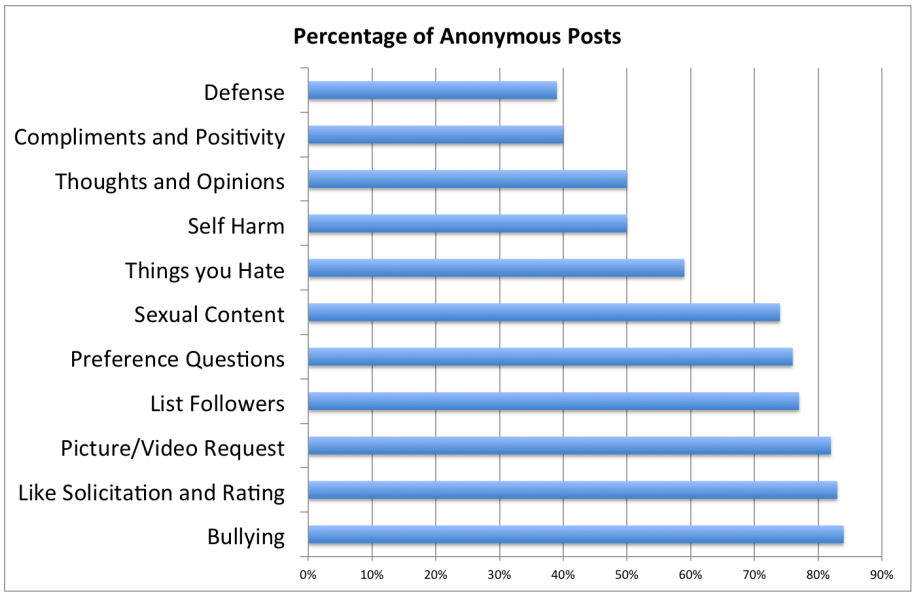


Figure 3.3: Percentage of Anonymous Posts for Each Mode of Discourse

3.4.2.2 Anonymity and Discourse Types

While messages sent on ASKfm are anonymous by default, senders may choose to make a message non-anonymous. To understand whether this choice was related to the type of discourse, I measured the fraction of each category for which the messages were anonymous. The results are shown in Figure 3.3. The majority of the exchanges were anonymous. While bullying may seem to be a natural byproduct of anonymity on ASKfm, other more positive discourse is also associated with anonymity. It is worth noting that the most anonymous categories include both healthy/fun things, such as "like" solicitation and picture requests, as well as bullying. It wasn't surprising that the positive discourse types such as *Compliments and Positivity Discourse* and *Defense Discourse* were more often less anonymous than their negative counterparts like Bullying Discourse.

3.4.3 Limitations

We acknowledge that the sampling choice is limited by the fact that the core of the sample are users who used a variation of terms "go kill yourself" and thus captures a facet of the ASKfm usership. The results show that despite the core of the sample, my snowballing was able to capture a wide variety of positive discourse types in addition to negative discourse types.

We discovered that despite the negativity associated with ASKfm [145], the existence of the other discourse types I discovered in this study shed light on the unique affordances the semi-anonymous social media platforms offer users who are

seeking social support or self-disclosing information on the website. I acknowledge that these aren't the only discourse types that occur on the website. Furthermore, observing different types of discourse and exploring correlations with anonymity still doesn't say anything about how users disclosure and interaction behaviors are shaped on askFM. To unpack disclosure and interactio behaviors on askFM, I describe Study II below.

3.5 Study II: ASKfm Use Motivations

To understand users' disclosure practices and interaction behaviors on askFM ASKfm (**RQ2**) and ways the site could be improved (**RQ3**), I conducted a survey in January 2015 of young adults (ages 18-20) who identified as active site users. In my survey, I asked participants about their personal experiences with bullying and cyberbullying, their question-asking practices, as well as demographic information and measures of personality [109] and self-esteem [222]. After receiving IRB approval, I first pre-tested the items with 50 Mechanical Turkers, then opened the HIT to include up to 250 responses. In total, I received 243 usable cases for analysis. I added my age restrictions to the Mechanical Turk HIT and participants selected a box confirming they were 18-21 years old. Though this is not foolproof, it is the most I could do given Mechanical Turk ToS restrictions.

3.5.1 Participant Demographics

In the full dataset, 35% of participants were female, and the average age was 19.6 (SD=.82). Two-thirds of participants were American, with the remaining participants representing 17 nations. The majority were enrolled in school full-time (60%) or part-time (15%) and lived at home with their family (57%). Participants reported spending just over five hours online per day (median=4.5 hours; SD=3 hours, 15 minutes), and said they used seven social media platforms on average (median = 6, SD = 3.76) from a list of 16 options.

3.5.2 Experiences with Bullying and Cyberbullying

Because popular media accounts have highlighted the prevalence of cyberbullying on ASKfm, I asked participants about their experiences with cyberbullying. Nearly three-quarters (73%) of participants reported that they had been bullied offline at some point, while 49.6% said they had been victims of cyberbullying. Conversely, when asked if they had ever participated in seven bullying activities (e.g., teasing, spreading rumors, name calling, threatening), 91.4% of participants said they had participated in at least one activity and 24.5% said they had engaged in all seven at some point. Females reported being victims of bullying (M=2.58, SD=1.12 vs. M=2.21, SD=1.09), $t(241)$=-2.52, p=.012, and cyberbullying (M=2.10, SD=1.20 vs. M=1.79, SD=1.02), $t(242)$=-2.15, p=.03, more often than men, but there were no gender differences in engaging in bullying activities. Females were also significantly more likely to experience verbal bullying than males

(M=3.77, SD=1.02 vs. M=3.24, SD=1.03), $t(172)$=-3.35, $p<.001$, but no differences were reported in experiences with physical bullying.

3.5.3 ASKfm Interactions

The aim of my survey was to understand how ASKfm users communicated through the site, given specific affordances like anonymity and high visibility of content. I asked participants how often they interacted with strangers on the site, finding that just 7.7% of respondents reported they never interact with strangers, while more than 60% said they interact with strangers with some regularity.

Given the ability to post questions anonymously on ASKfm, I asked participants if they had ever asked themselves a question anonymously, and 21% said they had. When asked about the reasons behind this practice,the most notable responses were that they did it to increase activity on their profile (71.4%), to make identity disclosures they wanted others to see (67.8%), and to cheer themselves up (54.4%). I also asked users how much they agreed with the following statement: *Posting anonymous questions on my page makes me feel better about myself*; 51.8% agreed or strongly agreed, suggesting that the simple act of posting and viewing content on their profile page–even when the content is self-written–can positively impact well-being.

It is important to note that ASKfm users receive all "questions" privately before choosing to answer questions publicly. If a user declines answering a question, only the person sending the question and the recipient know about the question. I asked respondents why they think people decide to answer questions that are

mean and hurtful, ultimately publishing the mean and hurtful comment to a wider audience; 54.8% said they posted the malicious comments they received because they wanted people in their network to comment on the malicious post to show support and defend them against the poster, while 52.5% said they are angry or upset and want to say that the comment is not true or they want to look like they don't care.

Finally, I asked participants if other users had ever posted malicious comments to their page. Approximately half (48.7%) said they had received negative comments about their appearance (weight, looks), 35.4% had received negative comments about their sexuality, 33.6% had received insulting comments about personal relationships, 21.4% had received "threatening comments," and 45.1% had received comments that made them feel excluded.

3.6 Discussion

3.6.1 Emergent Behaviors on ASKfm

The discourse types I discovered suggest that there are other interactions and behaviors on ASKfm that occur beyond cyberbullying – interactions that are afforded by the same designs that lead to cyberbullying (anonymity for example). I found that users (1) engage in self-disclosure practices and (2) seek social support. I observe these practices across many of the discourse types I discovered.

3.6.1.1 Self-Disclosure on Semi-Anonymous Q&A Websites

On ASKfm, I observed very revealing acts of self-disclosure as part of *Self Harm* and *Thoughts and Opinions* discourse. *Thoughts and Opinions* discourse allows users to openly state their opinions about people mutually known by the questioner and recipient know, i.e "thoughts on Sarah?". In my survey, one highly cited reason for anonymous self-questioning was to share information that a user wanted others to see (57.1%), suggesting that anonymous self-questioning lowers the barrier to disclosing sensitive information. Self-disclosing on a semi-anonymous social media platform can be cathartic and comforting. Disclosures can occur as part of a question (usually anonymous), or in the response. What these sites offer is the opportunity to discuss something without explicitly bringing up a topic. A user might anonymously self-question, to give themselves an excuse to respond publicly, or they might ask others anonymously to recruiting those others to join in. For example, a user who anonymously posts a "suicide-list" list question can find out if anyone else in the network is experiencing the same things without identifying himself/herself.

Disclosures need not be sensitive, as I saw with *Things you Hate* and *Preference Questions*. *Things you hate* questions typically asked a person "What are some things that annoy you?" and a user would respond about the things that the particular person disliked. *Preference Questions* were diverse in the subject matter. They ranged from but were not limited to favorite foods, bands, mode of entertainment, and mode of communication. Some example *Preference Questions* include: "Do you

prefer talking or texting?", "Do you prefer tea", "do you prefer waffles or chips ?", "How do you prefer to be awakened up in the morning?", "Do you prefer books or movies?" Questions and answers in *Things you Hate* and *Preference Questions* were innocuous invitations for users to self-disclose.

3.6.1.2 Social Support on Semi-Anonymous Q&A Websites

The results demonstrate that social-support seeking behaviors are common on ASKfm. my conceptualization of "social support" is in line with other computer-mediated communication research focused on social media and resource provision (eg, Ellison et al. found that liking behaviors were linked to bridging social capital perceptions) [92]. Support from social media can manifest through low-cost interactions (e.g., PDAs) depending on a sites affordances. In this section I discuss how users seek social support through: (1) self-questioning anonymously, (2) choosing to publish cyberbullying content to one's profile and (3) Like-Solicitation exchanges. The *Self-Harm Discourse* category included people seeking social support on taboo subjects like self-harm, self-injury and depression. In previous studies, users reached out for social support completely anonymously on social media platforms like YikYak or anonymous message boards [137, 273]. On ASKfm, while a topic is brought up anonymously through the question-asking format, users can identify that they need help, or support those who need help, through a low-cost interaction of liking the post.

A common example of *Self-Harm Discourse* are Suicide List Questions, where the questioner asks any readers who have considered suicide to identify themselves

publicly (by liking the post) in order to receive support. The poster and commenter(s) are working together to identify themselves by liking the question if they have ever been hurt in any of the ways specified by the answerer of the question. The answerer then promises to send *Compliments and Positivity* to the victims of these kinds of hate. A "Suicide List" can also qualify as a form of like bartering. Users who "like" the question are bartering for *Compliments and Positivity*. This kind of discourse demonstrates how users seek and receive social support on taboo topics.

A study looking at self-injurious behavior on message boards observed that these message boards provide essential social support, but also normalize such behaviors [273]. It is not surprising to discover taboo discourse like *Self-Harm Discourse* on a semi-anonymous social media platform. What is unique, however, is the transition from anonymous questioning to non-anonymous social support permitted by ASKfm–users can ask for help anonymously, but replies in support are shown visibly.

As described above, questions on ASKfm are invisible unless the recipient decides to respond. On first glance it may seem strange that a user would choose to publish a negative "question" they have received, so I asked the individuals in my survey why they would make the malicious content public by answering it. The top reason (54%) participants listed for responding to negative questions was that they wanted other people to comment in support of them. ASKfm affords recipients of such comments the *choice* to publish the content and seek social support, a choice not given on another non-anonymous social media platforms.

In my survey, 47.4% of self-questioners reported that they had asked themselves questions anonymously to respond to negative or harassing posts on their page, suggesting that some ASKfm users publish negative "questions" and respond by posting an anonymous "question" to their own pages as a form of support to imply they have a supportive network. Users may also be asking themselves questions in order to defend themselves when they experience cyberbullying or other negative comments. Sometimes requests for social support were more explicit (and less serious). *Like Solicitation* discourse involved users exchanging "likes" or "ratings". For example, an anonymous user posted, "likes for anyone who likes this post", and the recipient replied, "sure". In this exchange, the recipient would then have to return the "likes" for all the friends who "liked" that conversation. The same applies to "rating" for the likers of that conversation. A user would then post on all the likers of that particular piece of conversation a rating between 1 and 10. This category appears to be a way people can anonymously instigate interactions with others–the instigators were almost always anonymous (as seen in Figure 3.3), yet ASKfm makes the identity visible when a person likes something in response. Thus this can be a way of transitioning from anonymous to non-anonymous interaction, yet "Like Solicitation" also appears to be a way to exchange social support through liking the activity of individuals in a user's network.

Of the 11 categories of ASKfm discourse identified in this study, the most anonymous, Bullying, is unambiguously negative, while the least anonymous, *Defense Discourse* and *Compliments and Positivity Discourse*, were clearly positive. Yet there is not a simple linear relationship between the degree of anonymity and

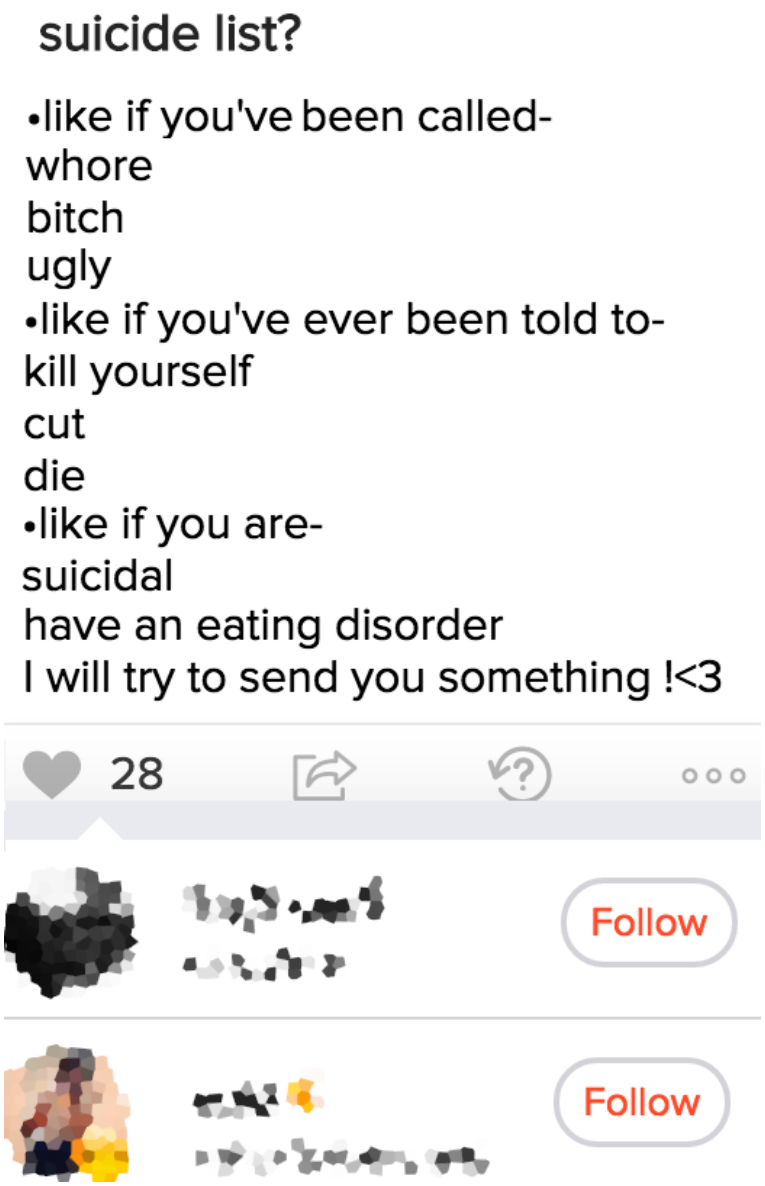


Figure 3.4: Example of a "Suicide List" question-answer pair on ASKfm. The question is asked anonymously. The 28 likers however are identified users. A preview of those who "liked" the question-answer pair appears at the bottom of the image. *Content has been changed to both reflect the reality of the content of these posts and to protect the identity of people involved in this post.*

the beneficial nature of the interaction. While *Bullying* was the most anonymous discourse category, the next four appear either innocuous or positive: *Like Solicitation*, *Picture Request*, *List Followers*, and *Preference Questions*. This supports my notion that positive and negative types of discourse co-exist on and are afforded by the anonymous interaction of ASKfm. I now discuss in more detail the affordances specific to semi-anonymous platforms, the behaviors that result from such affordances, and why people are motivated to use such platforms.

3.6.2 ASKfm Specific Interaction Types

The discourse types I discovered along with my survey responses demonstrate that ASKfm affords at least three specific types of interactions: (1) Anonymous Self-Questioning Practices, (2) Transitioning from Anonymity to Visibility, and (3) Built-in Filtering of Content. I discuss each of these in detail below.

3.6.2.1 Self-Directed Anonymous Questions

ASKfm's interaction model permits anonymous, self-directed questioning. The anonymous feature allows users to interact with themselves anonymously making it appear that they have more social support than what actually exists. This interaction is not much different than someone guilefully sending themselves flowers or gifts to their workplace anonymously to indicate to others more social approval than really exists. In my survey, 21% of the respondents said that they had asked themselves a question anonymously on their profiles, and their justifications for doing so included making identity disclosures that they wanted others to see, increasing

activity on their profile, and feeling better about themselves.

While research has found that users make identified disclosures on sites like Facebook as part of their "identity work" [268], guiding self-presentation anonymously may provide users with a greater perception of control. The potential for positive effects of self-posting on perceptions of well-being is also worth further investigation; for example, researchers have found that text-based interactions have a greater positive effect on well-being than face-to-face interactions [105] while public disclosures on social media serve a self-affirming purpose by satisfying needs for self-worth [259]. Likewise, decades of research extending back to the early bulletin board services (BBSs) supports the benefits of interacting through anonymous or pseudonymous channels [81, 217].

3.6.2.2 Transition from Anonymity to Visibility

The mix of anonymous and non-anonymous interactions on ASKfm provides the ability to transition from anonymity to visibility: while "questions" can be asked either anonymously or non-anonymously, "likes" are *always* visible. Users can broach a sensitive topic anonymously, safely, and only reveal themselves through a "like" if enough of their social circle does likewise. At first glance, the use cases for such a transition might seem trivial. However, after further examination of the discourse types like *Like Solicitation*, *Compliments and Positivity*, and *Self Harm*, this transition can be an important strategy garnering social support. For example, research has found that people with lower self-esteem consider Facebook a good place to disclose information; however, they also post more negative posts, which

receive fewer "likes." In turn, these users are less likely to obtain social resources from the site [92, 98]. This ability to transition from anonymity to visibility with a help of a friend allows individuals to share without regretting it later [269] and get social support on taboo subjects like self-harm.

3.6.2.3 Built-In Filtering

ASKfm differs from other more conventional social media platforms because of it's implicit built-in filtering mechanism. When a user navigates to another user's profile to "ask a question", the "question" is not automatically published on the recipient's profile. Instead, the recipient receives the question in a private inbox and then can *choose* whether to respond to the message. If a user declines answering a bullying question, only the bully and the recipient know about the question. my results reveal that users sometimes decide to answer questions even if they are hurtful or embarrassing, publicizing the hurtful question by answering it.

One of the reasons cyberbullying is so detrimental is because the nature of the internet makes it replicable, permanent, and searchable [40]. ASKfm's filtering allows users to reject cyberbullying and other malicious content and prevent it from becoming replicable, permanent and searchable. Users don't always reject such content, and it appears that the option to publish (or not) gives victims a degree of agency. I describe later how some users decide to publish content on their own terms, to gain social support after being bullied. This built-in filtering has implications for privacy as well. Studies have shown that youth control privacy by deleting content on their profiles that others have posted [215]. The built-in filtering allows users

to consider content that may breach their privacy before it is published to their profiles. The question-asking format implicitly gives users the ability to filter *who* and *what* is being posted to their profiles.

3.6.3 Design Recommendations

The results of my study demonstrate that while cyberbullying is a reality of the ASKfm platform, users utilize ASKfm's affordances to transition from anonymity to visibility on taboo subjects or self-direct anonymous questions for various purposes. I can use my results to help inform better features to minimize negative interactions and possibly highlight positive interactions. I make following design recommendations:

1. **Topic Model Filters** My topic modeling results revealed the various words and terms that appear in cyberbullying posts. I found that top words associated with cyberbullying included a range of words like: "hate", "ugly" and "gay". The various ways these words can be interpreted based on their respective contexts demonstrate that the degree to which words can hurt depends on many factors including the context in which the word was used. Furthermore, my my topic modeling results demonstrate that ASKfm users might use expletives affectionately and using only expletives as features in a filtering algorithm may lead false positives. For example, one document classified as "Compliments and Positivity" category was, "B****[redacted] u my bff", which was captured as this category correctly despite the fact that it included an expletive. Based on my results, I suggest that topic modeling be used to de-

termine categories of discourse that users can then choose to filter. Users can protect themselves from the stress of seeing cyberbullying related content or other categories of content which contribute to a deteriorating user experience by creating custom filters to avoid seeing certain categories. This approach to filtering is an alternative approach to previous automated methods of filtering that might not take into consideration context. I believe that this is the right approach to filtering content and can be adopted by all social media platforms that want to protect their users from cyberbullying. An example of a filtering page can be seen in Figure 3.5.

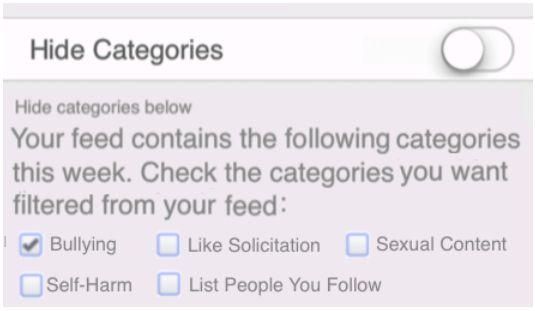


Figure 3.5: Example of filtering with topic modeling

2. **Introducing Paralinguistic Digital Affordances (PDA) for social support interactions:** My results show that many of the interactions on ASKfm involve sending "love" or "compliments" in exchange for a "like" on a question-answer pair. I observe this kind of bartering in the "Suicide List" questions, where users are asked to "like" a question-answer pair in exchange for support and kindness from the person who published the post. I also identified *Compliments and Positivity* discourse that involved users complimenting one

another and sending *love*. Paralinguistic Digital Affordances (PDA) are one-click social cues that provide meaningful interactions on social media. These types of interactions can be used to maintain relationships and even influence access to resources [120]. I recommend PDAs go beyond the "like" feature so users can express more nuanced emotions like empathy or sadness when self-disclosures are made or someone shares content that elicit such feelings. In these "like-bartering" exchanges for social support on ASKfm, visual PDAs that represent kindness, love, support, or compliments can be introduced so that users can engage in one-click interactions to continue to maintain relationships and provide one another with support, beyond the available "liking" mechanism.

3. **Increasing Accountability by Identifying Followers:** While ASKfm users are aware of *how many* people are following them, they do not know *who* is following them on the platform. While social media users generally do a poor job at estimating the audience of their posts on more popular social media sites like Facebook [31, 169], it more difficult to estimate who views one's content since users do not know who is "following" their posts. The audience is "invisible" [40] and the user does not have an accurate idea of who views their content. One of the discourse types that emerged in my first study was *Listing People You Follow*, precisely because people do not know who is "following" them. By making audiences visible to users, users will have a better understanding of who sees their content and they will be

held more accountable for the content that they post or send to one another. Studies show that perception of audiences influence content production and self-presentation practices [32]. Social media platforms that allow visible subscriptions and unsubscriptions give some degree of feedback about quality of content to users [148], but since ASKfm's subscriptions are unknown to the user, it is more difficult to understand how people react to the content beyond other cues like reshares [251] and likes. Giving this minimal level of audience transparency to users would increase accountability on ASKfm.

3.7 Conclusion

In this work, I discovered interactions occur alongside cyberbullying discourse on ASKfm, offered reasons why people use ASKfm despite the site's propensity for cyberbullying, and concluded with making design changes that can make ASKfm a safer space. I approached these questions through a data-driven approach coupled with a survey of active ASKfm users. I used topic modeling, and manual coding to derive discourse types on ASKfm and conducted a survey and about ASKfm user. I presented 11 discourse types on ASKfm that occur alongside cyberbullying discourse. I suggested that users engage in these discourse types to self-disclose and seek social support and used the affordances of ASKfm to engage in such behaviors using affordances on the platform like self-directed questions and transition from anonymity to visibility by liking anonymously-asked questions. I discussed how ASKfm's semi-anonymous and non-anonymous affordances impact the types

of discourse I observed, and the types of user behavior that emerged as a result. I made design recommendations that can potentially enhance user experience by decreasing the harm caused by cyberbullying, and enhancing its social support features. In the next chapter, I build on the findings in this chapter to explore user perceptions towards positive messages (*Cyberbully Reversal Pings*) received in the wake of cyberbullying incidents on anonymous social media platforms.

Chapter 4: Design Heuristics for Cyberbullying Mitigation on Self-Anonymous Websites

4.1 Chapter Summary

Chapter 3 focused on the potential of anonymous interactions to mitigate cyberbullying instances. In Chapter 4, I build on this finding to explore the notion of countering negative anonymous content on ASKfm with positive anonymous messages. Social media platforms that allow users to interact with one another anonymously have risen in popularity in recent years, especially among adolescents. The ability to interact anonymously has been tied to numerous incidents of cyberbullying, sometimes with tragic outcomes. I explore the notion of countering negative anonymous content on ASKfm with positive anonymous messages. I present two studies to first determine if users will be receptive to anonymous positive messages responding to bullying messages, then to administer positive messages or *Cyberbully Reversal Pings* to ASKfm users who have received bullying messages.[1] I discuss three iterations of *Cyberbully Reversal Ping* administration and conclude with a set of design heuristics for cyberbullying mitigation through this mode, as well as ethical

[1]This study was a collaborative project done with Jessica Vitak.

considerations for researchers studying similar populations. This study represents one of the first to empirically evaluate the feasibility of a cyberbullying mitigation solution "in the wild" and provides a useful case study for social media developers addressing this critical issue. The experimental design employed in this study allows control over the factors I evaluate and helps get insights into *how* users feels about different types of support, insights that cannot be garnered through strictly automatic methods.

4.2 Introduction

The affordances of social media [260] remove temporal and spacial constraints to all forms of interaction, including those that propagate harmful messages. These interactions persist over time and are visible to a much larger audience than those that occur in offline spaces, which may cause greater social and emotional damage. Some social media platforms (e.g., ASKfm, Formspring), further enable abusive behavior by allowing users to post messages anonymously. Because anonymity has been shown to lower people's inhibitions [252], it is not surprising that the anonymous nature of the platform has been used for cyberbullying [46, 112, 161]. Under the shroud of anonymity, users of these services can send abusive messages to one another without being held accountable for their actions. In this chapter, I address the challenges of anonymous cyberbullying by examining the effect of supportive messages to victims of cyberbullying on ASKfm. In the following sections, I highlight related literature on cyberbullying and adolescents' social media use before

presenting results from two studies. In the first, I evaluate the whether positive messages from an unknown user can mitigate negative effects of bullying messages. Based on the results of this pilot study, I describe the iterative process I employed to administer supportive messages to cyberbullying victims on ASKfm. Findings are discussed in terms of the potential for such a solution to be successful on social media platforms, including a set of design heuristics for cyberbullying mitigation in the realm of anonymous social interactions.

4.3 Related Work

In the following sections I present research on adolescents' social media use and cyberbullying experiences, highlighting the severity of this problem and the need for technical solutions. I also discuss the primary factors to consider in designing a system to mitigate the negative effects of cyberbullying.

4.4 Peer Support

Peer support has been found effective in motivating young people in a variety of contexts, including academic performance [271], substance use [276], depression [284], and exercise [210]. In times of social adversity, peer support can promote psychological well-being and can positively influence self-esteem [173]. Mentoring, befriending, conflict resolution, advocacy/advice-giving and counseling–based approaches are components of peer support systems and are effective of providing varying degrees of emotional support for someone in need, particularly in bullying

situations.

The affordances of social media provide additional avenues for young people to seek and exchange social support through both identified and anonymous channels. Research by Vitak and colleagues [90, 266] has found that even low-cost activities, such as "Liking" a post or wishing someone happy birthday on Facebook provide emotional support. Teens who often wish to hide their conversations from adult eyes, have developed workarounds, including social steganography [41], to express themselves and request support from peers while cloaking the meaning from outsiders (e.g., parents).

Diamanduros et. al identify ways that school psychologists can be effective in addressing the problem of cyberbullying in schools. They can promote awareness of cyberbullying and its impact. School psychologists are positioned to be able to assess the climate of cyberbullying in schools. School psychologists can play a role in prevention programs which are designed to address the problem of cyberbullying [77]. There have been no longitudinal studies specific to cyberbullying like there have for bullying [178] to identify the long term effectiveness of such efforts by school psychologists. Peer social support has three main components: feeling loved, belonging to a network, and feeling valued [167]. Mentoring, befriending, conflict resolution, advocacy/advice-giving and counseling- based approaches are components of peer support systems that have been found effective in providing varying degrees of emotional support for someone in need, particularly in bullying situations. For example, studies found that victims of bullying reported that they had been empowered to overcome the bullying problem when provided with one of

the of modes of peer support listed above [64, 189]. In real-life settings, befrienders are selected by teachers or facilitators of the support group for their friendly personalities [189]. In some ways, the proposed mitigation efforts on Instagram mirror the goals of the peer social support: feeling loved, belonging to a network, and feeling valued [167].

4.5 Well Being and Social Support

Researchers have long been invested in the use of technology and interfaces to promote the well-being of users. Previous researchers have employed user-centered design to create applications that promote the well-being of people suffering from mood disorders. For example, researchers created an application to "self-soothe" users who suffer from mood disorders. Self soothing is the act of calming down users with mental health problems. The application aims to create an atmosphere of positivity and achievement to help train those who suffer from mental health problems to self-sooth themselves. Furthermore, Good et al use reminiscent therapy (RT), a popular method used to promote positive moods to reduce feelings of loneliness that come with those who suffer from dementia or depression [106]. Reminiscent therapy involves using prompts that are meaningful to a user such as photos and music to aid in remembering positive life events; researchers have suggested it is effective in reducing depression [239].

Sa et al. present a software framework prototype for several case studies, one of which is psychotherapy. The low-fidelity framework allows researchers to address

usability concerns. Through the framework prototype, therapists and researchers used the prototypes on experimental therapy sessions to develop tools for depression, anxiety and various other disorders [72]. Exercise therapy may benefit those with mental health problems. In other work, researchers use a mobile device to deliver a exercise program to those who suffer with mental health problem ages between 15-24 [140]. De Chaudry stipulates that with the rise of widespread use of social networking platforms, researchers can extract valuable information about the mental state of social network users [69]. Textual information, like use of pronouns can inform us about the mental state of an individual [58]. This information can potentially prevent infliction of self-harm. There haven't been many efforts to leverage technology to mitigate depression and preserve mental health [69].

4.6 Experimental Design: Is Anonymous Support Effective?

A key component of sending *Cyberbully-Reversal Pings* is to counteract bullying posts and reduce the emotional distress victims may feel. Before I began to administer *Cyberbully-Reversal Pings* to ASKfm users, I first wanted to evaluate whether the source of the ping would affect how the message was received; specifically, I sought to identify if *anonymous* messages were viewed as beneficial.

To evaluate this, I developed a post-test only randomized experiment with three experimental conditions and one control condition. Each condition included screen shots of ASKfm pages that included a bullying message (followed by questions) and a message responding to the bullying message. The experimental con-

ditions included (1) a male profile with a male-focused cyberbullying comment, (2) a female profile with a female-specific cyberbullying comment, and (3) a gender-neutral profile with a gender-neutral cyberbullying comment. All bullying messages are variations of actual messages on the site. In each of these scenarios, a positive message responding to the specific category of cyberbullying message was also included. The fourth condition served as a control, featuring a gender-neutral profile that received a cyberbullying message and a neutral response (i.e., "Why do you think some people are so mean?"). All other profile information remained consistent across conditions. See Figure 2 for an example of what participants in the female profile condition saw, and Table 4.1 for details on the content of all scenarios.

After clicking through to the consent page on survey-hosting site SurveyGizmo, participants were randomly assigned to one of the four conditions and responded to a series of questions. Regarding the scenarios, participants were asked to rate the emotional impact of these messages by imaging they were sent by (1) a good friend, (2) an online-only friend, and (3) an anonymous user using a slider scale (range: 1-100, 1=Much Worse, 50=About the Same, 100=Much Better). In addition, participants provided data on their personal experiences with bullying and cyberbullying, as well as demographic information and measures of personality [110] and self-esteem [221].

After receiving IRB approval for this study, I pre-tested it with 50 Mechanical Turkers to verify the random assignment protocol, then opened the HIT to include up to 250 responses. The participant pool was limited to ASKfm or Formspring users between ages 18-21 to ensure that participants had experience with these types of

sites and because the majority of users of these sites are younger; this was enforced through disqualification logic in the survey. In total, I received 243 usable cases for analysis.

4.6.1 Results of the Experiment

In the full dataset, 35% of participants were female, and the average age was 19.6 (SD=.82). Two-thirds of participants were from the United States, with the remaining 1/3 of participants representing 17 nations. The majority were enrolled in school full-time (60%) or part-time (15%) and lived at home with their family (57%). Participants reported spending just over five hours online per day (median = 4.5 hours; SD = 3 hours, 15 minutes), and said they used nearly seven social media platforms on average (median = 6, SD = 3.76) from a list of 16 options.

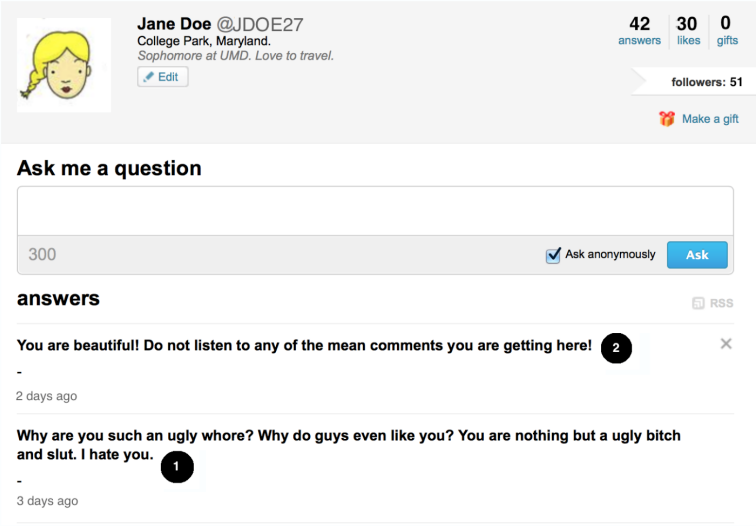


Figure 4.1: Sample Cyberbullying Scenario: Female User, Positive Response

Regarding bullying experiences, 73% of participants reported having been bullied at some point, while 49.6% said they had been victims of cyberbullying.

Table 4.1: Details of Experimental and Control Conditions

Scenario	Bullying Comment	Response to Bullying Comment
Male User, Positive Feedback	you're a fag. do u cut yourself?	hey, don?t call him a fag! John, you?re awesome don?t listen to the bully.
Female User, Positive Feedback	Why are you such an ugly whore? Why do guys even like you? You are nothing but a ugly bitch and a slut. I hate you.	You are beautiful! Do not listen to any of the mean comments you are getting here!
Gender-Neutral User, Positive Feedback	Go kill yourself. No one would even care.	You are amazing and do not listen to anyone who tells you to go kill yourself! Anon, why are you saying that? Who tells people to go kill themselves? That is plain mean.
Control: Gender-Neutral User, Neutral Feedback	Go kill yourself. No one would even care.	Why do you think some people are so mean?

Conversely, when asked if they had ever participated in seven bullying activities (e.g., teasing, spreading rumors, name calling, threatening), 91.4% of participants said they had participated in at least one activity and 24.5% said they had engaged in all seven at some point. Females reported being victims of bullying (M=2.58, SD=1.12 vs. M=2.21, SD=1.09), $t(241)$=-2.52, p=.012, and cyberbullying (M=2.10, SD=1.20 vs. M=1.79, SD=1.02), $t(242)$=-2.15, p=.03, more often than men, but there were no gender differences in engaging in bullying activities. Females were also significantly more likely to experience verbal bullying than males (M=3.77, SD=1.02 vs. M=3.24, SD=1.03), $t(172)$=-3.35, p<.001, but no differences were reported in experiences with physical bullying.

To evaluate the primary research question regarding the relationship between emotional impact and source of messages, a one-way repeated measures analysis of variance (ANOVA) identified significant differences in response to the bullying com-

ment based on the identity of that person (i.e., good friend, online-only, anonymous), Greenhouse-Geisser $F(1.86)=65.70$, $p<.001$, with messages from a good friend causing significantly greater negative response than the other conditions, and messages from an online-online friend causing a significantly greater negative effect than in the anonymous condition ($M=22.82$ v. 30.31 v. 36.37). Note that all means are significantly below the midpoint. When looking across conditions, a between-subjects ANOVA revealed just one significant difference, with the response to a bullying comment from a good friend in the female condition being significantly lower than in the male condition, $M=15.78$, $SD=18.47$ v. $M=28.96$, $SD=2412$, $F(3, 240)=3.54$, $p=.015$. Controlling for the sex of the participant produced non-significant results, $F(1, 240)=.18$, $p=.063$, suggesting that hurtful messages make people feel bad about themselves, even if there is no direct connection to the content of the message.

Next, I compared the extent to which participants said the response comment would make them feel. First, the mean scores across all scenarios revealed a positive effect on well-being regardless of the identity of the poster ($M=68.83$ v. 66.43 v. 67.30), with a one-way repeated measures ANOVA finding no differences between conditions, Greenhouse-Geisser $F(1.77)=2.57$, $p=.085$. When looking at responses across scenarios, significant differences emerged across scenarios for comments from friends, $F(3, 243)=4.34$, $p<.01$; comments from online-only friends, $F(3, 244)=5.46$, $p<.001$; and comments from an anonymous user, $F(3, 244)=2.78$, $p<.05$. Scheffe post-hoc comparisons revealed the only significant difference between groups involved the control condition; in other words, there were no significant differences in how people rated the effect of the positive comments across the three experimen-

Table 4.2: Results of t-tests comparing positive feedback to bullies in the experimental and control conditions

Responder (t-test)	Condition	N	Mean	Std. Deviation
Positive response from close friend, t(242)=-3.37, p<.001	Control	57	58.58	27.267
	Experimental	187	71.95	25.907
Positive response from an online friend, t(243)=-4.02, p<.001	Control	57	55.21	25.004
	Experimental	188	70.01	24.138
Positive response from an anonymous sender, t(243)=-2.89, p<.01	Control	57	59.68	23.016
	Experimental	188	69.56	22.697

tal conditions. Therefore, the experimental conditions were collapsed into a single group, and t-tests comparing the experimental groups to the control group found that the positive comments responding to the bully in the experimental conditions made participants feel significantly better than the more neutral response in the control condition across all categories of posters. See Table 2 for a summary of findings from the t-tests.

4.7 Main Study: Cyberbully-Reversal Ping Pipeline

The pilot study findings provide very useful data for administering *Cyberbully-Reversal Pings*, which enables the anonymous provision of support to victims of cyberbullying. The significant positive effect on well-being remained consistent across three types of posters: someone you know well, someone you know in a specific context, and someone whose identity is hidden. This answers an important question for sending *Cyberbully-Reversal Pings*: will people respond to anonymous help or does it need to come from someone they already know? The findings from this study

suggest that the identity of the poster is not nearly as important as the context of the message, which supports the structure for sending *Cyberbully-Reversal Pings* for mitigating negative effects of cyberbullying. Based on these findings, I now present results from a study identifying and responding to actual cyberbullying messages on ASKfm.

The premise behind *Cyberbully-Reversal Pings* is that if negative and malicious messages can make one feel depressed or anxious, positive messages can counter those feelings. The goal of *Cyberbully-Reversal Pings* is to emulate the feelings of social support typically generated through peer support systems. Naylor et al. [189] demonstrate that peer support in the real world helps targets of bullying feel empowered and able to overcome the challenges which result of bullying. Below, I describe the types of peer support I believe *Cyberbully-Reversal Pings* can facilitate. The *Cyberbully-Reversal Ping* pipeline attempts to transfer the benefits of this kind of support support into the online realm through provisions of support to (unknown) cyberbullying victims.

The data used in this study was collected from ASKfm. The data collection process is documented in Chapter 3. For clarity, I will recall the data collection process in this chapter. I collected data from ASKfm, a social media platform that allows users to ask questions and respond to other users' questions, with an option to interact anonymously. Drawing on established frameworks of cyberbullying detection [145], I queried ASKfm through Google for variations of the phrases "go kill yourself" and "go die". The search yielded 19 public profiles, from which I collected user information (username, biography, headline), and 100 most recent

questions with respective time posted, author information if applicable, and respective answer. Additionally, I collected the 25 "best" questions for each user, i.e., the questions posted by a user that receive the most 'likes'. In order to gain a better understanding of these nineteen users' networks, I collected user information from random sample of users who had 'liked' questions on the "best" question pages of the nineteen original profiles queried. I opted for this method of sampling because I wanted to collect data from ASKfm users who had some interaction and connection with traces of cyberbullying. I repeated this process of random sampling the networks through breadth-first search until I collected user data from more than 40,000 user profiles.

4.8 Construction of Cyberbully-Reversal Pings

Peer social support has three main components: feeling loved, belonging to a network, and feeling valued [167]. Mentoring, befriending, conflict resolution, advocacy/advice-giving and counseling-based approaches are components of peer support systems that have been found effective in providing varying degrees of emotional support for someone in need, particularly in the bullying situations. For example, studies found that victims of bullying reported that they had been empowered to overcome the bullying problem when provided with one of the of modes of peer support listed above [64, 189]. Below, I introduce *Cyberbully-Reversal Pings* modeled after (1) mediation, (2) advocacy and advice giving, and (3) befriending.

4.8.1 Mediation

Mediation can be defined as a process in which a bystander plays the role of a neutral third party to resolve the dispute between two people [68]. While the majority of cyberbullying occurs anonymously, a small percentage of cyberbullying posts occur non-anonymously on ASKfm. For these posts, I created "Mediation" *Cyberbully-Reversal Pings*, in which I tag the bully in hopes of calling them out for perpetration. In the examples below, "@bully" is replaced with the bully's screen name and "@victim" is replaced with the target's ASKfm handle. The pings were modeled after discourse on ASKfm in which actual users defended one another.

"Mediation Pings"

- *"Hey @bully.. you should leave @victim alone"*
- *"@bully you need to stop hatin on @victim! @victim is a great person"*
- *"Seriously @bully what are you even getting out of this leave @victim alone!"*

4.8.2 Preventative Advocacy and Advice Giving

Advocacy and Advice giving is valuable peer support that provides needed information and resources to targets of cyberbullying in which neutral third party bystanders provide support and materials for people seeking help [189]. For our purposes, I have defined two modes of counseling: Preventative Advocacy for Self-Harm and Advocacy for Bullying.

4.8.2.1 Advocacy and Advice Giving for Self-Harm

Preventative Advocacy and Advice Giving for Self-Harm provides resources for those users who are contemplating self-harm. Phone numbers are provided for those who are contemplating taking their own lives.

"Preventative Advocacy and Advice Giving for Self-Harm Pings"

- *"If you're not sure where to turn call the S.A.F.E. Alternatives information line in the U.S. at (800) 366-8288 for referrals and support for cutting and self-harm"*
- *"If your feeling suicidal and need help right now call the National Suicide Prevention Lifeline in the U.S. at (800) 273-8255"*

4.8.2.2 Advocacy and Advice Giving for Bullying

Advocacy and Advice Giving for Bullying provides resources for those users who have experiences bullying. Phone numbers and advice is provided for those who have experienced bullying.

"Advocacy and Advice Giving for Bullying"

- *"If you feel embarrassed to let your friends know you are being bullied remember that most young people disapprove of bullying behavior and will most likely be on your side"*

4.8.3 Befriending

Befriending *Cyberbully-Reversal Pings* can be defined as responses that mimic the language and culture on ASKfm to appear to come from a peer in the same age group as the victim. In real-life settings, befrienders are selected by teachers or facilitators of the support group for their friendly personalities [189]. The Befriending *Cyberbully-Reversal Pings* emulate same-age friendly befrienders in anonymous on-line setting. I mimicked "befriending posts" closely after discourse I witness on the ASKfm between users exchanging compliments and friendly banter to ensure that they appear that they were coming from the age group. In many of our Befriending *Cyberbully-Reversal Pings* I used words like "gorgeous" and "beautiful", mimicking the actual language existent on the website. Many of the cyberbullying posts also were insulting a victim's physical appearance, which is why our initial list of possible Befriending *Cyberbully-Reversal Pings* included words like "beautiful" to counter negativity towards an individual's appearance. I discuss the limitations and outcomes of using such gender-specific language in more detail in Section 4.10.3.

"Befriending"

- *"youre beautiful inside and out so dont let anyone tell u different"*
- *"you are a wonderful person :)"*
- *"Stay strong. Ignore haters because you are, beautiful, gorgeous, perfect, kind, and nice. No words can describe it!"*

4.9 Identification of Cyberbullying Targets

I explored two different classification methods for automatic detection of cyberbullying targets. The first looks at classifying self-harm discourse as well as bullying discourse and utilizing those discourse types in the heuristics for identifying targets, while the second classifier contains additional training data and features catered specifically at detecting cyberbullying discourse.

4.9.1 Method 1: Identifying Cyberbullying Targets with Self Harm Detection and Bullying Detection

My first method makes use of a classifier that detects Self-Harm and Cyberbullying. I include a Self-Harm as a category in this classifier because ASKfm has been associated with suicide in extreme cases [46]. I operate on the notion that if a user is discussing inflicting self-harm upon themselves, they are more likely to actually do so [135]. Self-harm posts inquire about someone's opinion on self harm and whether they participate in it. These questions also ask about the type of self harm in which the recipient engages (i.e. cutting, starving oneself, etc.). I have listed some examples of self-harm discourse in our data below. The first piece of discourse is an example of the type of behavior I were attempting to mitigate. The second piece of discourse listed, while relevant to self-harm is not the type of individual I were attempting to get in touch with through *Cyberbully Reversal Pings* The wording of the messages has been altered to protect the identity of the individuals who published this content.

At-Risk Individual: Qualifies for *Cyberbully-Reversal Ping*

Question *what kind of razor did you use to cut with*

Answer *a sharpener blade was the the easiest for me to get*

Not-At-Risk Individual: Does not qualify for *Cyberbully-Reversal Ping*

Question *what is your opinion on self harming? x*

Answer *i think its a horrible thing. for someone or something make someone feel like they should use their skin as paper. some people right now feel unwanted or ugly or fat or like they just dont belong with the world because theyre being bullied*

4.9.1.1 Classifier

I labelled approximately 2,200 instances of discourse as to whether they pertained to cyberbullying or self-harm using keywords described in the literature, including "suicide", "cutting" and "self-harm," [119] as well as expletives. I built a Multinomial Naive Bayes Classifier using the lexicon and keywords described above as features. The precision, recall, and f-measure of our classifier for Bullying/Inflammatory posts was 0.798, 0.753, and 0.775 respectively. I conducted 10 fold cross validation. The precision, recall, and f-measure score for Self-Harm classifier was 0.888, 0.807, and 0.845 respectively.

Table 4.3: Response Rate for Iterations

Iteration	Overall	Receptive Thankful	Receptive Confused	Sarcastic Non-Receptive
Iteration 1: Self-Harm/Bullying Classifier + Anonymous + All Posts	30.5%	82%	12%	6%
Iteration 2: Bullying Classifier + Anonymous + All Posts	19.6%	100%	0%	0%
Iteration 3: Bullying Classifier + Anonymous + Recent Posts	66.7 %	100.0%	0%	0%

4.9.1.2 Heuristics for Determining Target

I developed three different types of positive messages, which I refer to as *Cyberbully-Reversal Pings*: Advocacy and Advice Giving, Befriending, and Mediation. From the collection of posts I had collected for each user, I check for the occurrence of "Self Harm" discourse , "Bullying" discourse, and whether "Bullied" posts were posted non-anonymously. If a user has received a Self-Harm question, I include a random selection of Advocacy and Advice Giving for self-harm *Cyberbully-Reversal Ping*. If a user has received more than four bullying comments, I include a befriending *Cyberbully-Reversal Pings*, and if a user's collection of posts include a bullying post from a non-anonymous aggressor, I include a Mediation *Cyberbully-Reversal Pings*. I then select a post randomly from the selection of *Cyberbully-Reversal Pings* for which the target qualifies.

4.9.2 Method 2: Identifying Cyberbullying Targets with Bullying Detection

4.9.2.1 Classifier

I created a secondary classifier to focus solely on bullying detection. The heuristics and pings associated with the detection of self-harm discourse were high-risk. In the first iteration of messages, I noticed that some of the respondents were confused why there were receiving information on suicide prevention, possibly suggesting a high false-positive rate. For the second classifier, I supplemented our training set with 13,000 labelled formspring posts which had been given binary labels by Mechanical Turkers as to whether they were instances of cyberbullying. Each post was labeled by three Turkers and I considered posts to be instances of cyberbullying when all annotators agreed. I added additional lexical features to our classifier as added by Kontostathis et al. [145].

4.9.2.2 Heuristics for Determining Target

With respect to determining whether a target qualified for *Cyberbully-Reversal Pings*, I took a minimalistic approach to pair with this classifier. If one of a user's posts was classified as bullying, I sent a *Cyberbully-Reversal Ping*. I also limited our *Cyberbully-Reversal Ping* to Befriending *Cyberbully-Reversal Ping* due to reasons I have outlined later in this chapter.

4.9.3 Design Iterations

I conducted three iterations of automatically administering *Cyberbully-Reversal Pings* to determine which method is the most effective in reversing the effects of cyberbullying. In our *Cyberbully-Reversal Ping* iterations, I considered the following variables: classification method, guidelines for qualifying user for receiving a *Cyberbully-Reversal Ping*, and whether I consider all posts when sending a message or just the most recent posts on a user's profile. After each iteration I evaluated the response rates and types of responses received and made changes onto the next iteration accordingly.

4.9.3.1 The Challenge of Evaluating Cyberbully-Reversal Pings

Evaluating the outcome of sending *Cyberbully-Reversal Pings* was extremely challenging for several reasons. Firstly, an implemented cyberbully-mitigation system has not been evaluated with users before. Secondly, there is an element of deception involved in sending *Cyberbully-Reversal Pings* to victims of cyberbullying. By engaging with the victims after the *Cyberbully-Rerversal Pings* were sent to ask how they felt about the message which was sent to them, I could be causing them potential distress by letting individuals know I were researchers who were conducting the study. For this reason, I settled with evaluating *Cyberbully-Reversal Pings* with observable data: the response rate and types of responses received.

While there have not been previous studies that indicate dialogue is an effective outcome of cyberbullying mitigation, previous studies suggest that the response

rate of individuals in social media is related to traits in their personality, namely: extraversion, emotional stability and openness to experience [62]. Personality can be dynamic and situational [78]. While I cannot say conclusively that recipients of such messages experience a change in their emotional stability as a result of receiving a *Cyberbully-Rerversal Pings* and thus were more likely to respond, I consider the possibility of such an occurrence in this study, as the only observable data are the response rates and types of responses.

4.9.3.2 First Iteration: Method 1 Anonymous + All Posts

In the first iteration, I used the "Method 1" classifier to send out *Cyberbully-Reversal Pings*. Pings were sent anonymously and I considered all of a users posts before sending out positive messages. The "Method 1" classifier is important because it helps classify "Self-Harm" discourse which is essential to administering "Advice Giving and Advocacy for Self-Harm" pings. I sent out 59 *Cyberbully-Reversal Pings* and received 18 responses (a response rate of 30.5%). The number of *Cyberbully-Reversal Pings* was highest in the first iteration compared to the later iterations because the heuristics (i.e. stricter classifier, considering only recent posts) for identifying targets is stricter in later iterations and thus finding users who qualify from our sampled data for *Cyberbully-Reversal Pings* was more challenging in later iterations than in this iteration.

4.9.3.3 Second Iteration: Method 2 Anonymous + All Posts

In our second iteration, I focused our detection more closely on bullying posts by using the "Method 2" classifier. In the first iteration, some of the respondents were confused about why there were receiving information on suicide prevention or advice referencing bullying. These confused responses include responses to "If you feel embarrassed to let your friends know you are being bullied remember that most young people disapprove of bullying behavior and will most likely be on your side", such as "idk where this came from sorry ahahah but cheers for the advice". Only the Befriending *Cyberbully-Reversal Pings* were received warmly by all recipients. Responses to "youre beautiful inside and out so dont let anyone tell u different" were met by "Luv u anon x" and "Im really not, but thank you for the thought. It means a lot." Responses to the cyberbullying reversal pings in Iteration 1 taught us to frame our language as less formal in the second iteration and not to reference previous negative posts, just focus on positivity. In this iteration, I sent out 46 *Cyberbully-Reversal Pings* from the list of pings I had identified and received 9 responses, a response rate of 19.6 %.

4.9.3.4 Third Iteration: Method 2 Anonymous + Recent Posts

While the response rate decreased from Iteration 1 to Iteration 2, the types of responses became more "receptive and thankful" in Iteration 2. A table of response type definitions and examples can be found in Table 4.4. It is important to note that these response rates (19.6% and 30.5%) are comparable or higher than

findings in studies exploring engagement with strangers through social media [205]. At the same time, I wanted to see whether the response rate would increase if I only considered the most recent cyberbullying posts when sending messages. In our third iteration, I only classified bullying posts of users who had posted recently (less than 30 days). I sent Befriending *Cyberbully-Reversal Pings* to identified targets. I decided to analyze only the 25 most recent posts on a user's profile before determining that they qualified for a *Cyberbully-Reversal Ping*. In this iteration, I sent 45 *Cyberbully-Reversal Pings* from the list of pings I had identified and received 30 responses, a response rate of 66.7% percent. This was higher than previous iterations, suggesting the additional restrictions reduced discrepancies between classification and *Cyberbully-Reversal Pings*.

4.9.4 Evaluating Success of Cyberbully-Reversal Pings

In order to evaluate the effectiveness of the responses, I compare response rates and the types of responses of the three iterations, as well as individual variables (type of ping, anonymous vs non-anonymous, the method of identifying targets). Evaluating the effectiveness of *Cyberbully-Reversal Pings* without asking users how they felt about them after they received them was challenging. In order to evaluate the effectiveness of *Cyberbully-Reversal Pings*, I considered response rate and the response types to the *Cyberbully-Reversal Pings* that I sent.

4.9.4.1 Response Rate

I looked at the response rate for each of our iterations as well as each of the individual variables I tried in those iterations. The response rates are listed in Table 4.3. In terms of the quality of responses garnered, the second and third iterations yielded the most "Receptive and Thankful" responses. Befriending *Cyberbully-Reversal Pings* yielded higher overall response rates than Mediation and Advocacy and Advice Giving.

4.9.4.2 Types of Responses

I have identified three types of responses: "Receptive and Thankful", "Receptive and Confused", and "Non-Receptive or Sarcastic". These types of responses were used to evaluate the quality of the messages I sent. The different types of responses and examples are in the Table 4.4. In Iteration 3, which was the only non-anonymous response iteration, I observed differences in responses to the *Cyberbully-Reversal Pings*, with one of the respondents privately instead of posting it publicly on their page.

4.10 Discussion and Design Implications

My automation of *Cyberbully-Reversal Pings* informs design considerations for systems aimed at mitigating cyberbullying through positive messages. I propose three design heuristics, which are discussed in more detail below.

1. Most recent interactions should be considered when determining cyberbullying

Table 4.4: Response Types to *Cyberbully-Reversal Pings*

Response Type	Positive Ping	Response
Receptive and Thankful	youre beautiful inside and out so dont let anyone tell u different	Thank you for those words! But who are u? come off anon!
Receptive and Confused	If you feel embarrassed to let your friends know you are being bullied remember that most young people disapprove of bullying behavior and will most likely be on your side	idk where this came from sorry ahahah but cheers for the advice
Non-Receptive or Sarcastic	hey you are awesome. Dont listen to what other people are saying!	What are other people saying? Hahaha

victims and those who qualify for *Cyberbully-Reversal Pings*

2. Befriending *Cyberbully Reversal Pings* are the most effective types of pings
3. Language in Befriending *Cyberbully-Reversal Pings* should be gender-neutral

4.10.1 Considering Recent Interactions

In our third iteration, I applied our cyberbullying classification algorithm only to the most recent activity of a user, so that the *Cyberbully-Reversal Ping* appears to be timely and appropriate. My response rate and quality of response rates were better in the final iteration than when compared to the average response rates and quality of response rates in the latter iterations. All of the responses in the fourth iteration, in which the *Cyberbully-Reversal Pings* were only on the most recent pings in the fourth iteration, the percentage of "Receptive and Thankful" pings increased to 100% with no occurrences of "Non-Receptive or Sarcastic" responses. My overall response rate also increased from an average of 19.6% in the previous iterations to 66.7% in the third iteration.

4.10.2 Befriending Pings are the Best type of Pings

Of the responses for befriending pings, 94.3% of them could be classified as "Receptive and Thankful," none could be classified as "Receptive and Confused," and 0.3% percent could be classified as "Non-receptive or Sarcastic." Furthermore, since most bullying messages are posted anonymously, I sent out very few Mediation *Cyberbully-Reversal Pings*. The few Mediation *Cyberbully-Reversal Pings* that I sent recieved no responses. Additionally the Advocacy and Advice Giving *Cyberbully-Reversal Pings* received no responses. One potential reason is that the language in the Advocacy and Advice Giving *Cyberbully-Reversal Pings* may have been "too adult" for the given population, and were inconsistent with the norms of interaction on the site. In addition, mediation pings may have been deemed too controversial because perpetrator of the bullying was tagged in the post and the ASKfm user may not want to get involved in such controversy. On the other hand, the Befriending *Cyberbully-Reversal Pings* mimicked the style of language used on the website, which may have contributed to the higher response rate. For this reason, I believe Befriending *Cyberbully-Reversal Pings* are most effective in eliciting a positive response. According to our survey data, posts that are most likely to illicit a response also made the recipient of the post feel happier upon receiving it.

For the first iteration, if a user had a self-harm post, they would qualify for a Advocacy and Advice *Cyberbully-Reversal Ping* that provides the cyberbullying victim with advice and relevant information on how to seek help. My results demonstrated that I must keep in mind that no classifier is perfect and that there is a risk

associated with sending suicide hotline information to someone who is not actually at risk for suicide. If there is a chance that the classifier might be incorrect, such sensitive information must not be sent. Furthermore, the language in the Advocacy and Advice Giving *Cyberbully-Reversal Pings* may have been too formal (when compared to the language on the website) and thus may have alienated ASKfm users from responding to them.

Furthermore, if a post is falsely identified as a cyberbullying post and the author is non-anonymous, there could be potential harm in the network by tagging the bully in a Mediation *Cyberbully-Reversal Ping*. Even if the classifier accurately identifies a bully and victim pair, the victim may not want to post information on their own profile in which they call out the bully in fear of increasing the actual bullying.

4.10.3 Gender-Neutral Language is Important

In many of our Befriending *Cyberbully-Reversal Pings*, I used words like "gorgeous" and "beautiful" to mimic the actual language existent on the website. However, these words may not be encouraging words for boys. From the pings I sent out to the user profiles, I can infer from profile pictures and names listed on the site, that 67% of them are female and 33% are male. Many of the cyberbullying posts also were insulting a victim's physical appearance, which is why our initial list of possible Befriending *Cyberbully-Reversal Pings* included words like "beautiful." Dinakar et al. [79] stipulate that certain gender-specific words can be used to detect cyberbullying in the LGBT domain. For example, a comment like "you looked

gorgeous today! What beauty salon did you go to?" might be a positive comment when directed to a female, but given certain stereotypes about the LGBT community and gender norms, a comment like this could be perceived as maligning male sexuality and thus be perceived as an insult. For this reason, I must consider the language that I use when sending *Cyberbully-Reversal Pings* and focus on sending gender-neutral pings. In a future iteration of this system, *Cyberbully-Reversal Pings* will only include gender neutral language.

4.11 Ethics

With increasing use and analysis of big data, ethical considerations have received significant attention [43]. Ensuring the confidentiality of data and anonymity of any participants is especially important when (1) analyzing adolescents and (2) analyzing negative events such as self-harm, bullying, and suicide. All scraping and data analytics in this study were administered on public profiles. In line with work by Goode [107], I argue that deception is an an integral part of testing the iAnon system; in order to effectively measure the impact of positive messages on a victim of cyberbullying in a natural (i.e., not lab) setting, recipients of positive messages must believe they are receiving positive messages from someone who cares about them and not from an automated system or researchers.

Thinking more broadly about the challenges of ethically working with this kind of data, it is important to recognize that ethics in social research must be *situational ethics* in that there are exceptions for certain situations and scenarios [130]. Many

users of these technologies have either a low understanding of how privacy settings work or have a (somewhat incorrect) assumption about privacy of the content they post. Researchers should not take advantage of public content and should instead consider ways to consent users or engage in other ethical practices.

4.12 Conclusion

Responding to the increasing prevalence of cyberbullying among adolescents, sending *Cyberbully-Reversal Pings* respond to a societal-level problem by creating a space for users to provide peer support to victims of harassing messages. Positive messages may mitigate the feelings of loneliness, depression, and anxiety that result from cyberbullying posts, and messages originating from peers, rather than from adults or other figures, may be especially powerful in helping adolescents overcome negative feelings resulting from being bullied. In this chapter, I (1) provided experimental support for the positive effect of anonymous posts to combat cyberbullies, (2) evaluated different potential modes of interaction to be incorporated into sending *Cyberbully-Reversal Pings* to determine the most effective way of mitigating cyberbullying detection, (3) introduced three design heuristics that should be incorporated into the design of such a system, and (4) evaluated the ethical considerations of our designs and similar tools that scrape user data. Findings from this research provide a foundation through which new tools can be developed and customized to address a variety of threats technology users may encounter when online.

Chapter 5: Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers

5.1 Chapter Summary

While the previous studies focused on ASKfm, a popular teen social media platform characterized by anonymous interactions, research suggests the majority of teens are spending the majority of their time using platforms like Facebook, Instagram, and Snapchat to interact with friends and are often identified by their real name. Therefore, this chapter expands my exploration of cyberbullying to consider these other platforms as well as solutions teenagers think would be most successful in curtailing cyberbullying. Teenagers represent an especially vulnerable population for negativ%e emotional responses to cyberbullying [156]. At the same time, attempts to mitigate or prevent cyberbullying from occurring in these networked spaces have largely failed because of the complexity and nuance with which young people bully others online. To address challenges related to designing for cyberbullying intervention and mitigation, I detail findings from participatory design work with two groups of high school students in spring 2015. Over the course of five design sessions spanning five weeks, participants shared their experiences with cy-

berbullying and iteratively designed potential solutions [18]. [1] I provide an in-depth discussion of the range of cyberbullying mitigation solutions participants designed. I focus on challenges participants'= identified in designing for cyberbullying support and prevention and present a set of five potential cyberbullying mitigation solutions based on the results of the design sessions. The findings in this chapter will later be used to inform the design of a cyberbullying mitigation tool to combat negative experiences and promote well-being for individuals who have experienced cyberbullying (see Chapter 6).

5.2 Introduction

With the rise of mobile phones and social media, teenagers have become a driving force in the development and design of new tools for interaction with their friends, and are ardent adopters of sites including Facebook, Ask.fm, and Twitter [159]. More than 90% of all U.S. teens 13-17 own a cell phone, while nearly three-quarters (73%) own a smartphone [157]. In recent years, researchers have documented shifts in social media use as teenagers seek "parent free" spaces; while Facebook no longer holds the same appeal for this younger demographic, Snapchat, Twitter, and Instagram are seeing burgeoning growth [150,157]. On these platforms, teenagers can continue to engage in innocuous correspondences and exchanges. At the same time, many of these spaces have become grounds for a range of cyberbullying activities, which can have serious negative effects on young people's mental

[1] This study was collaborative project done with Jessica Vitak and was presented at the 2016 ACM CHI Conference on Human Factors in Computing Systems.

health [16].

Research is mixed regarding potential outcomes of social media use by adolescents. Several studies have highlighted both positive and negative correlations between specific behaviors (e.g., frequency of use, engagement in social compensation) and adolescents self-esteem [23, 262, 263]. These spaces also enable posting of mean, unflattering, and bullying content. In a large study of adolescents by Patchin and Hinduja [124], they found that 20% of youth in reporting school districts reported being victims of cyberbullying, and 20% reported engaging in cyberbullying at some point in their lives. In a separate study, the Pew Internet Project [158] found that nearly all teens had witnessed someone being cruel online, while 15 % said they had been the recipient of mean comments in the last year. Victims of cyberbullying may experience significant emotional problems, including anxiety and depression [123, 141] and, in extreme situations, they may commit suicide.

Designing for teenagers is challenging, as their motivations for using these spaces and the emergent norms that develop among teenagers are often significantly different from their adult counterparts [96]. Because of this discrepancy in social media motivations, designers may overlook factors and features that facilitate cyberbullying or harassment. In recent years, the few attempts of designing for cyberbullying prevention have not included the perspective of those most affected: young people [79].

Participatory design represents a potential solution to this gap in understanding and may help designers build novel tools [282]. In this study, I extend existing participatory design techniques with adolescents [82–84, 115, 282] to the cyberbul-

lying mitigation domain. I share our methodology and findings from participatory design sessions conducted with ninth and twelfth graders at a private high school in a large metropolitan region in the U.S. In working with these two groups of teenagers, I systematically sequenced a range of design techniques over five sessions with the goal of prompting participants to both explore potential solutions for the larger problem at hand (cyberbullying across all social platforms) and smaller sub-problems (specific types of cyberbullying across specific platforms). These findings are synthesized in the discussion section, with a focus on challenges to designing for support and potential avenues for future design work in this area.

5.3 Related Work

Increasingly, social media plays an important role in the social lives of children [82,197]. While interacting with peers contributes to a child's growth, development, and well-being [263], directed malice and cyberbullying are stark realities of using social networks. Children who are experiencing and engaging in cyberbullying can be viewed as domain experts of cyberbullying. Currently, however, there has been little published research that involves children and adolescents, those most affected by the malice of cyberbullying, in designing and building technology to mitigate the effects of cyberbullying.

5.3.1 Participatory Design and Youth

Partnering with children as design partners has led to technologies being better suited for the needs of children. There have been quite a few participatory design techniques for children implemented and evaluated to design new technologies. Druin et al. introduced cooperative inquiry by interpreting participatory design contextual inquiry methods for children as partners in the design process [83]. The Mixing Ideas technique is used to encourage collaboration during the design process [115]. Comics and roleplaying through techniques like KidReporter demonstrate diverse ways to elicit information from children [29].

With the burgeoning use of social media platforms by teens, the importance of incorporating teens in the design process is being realized. Fitton et al. [96] included several ongoing studies that are involving teenagers in their design processes in order to yield better interactive products. A recent study on Snapchat highlighted the importance of including teenagers in the conversation since they are domain experts on some of these social media platforms [36]. Similarly, another study found it beneficial to involve teens in the design process for developing applications that aim to prevent unhealthy habits [191].

5.3.2 Designing Cyberbullying Mitigation Tools

Until this point, there has been very little work on the design of technological solutions for cyberbullying. For instance, there has been work on cyberbullying that focuses on community involvement and parental responsibility to address the

problem (i.e., education) [74]. Dinakar et al. [79] introduced "reflexive interface" prototypes as a means to prevent cyberbullying across a limited range of subjects, including appearance, intelligence, racial and ethnic slurs, social acceptance, and rejection. The reflective interface encourages positive digital behavioral norms and consists of the following interactions in order to deter malicious behavior: notifications, action delays, displaying hidden consequences, system-suggested flagging, and interactive education. The reflective interfaces to mitigate cyberbullying did not involve youth in the design or evaluation processes.

5.3.3 User-Centered Design and Cyberbullying

Traditionally, users have been placed in a reactionary role in the design and development process of technologies [186]. The pitfall of such an approach is that users are only reacting to what designers are creating and are not offering their own designs or solutions. User-centered design seeks to introduce users to the design process in the earlier stages so that they can influence the design of the tool. Many methodologies bring users of technology into the design and development process. Users have been involved in all stages of development including co-designers, testers and subjects.

Children provide valuable insight in the design process. Increasingly over the past decade, researchers have included children more actively in the design process, successfully demonstrating that children offer fresh perspectives that lead to innovative designs in technology [84,231]. However, in the realm of cyberbullying, very few design innovations have been introduced through participatory design. Bowler et al.

conducted a narrative-inquiry based study in their participatory design sessions [39]. The narrative inquiry was the only participatory design methodology employed to develop their seven emergent themes for cyberbullying mitigation: design for reflection, design for consequence, design for empathy, design for personal empowerment, design for fear, design for attention, and design for control and suppression. While this methodology is valuable, it is only an initial step in the right direction. To move this area of research forward, I conducted five participatory design sessions with ninth and twelfth graders. I employed multiple participatory design techniques in our participatory design sessions including: Focus Groups, "Bags of Stuff" and Mixing Ideas [83, 115, 282]. A narrative-inquiry based study limits participants to a singular narrative and thus confines their capacity for introducing novel solutions. My diverse techniques allowed our teen co-designers to consider all types of cyberbullying, and how the various types are enacted the context of different social network platforms.

5.4 Method

To enable teenagers to become a part of the design process, the research team collaborated with a local private high school during the 2014-2015 academic year. The research team worked closely with school administrators to develop a survey instrument and identify classes that would be able to participate in multiple sessions during the spring semester. The team chose to work with high school age youth (14-17 years old) because they are legally allowed to use social media applications [211]

and are more likely to own smartphones. In addition, 9th and 12th graders are at different life stages and likely have different world views toward technology.

After receiving informed consent from parents and child assent forms, I began running design sessions during participants' "open period", averaging one session per week. I worked with fourteen 9th graders and seven 12th graders over the course of five weeks, using a range of participatory design techniques, such as "Bags of Stuff" and Mixing Ideas [83, 115, 282]. I purposefully chose a diverse set of PD activities to encourage the participants to consider multiple aspects of cyberbullying and mitigating solutions.

Below, I provide additional details on the descriptive data collected from the participants through a survey, as well as a breakdown of each of the sessions.

5.4.1 Survey Details

Before the PD sessions began, I surveyed our 21 prospective participants to understand how they interact with social media and different technologies, and to provide contextual information to tailor our design sessions to their personal experiences.

The participants were heavy technology users. Everyone owned a cell phone and all but one (a 9th grader) owned a smartphone, and they spent a large portion of their day using their phones–eight participants said they spend more than two hours a day using their cell phone. Regarding other technologies, they spent, on average, about one hour per day using a computer or laptop and an additional 1-2 hours per day using a tablet (which they used in school). Only six participants

reported that they played games regularly. Looking at their use of various social media platforms, Snapchat (80%) and Instagram (76%) were the most popular by far among overall, while less than half said they used Facebook, Google Plus, Tumblr, Vine, or Twitter. Everyone in the 9th grade group used Snapchat (compared to just three of the 12th graders); on the other hand, 12th graders were significantly more likely to use Facebook than 9th graders (71% vs. 36%).

Regarding past experiences with bullying and cyberbullying, five of the 21 participants (24%) said they had personally experienced cyberbullying, while eight participants (38%) said they had at least one friend who had been cyberbullied. In total, just under half (48%) of participants had personal experience with cyberbullying. For the five students who had personally experienced cyberbullying, one said there had only been one incident, while the other four participants said they had been cyberbullied a couple times. Cyberbullying incidents reported by these participants occurred most frequently through text messages (three participants reporting this), with Twitter, YouTube, ASKfm, Instagram, Kik, and online games each being selected by one participant. In dealing with these incidents, the participants said they either did nothing or turned to a friend for support. Below, I show a sampling of comments participants provided when asked to describe a time when they or their friend had been cyberbullied.

- *"People took pictures and made fun of my friend. Also bullied her on Twitter."*
- *"Someone called my friend fat, and mean words."*
- *"I was online playing Minecraft and this kid just starting cursing at me because i won the game and he lost and now i go on 'no cursing' servers on Minecraft."*

Item	Mean	SD
Many people in my school engage in cyberbullying.	2.33	1.49
Many people in my school have been victims of cyberbullying.	2.95	1.77
People in my school don't step in when someone is being cyberbullied.	3.19	1.87
When a cyberbully has been caught, he/she was punished by the school.	4.35	1.60
Most of the cyberbullying I've seen occurs anonymously.	4.10	1.22
People in my school are quick to report cyberbullying to parents, teachers, coaches, etc.	4.24	1.55

Table 5.1: Participants' attitudes toward cyberbullying at their high school (Response scale: 1=Strongly Disagree, 5=Strongly Agree).

- *"I was at a different school and people were making fun of me on twitter and with their other friends. People took pictures of me and made fun of me."*

In general, the participants who participated in this study reported that cyberbullying was not a problem at their school, but when it did occur, participants reported the incident and the school administration was likely to step in. Table 5.1 includes responses to five Likert-type items about their perceptions of cyberbullying at their school (Scale: 1=Strongly Disagree, 5=Strongly Agree). There were no significant differences between responses from the two groups.

5.4.2 PD Sessions: My Design Partners

I conducted a total of ten sessions, five each with two classes at a local high school. The ninth graders were ages 14 and 15, while the twelfth graders were ages 17 and 18. Sessions were constrained to the times during which participants had "free periods," which typically lasted 45 minutes. Since cyberbullying is common in high school settings [235], I chose freshmen and seniors as are co-designers; this allowed us to gain insights into how perspectives vary between younger and older adolescents,

who have different degrees of access to technology and social media [165]. My ninth grade sessions consisted of 14 participants, 10 of whom were female and four of whom were male. My twelfth grade class consisted of seven participants, five of whom were female, two of whom were male. All of the participants reported through our surveys that they had been active on some social media platform. I applied the same structure to both sets of participants across the five sessions. Three adult researchers were present at each session to facilitate discussion and collaborate with the participants in creating new design ideas.

5.4.3 The Design Activities

Each of the five sessions conducted with the participants focused on specific aspects of participatory design. These were: (1) Focus Groups, (2) Scenario Centers, (3) Bags of Stuff, (4) Mixing Ideas, and (5) Evaluating Prototypes.

5.4.3.1 Session I: Focus Group

In our initial session, I held a focus group to familiarize ourselves with the participants' environment and to familiarize them with the researchers. In these sessions, I stressed that the participants would not be graded on their performance in the sessions. Additionally, I explained that I were studying online harassment and I wanted to work with the participants for the best solutions to fix the problem of online harassment. My goal was to understand how the participants interacted with social media platforms and how these platforms might be being used for cyber-bullying. Participants sat in a large circle and the moderators of the sessions asked

questions about social media practices.

5.4.3.2 Session II: Scenario Centers

In the second session, the researchers developed a set of "Scenario Centers" to help participants begin to think about the specific scenarios they would be designing for. The concept of "Scenario Centers" stems from the childhood experience of *center time*, in which participants learn and engage in different experiences in "centers" [115]. Participants in each grade were presented with scenarios from the different social media platforms based on the themes that emerged during the focus group session. Participants got into groups to discuss each scenario, then were asked to think critically about possible technological and non-technological solutions to mitigate the negative behaviors. The goal of this question was to prompt them to begin thinking about technological mechanisms (both existing and non-existing) on the social media platforms that would aid in helping the victim or prevent the bully in the situations presented them in the centers.

My scenarios focused on the social media platforms participants reported using most frequently and represented several types of cyberbullying, including Flaming, Harassment, Cyberstalking, Denigration, Outing and Trickery, and Exclusion [275] and to include the social media platforms on which participants indicated they were most active. Below I describe a subset of the scenarios.

1. **Facebook (Denigration)**: Sara is a new girl at school who dresses differently than the other kids. She is quiet and introverted so she has had trouble finding friends. Another kid at school starts taking pictures of Sara and posting them

on a group on Facebook, “Sara’s Weird Outfits.” The page has over 1000 likes and people start to comment on the strange clothing Sara wears to school. The comments seem to keep getting more malicious and personal. Recently, someone wrote: “She’s so ugly and her style sucks. She needs to die.”

2. **Snapchat (Flaming)**: Kyle keeps receiving repeated snapchats from Tom and Jake calling her names. They update their public stories which video messages of themselves saying “Kyle is ugly” or ”Kyle needs to die.” These videos are sometimes coupled with captions. Tom and Jake also send direct snapchats to Kyle.

3. **Ask.fm (Harassment/Cyberstalking)**: Jenna keeps receiving repeated anonymous messages on her ask.fm account: ”Go kill yourself” and ”No one likes you.” She responds to these messages to show that they aren’t affecting her. Because the messages are anonymous, she doesn’t know if they are coming from multiple people or just one person.

4. **Instagram (Exclusion)**: Jenny, Kayla, Sara and Felicia are all very good friends and have lunch together at school daily. Recently, however, Jenny has been avoiding Felicia. Jenny begins posting multiple pictures on Instagram in which she crops out Felicia and only tags Kayla and Sara. Felicia is feeling sad about how Jenny is excluding her and does not know how to react. She does not know if it was something that she did to make Jenny feel and act this way.

5. **YouTube (Outing and Trickery)**: Frank secretly records video of Sara and Ben getting intimate at a party, then posts it to YouTube the next day. The video goes viral within the school and is shared on all the social networks. The situation escalates when people from school begin commenting "sl*t"and "wh**e" on the video.

5.4.3.3 Session III: Bags of Stuff/Low-Fidelity Prototyping

In our third session, I employed a "Bags of Stuff" [85] design method to allow participants to design low-fidelity prototypes to address a specific cyberbullying issue. The researchers presented each group with a broad description of a cyberbullying even and instructed participants to utilize the provided art supplies (markers, white paper, construction paper, pipe cleaners, and stickers) to design a solution. The goal of "Bags of Stuff" is to allow children to feel that *anything* is possible in design. In the previous sessions, where participants considered different scenarios, it became clear that participants were limiting their discussions of potential solutions to tools that already existed. Thus, the research team that "Bags of Stuff" would encourage the participants to stretch their thinking and be creative in their solutions.

The scenario presented to both groups of participants included the range of cyberbullying activities outlined in previous research [275] and incorporated the various social media platforms identified by the participants in previous sessions.

"Bags of Stuff" Scenario

John is a victim of various types of cyberbullying on all of the social media

platform of which he is a part of. He keeps receiving repeated vulgar, offensive, or insulting messages/comments/snaps/posts on Snapchat, Facebook, Twitter, Instagram (respectively). John feels sad and depressed from all of the negative messages that are being received. John fears for his safety because he is repeatedly receiving threats online.

People are also posting cruel gossip or rumors about John to tarnish or damage his reputation. Students in his school have created a group on Facebook, a standalone website, or page dedicated to insulting him. Students in his school have also hacked his social media (Twitter, Facebook, Instagram) and are posting as him in an attempt to get John into trouble or make him look bad. John has also been tricked into giving personal or embarrassing information and pictures to Mike who shared it with everyone on Instagram, Facebook and all other social media. John is intentionally being excluded from his friends' online chat group on Facebook.

Participants were instructed to create a tool or application with their available materials that would address one or more of the following: 1) Prevent the cyberbullying behavior from happening, 2) Ease the emotional pain for the cyberbullying victim, 3) Stop the cyberbullying behavior from happening again (once it has already happened), and 4) "Solve" this problem in another way. They were told that the tool they created could be a separate application/website not associated with any particular social media platform AND/OR it could be incorporated within any or all of the social media platforms previously discussed.

5.4.3.4 Session IV: Mixing Ideas

In the fourth session, I reviewed the prototypes that participants had created and used the design methodology of "Mixing Ideas" to create new solutions. "Mixing Ideas" further fosters collaboration between participants by encouraging them to think about common themes between their solutions to create better solutions and prototypes [115]. With the researchers acting as discussion moderators, the classes then identified a set of prominent themes across groups, then got back into groups and spent the remainder of the session to further refine their prototypes.

5.4.3.5 Session V: Evaluating Prototypes

In the final session, the participants discussed the feasibility of each of the designs they had created by addressing strengths and limitations to implementation. While discussing each of the prototypes, the research team posed the following questions: "How would this work in real life?", "Is this implementable and useable?", and "Is this solution ethical?" While the participants were encouraged in the previous sessions to think "outside of the box with an open mind", the goal of this last session was for the designers to think about the *implications* of their designs.

5.5 Results

Across our two groups of co-designers, differences in social media platform use affected the perceived "coolness factor" of various platforms. In sessions, participants unanimously echoed their classmates (and the results of the preliminary

survey) by focusing the most attention on Instagram and Snapchat. The participants responded that Facebook and LinkedIn were the most "uncool" social media platforms. The appeal or "coolness" of a particular application is relevant when designing applications because of the bystander role in a bullying scenario could be played by an adult. One student said, "I go on Facebook because these old people relatives don' have any other social media. So I go to wish them happy birthday on there." Another student said, "Google Plus–who even uses that anymore?" The co-designers' responses to the different types of social media demonstrated that our design focuses for cyberbullying mitigation should focus on the affordances of social media platforms that are more widely used by demographics who are most affected by cyberbullying.

5.5.1 Negative Experiences on Social Media Among Design Partners

One of the main takeaways from the focus group sessions was understanding the co-designers familiarity with various social media platforms, the norms of interacting in these spaces, and what they perceived to be bad behavior. When asked about negative experiences on Snapchat, one student expressed that it is concerning when the recipient of a "snap" screenshots the conversation because Snapchat, unlike many other social media sites, affords ephemeral interactions. When someone takes a screenshot of a post, they violate the sender's trust by breaking the "unwritten rules of Snapchat." For Instagram, one student said users will create "hate pages" by posting screenshots of someone's social media posts and "making fun of the person." With regards to malicious behavior, the participants mentioned

that the silencing of dissenting opinions on Tumblr could occasionally "get out of hand." One student reported that some trending topics like "#stopblackpeople" or "#stopwhitepeople" start out as a joke and then just turn into racist posts as they spiral out of control. Regarding Tumblr, one of the participants said, "I hate all the racism and stereotyping that happens on there."

5.5.2 Design Applications

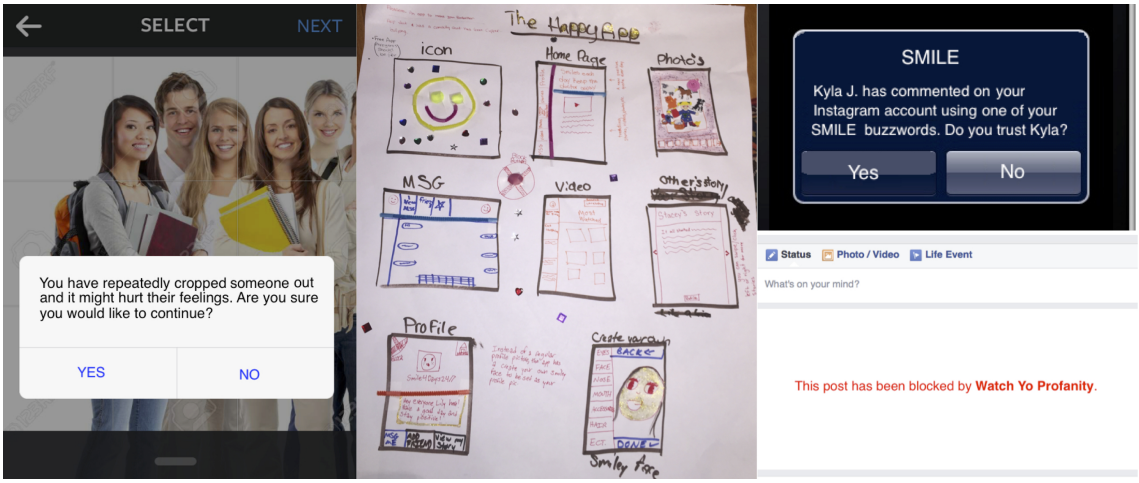


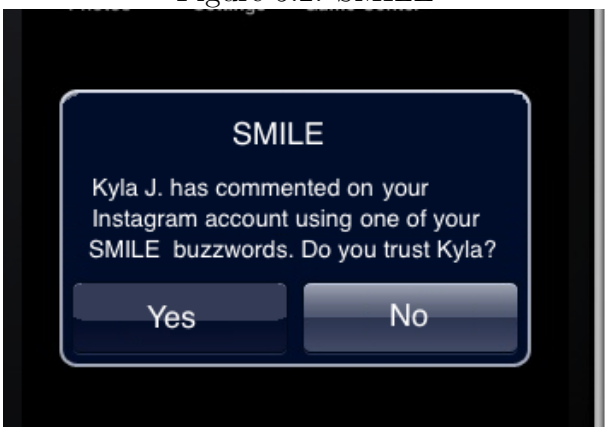
Figure 5.1: Application prototypes from sessions with participants, including Exclusion Prevention, Happy App, SMILE, and Watch Yo Profanity

Below, I describe the nine design solutions created by participants as a result of the prototyping sessions. The first set of seven were developed during Session III. Figure 5.1 includes visualizations of four of the prototypes.

1. **SMILE: Social Media Informative Life Empowerment**: SMILE is a third-party application that addresses the lack of control social media users have over what content *other users* can post to their pages. Users first create a

"buzzword" list containing potential sources of bullying or harassment. When a user receives a comment/post on their profile that includes one of their identified buzzwords, they receive a notification and are given the option to accept or reject the post on their profile. This particular social media application enhances the user experience by allowing the user to choose the content that becomes visible to other users. Furthermore, in the realm of automatic detection, detecting cyberbullying content just using expletives is often ineffective because youth will use expletives emotionally [79]. One of the participants pointed out that if her best friend used an expletive in a post, she would have no concerns about accepting it because she trusts the source of the content. In this sense, SMILE offers an innovative solution to the automatic detection cyberbullying problem.

Figure 5.2: SMILE



2. **"Happy App"**: Participants designed the Happy App for individuals who are upset or depressed due to cyberbullying. Users create their own profile and can share their experiences with cyberbullying. Interactive features allow users to connect with other cyberbullying victims and obtain peer support, which

has been shown to promote psychological well-being and positively influence self-esteem [173]. The designers also suggested having a positive quote of the day on the app's homepage.

3. **"Fight Back"**: This application is designed to help people who have been cyberbullied. Features include a chat room to connect with trained therapists or a friend of your choice to speak with about your experiences and get advice on responding to the harassment. Users can block people they don't want to connect with through the app. The app also has a "happy room" that includes positive messages to mitigate cyberbullying harm (similar to the Happy App's homepage).

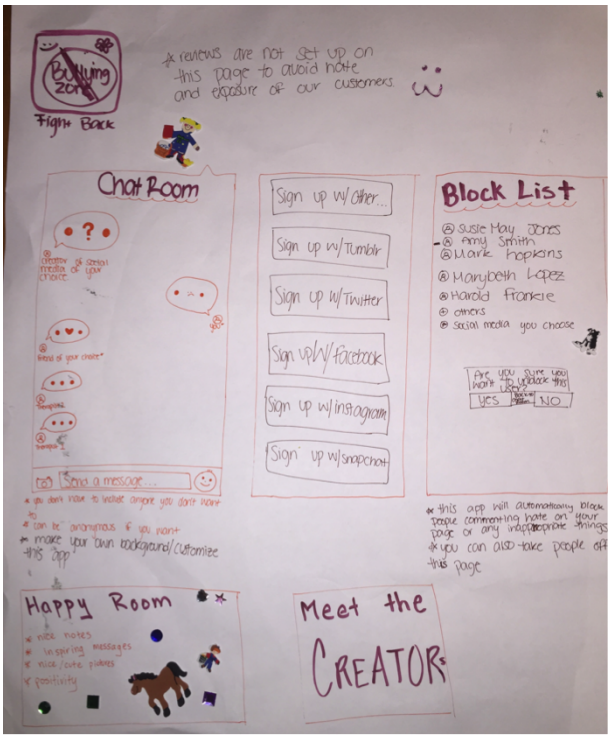
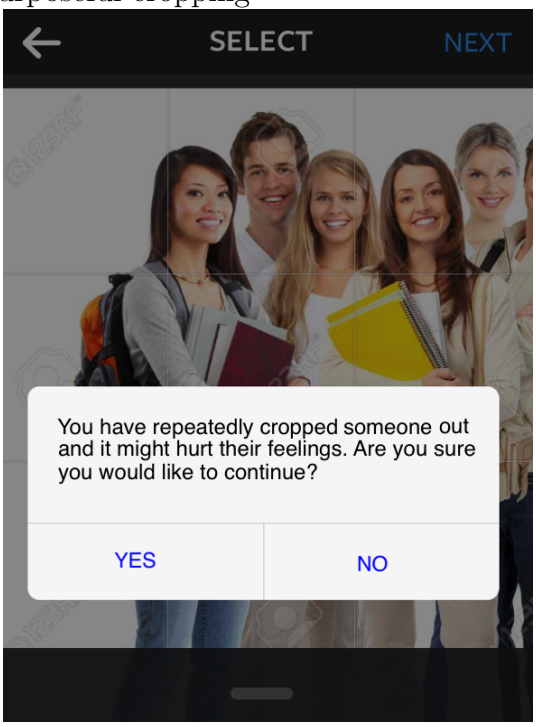


Figure 5.3: "Fight Back" Application

4. **"Exclusion Prevention"**: Repeatedly excluding someone is a form of bul-

lying [275]. Exclusion can manifest in multiple forms on social media. Specifically, repeatedly excluding someone by cropping them out of a picture before posting it may make someone feel excluded and hurt their feelings. Using face-detection technology, the teens created an application that alerts a social media user when they using the cropping feature on sites like Instagram and crop out one or more people in the picture. A message pops up before the picture posts to the site, telling the user that cropping people out of the picture could hurt their feelings. The user then decides whether they want to continue with posting the picture.

Figure 5.4: Exclusion Prevention design, aimed at preventing individuals from feeling excluded due to purposeful cropping



5. **"Watch Yo Profanity"**: This plugin features a filter that blocks out expletives and vulgar phrases from appearing on social media using an existing

dictionary of words and phrases. Users can further customize the filter to block additional words and phrases that are not in the plugin's database. Users can also block specific people in social media, hiding all content from those users. For example, if someone keeps posting inappropriate content on social media, a plugin user can opt to block that person.

6. **"The Broiler":** One student indicated that he had already chosen not to go to a specific university because of a campus-based website where participants could anonymously post mean comments about other participants at the school (in the same vein as the short-lived "Juicy Campus" website). As a means to fight it, he proposed an application or Twitter add-on called "The Broiler," that "would roast whoever is roasting others" on social media.

7. **"Reporting Bullies With Feedback":** Participants discussed their dismay at the lack of feedback they receive when reporting abuse on various social media platforms. When they alert the site about negative content, they want to be notified not only of the abuse, but also receive feedback about how the situation was being handled and additional information about the victim post-abuse. They designed a feedback tool that reported back to a user that had reported malicious content about the current status of bullying with the user who may have been effected and if that user requires additional emotional support.

During the "Mixing Ideas" sessions, two additional prototypes were generated:

1. **Positivity Generator**: This application was motivated by the "Watch Yo

Profanity" application designed in the third session. While "Watch Yo Profanity" merely censors potential cyberbullying content, the Positivity Generator *replaces* instances of negativity with positive and uplifting quotes from a selected celebrity. During the session, users chose Kanye West, a celebrity renowned for his self confidence [48]. In the sessions, the participants introduced the notion of expanding the "Kanye West Self-Confidence Generator" [1] to include self-confidence enhancing quotes by popular celebrities to boost the self-esteem of a cyberbullying victim. The participants suggested that a user's favorite artist could be inferred from their social media activity and that celebrity's most uplifting and encouraging sayings could be used replace negative content.

2. **"Hate Page Prevention":** The participants mentioned that the same face detection technology used for the "Exclusion Prevention" feature could also be leveraged on Instagram to automatically discover and report "Hate Pages". "Hate Pages" can be defined as pages where a cyberbully posts screenshots of another user's pictures and uses malicious captions under those photos to harass the person. The co-designers had a keen understanding of facial recognition software available through platforms like Facebook through their experience with tagging [28]. They proposed a monitoring system to prevent hate pages. If a user's photos looked too similar to another's then the page would be automatically flagged and investigated by platform administrators.

5.6 Discussion

The wide range of cyberbullying scenarios I discussed with our design partners prompted them to consider forms of malicious online behavior that may not traditionally be deemed as cyberbullying. Below I discuss in more detail how our participant designers conceptualized cyberbullying and how their proposed solutions may be enacted in meaningful ways.

5.6.1 Defining Cyberbullying

In order to start designing for cyberbullying, the research team explored whether our co-designers were in agreement about the definition of cyberbullying. Participants unanimously agreed that all of the scenarios constituted cyberbullying, except for *Instagram (Exclusion)* in which a girl is continuously cropped from photos on Instagram. A heated discussion emerged among the participants as to whether cropping a user out of photos "is just rude and not targeted enough" or if it is more severe. The participants who argued that this scenario did not constitute cyberbullying claimed that in order for something to constitute as cyberbullying, 1) it must go "viral" so as to include a wide audience and 2) it must be directly targeted. For the male participants, the *Instagram (Exclusion)* scenario was missing these two components. One male student said, "Do you know why it's not bullying? Because [the girls] are still sitting with her when they take the photograph. It is still peaceful." One female student who adamantly believed the scenario constituted as bullying countered with, "Making someone feel insecure about themselves

is bullying too. Bullying isn't just physical."

While many researchers have surveyed youth to understand the climate of cyberbullying activities [74], exclusion cyberbullying has been largely ignored in cyberbullying research. My study sheds light on the emotional trauma this kind of cyberbullying causes since one of our participants claimed that she had experienced exclusion repeatedly online.

5.6.2 Designing for Support

Many of the prototype designs fall under the emergent themes described by Bowler et al. [39]. For example. "Hate Page Prevention", "Reporting Bullies with Feedback", "SMILE", and "Watch Yo Profanity" fall under "Design for Control and Suppression," a theme that involves controlling content through a social media platform's administrators or a third party algorithm. "Exclusion Prevention" creates a pause in the cyberbullying process by asking users if they want to proceed by excluding someone from a picture, which is in line with Bowler et al.'s [39] "Design for Reflection" theme. Finally, "The Broiler" (though ethically questionable) and "Reporting Bullies With Feedback" are designed for consequence, because they ensure that their are consequences for bullying behavior.

While Bowler et al.'s [39] themes accurately describe most designs for cyberbullying mitigation, the design sessions highlighted a fourth critical theme: Designing for Support. Three of our designed prototypes–"Positivity Generator", "Happy App", and "Fight Back"–recognize that a bully's actions cannot always be controlled on social media. A bully may face negative consequences for their behaviors, but

post-bullying solutions cannot *prevent* bullying from occurring. The seven emergent themes described by Bowler et al. [39] are all bully-centric. They focus on instilling fear or engendering empathy for the victim. Therefore, it is important to have tools to provide cyberbullying victims with emotional support and positivity *after the fact*. Based on the conversations and design sessions with participants, mitigation and support after the cyberbullying occurs is vital part of the mitigation process. This focus on mitigation after the cyberbullying is reflective of the lack of control over the bully and over the social media applications which afford bullying behavior.

I should note that while the "Designing for Empowerment" theme aims to redress the balance of power on social media by asking adults to play a more active role in intervention [39], student-designed automated and peer-focused solutions such as the "Positivity Generator" and "Happy App" may also potentially play an important role in empowering and supporting victims of cyberbullying; this should be outlined specifically in the list of design themes for cyberbullying mitigation.

5.6.3 Designing for Prevention

The prototypes generated from our design sessions varied in terms of who held control in either preventing a bullying scenario or mitigating it after it occurred. In addition to bullies and their victims, bystanders play an important role in bullying scenarios, often offering implicit or explicit encouragement or discouragement of the bullying [265]. Furthermore, bullying roles are not always dichotomous, with individuals roles based on contextual factors.In our sessions, sample prototypes were generated in which all three actors (bullies, victims, and bystanders/systems) poten-

tially had control over preventing or mitigating the bullying scenario. This particular taxonomy of cyberbullying solutions complements regarding power dynamics in bullying situations and identifying who–the victim, aggressor, and/or bystander–has the most power to mitigate the situation.

5.6.3.1 Cyberbullying Prevention by the Bully

While literature has discussed how to approach denigration and flaming [79], I have found no academic research discussion issues related to exclusion online. A *New York Times* parenting blog noted, "To be in a photo and to not be tagged is to be rendered socially invisible. Commenting on a party photo, my untagged daughter wrote, 'I was there too!"' [180]. The "Exclusion Prevention" application aims to remedy the potential emotional damage of exclusion-based cyberbullying by presenting the potential bully with a reflective notification. In "Exclusion Prevention", the bully decides whether she wants to continue with publishing content after the system warns the [potential] bully that they may be hurting someone by continuously cropping them out of photos. Ultimately, the decision of publishing the content lies with the potential bully. Dinakar et al. [79] have shared examples of preventive measures when discussing reflective interfaces, which ask users to reflect on their behavior before publishing malicious content online.

From an implementation standpoint, aiming to prevent cyberbullying by focusing on the [potential] bully requires some monitoring since it is attempting to prevent the cyberbullying before it occurs. While privacy advocates may find this monitoring particularly troubling, many parents believe they have the right to ac-

cess and monitor their children's online activity [24]. There are three notions of a reflective practitioner: "reflection in action", "reflection on action", and "ladders of reflections" [237]. Reflective user interfaces aim to prevent cyberbullying by asking the aggressor to reconsider their actions and reflect on them through showing potential consequences of their actions, flagging their content and notifying them of the potential harm they can cause.

5.6.3.2 Cyberbullying Prevention by Victim

In the realm of cyberbullying prevention, cyberbullying applications that filter or report content can aid the victim in preventing further occurrences of bullying. A victim can choose to filter bullying content so they never see it (and subsequently experience negative emotional consequences). In "Watch Yo Profanity" and "SMILE", the victim decides if she would like some degree of filtering to be happening on his profile. Depending on the individual, these applications can serve a more proactive approach, whereby the individual chooses filters prior to negative events, or a more reactive approach, in which victims take apply filters to prevent future instances of cyberbullying content.

5.6.3.3 Cyberbullying Prevention by Automated Systems and Bystanders

From the suite of solutions produced in our design sessions, many attempted to mitigate negative emotional outcome of cyberbullying by sending positivity. This kind of prevention can be initiated by bystanders or automated systems. In the cyberbullying domain, the "Positivity Generator" allows victims to replace malicious

content on their profiles with uplifting quotes from their favorite celebrities. This particular solution aims to do more than just filter negative content, but provide support and encouragement to counter the negative cyberbullying content they have experienced. Likewise, for "Hate Page Prevention", a bystander or a third-party automated system has the ultimate control over the cyberbullying content being published.

5.6.4 Limitations of Emergent Solutions

In our design sessions, all participants were encouraged to think outside of the box. In Session V, however, the participants explored the feasibility of their design solutions. For example, there was a discussion that "Exclusion Prevention" may be disruptive to a user's experience on Instagram if a user was prompted that they may hurt someone's feelings every time they tried to crop someone out of a picture. One participant said, "What if my friend doesn't look good in a picture and she would actually prefer for me to crop her out?" These questions prompted them to begin to think of allowing a user to opt-in/opt-out of using their application. The participants questioned the accuracy of filtering algorithms. To counter this limitation, they introduced the notion of letting user decide whether they wants to see the content based on the person who is sending the content. One student said, '1I know that [close friend] would never send me anything malicious, so if I was notified that 'Watch Yo Profanity' filtered somethings she sent me, I would know to undo it." When reflecting on "The Broiler", the participants decided that "bullying the bullies" was not a ethically sound solution; they ultimately decided that such

an application was counter-productive to the cyberbullying mitigation movement.

5.6.5 Technologies and Tools for Implementation

In the design sessions, participants were encouraged to think out of the box regarding what would be technologically possible. That said, many of the solutions they designed are implementable. Below I discuss the technologies required to implement specific cyberbullying solutions.

1. **Application Programming Interfaces.** Many of our teen co-design partners' design solutions, whether standalone applications or browser plugins, required some degree of interaction with social media platforms (e.g., Snapchat, Facebook, Instagram). Just as social media platforms provide Application Programming Interface (API) services that developers and researchers use to expand our understanding of other technology-mediated social interactions (e.g., trending topics), they are important for developers and researchers who aim to design and implement cyberbullying solutions.

2. **Image Recognition Technology.** Image recognition technology plays a vital role in two of the design solutions: "Hate Page Prevention" and "Exclusion Prevention." The participants described the ease with which social media platforms detect faces when making "tagging" suggestions. They noted that such face detection technologies could be leveraged to prevent "hate pages" which involved the use of screen-shot photos. Screen-shot photos were a re-occurring topic of discussion in our sessions. Many participants expressed

that while screen-shotting a Snapchat photo was not necessarily cyberbullying, they would feel threatened if someone screen-shot their photos because it would be both be a violation of privacy and social media etiquette. Their solution to this violation was embedding image recognition technology into social media to detect when a photo was being re-used, and the using primary prevention tactics via reflective interfaces to prompt the aggressor to reconsider re-posting someone else's photo if it was indeed for malicious reasons. Furthermore, the participants spoke about the importance of face recognition technology to prevent exclusion via cropping on applications like Instagram.

3. **Automatic Detection of Malicious and Vulgar Content.** Within the applications that filtered malicious content, participants expressed that there should be filtering of some kind employed on the website. While most of the filtering solutions included use of "negative buzzwords", participants references many of the same challenges researchers face in the automatic detection realm, especially false-positives in the cases where a negative word is used but the overall content of the post is not negative. To solve this problem, participants said they allow users to play a decision-making role in filtering items. A victim would be notified if a buzzword was used, but would then get to evaluate whether the post was positive or negative based on the person sending the message. According to the participants in the sessions, the likelihood of being cyberbullied by someone you trust is not high so allowing users to evaluate content based on the person sending it could be a viable solution. While there

have been many attempts to accurately identify expletives and negative words in the cyberbullying domain through sophisticated classifiers [80, 145], none has ever attempted to give the victims the power to choose what is acceptable or unacceptable to be posted on their profile. The challenge with automatic detection of cyberbullying is that often expletives can be used affectionately in this particular domain, so it would be counter-productive to filter those cases [145].

4. **Collaborative Filtering: Inferring Favorite Artists from Social Media Data.** With the "Positivity Generator", participants discussed the possibility of an algorithm inferring a user's favorite artists. On many social media platforms, it is possible to follow/like celebrity pages, and collaborative filtering can be leveraged to infer a user's likes and interests [229]. Those likes could be used to provide support for a cyberbullying victim. Collaborative filtering and similar methodologies in recommendation system research could be leveraged in an application like the "Positivity Generator" to provide targeted support for a victim of cyberbullying. Previous solutions in the non-academic sphere have sent out targeted song lyrics to a victim of cyberbullying based on the artists the cyberbullying victim enjoys. The design participants expressed that this kind of targeted support, or the sophisticated understanding of a system of what things a victim enjoys would be particularly valuable in providing support to counter the negativity of cyberbullying.

5. **Networks.** Prototype solutions like "Exclusion Prevention" and "Hate Page

Prevention" can be implemented with higher granularity if they make use of the available information provided by a user's network on social media platforms. For example, the formation of "cliques" based on tagging behavior on Instagram and the way that they change coupled with face detection technology can predict more accurately if someone is indeed being excluding maliciously. By analyzing a user's network, it is possible to discover individuals who play an important role in someone's life based on common connections and the clusters within a user's network [102]. Leveraging this existing technology can be helpful in applications that require someone close to the victim, a peer, to provide support for a victim.

5.7 Self-Evaluation of Co-design of Researchers and Teenagers

In our last participatory design session, I asked participants to reflect on their experience doing participatory design. From these discussions, three primary themes emerged: 1) appreciation of opportunities to discuss an important issue affecting participants and their peers, 2) surprise over discrepancy of opinion of definitions of cyberbullying and 3) excitement about collaborating with adults as equals. Ihave expanded on these three themes in the list below.

1. Participants appreciated inclusion in an initiative to prevent cyberbullying since many knew of the aftermath of such incidents, and participants were eager to have the designed solutions implemented. While many of the participants did not personally experience cyberbullying, they knew peers who had

experienced cyberbullying and the resulting negative repercussions. They expressed their excitement of being involved in an initiative that affected people around to them and that they considered a real-world problem.

2. Participants expressed surprise over the diversity of opinions within their peer group over the definition of cyberbullying (e.g., whether cropping an individual from a photo was cyberbullying). These discussions fostered mutual appreciation between design partners and allowed them to consider nuances of these differences when designing mitigation tools.

3. Participants expressed excitement over the novelty of collaborating with adults for a shared cause and said that the process fostered communication skills with collaborators who were older. Many of the participants had not had the opportunity work with adults as equal design partners and this experience was novel for them. They enjoyed the dynamic of working with adults as equal design partners.

5.8 Conclusion

While this study resulted in potential solutions for cyberbullying mitigation, much work lies ahead. I proposed a number of potential mitigation solutions and the technologies required to implement these solutions. Future research should implement and evaluate these solutions with users through longitudinal studies to evaluate the behavioral impact they have on bullies, victims, and bystanders. In the next chapter, I leverage the existing technologies to implement an extension of one of the

proposed solutions resulting from co-design between researchers and adolescents.

My analysis and categorization of the different preventative types allows us to consider additional research questions, such as which preventative solution is most effective for cyberbullying prevention and how can we accurately measure this effectiveness. Until this point, technological cyberbullying prevention mechanisms have not been evaluated for effectiveness. The framework presented in this paper provides a straightforward way to begin to consider how one would compare different solutions. The ethical challenges of such a study are daunting, but would provide critical insights to preventing cyberbullying.

This paper presents solutions to cyberbullying that were designed by the users most vulnerable to it: adolescents. Specic ways in which this study contribute to HCI are: 1) extending existing cyberbullying intervention design themes(specifically, Bowler et al. [39]) through the analysis of solutions designed with teenagers; and 2) implementing new techniques within the participatory design process to generate cyberbullying solutions from teens perspective (as compared to implementing teen feedback on designs rst created by adults). Finally, the study demonstrates that participatory design using teenagerswho have a vital stake in cyberbullying prevention and mitigationprovides novel insights and solutions. In the next chapter, I implement a cyberbullying mitigation tool based on the findings in this study and evaluate its effectiveness in a longitudinal study.

Chapter 6: Mitigation of Negative Experiences on Social Media through Curated Technology Mediated Memory

6.1 Chapter Summary

In this chapter, I draw from the design recommendations in Chapter 5 to design and evaluate a cyberbullying mitigation tool. The study described in this chapter addresses cyberbullying and online harassment among young adults and investigates the design and effectiveness of technological mechanisms to mitigate sadness and decline in well-being caused by negative online experiences and cyberbullying. I administer cyberbullying mitigation through Curated Technology-Mediated Memory (CTMM); in other words, I curate positive posts and images participants have previously shared on social media to remind participants of existing social support in users social networks. Based on the results of this study, I offer design recommendations for creating and administering cyberbullying mitigation as well as recommendations for designing a study to evaluate the effectiveness of cyberbullying mitigation tools.

6.2 Introduction

In this section, I provide an introduction of Curated Technology Mediated Memory (CTMM), how it is defined, and the motivation behind studying it. I provide a review of relevant literature that discusses reminiscence as a mediating factor.

To avoid the word bully, some employ synonyms (e.g., mean things) [1, 2, 4, 6] or omit it from the definition entirely [20]. In a 14-country comparison of 67 words and phrases used to describe bullying, Smith and colleagues report that the terms bullying and picking on cluster together, whereas the words harassment, intimidation and tormenting relate to each other in a different cluster [22]. Thus, the use of synonyms may not always connote bullying.

In Chapter 5, I discussed different prototypes that resulted from the participatory design techniques administered with teens at a local high school. One of the designed prototypes - "Positivity Generator" replaces instances of negativity with positivity on a user's profile. In the design sessions, participants suggested using existing content on a user's profile to infer positive aspects of an individual's life that might potentially mitigate cyberbullying. The "Positivity Generator" falls under the theme, *Designing for Support*. The prototypes that *Design for Support* focus on providing victims with emotional support and positivity *after they have been bullied*. Based on the conversations with the participants in the study, providing support after a cyberbullying episode is an important part of the mitigation process.

A mitigation theme identified by Bowler et al., *Designing for Empowerment*

addresses the balance of power on social media by asking adults to play a more active role in intervention [39]. However, prototypes like the "Positivity Generator" and Curated Technology Mediated Memory (introduced in this chapter) have the potential to play an important role in empowering and supporting victims of cyberbullying.

In this study, I use Curated Technology Mediated Memory (CTMM) to mitigate cyberbullying. Evaluating the effectiveness of cyberbullying mitigation tools poses many logistical and ethical challenges to HCI researchers. While there has been much research on *detecting* cyberbullying [79, 80, 279] and less work on *designing mitigation tools* [39] for cyberbullying, there have been no studies that *evaluate* the effectiveness of a cyberbullying mitigation tool. Cyberbullying mitigation is difficult for many reasons. Firstly, finding populations affected by cyberbullying who are willing to participate in longitudinal studies that measure the effectiveness of a cyberbullying mitigation tool is difficult. Secondly, cyberbullying and the negative effects of cyberbullying are deeply contextual; individuals are affected by malice differently based on their existing social support systems and self-esteem. Thirdly, as demonstrated in the participatory design study in Chapter 5, there are many different types of cyberbullying and ways they can be potentially mitigated or prevented. In this preliminary work, I focus on promoting well-being after an individual has already experienced cyberbullying. Using existing sentiment analysis technologies, I curate positive memories for individuals who have been previously affected by cyberbullying. I monitor participants for four weeks and measure well-being before and after I administer the cyberbullying mitigation. CTMM supports

well-being after individuals have had negative experiences on social media. I also probe about participants' positive and negative weekly experiences on social media. At the completion of the longitudinal study, I ask for participant feedback on the design and delivery of CTMM. The research questions driving this study are listed below:

RQ1 In addition to cyberbullying, what kinds of positive and negative experiences are users having on social media?

RQ2 Can Curated Technology Mediated Memory *mitigate* the negative emotional effects of cyberbullying and other negative experiences on social media?

RQ3 How can the design and curation of Curated Technology Mediated Memory better promote well-being for victims of cyberbullying?

6.2.1 Defining Cyberbullying in this Study

Definitions of cyberbullying vary [258], leading to lack of consensus on the definition and thus a lack of progress in research. In this study, I treat cyberbullying as a mode of communication and not as a type of bullying. If cyberbullying is considered a type of bullying, research on the subject is at risk of double counting instances of bullying (online versus instances of bullying at school) [170]. When measuring cyberbullying researchers operationalize it differently. Some may ask about "cyberbullying" experiences without offering a definition, while others may use an explicit definition. Defining "cyberbullying" may limit respondents' to the researchers' definition and may not wholly capture experiences that differ from the

experience provided. Other researchers have used synonyms like "mean things" to equate bullying [281]. However, these synonyms may not always be equal to bullying since studies have found that words like "harassment" and "intimidation" are clustered together while "bullying" and "picking on" are often clustered together. One study found that studies should include the word "bully" when conducting studies on cyberbullying [280]. For this reason, both the preliminary survey and the exit survey in this study include this word. Providing a definition of "cyberbullying" results in experiences that are beyond the scope of cyberbullying and for this reason, I included this word in both the preliminary and exit survey in this study.

6.2.2 Defining Curated Technology Mediated Memory

Curated Technology Mediated Memory (CTMM) can be defined as a collection of curated digital artifacts that have are delivered to a recipient to trigger positive memories. Digital artifacts can be represented in the form of text published by the subject, interactions between the subject and online connections, or visual content like photos and videos. In the implementation of CTMM presented in this study, I employ (unsupervised) sentiment analysis, checking for digital artifacts that encompass cues and features that indicate they are positive aspects of a user's experiences. While digital artifacts are complex and inclusive of events or representative of individuals that may trigger mixed feelings, the curation process is a first step towards filtering memory triggers that may not be positive. In the next section, I discuss additional heuristics used to create CTMM in this study.

6.2.3 Curated Technology Mediated Memory on Facebook

The CTMM in this study presents participants with a collage of curated Facebook friends' photos and a collection of positive timeline exchanges. Interactions include timeline posts, mutually tagged photos, as well as likes and comments on content. In order to ensure only photos that document positive events are included in the collage, I used an unsupervised sentiment analysis method to classify posts as positive, negative or neutral comments [201]. CTMM administered to the users in this study aims to leverage bonding social capital from close personal relationships [212].

To identify strong ties, I chose individuals who are tagged in photos more than two times. This heuristic ensures that the individuals have been co-located with the person indicating that bonding social capital can be potentially redeemed. Facebook fosters many different types of interactions, from online connections of very close friends to strangers who have not met [91]. The Curated Technology Mediated Memories administered include a collage of the most liked photos of two individuals (the participant and their close friend), and the positive comments or content shared on a user's timeline and photo comments.

6.2.4 "See Friendships" as a type of CTMM

"Facebook Friendships" are a type of CTMM since the content presented in "Facebook Friendships" are curated memories. Specifically, the differences between "Facebook Friendships" and Curated Technology Mediated Memory in this study

are two-fold:

1. Curated Technology Mediated Memory is directly delivered/pushed to users

2. Sentiment analysis is used on Curated Technology Mediated Memory to ensure only positive content is sent to users

It is important to note that CTMM in this study is "pushed" to users, as opposed to "Facebook Friendships" which users must seek out in order to access. Reflection on friendships is spontaneous and intentional and is triggered by random cues (common interest, related events or online/offline co-presence with a friend). Participants in a study exploring the motivations behind why people reflect on friendships and the benefits of reflecting on friendships reported that reflecting on friendships helps people value their friendships more and increases their trust towards each other. Participants reported that reflection on friendships leads to more happiness when they reflect and value their friendship. Despite these known benefits of increased happiness and increase of trust between friends, participants in this study reported that they "almost never" or "rarely" used the "See Friendship" page even though they were aware of it [238]. CTMM is delivered directly to users and thus directly delivers the benefit of providing these feelings of increased trust, value towards friends, and happiness to users.

Furthermore, sentiment analysis was used to only choose photographs and memories that were classified as positive by the algorithm. While classification is not perfect, these methods ensure that users are only reminded of memories that may be deemed positive to avoid reminding users of memories which may not be

favorable. Sentiment analysis has its obvious limitations. For example, based on the positive comments on a photograph of two individuals in a romantic relationship, a photo may be classified as positive. However, sending that photo to a user at a later time when the romantic relationship has dissolved may trigger painful emotions, accomplishing the opposite of that which was intended. Sas et. al describe how users who have dissolved romantic relationships participate in disposal strategies of digital possessions in order to forget the romantic relationship [230]. While sentiment analysis is an initial step to finding positive memories, the dynamic nature of human relationships as well as cues from social media interactions that reflect the current state of the relationship should be considered when curating digital artifacts and memories aimed at promoting well-being. In the CTMM administered in this study, photos that included comments classified as "negative" were omitted from collages. Furthermore, only interactions classified as "positive" were presented to the recipient of the CTMM.

6.3 Related Literature

My prior participatory design work, while yielding interesting findings requires additional empirical foundations in order to support the direction of my study. Below, I present a short review of reminiscence as a mediating factor to support the direction of my study.

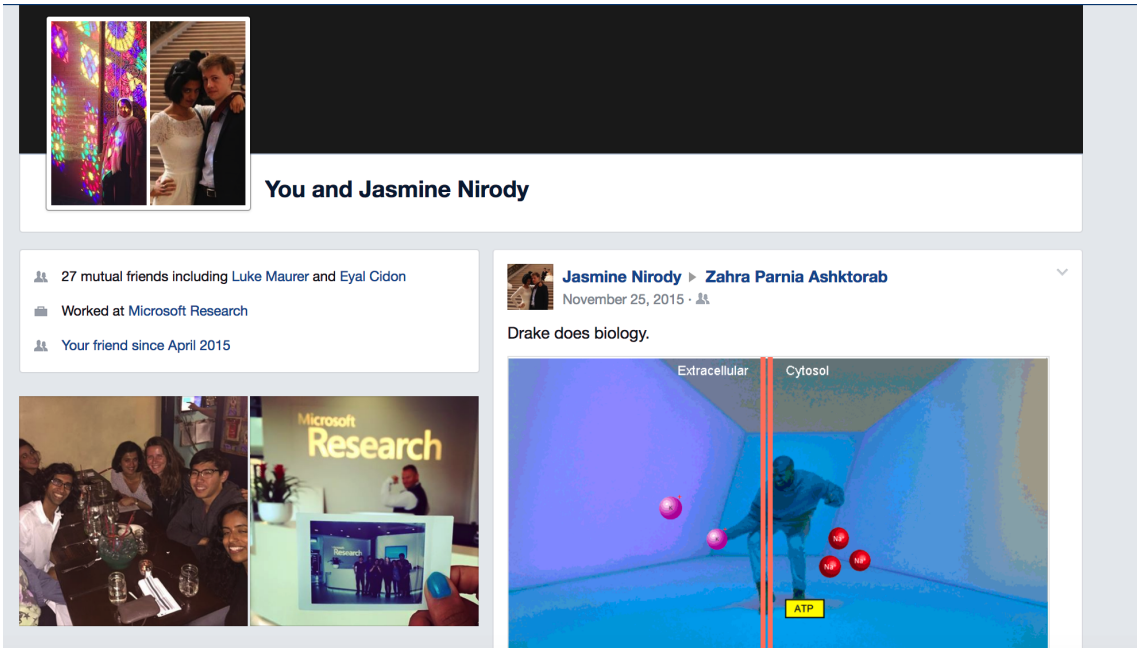


Figure 6.1: Facebook "See Friendship" Option

6.3.1 Reminiscence

Positive reminiscence has been used in various contexts to boost feelings of well-being and happiness. Bryant et al. demonstrate that positive reminiscence through cognitive imagery can boost happiness [49] among young adults. Some theorists stipulate that positive reminiscence can be used as an adaptive coping response [52] that provides comfort and reduces negative affect [94] for older adults. Habermas et al. conduct a study that found that the use of memorabilia to aid reminiscence and promote adjustment to college life in university students elicited feelings of "fun and enjoyment" [116]. Pasupathi et al. investigate social reminiscence (participating in positive reminiscence with another individual) and found that enhanced positive emotions and was a compelling emotional management strategy [203].

Furthermore, with the introduction of technology in our daily lives, people can move beyond memorabilia and use technology-mediated-memory (TMM) to capture moments in their lives and revisit it at a later time [131]. A study on TMM, found that technology does not disrupt adaptive biases and edits which often occur in memories to complement human well-being. The study found that technology mediated memory in fact helps promote sustained well-being [144].

Reminiscence helps maintain and strengthen relationships and helps individuals to make sense of one's own identity by considering the past [162] [270]. Peesapati et al. built *Pensieve*, an application that emailed memory triggers of content from individuals' social media platforms and found that sending memory triggers to individuals improved participants' moods [206]. Furthermore, rosy retrospection is the notion that individuals remember an event more fondly than their actual experience of the event. Mitchell et al. describe the "rosy view" phenomenon, that negative thoughts during an event caused by distractions and disappointments dissipate days after the event leaving much more positive memories of the actual event [182]. Reminiscence intervention has been used to to treat elderly adults in order to increase life satisfaction. Studies have found that reminiscing about positive life experiences lead to an increase in life satisfaction [60].

Participants in a study on reminiscence conducted by Cosley et al. reported that the experience of reminiscence was spontaneous and beneficial; that they would be triggered by external causes to think about positive memories. Participants in this study also reported that people were a central focus of reminiscing and that the people in their lives often triggered reminiscence. The nature of reminiscence

also varies based on the nature of the relationship with the people who are central to the reminiscence process. Physical mementos were also triggers in the reminiscence process. The individuals in the study enjoyed being prompted to reminisce because they enjoyed the process. The participants also liked memory triggers to be randomly selected [63]. In this study, I took into consideration this feedback on the reminiscence process when building the CTMM.

6.3.2 Determining Tie Strength

The strength of a tie between two individuals is "a combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services with characterize the tie" [111]. Granovetter characterized two types of ties: strong and weak. *Strong ties* are people whose social circles overlap with your own and you highly trust while *weak ties* can be defined as mere acquaintances [102,111]. Weak ties provide access to new information that is not existent in the network of strong ties. Young, highly educated and metropolitan individuals have a diverse network of strong ties [168]. Social support provided by strong ties can improve mental health [232]. Many dimensions have aimed to define tie strength: recency of communication [163], communication reciprocity [99], and interaction frequency [111].

Gilbert et al. present seven dimensions of predicting tie strength: *Intimacy* (days since last communication, intimacy words, number of friends), *Intensity* (wall words exchanged, outbound posts, inbox thread depth), *Duration* (days since first communication), *Social Distance* (educational differences, political differences, occu-

pational differences), *Services* (links shared, applications shared), *Emotional Support* (positive emotion words in inbox, positive emotion words on wall), and *Structural* (mutual strength, interest overlap, common groups) [102].

Bapne et al. conducted an exploratory study to understand how social ties are linked to economic measure of trust. They analyze three tie measures: 1) interactions between friends on a Facebook "wall" (now referred to as a Timeline) 2) number of mutual friends and 3) being tagged in a photo, which would be indicative of a real-world interaction. Bapne et al. found that for users with a larger amount of Facebook friends, the only measure that was associated with trust was whether users were tagged together in a photo. For users who had less Facebook friends all three measures were correlated with degree of trust. [22]. In another study exploring the relationship between strong social ties and online interactions on Facebook [134], found that the value and confidence interval for appearing in the same photo (photo tags) along with comments, messages, wall posts, pokes, family members, and same-sex friendship were all positive and significant.

In this study I am interested in promoting well being through the reminiscence of memories with strong ties. For this reason, I use the heuristic of capturing individuals through which the individual appears in photos, which is indicative of a real-life relationship. Previous studies [22, 134] demonstrate that appearing in photos together on Facebook is indicative of a stronger social tie and the presence of more trust between individuals.

6.4 Methods

In this section, I describe the methods I used in this study. A visualization of the flow of methods used in this study can be seen in 6.2. I begin by describing recruitment of participants, a description of the survey instruments administered, how users were selected to participate in the longitudinal study, and how check-ins were administered throughout the study.

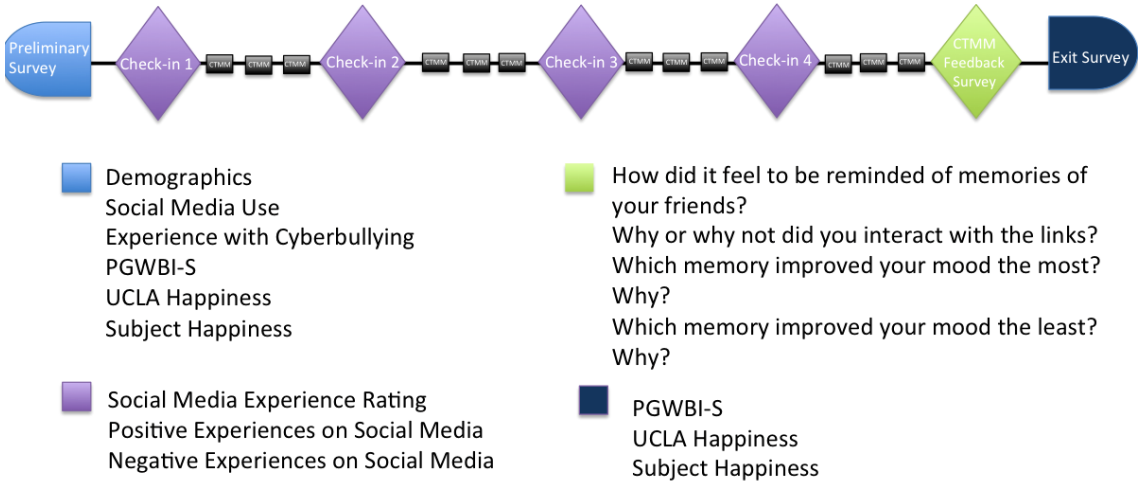


Figure 6.2: Pipeline of the various steps in this longitudinal study.

6.4.1 Participants

A random sample of 3000 incoming freshmen was obtained from the university's Registrar's Office and were sent an email invitation to participate in the study [204]. The recruitment email described the study and included a link to the consent form and survey, hosted on SurveyGizmo. The survey was short (completion time of approximately 8-10 minutes) and asked participants a number of

questions about their use of social media, positive and negative experiences online, and their perceived well-being and access to resources. At the conclusion of the survey, participants were invited to enter their email address to be considered for a four-week follow-up study with the ability to make up to $10.00 in compensation. Participants were also able to enter their email address to be entered into a drawing for one of five $20.00 Amazon gift cards.

In total, 200 participants completed the preliminary survey. However, 29 participants completed the entire study (initial survey, four check-ins, final survey), 19 of whom identified as female and 10 of whom identified as male. Participants were told that the study would address their technology use, social support, and well-being.

The sample is made up of freshmen college (18-19 year olds) on Facebook in the United States. While it is true that studying 18-19 year olds is certainly convenient for logistical purposes (no parental assent required for the Internal Review Board), the motivations for choosing this sample is supported by previous studies. A study across age groups found that both adolescents (12-19 years) and young adults (20-26 years) were more often targets of online aggressive behavior compared to older respondents [240]. Another study surveyed 799 college students and found that 8.6% of college students had reported that they had experienced cyberbullying and those that did reported scoring higher on depression, anxiety, phobic anxiety, and paranoia [234]. Arnett discusses*emerging adulthood*, a period between the ages of 18-25 which is distinct from both adolescence and young adulthood in that it is characterized by independence from social roles and normative expectations. Emerging adulthood

is a key period for identity exploration [14]. Furthermore, the typical American emerging adult is engaged with some form of media at least 12 hours a day [15]. These studies demonstrate that not only does the demographic of the selected sample experience cyberbullying, but they are seriously affected by behavioral symptoms when they do.

Table 6.1: Descriptive statistics for full sample (N=200)

Item	Mean
Gender	
Male	36.6%
Female	63.0%
Other	0.4%
Race	
Caucasion or White	51.2%
African American or Black	12.3%
Asian	23.1%
Hispanic or Latino	5.0%
Multiracial	6.5%
I don't want to disclose	1.9%
Social Media Use	
Ask.fm	18.1%
Formspring	7.7%
Facebook	99.6%
Twitter	68.0%
LinkedIn	17.4%
Google Plus	19.7%
Personal blog	7.3%
Instagram	84.6%
Tumblr	38.6%
Snapchat	91.9%
Whatsapp	25.9%
Tinder	20.1%
Yik Yak	27.0%
Vine	32.4%
Whisper	6.9%
Others	13.1%
Experienced Cyberbullying	
Never	52.3%
Rarely	33.7%
Sometimes	13.0%
Often	1.0%

Following the collection of the preliminary survey data, criterion sampling

was employed to identify participants who met the minimum criteria for inclusion in the longitudinal study and who had indicated their interest in participating in the longitudinal study. This criteria included regular use of social media, regular use of Facebook specifically and at least some reported negative experiences online. This latter criterion was included to ensure the potential for data collection during the longitudinal study. Students who met the minimum criteria (N=98) were sent a second email inviting them to participant in the longitudinal study. The email included a detailed overview of the study components, directions on how to access to instruments used in the longitudinal study, and access to a link to confirm their participation.

6.4.2 Measures

In this section, I describe the measures collected throughout this study. These measures aim to evaluate perceived well-being, perceived social support and mood throughout the longitudinal study.

6.4.2.1 Platform-specific experiences

In the pre-test survey I asked about the frequency of use of different social media platforms using a 10-point sliding scale ranging from 1-10 (Less Than Once a Week -Multiple Times Per Hour). Participants were asked a series of questions about negative experiences for each social media platform they used, such as "How frequently have you had interactions or seen content on [site] that made you upset or uncomfortable?" Participants could respond using a 10-point slider scale ranging

from “Never” (value=1) to “Very Often” (value=10).

6.4.2.2 Measuring Perceived Well-being

A central construct to this study is participants perceived well-being and assessing whether interventions induce positive changes in participants' well-being over time. Three validated and frequently used scales were included in the study: Psychological General Well-being index (PGWBI), the UCLA Loneliness Scale, and Subjective Happiness Scale (SHS). Researchers have long been interested in measuring individual happiness and well-being [13, 44, 50]. Happiness includes factors about individual's judgements about their personal well-being. Multi-item scales of well being are more reliable than single-item indicators. Well-being measures often revolve around positive and negative affect and life satisfaction. Well-being can be operationalized into six dimensions: self-acceptance, positive relations with others, autonomy, environmental mastery, purpose in life, and environmental growth. An individual's sense of positive feelings towards oneself is a central theme of positive psychological well-being. The existence of warm, trusting interpersonal relationships in one's life are also a reflection of an individual's well-being. Environmental mastery, or the ability to mold environments to suit an individual's psychic conditions is an important facet of well-being. An individual's sense of purpose and attitude towards the meaning of life influences his/her overall sense of well-being. Personal growth, or a person's ability to to be open to new experiences, adapt and grow as a result is also a central tenant of well-being [226]. In this study, I use multi-item scales that capture the facets (autonomy, environmental mastery, purpose in life,

and personal growth growth) of percieved well-being.

Below, I describe the different scales employed in this study to measure well-being.

1. *Psychological General Well-being Index (PGWBI-S)* : The Psychological General Well-being index is a 22-item evaluation of perceived well-being. In this study, I used a shorter shorter six-item validated version of the questionnaire (PGWBI-S), which is made up of the following items, "Have you been bothered by nervousness or your 'nerves' during the past month?" and "How much energy, pep, or vitality did you have or feel during the past month?" Each item was ranked a scale of 1-5 based on whether the item measured positive affect or negative affect. The sum of the scores for each item yielded the PGWBI-S score [113].

2. *UCLA Loneliness Scale*: The UCLA Loneliness scale is a 20-item scale designed to measure feelings of loneliness and social isolation. Items include the following: "I am unhappy doing so many things alone", "I have nobody to talk to", and "I cannot tolerate being so alone". Participants rated each item on a scale from 0-3: "I often feel this way" (3), "I sometimes feel this way"(2), "I rarely feel this way" (1), and "I never feel this way" (0) [225].

3. *Subjective Happiness Scale (SHS)* The Subjective happiness scale is a four-item scale designed to measure subjective happiness. Each item asks a participant to finish a sentence clause along a seven-point scale. The items consist of: "In general, I consider myself" (not a very happy person- a very happy person) and

"Compared to most of my peers, I consider myself" (less happy-more happy).

6.4.3 Facebook Application and Data Collected

To collect data to curate the CTMM that would later be administered during the longitudinal study, I used the Facebook Developer platform to build an application that collects data from users timeline, friend list, and tagged photos. The data collected from the application identifies ways to present mitigation techniques to promote well-being. A long-lived access token is created so that data can be collected throughout the entire period of the longitudinal study. I collected all timeline content and associated meta data: like count/reaction count of content, date content was published, any associated attachments (photos, url, etc.), privacy status, and associated comments and sub-comments.

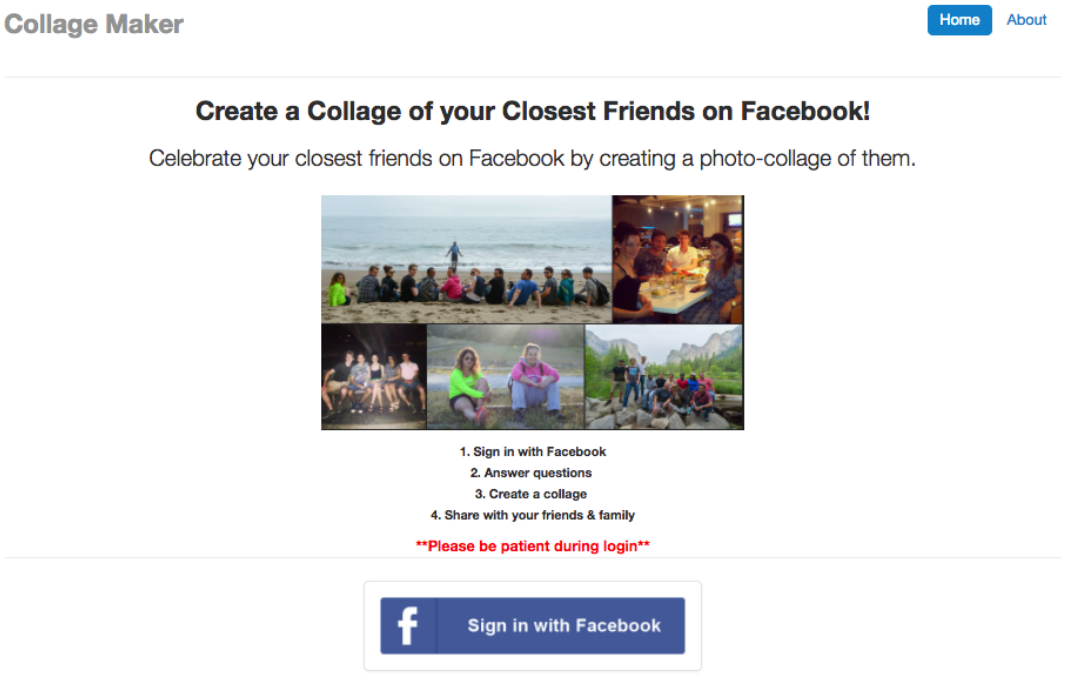


Figure 6.3: Initial Application for Data Collection

In order to publish an application to Facebook, a application developer must justify the collection of every data point by describing how collecting that data would ultimately contribute to a better user experience for the Facebook user. For this reason, the initial Facebook application was used to collect data created an initial collage of *all* friends (not a CTMM) whom users indicated they 1) can really count on to be dependable when they need help 2) can really count on to get help to feel more relaxed when they are under pressure or tense 3) accept them totally, including both their worst and best points 4) can really count on to care about them, regardless of what is happening to them, 5) can you really count on to help you feel better when they are feeling generally down-in-the-dumps and 6) can count on to console them when they are very upset. These heuristics are based on social support questions [228]. Photos of friends were selected based on the responses and a collage was created to share on Facebook. While my goal was to collect data in the initial phase of the study to curate CTMM later in the study, the initial Facebook application's goal was to enhance user experience to abide by the ToS of application publication on Facebook. For this reason, a collage was offered to participants that they could share on their personal Facebook pages.

6.4.4 Experience Sampling with Weekly Check-ins

Experience sampling is a research procedure that involves asking participants to self-report during various times of a study. Experience sampling captures both private and public aspects of a participants' life and reduces cognitive load by allowing participants to report as events occur and thus rely less on memory [153]. The

Figure 6.4: Example of initial collage shared on Facebook Application

Experience-Sampling Method (ESM) provides a valid instrument to help participants self-report various mental processes in every day situations [67]. It has been used to collect information about the pattern of and changing of thoughts [171], location, social interactions [122], and emotional cognitive dimensions [286]. Csikszentmihalyi provides data and empirical support for why ESM is a reliable and valid instrument for assessing these variables [67].

In order to capture online experiences, I used experience sampling [153] in the form of weekly check-ins and asked open-ended questions about respondents' experiences that may have angered them or hurt their feelings, including, "Describe a moment/event this week when something on the social media platforms that you use that upset you." Conversely, participants were asked to "Describe a moment/event this week when something on the social media platforms that you use that cheered you up and improved your mood". Participants were also asked to rate their social media experience on a scale of "Very Unpleasant" to "Very Pleasant".

6.4.5 Feedback on Curated Technology Mediated Memory

At the conclusion of the study, I surveyed participants about their perceptions of CTMM, specifically why or why not they chose to interact with the weekly links. The questions I asked at the completion of the study were: "Throughout the course of this study, you were emailed memories (pictures and comments from your Facebook friends). How did it feel to be reminded of memories of your friends?", "You received weekly reminder emails about memories with friends. In the weekly reminder emails you received about memories with friends, why or why not did you interact with the links?", "How did being reminded of the memories?"

6.4.6 Timeline

Every week following the pre-test, participants were sent three curated technology mediated memory messages via email. They were asked to check-in weekly and report any possible negative interactions or experiences they have had on social media and how it affected their overall well-being. Each participant received four check-in requests during the first, second, third, and fourth week of the study respectively. In the check-in, participants were asked about their experiences with social media during that particular week. At the completion of the study, participants were asked to take a post-test survey that included the same measures of well-being as those included in the pre-test.

6.5 Results

The data collected throughout this study includes both open-ended qualitative data (questions about cyberbullying experiences, perceptions of curated technology mediated memory) and quantitative measures (perceived well-being collected at the start and completion of experience sampling and weekly social media experiences).

6.5.1 Open-ended questions: Cyberbullying Types, Feedback on CTMM, and Weekly Experiences

For the open-ended questions, two researchers independently reviewed the corpus and created a set of codes to apply to the corpus. After three iterative rounds of comparison, all researchers agreed on a final codebook for the different data sets. With the exception of analysis of types of cyberbullying, all responses were approached with the inductive approach. The responses were read several times to identify themes and categories by both researchers. After an initial discussion, codes were discussed to identify a coding frame. If new codes emerged throughout the analysis, a discussion would occur between the researchers, and codes were added to the existing coding frame. Below, I discuss the emerging codes for each of the open ended questions we asked throughout the study [256].

6.5.1.1 Cyberbullying Examples

For the open-ended question in the preliminary survey which asked respondents about the types of cyberbullying they have experienced, I along with another

Table 6.2: Cyberbullying Types

Cyberbullying Type	Definition	Example
Flaming	Sending angry, rude, vulgar messages	*Some douche told me to kms (kill myself)*
Harassment	Repeatedly sending offensive messages	*MySpace and Formspring were the go-to's. So, I was in 7th grade getting TONS of people bullying me, calling me names, calling me rude things, anonymously of course.*
Cyberstalking	Repeatedly sending threats of harm or highly intimidating messages	*In middle school i had a girl email me telling me everyone hated me and was going to come to school and beat me up the next day it was horrific*
Denigration	Posting untrue or cruel statements	*On yik yak, there were rude rumors spread about several people and there was no way for the victim to respond and defend themselves*
Impersonation	Pretending to be someone else to make that person or place in danger	*My classmates have made a fake Facebook account for my close friend and demonized him (i.e. ridiculed him) by posting content that does not even come close to what represents my friend.*
Outing and Trickery	Posting material that contains, sensitive, private information about another person or forwarding private message	*My friends ex sent her nudes into a group chat after the broke up.*
Exclusion	Intentionally excluding a person from an online group	*I have a friend who has been excluded from posts about friendships.*

researcher labeled each response with the type of cyberbullying (based on Willard et. al [275]) and the social media platform on which the cyberbullying occurred. The results are presented in the sections below. In the preliminary survey, I asked users to detail a time in which they or a friend experienced cyberbullying. A total of 52 users responded. Of the respondents, 60% detailed a time they had personally experienced cyberbullying, 19% percent detailed a time when their friends had experienced cyberbullying, and 21% said they had not been cyberbullied. The cyberbullying examples were coded by myself and one other researcher based on the existent types (flaming, harassment, denigration, cyberstalking, outing and trickery, impersonation, and exclusion) of cyberbullying by Willard et al. [275]. The types and definitions are in Table 6.2. Additionally, items were labeled with more granular predominant themes. These codes can be seen in Table 6.3.

The most prominent type of cyberbullying was harassment (46%), followed by flaming (29%), outing (10%), impersonation (10%) and exclusion (5%). One user described a particular disastrous experience, a part of which I have detailed below:

I was on Formspring when I was young, about 12-13. Big thing in [hometown] at the time, especially my middle school. MySpace and Formspring were the Go-to's. So, I was in 7th grade getting TONS of people bullying me, calling me names, calling me rude things, anonymously of course.

6.5.1.2 Granular Cyberbullying Sub-themes

While I identified the seven different types of cyberbullying defined by [275], we also looked for other themes in the instances of cyberbullying. Furthermore, the

granular cyberbullying themes address **RQ1** revealing the state of cyberbullying more in depth. I identified three reoccurring themes: 1) Self-harm and Suicide, 2) Anonymity, and 3) Online harassment resulting from the termination of a romantic relationship.

Table 6.3: Granular Cyberbullying Themes

Code	Definition	Example
Self Harm and Suicide	Cyberbullying that explicitly instructed victims to harm or kill themselves	*Some douche told me to kms [kill myself]*
Anonymity	The cyberbully was anonymous	*Uncomfortable message sent by unknown person*
Former Relationship	Cyberbullying perpetrated by someone whom victim had a romantic relationship with	*My friends ex sent her nudes into a group chat after the broke up. I confronted him but he didn't change.*

Self Harm and Suicide The participants described instances where either they or someone they knew were told to kill themselves. This particular theme re-emerged multiple times. A female participant said, "Often times it's very difficult to validate what's considered *cyberbullying.* Somebody from my school once tweeted a very racist and horrible post regarding African Americans and slavery. The tweet ended up going semi-viral, and he was getting tweets from every direction saying to *kill himself* or criticizing his appearance in his icon photo. Those people were obviously defending the more ethical side to the debate, but it can be argued that their reactions were just as bad."

A male participant said, "Some douche told me to kms". In this case, kms is abbreviation for *kill myself.* The very fact that this abbreviation is used is an indicator that this type of message and term is reoccurring.

Anonymity Another reoccurring theme among users who described instances where they experienced online harassment or cyberbullying was anonymity. One participant described a time she received an "Uncomfortable message sent by unknown person". The anonymous cyberbullying occurred on social media accounts that afforded anonymous interactions (Formspring, askFM, Yikyak).

One participant noted that "Hurtful ask.fms" was an example of a time they had experienced cyberbullying. ASKfm is a platform that allows anonymous communication. One participant described a time where she was cyberbullied anonymously, "It was a long time ago, but when I was a freshman in high school someone posted on my ask.fm asking why I was such a snob. I stopped using it after that." Another participant noted that she had experienced, "sexual harassment on anonymous platforms such as ask.fm (mostly inappropriate questions/comments)". One participant wrote, "I have experienced it on Yik Yak at my high school." All of these participants reported that they had experienced cyberbullying on platforms that allow anonymous communication.

Dissolution of Romantic Relationships The dissolution of romantic relationships was another theme I identified among the instances of cyberbullying referenced in the preliminary survey. One participant wrote, "My friends ex sent her nudes into a group chat after the broke up. I confronted him but he didn't change." In this instance, a shared digital artifact is being shared without the consent of the victim. "Revenge Porn" is the non-consensual publication of sexually graphic images and can lead to emotional harm and even financial repercussions [59]. Another

participant wrote, "My ex boyfriend was being very rude to me after we broke up." While it is unclear on which platform the "rude" behavior was occurring, this quote is another example of negative experiences in an online space due to the dissolution of romantic relationship.

6.5.1.3 Weekly Check-ins

Similarly to open-ended questions in the preliminary survey, two researchers agreed on two codebooks for the open-ended questions in the weekly check-ins for positive and negative social media experiences respectively. Two researchers labeled the weekly-checkin corpus with the 9-factor code book for negative social media experiences and a 6-factor code book for positive social media experiences. We discussed the codes until we reached a consensus.

I launched four weekly check-ins throughout the course of this study. The very first one was sent before any technology mediated memory messages were sent. The next three check-ins were sent weekly during the course of the study. During the study, participants were receiving three technology mediated pings weekly. Every week, I asked participants open-ended questions about both the negative and positive experiences they had that week on social media, SMS, or other modes of communication/media. The 9-factor coding scheme for the negative experiences that respondents experienced weekly emerged is visible in Table 6.5. The 6-factor coding scheme for the positive experiences that respondents experienced weekly emerged is visible in Table 6.6.

The response rate for the weekly check-ins can be seen in Figure 6.5. As

expected, participation decreased at each step of the study (pre-test, check-in 1, check-in 2, check-in 3, check-in 4, post-test). Because of the payment incentive for the final response, there was an increase in responses for the final post-test survey.

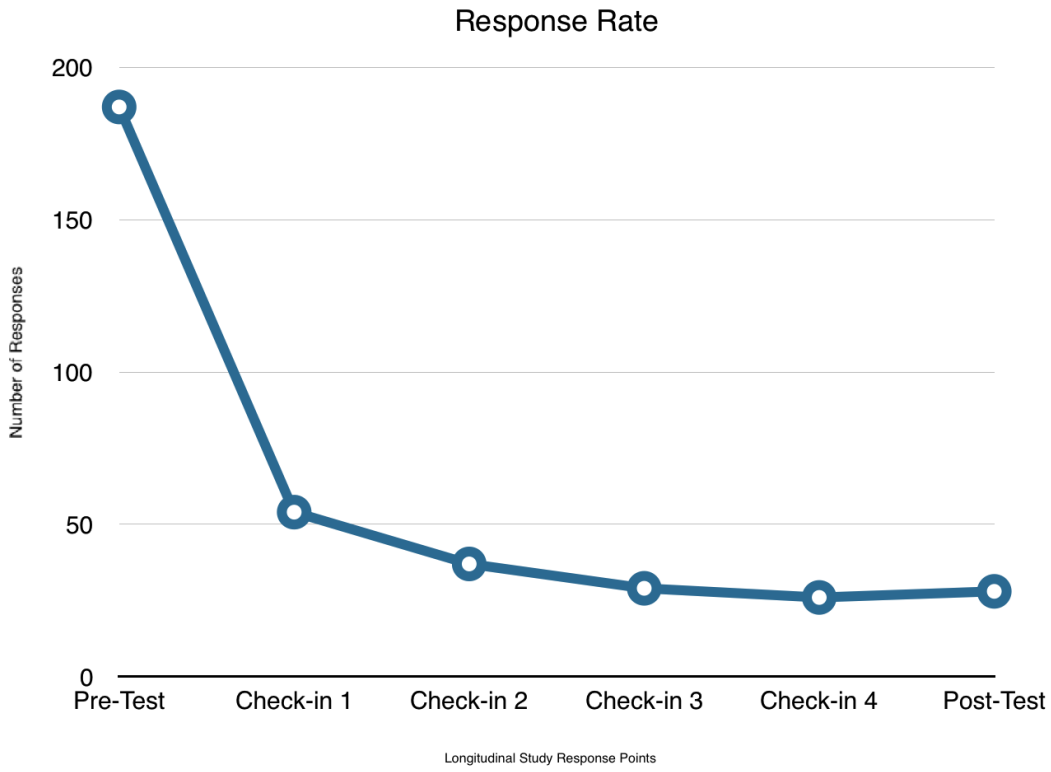


Figure 6.5: Response rate at each point of data collection

6.5.1.4 Exit Survey

At the end of the study, I asked participants open-ended questions probing into their experiences and perceptions of the CTMM they received throughout the study. The exit interview aimed to address **RQ2**, whether CTMM are successful in mitigating negative emotional effects of cyberbullying and other negative experiences on social media. The exit survey was administered through a online survey. For each

Table 6.4: Emergent themes from qualitative responses regarding participants' perception of CTMM

Code	Definition	Example
Genuine Interest	Recipients were genuinely interested in seeing and interacting with the photos and content about the subject of the ping	*I clicked on the links because I was interested in finding out what memories were chosen for me to look at.*
Not Close to Subject	Recipient was not close to subject of ping	*I did not interact with some of the links because I was never good friends with that person, and there was only a few memories that involved other, closer friends as well.*
Busy	Recipient was busy with school	*I usually did not because I was busy with schoolwork.*
Location Dependent	Interacting with the ping was location dependent	*Really just depended on where i was when i opened the email or if i felt like getting into anything at that time.*
Spam	Links looked like spam	*I kinda thought it was spam*
Repetitive	Content was repetitive and seemed too similar to other social media features like "See Friendship" on Facebook	*I kinda thought it was spam*
Too much Information	Some respondents felt eery about the amount of information that was presented to them in the ping	*I was a little creeped out that I was being emailed compilations of photos.*

Table 6.5: Emergent themes from qualitative responses regarding participants' weekly negative experiences on social media

Code	Definition	Example
Election-Related Content	Discussions, arguments or content that is election related that bother the respondent	*There is still residual anger and discontent about the elections and it still bothers me that people keep attaching their opponents with words rather than working it out with a level head*
Fights with Friends	Arguments with friends that are not election related	*Fight with friends through iMessage*
Current Events (Catastrophes)	Current events that made respondents feel helpless. Syrian Civil war [151] and Ohio shooting [12] were mentioned during the course of this study	*I saw a video about the Syrian civil war and I felt so helpless. It's not that I think there should be less coverage it just made me sad*
Social Media Envy	Instances in which respondents felt envious of others' lives as they appeared on social media or felt insecure as a result of showing off their own lives on social media	*Felt bad about my body and life by comparing them to other peoples'.*
Disenfranchised Joy	Instances where users felt they could not share positive experiences online due to insecurities or other factors	*I had a glamorous weekend this week, and as much as I wanted to show it off, showing it off also made me feel insecure, so it was a double edged sword*
Racist and Sexist Posts	Complaints about content that is sexist or racist.	*Racism, sexism. the usual*
Personal	Personal events that spilled onto social media	*Finding out my friend went missing*
Fake News	Complaints about the influx of fake news on social media and irreputable sources	*The constant flow of fake or stupid news coming from sources that once used to reputable is kind of sad to see.*
Exclusionary Behavior	Interactions that made the respondent feel excluded.	*Seeing snapchat stories of people and events that I was not invited to*

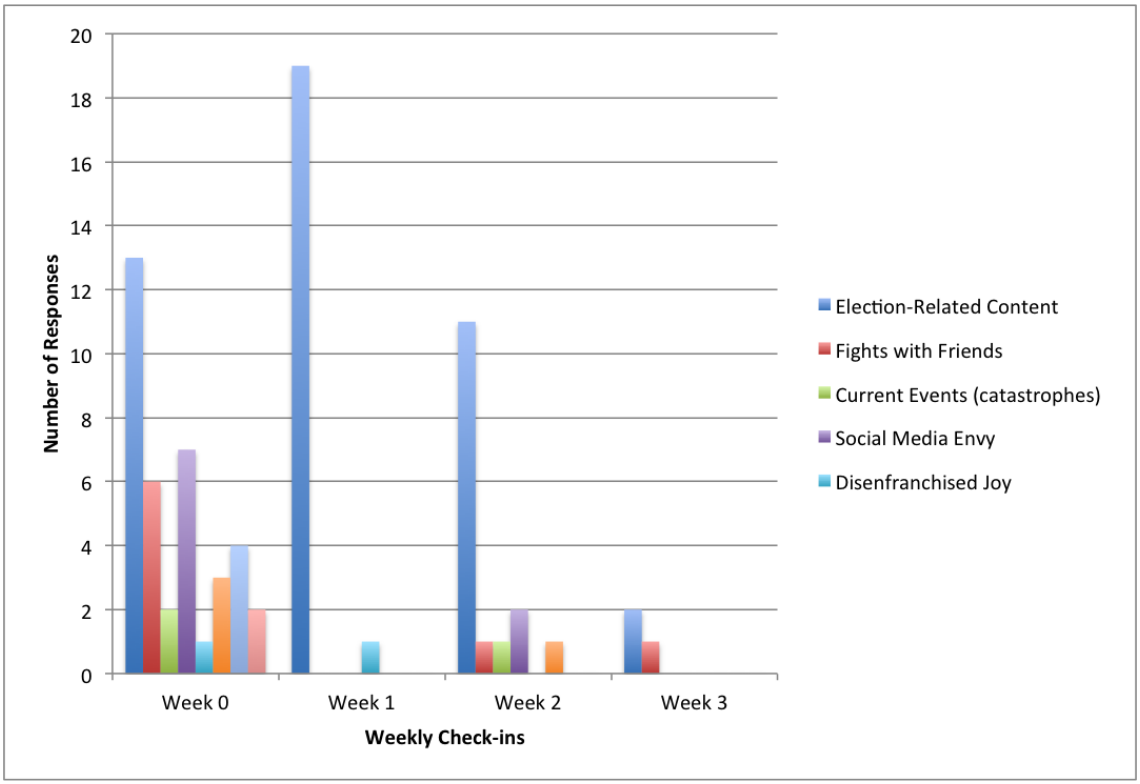


Figure 6.6: Negative Emergent Themes Across Weeks

Table 6.6: Emergent themes from qualitative responses regarding participants' weekly positive experiences on social media

Code	Definition	Example
Election-Related Content	Discussions or content that is election related that delight the respondent	*Other people supported me in my fight against a Trump supporter.*
Entertainment	Content (Videos, Memes, etc), that entertains the respondent	*Funny videos helped me destress*
Connecting with Friends	Connecting with friends through social media	*I got to see family and friends who I love and miss, I got to know they are doing well and their smiles made me smile.*
Inspiration	Content that inspires respondent to achieve their goals	*Got inspired to do great things and do well in school to achieve those things*
Receiving Support	Support in the form of text or Paralinguistic Digital Affordances (PDA) [120] through likes and other reactions	*The whole "like my status" and I'll say something nice about you. Those are cute and friendly. Also comments on Instagram of like "wow you're so beautiful!" etc.*
Getting Informed	Getting informed through news and other content	*Learning about the election live helped me as a journalism major*

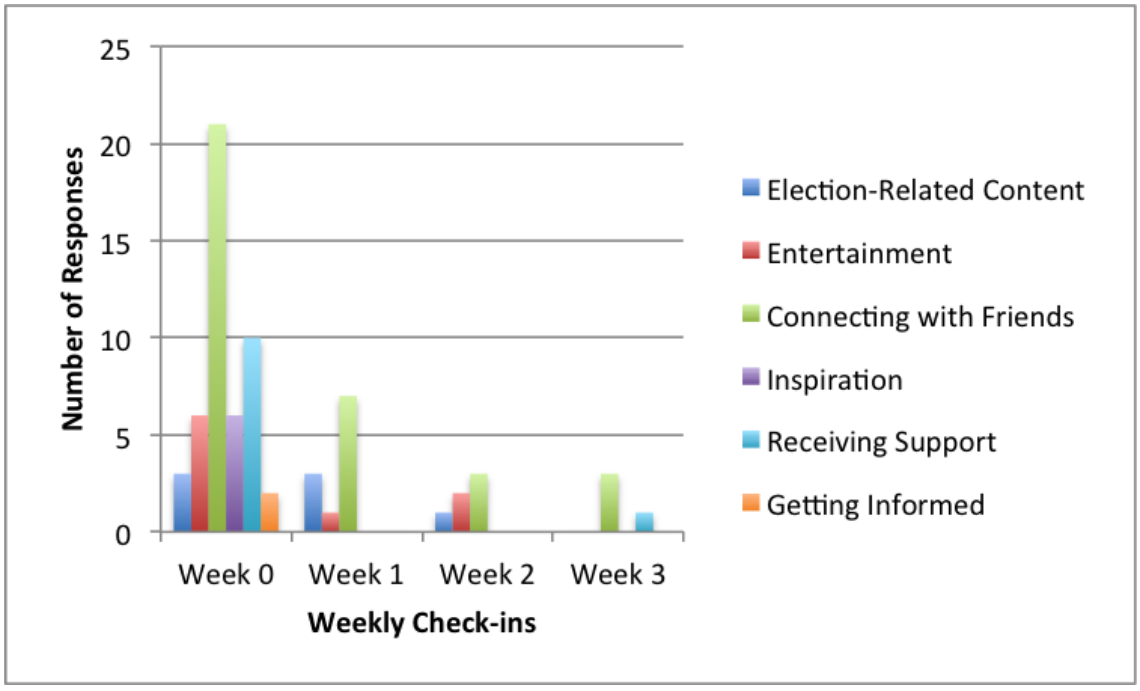


Figure 6.7: Positive Emergent Themes Across Weeks

question, I used the same coding analysis method detailed for the other open-ended questions in which two researchers (myself included) agreed on a codebook for the responses to each question. Two researchers labeled the corpus with the code book iteratively until all labels were agreed upon. For the first question about thoughts towards CTMM, the researchers also coded for valence towards the pings (positive, negative, neutral). Below, I have listed main findings.

Positive Reactions Towards CTMM 52% of participants reacted positively to the curated technology mediated memory pings. The primary reasons for positive reactions to the curated technology-mediated pings was: **Good memories improved users' moods and distracted from negative current events**. All participants who reacted positively noted this reason for reacting positively. Respondents re-

ported that they liked interacting with photos if **they were genuinely interested in seeing and interacting with the photos and content about the subject of the curated technology mediated memory ping**. A CTMM cheered a user up if the subject was **close to the participant** and was **associated with good memories**.

Negative Reactions Towards CTMM 14% of the participants who participated in the exit survey reacted negatively to CTMM. Respondents felt that **the CTMM were unnecessary since the pictures and content can be found on Facebook** and felt that the content of the CTMM was redundant. Furthermore, if **subjects were no longer close with the subject of the CTMM** they would have a negative reaction to the CTMM. All participants who had a neutral or a mixed reaction to the curated-technology-mediated-memory pings (34%) noted that **the memories were "bittersweet" if the person in the photos was someone they missed and had not seen in a long time**. These respondents described feeling nostalgic and both sad and happy simultaneously.

Reasons participants reported for not interacting with the CTMM were: **if they were not close or good friend with the subject of the curated technology mediated memory ping** (33%), **if they were busy with school work**(11%). Interacting with the links was also **location dependent** and dependent on where they were when they received the email. 11% of the respondents to this question **likened the links to spam** or were **creeped out by the amount information captured in the links** (11%). Participants noted that a CTMM did not cheer them

up if the subject was **not close or no longer close to the participant**. This theme dominated all of the responses when participants were reflecting on CTMM that had little to no effect on their moods.

6.5.2 Curated Technology Mediated Memory, Well Being, Loneliness, and Happiness

Three validated and frequently used scales were used to measure well-being in this study: Psychological General Well-being index (PGWBI), the UCLA Loneliness Scale, and Subjective Happiness Scale (SHS). Below, I describe the results of the statistical analyses conducted on the measures collected before CTMM were administered and after CTMM were administered.

6.5.2.1 Comparing General Well-Being, Happiness, and Loneliness Before and After Study

To test the hypothesis that this cyberbullying mitigation tool would improve perceptions of well-being (UCLA-Loneliness Scale, PGWBI-S, and Subjective Happiness Scale (SHS), I conducted a Mann-Whitney U test [175] to compare the difference between measures before entering the study and measures after entering the study. Also known as the Wilcoxon rank sum test, the Mann-Whitney U test tests for differences between two groups on a single, ordinal variable with no specific distribution. To address **RQ2**, all of the variables for the different items in the different measures (PGWBI-S, UCLA Loneliness Scale, and the Subjective Happiness Scale) were compared for before and after the completion of the study.

The statistical analysis for the Psychological General Well Being Scale shows that the total PGWBI-S score was statistically significant when comparing well-being before launch of study and after completion of CTMM administration M=14.07, SD=1.80 v. M=15.27, SD= 2.04, α = 0.78, p=0.04. There were no statistically significant changes for the Subjective Happiness Scale or the UCLA Loneliness scale. One reason for these non-significant findings may be that a larger sample size was required to reflect significant findings for Subjective Happiness and UCLA Loneliness Score. Furthermore, it can be argued that CTMM effects are in the moment (at the receipt of the CTMM) and have less of a long-term effect. The weekly measures more accurately measure the immediate CTMM effects of positive social media experiences and significant improvements were discovered between Week 1 and Week 2 in the weekly check-ins (discussed in the next section).

6.5.2.2 Weekly check-ins: Social Media Experiences

For the weekly check-ins, I used a one-way ANOVA test to compare overall Social Media Experiences across each week [61]. Social Media Experiences improved from Week 1 and Week 2 M=5.78, SD=2.13 v. M=7.03, SD=1.67, p= 0.001. For Weeks 1,2,3 and 4, participants reported Social Media Experiences improving as well M=6.66, SD=0.77 v. M=6.0, SD=1.87 v. M=6.33, SD=1.17 v. M=7.14, SD=1.25, p=0.04. Comparing all four weeks in the study, participants reported an improvement of Social Media Experiences.

6.6 Discussion

In this work, I use a human-centered approach to mitigate the effects of cyberbullying and other negative experiences on social media. In this preliminary study, I demonstrate that positive memories, when curated and delivered to social media users, have the potential to improve well-being. Such improvement can be applied to improve well-being during times when individuals might be affected by cyberbullying and are feeling isolated and alone. This study is the first of it's kind to test the effect of a cyberbullying mitigation tool on users over a long period of time. This study provides a preliminary approach to resolve negative emotions that are associated with cyberbullying by pushing positive memories to users. The results of this study demonstrated that participants in the study indicated that their social media experiences improved between the four weeks in the longitudinal study. There was no statistically difference for comparison of social media experiences across other weeks. In the section below, I discuss further in detail the implications of the study. At every step during the administration of CTMM, design decisions were made (from their design to their curation). The feedback received by participants can aid into bettering both design and curation of CTMM to be more effective.

6.6.1 Captured Online Experiences

The first research question in this study, **RQ1**, inquired about the various experiences of social media users. Users reported on weekly positive and negative experiences throughout this study (while they were receiving CTMM). The results

reveal that beyond cyberbullying, many of the content and events on social media that contributes to negative weekly experiences are highly contextual and often based on current events.

6.6.1.1 Contextual Factors: Elections and Bullying

The data collected in this study was collected during the 2016 presidential election and thus contextual factors influenced the outcome of the results. Specifically, election-specific themes emerged when participants were asked about both weekly positive and negative experiences. A national survey of 50,000 teens reports a surge in abusive online behavior since the 2016 election. The survey linked race, sexual orientation, and immigration status to bullying and social marginalization [257]. In the weekly feedback collected, elections emerged as a part of both negative weekly experiences and positive experiences for participants. In this particular election, participants reported the negative experiences being tied to "hate": *I saw a lot of hate related to the election and it really brought my mood down thinking about how some people can never accept others.* and "racism" *nothing too bad ive been working out so tiny things dont get to me seeing reactions and racism in response to Trump.*

Others discussed feeling overwhelmed by the amount of Trump-related or election related content on their newsfeeds, *The amount of buzz around Trump getting elected. All the talk and controversy on the election has provided a very negative experience for me.*, *The election is a mess, and everything on social media is only about the election.*

Negative Election Related Experiences From the negative social media experiences reported during weekly-check-ins, a large portion of users reported that content posted about the election was their most negative social media experience. Some election-related content was reported to even escalate to online arguments. In some cases, participants described that election-related elections led to different types of cyberbullying, including harassment and exclusionary behavior (defriending on social media).

One participant reported that there had been an increase in bigotry and hatred as result of the recent elections, "I've observed a lot of arguments and uneducated comments from Trump supporters on social media, a lot of bigotry and hatred, a lot of aggression."

Another participant reported being sad and worried about the future as a result of the interactions witnessed on social media following the elections "After the election of Donald Trump to the presidency of the United States, many people took to social media to express their hatred or destress at his election. People, as they have been the entire election cycle, were rude and belligerent and violent and it made me sad and scared and worried about the future of our country and the world."

Other participants reported that their friends had deactivated or left social media as a result of the escalated debates they witnessed following the elections, "My cousin deleted her Facebook profile due to debate on the election."

Positive Election Related Experiences While 36% of the reported negative weekly feedback was related to the elections, the elections appeared as a theme in the weekly positive social media experience feedback as well. Election-related positive experiences were related to Facebook friends supporting one political faction or group. One participant reported, *I have seen some of my friends who are white stand up for those in minority groups. It was really nice to see.*

Another participant felt positively after receiving support against an individual supporting an opposing political party, *Other people supported me in my fight against a Trump supporter.* Some participant reported that they felt a sense of unity and togetherness because of the election results, *People bonding together over the election.* Overall, 6% of the positive weekly experiences were related to the 2016 election.

Other Contextual Factors: Traumatic Events One theme that emerged among the weekly negative experiences reported by participants were current events and catastrophes that made users feel helpless. During the time of data collection, two instances of such events were videos and content surround the Syrian civil war and the Ohio shooting.

A traumatic event typically involves exposure to "death, threatened death, actual or threatened serious injury, or actual or threatened sexual violence" [108]. Traumatic events can provoke fear or horror, feelings of helplessness, and have both short and long term psychological and physiological consequences [264]. Affected communities of mass violence events differ on two factors: directness of threat to

loved ones and geographic proximity [38]. Mass violence events affect communities more acutely than natural disasters or oil spills [194].

Participants in this study reported that viewing content about traumatic experiences were the most negative part of their social media experiences that week. One participant reported, "I saw a video about the Syrian civil war and I felt so helpless. It's not that I think there should be less coverage it just made me sad."

Another participant reported that the most negative social media experience of her week was that she "Learned about the attacker in Ohio".

While prior research on such disasters has used standard psychological instruments to measure the subjective stress caused by traumatic events, [97, 194], these instruments exhibit limitations: they are completed by respondents retrospectively and not immediately after the traumatic event occurs; there is no pre-event symptomology baseline available, and respondent numbers are commonly are not high. The vast amount of social media content available during traumatic events can serve to complement these existing such instruments because social media content does not encompass such limitations [104].

6.6.1.2 Disenfranchised Joy

Additional themes discovered in the data, like *Disenfranchised Joy* address **RQ1** by revealing the kinds of positive and negative experiences participants in the study were having on social media. One theme that emerged when I asked about participants' negative weekly experiences was the notion of *Disenfranchised Joy*, that a user wanted to share the positive things that happened to him/her on social

media but felt that followers would not be receptive to it. One participant said, "I had a glamorous weekend this week, and as much as I wanted to show it off, showing it off also made me feel insecure, so it was a double edged sword". Having a collection of this users' Facebook posts reveals that she indeed opted not to share anything about her weekend. Later, I discuss how users can leverage CTMM to alleviate *Disenfranchised Joy*.

Disenfranchised Joy can be described as the withholding of publishing positive content on social media over fear that followers and friends would feel that one is lacking humility or trying to "show off". Many studies have demonstrate that sharing content and receiving positive feedback on published content on social media has many benefits, including the reaping of social capital [91] and maintaining and strengthening online relationships [266]. Brandtzaeg et al. explore the relationship between content sharing sociability and usage behavior and privacy experience on Social Network platforms. They report that having too many friends on Facebook and access to different groups of people (social capital) can disrupt sharing patterns due to social surveillance and social control. Young people may use conformity to preserve their privacy [45]. While the intent of the CTMM in this study are to promote well-being, they are also reminding users of the existing social support in users' social networks, a reminder that may in turn ameliorate or alleviate "Disenfranchised Joy" by reminding individuals of the close ties within their network and the shared positive memories. By reminding users of close friends in their networks, CTMM may even help users to manage their privacy settings in a manner so that they would be more comfortable to share content with their peers and reap the

benefits of sharing and publishing content on social networks.

Previous research [26] draws a connection between users' personality traits, specifically extraversion and narcissism and their sharing practices. The decision to share information about one's private life is a form of exhibitionism [21]. Individuals who are overtly narcissistic are more likely to be concerned with the attention of others' and admiration. Studies demonstrate that beyond personality, the public vs. private nature of a platform as well as directed vs. undirected forms of communication lead to different strategic goals for sharing. Social validation and self-expression make up a primary reason for why users share publicly, while relational development is the primary goal of sharing and exchanges in private channels. While the affordances of social media platforms and personality traits may explain why users choose to self-disclose on social media platforms, the *Disenfranchised Joy* theme discovered in this work presents a greater depth to the struggle of those users who yearn the validation of sharing and self-disclosure on social media but their fear of sharing. Exposure to highly idealized versions of individuals and peers leads to envy and the distorted belief that these individualized lead happier lives [147]. Such distorted beliefs have the potential to lead to feeling inferior over time [247]. *Social Media Envy* was another negative theme that emerged through our study. Individuals who experience *Social Media Envy* in the past and may be reluctant to inflict similar feelings to their peers, thus experiencing *Disenfranchised Joy*. While the aim of the CTMM in this study are to mitigate the negative social media experiences (like *Disenfranchised Joy*), further research can be done to unpack this struggle and provide impression-management design recommendations to make the

self-disclosure and sharing process for such users less difficult so that such users can benefit from the social gratification of online self-disclosure practices.

6.6.1.3 Social Media Envy

"Social Media Envy" describes envy by social media peers. This theme addresses **RQ1** and emerged from the weekly check-ins that probed into participants' weekly negative experiences. Aksoy et al. discuss the implications of their research; that some social media experiences lead to jealousy and envy [7], which in turn can lead to ramifications in offline behavior [34]. Muise et al. reveal that social media affordances like revealing romantic partners to ambiguous information that they otherwise would not have access to leads to jealousy and suspicion in romantic relationships [185].

Dissatisfaction with bodies is not limited to women. Previous studies have revealed that exposure to muscular media images caused men to feel dissatisfied with their bodies [100]. Perlof et al. describe ways to potentially ameliorate dysfunctional habits of disordered eating and influence beliefs and attitudes. Media interventions and persuasive regimen can positively influence appearance based perceptions [160, 208]. Perloff et al. note knowledge bias as a potentially effective starting point for social media campaigns. Knowledge bias is when someone is seen through more credible eyes due to their background religion, age or other factors [89]. Perlof et al. stipulate that knowledge bias in this domain, or leveraging women who have experienced dysfunctional eating habits or unhealthy appearance perceptions to lead social media campaigns.

6.6.1.4 Receiving Support through Likes and Like Solicitation

In **RQ1**, I was also interested in discovering the positive experiences of participants. When prompted about positive social media experiences, participants described *Receiving Support* (Support in the form of text or Paralinguistic Digital Affordances [120]) as a positive type of interaction that they experienced on Facebook. One participant noted about the positive experience "The whole "like my status" and I'll say something nice about you." This type of interaction emerged in positive weekly experiences as well with one participant reporting, "I got to post some stuff on Instagram and I got a lot of likes and compliments." Another participant reported, "A lot of people liked a selfie I posted." These results demonstrate that small "low-cost" [51] interactions like "liking" can contribute to positive experiences on social media.

6.6.2 CTMM and Mitigation of Cyberbullying and Other Negative Experiences

When reporting feedback on CTMM, users reported that the "good memories" helped to distract from the negative events they noted in the check-ins (election results, shootings, and Syria). Furthermore, the repeated ANOVA measures for social media experiences revealed an increase in social media experiences across all four weeks throughout the study ($p < 0.04$). Additionally, there was an increase in the PGWBI-S score for before and after the study ($p < 0.05$).

6.6.2.1 The Benefits and Drawbacks of CTMM

RQ2 probes into whether CTMM can indeed promote well-being and mood. Much of the feedback in the exit survey revealed that participants were happy to receive and interact with CTMM if they felt *close* to the subject of the CTMM. In other words, if positive memories were triggered by the CTMM, users enjoyed the experience of receiving and interacting with it. However, users also noted no changes in mood if the subject of the CTMM were individuals with whom they felt no particular kinship towards. Furthermore, if the dynamic of the relationship with the subject of the CTMM had changed and the participant *no longer* felt close to the subject of the CTMM, participants reported mixed feelings, melancholy nostalgia, or feeling "bittersweet". These results demonstrate that Curated Technology Mediated Memory, if curated correctly, has the ability to improve well-being in the face of online negative experiences.

The importance of effective curation was reflected in the results. The CTMM must trigger positive memories in order to be effective. In fact, if CTMM trigger negative memories they might have the opposite effect of their intent. While sentiment analysis was used to curate the memories, the limitations of sentiment analysis contribute to some of the drawbacks of CTMM in the way that they were implemented in this study. While social network interactions can give cues and insight into the nature of relationships between individuals, sentiment analysis may not be able to accurately detect that two individuals are no longer close to one another and reminding them of their past friendship might even lead to "bittersweet" or even

negative feelings. For this reason, CTMM must be carefully curated in order to be effective. In the next section, I address **RQ3** and make design recommendations for the curation and delivery of the CTMM.

6.6.2.2 How to Improve Curated Technology Mediated Memory

The results of the study evaluating the effectiveness of CTMM demonstrates that reminders of one's past positive memories have the potential to improve social media user experience. **RQ3** addresses how the design and curation of Curated Technology Mediated Memory be optimized to promote well-being for victims of cyberbullying? The results from this study can help inform better features for a cyberbullying mitigation system. In the following sections, I make the following design recommendations for CTMM: 1) Use multiple sources for curation of CTMM 2) Identify alternative curation strategies, and 3) Consider alternative modes of delivery.

Use Multiple Sources for Curation of CTMM *S*ome of the feedback from the participants in the study described in this study revealed that participants felt that the CTMM were redundant with respect to the "Facebook See Friendship" option. Participants identified that one reason that they did not interact with content was because they felt that it was redundant. They felt that the content in the CTMM was unnecessary since the pictures and content can be found on Facebook in the "See Friendship" functionality. Increasingly, visualization tools like "See Friendship" are giving users the opportunity to reflect on their friendships. Sosik et al. present design considerations for different types of reflection for the "See Friend-

ship" functionality on Facebook [238]. Their results demonstrate that encouraging recall of events through various imagery helps trigger the same emotions felt at the time that the memory was formed. While I note two key differences in CTMM when compared to "See Friendship" (use of sentiment analysis to curate as well as direct delivery), the feedback from the participants indicates that diversification of content would be more effective. Participants in the study indicated that they use multiple social media platforms. Future iterations of CTMM could include content from other social media platforms (Instagram, Snapchat, etc.) so that it does not appear to be redundant to users. Furthermore, studies reveal that different social media platforms serve different needs for users. For example, one person may choose to interact with only close ties on a social media platform like Snapchat [278], but interact more widely through low-cost interactions like "liking" statuses of weak ties on Facebook to ultimately invest in the existent social capital [51].

Utilize Alternative Curation Strategies. When providing feedback on CTMM, some participants reported that they were unaffected by this mode of *Tertiary Prevention* since they were no longer "close" with the subject of the CTMM. Furthermore, upon providing feedback about why participants interacted with the links containing the CTMM sent to them, some lamented that they did not interact with the link if they were no longer close to the subject (indicated in the subject of the email). When identifying the CTMM that cheered them up the least, participants identified the CTMM whose subject was no longer close to the participant. Conversely, users identified the CTMM that cheered them up the most whose subject was still close to the participant. These results exhibit that varying heuristics

need to be considered to curate CTMM that will actually promote a users' well-being. When curating CTMM, I used the heuristic of being co-tagged in a photo to determine closeness of a relationship. The premise behind this design decision was to present the user with the potential social support existent in their feeds. There are alternative methods of finding strong ties to presenting participants with social support memories.

While the CTMM are curated to revolve around the memories associated with a particular person to remind a participant of their existing social support, in future iterations and work, CTMM can focus on positive memories through a different lens. Perhaps highlighting a particular person in the CTMM is less effective than highlighting positive memories from many friends overall. Previous research reveals that posting on Facebook and receiving feedback can decrease an individual's sense of loneliness [76]. The loneliness is decreased as a result of friends' interactions (likes and comments) with content [76]. Coupled with sentiment analysis to avoid reminding a user of a negative past event, memories can be curated to focus on important events that received relatively higher number of responses (likes and comments) from friends. Conversely, curating CTMM and other modes of *Tertiary Prevention* with posts which may not have a high number of responses from friends but still hold importance to a user could potentially boost mood. Exploring alternative curation strategies for CTMM and other *Tertiary Preventions* can ultimately lead to the best curation and intervention strategies for promoting well-being after a user has experienced cyberbullying.

Consider Alternative Modes of Delivery. I asked users how they felt

about interacting with the links in the survey and the mode of administration. Participants gave various responses as to why they did not interact with the links. Reasons include: being busy with school (33%), location dependent (11%), thought the links were spam (11%), or creeped out by the amount of information captured in the links (11%). This data gives insight into ways cyberbullying mitigation pings can be delivered to users. Pings were delivered via email. Some recommendation for *Tertiary Prevention* tools that deliver content to users include:

1. Many users indicated that whether they checked email depended on where they were when they opened the email. Using location-services like GPS to determine location of user before sending pings.

2. Participants in the study reported that they likened the pings to spam. The pings were sent from the email: which may appear to spam. This calls into question the medium I used to deliver pings in the study and that perhaps delivery through an application would be deemed as more trustworthy by recipients.

6.6.3 Limitations

I acknowledge that the results in this study are limited by the sample size. A larger sample would yield more sufficient power. This work is the first of its kind to evaluate a cyberbullying mitigation tool. Despite the small sample size, the qualitative results compliment the findings in my analysis of well-being comparison before and after the launch of CTMM. I acknowledge that a larger sample size would

contribute to a better understanding of the affect CTMM have on individuals who have and continue to experience cyberbullying.

6.6.3.1 Historical Factors

Furthermore, the data in this study was collected during the historical 2016 elections. This historical event could have very well affected the outcome of well-being measures and social media experiences. To control for the historical election, a control group could have been done with a pre/post test comparison to understand whether CTMM specifically influenced well-being. However, given that the the contributions of this study are design-based, and the data collection methodology is a contribution to practice, a control-group would have strengthened the statistical results, but not expanded the overall design contributions of this study.

6.6.3.2 Broad Definition of Cyberbullying

The research questions addressed in this study address cyberbullying and other negative experiences. As in Chapter 5, I define cyberbullying incidents through Willard et. al's framework (flaming, cyberstalking, denigration, outing and trickery, exclusion, harassment, and impersonation) [275]. The preliminary survey at the onset of this study revealed that at least one participant in the sample had experienced each cyberbullying type. However, throughout the course of the study, negative experiences (escalation of political arguments, disenfranchised joy, social media envy) on social media were addressed as a part of the mitigation objective as well as the observed cyberbullying scenarios. As a researcher, I can not guarantee

that my subjects would experience a specific type of cyberbullying throughout this study nor did I wish this upon them. Throughout the experience sampling, I merely observed their experiences and administered CTMM throughout the process. The results of these experiments would be much stronger if the sample had consistently experienced severe/repeated forms of cyberbullying and online harassment and this is a logistical and ethical challenge of cyberbullying mitigation research. However, participants were chosen to participate in the study if they had experienced cyberbullying in the past and many of the experiences throughout the study are defined as cyberbullying scenarios (racism and bigotry, exclusionary behavior). Thus, the implications of this study are inclusive of negative experiences but these negative experiences tie in directly with cyberbullying experiences and in many cases are not mutually exclusive.

6.6.4 Thinking Ahead: Prevention, Mitigation and Beyond

The work in this chapter signals a new area of work *evaluating* the effectiveness of cyberbullying mitigation tools. It is important to note that CTMM address cyberbullying *after* it has already occurred, and that the types of cyberbullying are vast, both across platforms and intent. CTMM focuses on mitigating cyberbullying *after* it has already occurred. However, across the studies presented in this dissertation, I discussed many types of cyberbullying mitigation techniques that aim to *prevent* cyberbullying before it occurs. For example, users mentioned the spread of private photos without consent after the dissolution of a romantic relationship. Changes to the design of existing social media platforms can be considered to pre-

vent these types of cyberbullying before they occur. In the next chapter, I introduce cyberbullying events through the *Continuum of Harm* framework, which considers online interactions before cyberbullying events occurs, once they occur, and if they reoccur repeatedly. Three different types of prevention are introduced, *Primary Prevention*, to prevent cyberbullying before it occurs, *Secondary Prevention*, mitigation of cyberbullying once it occurs, and *Tertiary Prevention*, mitigation of cyberbullying once it is reoccurring. CTMM is framed as an example of *Tertiary Prevention*, a mitigation method aimed at improving well-being and perception of social support once cyberbullying incidents negatively effect the victim. In the next chapter, the *Continuum of Harm* framework is discussed more in detail.

Chapter 7: Conclusion

7.1 Chapter Summary

In Chapters 3-6, I discussed a series of iterative studies that explored some of the key cyberbullying challenges adolescents and emerging adults face. In this final chapter, I use these findings to present a series of cyberbullying mitigation solutions for designers of social media and other interactive platforms to consider as they continue to address the bullying, harassment, and other negative content that plagues the Internet. The solutions presented throughout these studies are diverse and consider mitigation at various points of the "Continuum of Harm" of cyberbullying. I consider cyberbullying mitigation solutions through the "Continuum of Harm" framework, which consists of three types of prevention mechanisms: 1) **Primary Prevention**, in which the cyberbullying incident is prevented before it starts; 2) **Secondary Prevention**, where the goal is to decrease the problem after it has been identified, and 3) **Tertiary Prevention**, when intervention occurs after a problem has already caused harm. This chapter discusses the design of technological mechanisms to mitigate cyberbullying through Primary, Secondary, and Tertiary prevention based on the findings of the studies in this dissertation. I conclude this chapter with a discussion of potential future research directions in the

area of automatic detection and evaluation of cyberbullying mitigation tools.

7.2 Iterative Succession of Cyberbullying Mitigation Studies

Each chapter in this study built on the efforts and findings from the previous study. I began with a data-centric approach, attempting to find the motivation of the use of a social media platform associated with cyberbullying and ultimately building a classifier to detect it and various types of discourse. I evaluated the effect of *cyberbully-reversal pings* and found that sending victims of cyberbullying positive messages has the potential to mitigate cyberbullying. I then conducted participatory design which teens that ultimately led to a suite of cyberbullying mitigation prototypes. I finally evaluated CTMM as a cyberbullying mitigation tool. From each study, I gained an understanding of human behavior in the context of cyberbullying and introduce design recommendations for various cyberbullying mitigation tools. Cyberbullying is an all-encompassing term inclusive of various types of malicious interactions on various social media platforms. Some of the mitigation tools introduced in this thesis involve the prevention of cyberbullying - preventing aggressions before they even occur, while others focus on mitigating and promoting well being once cyberbullying occurs. In the next section, I present the findings of the studies in the previous chapters through the "Continuum of Harm" framework for cyberbullying, a framework that considers the various stages of cyberbullying aggressions and prevention and mitigation during these different stages.

7.3 The Stages of Cyberbullying and it's Prevention: Continuum of Harm

Cyberbullying has a variety of negative influences on the emotional and physical well-being of victims. Cyberbullying has the potential to affect individuals at all times of day regardless of an individual's location. Victims of cyberbullying are at higher risk for depression, anxiety, and suicidal ideation [65]. Similarly, victims of cyberbullying have faced psychosomatic symptoms like sleeplessness and abdominal pain [249]. Victims of cyberbullying are also more likely to be involved in anti-social behaviors like alcohol consumption and drug-misuse [125]. While there has been no in-depth longitudinal study of the relationships between the different symptoms of cyberbullying, literature informs us that emotional disturbance precedes behavioral and psychosomatic symptoms [5, 54].

Similar to cyberbullying, domestic violence involves different stages by which the victim is affected. Prevention mechanisms have been outlined very clearly in domestic violence prevention. Wolfe et al. identify a health model which can be used to identify opportunities for domestic violence prevention along a "Continuum of Harm" [277]. At one side of the domestic violence "Continuum of Harm" lies gender-focused jokes and vulgarity, while physical force and rape lie at the other end of the spectrum. Similarly, I model a "Continuum of Harm" specific to cyberbullying. The continuum is triggered by cyberbullying. At one end of the continuum, there is damage to self-esteem, while suicidal ideation lie at the other end of the

continuum. I use this continuum to evaluate the best point of entry for each of the intervention mechanisms we discover in our design sessions. For domestic violence, the "Continuum of harm" prevention includes a three-pronged approach: **Primary Prevention**, **Secondary Prevention**, **Tertiary Prevention**. Each of the design recommendations in the next section can be classified into one of these approaches with respect to the Cyberbullying "Continuum of Harm".

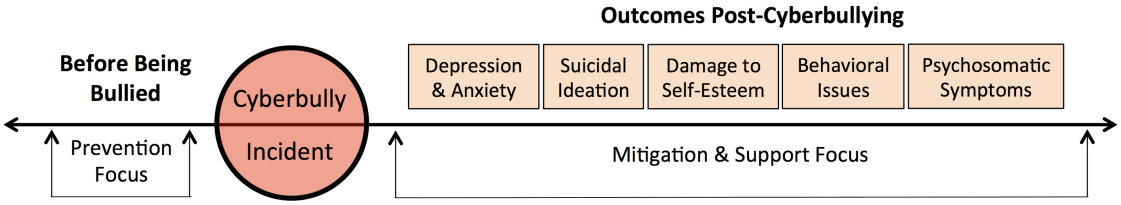


Figure 7.1: The Cyberbullying Continuum of Harm describes the different types of emotional distress may follow cyberbullying.

Cyberbullying mitigation solutions can be analyzed through a framework that considers the different stages of cyberbullying symptoms and is based on preventative measures aimed at mitigating the "Continuum of Harm" in domestic violence [277] (see 7.1). Through this framework, the technological solutions resulting from the design sessions as well as design recommendations derived from the longitudinal study in Chapter 6 can be categorized through a three-pronged approach: 1) **Primary Prevention**, in which the cyberbullying incident is prevented before it starts; 2) **Secondary Prevention**, where the goal is to decrease the problem after it has been identified, and 3) **Tertiary Prevention**, when intervention occurs after a problem has already caused harm [277]. Two researchers who were involved in designing the participatory design sessions coded each resulting solution based on this

framework individually. Below we describe the solutions on which both researchers agreed regarding the prevention category in which they fell.

7.3.1 Primary Prevention

The goal of primary prevention is to stop cyberbullying before it happens. In the traditional non-technical realm, such a solution would include school-based programs that warn of the harms of cyberbullying (e.g., [27]). However, "Exclusion Prevention" application, designed in the Participatory Design sessions to prevent purposeful repeated cropping on an individual on Instagram, is a strong example of a type of a cyberbullying preventive measure that aims to stop the incidence of cyberbullying before it can occur. While literature has discussed how to approach denigration and flaming [79], there are no academic research discussion issues related to exclusion online. In a *New York Times* parenting blog, the author said, "To be in a photo and to not be tagged is to be rendered socially invisible. Commenting on a party photo, my untagged daughter wrote, 'I was there too!"' [180]. The "Exclusion Prevention" application aims to remedy the potential emotional damage of exclusion-based cyberbullying by presenting the potential bully with a reflective notification.

The main aim of primary prevention is to raise awareness about the potential harm that could be caused as a result of someone's actions. In "Exclusion Prevention", the bully decides whether she wants to continue with publishing data after the system warns the [potential] bully that she may be hurting someone by continuously cropping them out of photos. Ultimately, the decision of publishing the content lies

with the potential bully. Dinakar et al. [79] share examples of primary preventive measures when discussing reflective interfaces, which ask users to reflect on their behavior before publishing malicious content online.

From an implementation standpoint, preventative prevention requires some degree of monitoring since it is attempting to prevent the cyberbullying before it occurs. While privacy advocates may find this monitoring particularly troubling, many parents believe that they have the right to access and monitor their children's online activity [24]. There are three notions of a reflective practitioner: "reflection in action", "reflection on action" and "ladders of reflections" [237]. Reflective user interfaces aim to prevent cyberbullying by asking the aggressor to reconsider their actions and reflect on them through showing potential consequences of their actions, flagging their content and notifying them of the potential harm they can cause. Since the main goal of Primary Prevention is to prevent the cyberbullying narrative from taking place, and the initiator of the cyberbullying narrative is the perpetrator, the perpetrator holds primary control over initiating the bullying after being presented with primary preventative measures.

7.3.1.1 Mitigation for Exclusionary Behavior

One common type of cyberbullying described by Willard et al. is "Exclusion", the act of purposefully excluding individuals through social media. Many respondents described instances of such behavior contributing to experiencing a negative social media experience for the week. This type of behavior occurred through various mediums: Snapchat and exclusionary Facebook posts. One participant reported,

"Good friends always post exclusive pictures and it's annoying but not a big deal." In this particular case, the exclusion is occurring through photographs. Another participant described the exclusionary behavior as transcending beyond the online realm and translating into the offline realm, "seeing snapchat stories of people and events that I was not invited to". Sometimes the exclusionary behavior revealed a level of deception in which the subject was deceived about friends attending particular event and later on discovered the truth through social media content, "Friend said they weren't going somewhere and they went."

One type of "negative social media experience" that continuously reappeared in the weekly check-ins was related to exclusionary cyberbullying, cyberbullying that involves purposeful directed exclusion of individuals [177]. This cyberbullying manifested itself through different formats and on various platforms. Participants in the CTMM study reported being excluded on Facebook statuses, and purposefully deceived and uninvited to events later discovered on Snapchat. While other types of cyberbullying are more discerning and can be more inflammatory, exclusionary cyberbullying is less obvious and thus creating mitigation tools to address this kind of cyberbullying are more challenging. In the prototype solutions in the participatory design study in Chapter 5, clique-detection was suggested for victims of cyberbullying that were repeatedly cropped out of photos. The participants in the participatory design study suggested creating automated methods to detect exclusionary cyberbullying and present reflective interfaces to facilitate perpetrators of exclusionary behavior to rethink their actions.

7.3.1.2 Escalation of Political Discourse

The data collected during the longitudinal study evaluating the effectiveness of CTMM was collected during the 2016 Presidential elections [8]. For this reason, many of the negative experiences reported during the weekly were escalations of political debates into racism or sexism. Individuals reported that the most negative aspects of their weekly experiences were, "I saw a lot of hate related to the election and it really brought my mood down thinking about how some people can never accept others.". Another participant reported, "seeing reactions and racism in response to Trump" as a part of the the most negative experiences on the weekly checkin. In primary prevention, reflective interfaces can be used during such events to prevent the escalation of racism, bigotry from political differences. With a combination of classification techniques and reflective interfaces, users can be asked to reconsider in engagement of posts that promote bigotry and racism.

Some participants in the CTMM study (Chapter 6) deactivated their Facebook posts due to online fights and debates. One participant said, "My cousin deleted her Facebook due to debate over politics". Another participant retorted that the online debates and fights were leading to friends "unfriending one another on Facebook, "This week with the trump situation, there has been a lot of negativity about it and my friends unfollowing other friends for being trump supporters. While I wasn't involved in these online feuds, I was still indirectly affected." As part of primary prevention, reflective interfaces can *prevent* online fights [79] and ultimate dissolution of friendships before they start by prompting users to reconsider participating

in a futile political debate that may escalate. Figure 7.2 is an example *primary prevention*, a reflective interface prompting an individual to rethink posting a comment on a polarizing Facebook post. Furthermore, the reflective interface encourages the user to "hide" the post. Such a step would furthermore prevent the dissolution of online friendships since it encourages the user to "Hide Post" instead of unfriending or the act of terminating a digital friendship can lead to both negative emotional and cognitive consequences [34].

7.3.1.3 Cyberbullying and the Dissolution of Romantic Relationships and Contextual Integrity

Many of the participants in the CTMM study (Chapter 6) reported cyberbullying or online harassment occurring as result of the dissolution of a romantic relationship. Sas et al. make recommendations for managing the process of managing digital possessions after dissolving a romantic relationship: creating digital spaces for shared possessions, artifact crafting as sense making, incorporating tools for self control and harvesting digital possessions [230]. In the types of cyberbullying listed in the preliminary survey, many participants in the CTMM study reported experiencing cyberbullying as a result of the dissolution of a relationship. One participant wrote, "There were a lot of subtweets about me after I broke up with my most recent relationship." Another participant said, "My friends ex sent her nudes into a group chat after the broke up. I confronted him but he didn't change."

Sas et. al recommend creating shared spaces for digital possessions. For example, a *relationship profile* would allow a couple to both celebrate a relationship

Figure 7.2: Example of Reflective interface to prevent escalation of politically polarized Facebook posts that may lead to contentious arguments.

and in the case of dissolution, delete content that would be painful to reflect on after the end of the relationship. While this method might be useful in aiding to forget a memory, such a design does not protect against more nefarious and malicious interactions that may occur after the dissolution of a romantic relationship, like the widespread publishing of content like explicit photos that were not meant for specific private audiences. In some instances, this type of sharing content is referred to as "Revenge Porn". "Revenge Porn" constitutes a violation of sexual privacy and involves the publication of non-consensual graphic images. The publication of such images can lead to emotional harms and even increase the risk of physical assault [59].

One design recommendation which constitutes as Primary Prevention for social media platforms is adopting ephemeral interactions to prevent such interactions. Beyond ephemerality, social media platforms should opt to *notify* users if photos were screenshot on ephemeral technologies. Teens have turned to ephemeral communication on social media technologies [56]. Snapchat and Instagram allow the sharing of temporary photos and videos. Recently, iMessage and Facebook messages rolled out temporary messages [143].

In these particular scenarios, *screenshot* detection serves as *primary prevention*, mitigation of potential cyberbullying. Snapchat allows users to circumvent automatic deletion of content by allowing screenshot-ing. However, users are notified if their photos have been screenshot [278]. Xu et al. describe emerging norms on ephemeral communications that allow saving with a notification. To screenshot content that the poster would not like to be distributed would be violating the norms

Figure 7.3: Example of current notifications for Snapchat replays and screenshots

of snapchat. Nissenbaum describes the framework of contextual integrity. "Distribution" which refers to the movement of information depends on three factors: "actors (subject, sender, recipient), attributes (types of information), and transmission principles (constraints under which information flows) [192]. Since the default norm in ephemeral communications is ephemeral, Snapchatting and saving information is a violation of that norm. For this reason, social media platforms must notify users when Contextual Integrity is violated.

Screenshot detection and notification technologies in ephemeral communications serve as primary prevention of cyberbullying by notifying a user that their photo has been captured permanently by the recipient. Coupled with a reflective interface to reiterate that the once-ephemeral content has now been saved by the recipient, such a notification would make users reconsider sending explicit photos to a particular recipient who has violated the Contextual Integrity of the ephemeral communication platform, which might ultimately prevent any widespread publishing of such photos and further damage. Figure 7.3 demonstrates screenshot notifications in various ephemeral platforms. Figure 7.4 demonstrates a potential screenshot notification design for "secret" Facebook conversations.

Beyond screenshot notification, reputations can be assigned for those who vi-

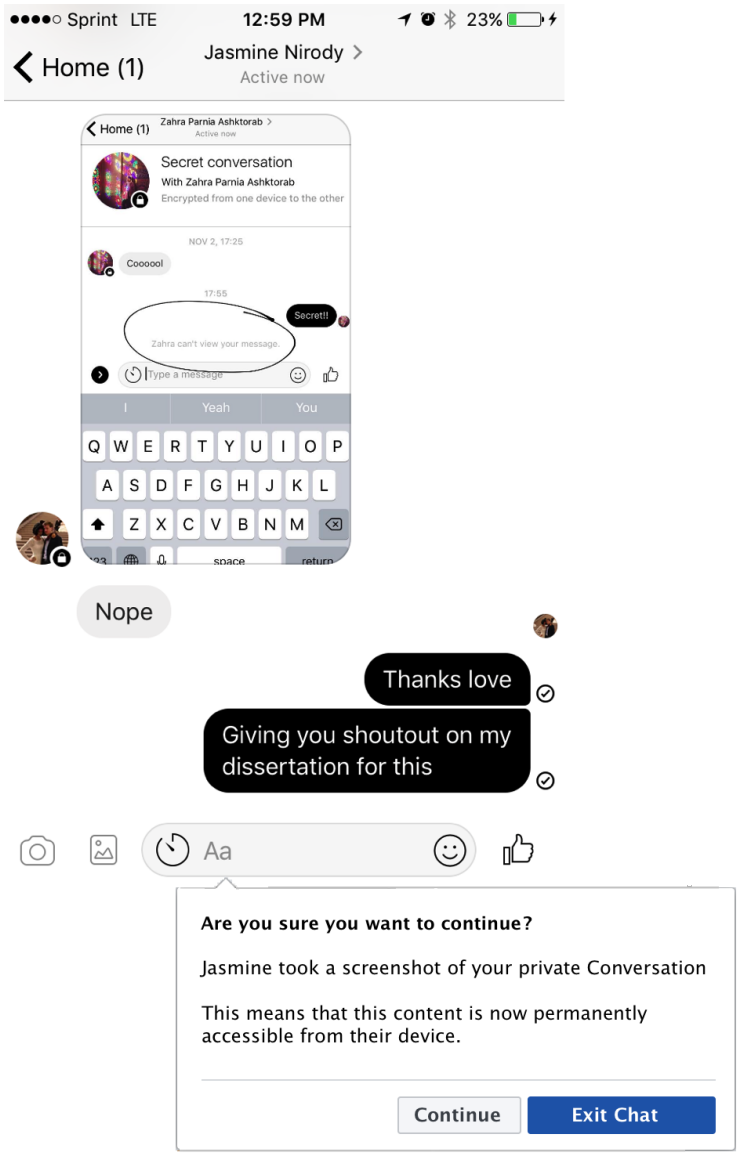


Figure 7.4: Prototype of reflective interface after screenshotting a "secret" Facebook chat.

olate the Contextual Integrity of social media platforms. Reputation allows users to identify the standing of others and themselves on social media platforms. Keitzmann et al. describe "reputation" as one of the building blocks of social media [139]. Reputation depends on aggregated measures of trust-worthiness by users. For example, LinkedIn assigns reputation based on endorsements [25]. StackOverflow assigns reputation based on up-votes on questions and other forms of interaction that contribute to the StackOverflow community [188]. Such crowd-sourced ratings signal the trustworthiness of an individual and in turn influence how individuals interact with one another. As a *primary prevention* cyberbullying mitigation design recommendation, ephemeral communications should crowdsource reputations to measure users' adherence to the contextual integrity of the social media platform. In Figure 7.5, symbols have been added next to Snapchat contacts in a prototype to demonstrate that an individual tends to screenshot photos. Such symbols would influence how users interact with one another.

7.3.2 Secondary Prevention

The aim of *Secondary Prevention* is to decrease the incidents of cyberbullying once it has already started. In the original "Continuum of Harm" pertaining to domestic violence, secondary preventative measures include home visits for high risk families to raise awareness of the harms of domestic violence. In the realm of cyberbullying prevention, *Secondary Prevention* manifests through cyberbullying

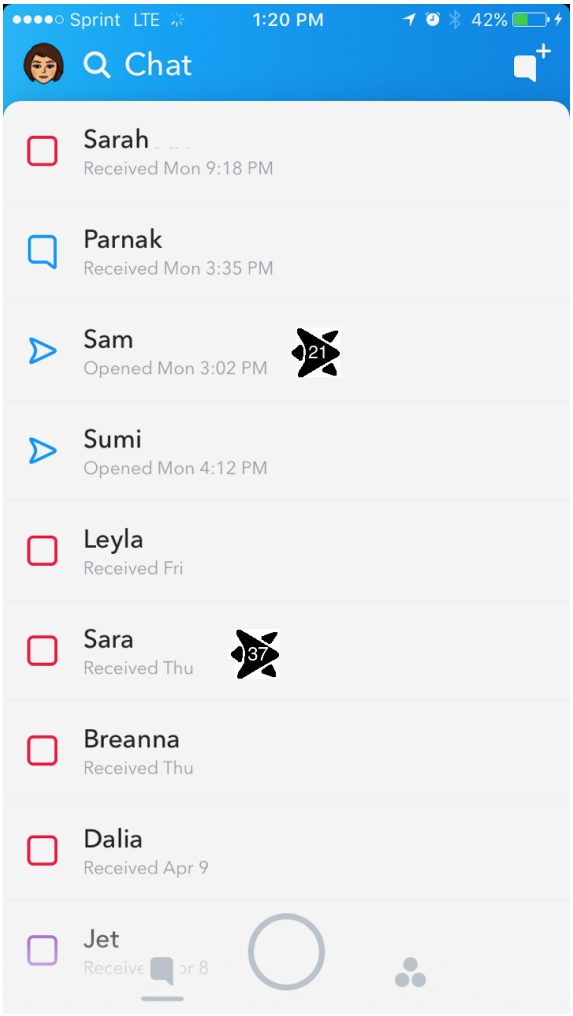


Figure 7.5: Prototype of visible reputations of Snapchat contacts denoted with black symbol with the number of screenshots taken in the past week.

applications which filter content or reporting content. A victim can choose to filter content once becoming aware of cyberbullying content. In the participatory design sessions, prototypes like "Watch Yo Profanity" and "SMILE," the victim decides if she would like some degree of filtering to be happening on his profile. In "Hate Page Prevention", a bystander or an automatic system flags the page and ultimately makes the decision to get rid of the data. In these types of cyberbullying design solutions, either a bystander or a third party automated system has ultimate control over the cyberbullying data being published. Control of solutions presented as a part of *Secondary Prevention* are held by the victim or a bystander of the cyberbullying since the victim is on the receiving end of the bullying.

In a study I conducted about young adult women's online harassment experiences [267] [1], I recommended the use of custom filtering, since participants in this study reported being called names or receiving unwanted content online [267]. While language and machine learning tools are constantly improving to detect online harassment, language continues to evolve and online harassment can be contextual. Furthermore, cyberbullying and harassment might lack key features (e.g., expletives) that are required to automatically identify harassment and other malicious content [80]. The prototypes resulting from the Participatory Design sessions described in Chapter 6 recommended the notion of user-centered custom filtering that allow users to identify the words that should be omitted from a user's timeline. After the publication of this work [18] both Instagram [172] and Twitter [190] adopted custom filtering, giving power to users to choose to identify words to be filtered

[1]Study conducted in collaboration with Jessica Vitak, Kalyani Chada and Linda Steiner

from their social media platforms. Figure 7.6 and Figure 7.7 are two *Secondary Prevention* prototypes co-designed with teens in participatory design sessions that allow users to identify words that should be filtered from their platforms.

7.3.3 Tertiary Prevention

Tertiary Prevention is a preventative measure that occurs when the problem has already caused visible harm. From the suite of solutions produced in the participatory design sessions, many attempted to mitigate negative emotional outcome of cyberbullying by sending positivity. Since visible harm on the Cyberbullying Continuum of Harm is inclusive of behavior ills, psychosomatic symptoms, and suicide ideation, *Tertiary Prevention* can be initiated by bystanders or automated systems. In the cyberbullying domain, the "Positivity Generator" allows victims to replace malicious content on their profiles with uplifting quotes from their favorite celebrities. This particular solution aims to do more than just filter negative content, but provide support and encouragement to counter the negative cyberbullying content they have experienced.

7.3.4 "Continuum of Harm" and Prevention

In the domain of domestic violence Primary prevention aims to reduce the incidence of the problem even before it occurs. For different age groups, primary prevention for domestic violence looks different. Ultimately, the aim is to *educate* individuals about the harms of domestic violence and conflict resolution. Wolfe et al. describe the different types of public education of diminishing cyberbullying before

Figure 7.6: Example of SMILE application that reacts to cyberbullying once it has already occurred by omitting posts including user-defined words from a users' social media timeline.

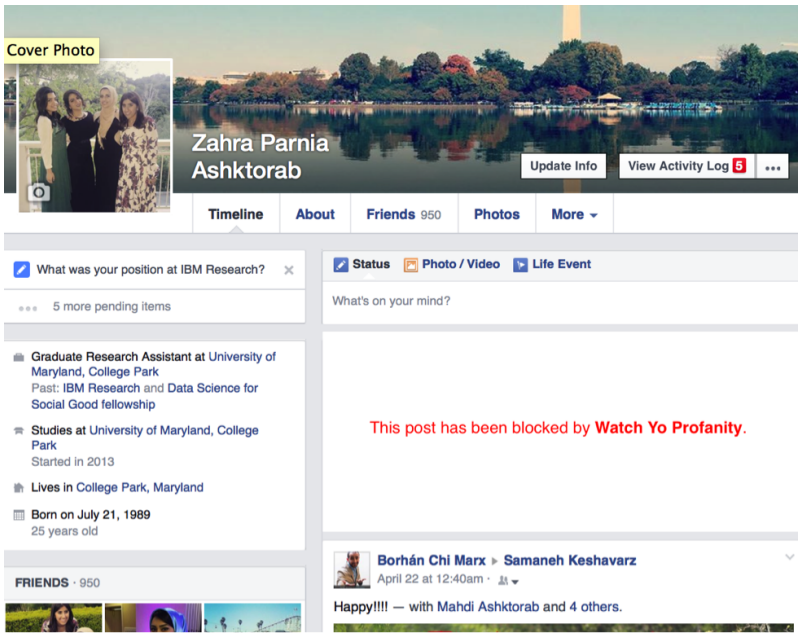


Figure 7.7: Example of "Watch Yo Profanity" application that reacts to cyberbullying once it has already occurred by omitting posts including user-defined words from a users' social media timeline.

it occurs [277]. For adolescents and high-school age youths (13-18 years), Wolfe et al. recommend school-based awareness and skill development. Communities should make a collaborative effort to teach awareness about violence and conflict-resolution skills [277]. Issues related to dating violence and forming healthy intimate relationships should be emphasized. For adults (18 years and older) Wolfe et al. recommend public education, media campaigns to promote awareness about domestic violence. For the primary prevention in the context of cyberbullying, we approach the prevention in terms of design and education. As primary prevention for domestic violence functions, schools should educate children about the harms of cyberbullying and online harassment in order to prevent such incidences.

For the domestic violence continuum of harm framework, secondary prevention which is targeted to individuals following early signs of domestic violence are offered community-based early intervention. For individuals exposed to violence aged 13-18, crisis support, individual counseling, and educational groups are offered with an emphasis on intimate relationships. For adults, individuals exposed to domestic violence are provided with coordinated services [277].

7.4 Using "Boosting" Policies for Ethical Cyberbullying Mitigation

Nudge theory is the notion of utilizing positive reinforcement and indirect suggestions to influence decisions [146]. It has been used in law and policy [10] and to influence behavior change [155]. Thaler and Sustein describe "libertarian paternalism", also described as "nudging" which helps people make decisions without

compulsion [255]. The "nudged" individual is given choices to move forward with his/her decision making. Many of the design recommendations for cyberbullying mitigation throughout this thesis tie in with "nudge theory", giving individuals the freedom to ultimately choose their course of action but also giving recommendations that may prevent cyberbullying along the "Continuum of Harm" described below.

Many of the design recommendations for cyberbullying mitigation rely on *nudging* policies, or creating a social environment within a platform to influence the behaviors of a potential bully or victim. Some critics and adversaries of nudging behavior conjure that nudging undermines human sovereignty since it exploits human weakness to influence behavior on social media platforms. Nudge policies undermine autonomy [274] since such policies change and alter contextual factors in order to influence decision-making. Critics stipulate that only "rational persuasion" can respect the sovereignty of individuals when they make choices [118], and nudging policies, or in this case, nudging design mechanisms do not constitute as "rational persuasion".

Nudging and *Boosting* are two varying approaches aimed at influencing individuals to make better decisions that lead to a better outcomes [114]. While nudging aims to co-opt systematic biases to influence behavior, boosting policies are more targeted to individuals who are competent and make informed decisions while increasing their skills [114]. At this end, critics have rendered some nudging policies manipulative [118].

Furthermore, Grne-Yanoff et al. describe the differences between nudges from boosts. Nudges and boosts differ in the "(i)immediate intervention targets, their

immediate intervention targets, (ii) their roots in different research programs, (iii) the causal pathways through which they affect behavior, (iv) their respective assumptions about human cognitive architecture, (v) the reversibility of their effects, (vi) their programmatic ambitions, and (vii) their normative implications" [114].

Nudges operated under the heuristics and biases (H&B) research program [103] that concludes that human biases are flawed as are motivations which leads to poor choices and decisions. Boosts, according to Grune, operate under the simple heuristics program (SH) [101], which argues that humans are "boundededly rational decision makers" and given the tools and skills, can make "good enough decisions". The design recommendations made in this dissertation to *prevent* and *mitigate* cyberbullying fall under the category of "boost" policy, not undermining the autonomy of users, but giving boundedly rational decision-makers the tools to make rational decisions. Below, I demonstrate how the various design recommendations in this dissertation keep users informed and thus utilize "boosting" design mechanism to ultimately help make users sound decisions along the cyberbullying *continuum of harm.*

7.4.1 Boosting Policy and Cyberbullying Detection

When describing potential *primary prevention* mechanisms, I recommend screenshot detection as well as escalation detection, pictured in Figures 7.2 and 7.5. These mechanisms use boosting policies to inform users to make rational decisions about moving forward. Screenshot detection, both in the example described on Snapchat as well as Facebook messages, merely inform a rational user of an action that may

have violated the contextual integrity of the ephemerality of the social media platform, and allow the user to proceed accordingly.

In *secondary prevention* mechanisms, *SMILE* and *WatchYoProfanity* are both human-centered mechanisms that depend on rational individuals to pre-determine a list of words that may be perceived as harmful towards the recipients. These mitigation mechanisms rely on the contextual nature of cyberbullying. Furthermore, by giving users the choice to determine the words that are being filtered, these mitigation mechanisms respect the individual sovereignty of users and treat them as rational individuals.

7.4.2 Nudging Policy and Cyberbullying Detection

In *tertiary prevention*, I recommend mitigation systems that remind users of the positivity or the existing social support on their social media platforms either by leveraging existing content on their social media profiles as is in the case of CTMM described in Chapter 6, or collaborative filtering as is the case in the Positivity Generator application. These systems fall under nudging policy, as they make assumptions about users' lack of knowledge of existing social support or ways they can promote their personal well-being. These *tertiary prevention* for cyberbullying mitigation make an assumption that the current state of the victims decision-making is mindless and passive [254].

While a nudge is intended to steer someone's decision making and behavior in a particular decision, these tertiary prevention systems use positive memory to influence mood which in turn may influence behaviors that result from in-

tense repetitive cyberbullying. A nudge utilizes empirically documented knowledge about human's cognitive capabilities and weakness and changes context to influence decision-making. While in the mitigation systems described above, users explicitly make decisions immediately when interacting with reflective interfaces, these tertiary prevention systems may be more subtle in their influence on an individual. While the goal is to improve mood, research shows as well as the data collected in this study that being reminded of dear friends and memories may cause an individual to communicate and seek social support from friends which will in turn ameliorate the negative effects of cyberbullying on well-being.

7.5 New Directions for Automatic Detection of Cyberbullying

Human Centered Machine Learning is an emerging field that incorporates the knowledge garnered from ethnographic studies into machine learning algorithms and techniques. Cyberbullying detection is a vital component of cyberbullying mitigation [17, 18].

Many participants across both the study conducted in Chapter 6 (Curated Technology Mediated Memory), and Chapter 5 (Participatory Design) cited exclusion as a type of cyberbullying they had experienced. Exclusion can be defined as directed repeated exclusionary behavior [275]. Participants reported experiencing exclusionary cyberbullying on Snapchat, Facebook and Instagram. In this section, I give examples of different types of exclusionary cyberbullying and provide recommendations for detecting exclusionary cyberbullying.

7.5.1 Exclusion through Photo Cropping

Exclusion by way of photo cropping was discussed extensively in Chapter 5. Through the participatory design sessions, I provided a prototype for *Exclusion Prevention* as a cyberbullying mitigation prototype. *Exclusion Prevention* alerts a social media user when using the cropping feature on sites like Instagram and crop out one or more people in the picture. When considering technologies required to implement this feature, face-detection-technology [223] and other image processing tools can help detect whether someone has been cropped out of a photo. In the reflective interface designed for *Exclusion Prevention*, (as seen in Figure 7.8), a reflective message prompts the user to reconsider posting since exclusion has been detected. The user then decides whether they want to continue with posting the picture. While crop detection has not been an explored area of research, far more computationally difficult automated image processing tasks like achieving high accuracy when detecting the optimal photo crop based on subjects' facial gaze have been explored [227]. A supervised image processing algorithm considering features like color, pixels, photo size, and number of people can potentially discover whether a photo has been cropped to exclude someone [35].

Additionally, exclusion prevention can be aided by social network analysis methods of groups and cluster detection [86]. The changing dynamics of visible social media interactions (likes, comments) of individuals in groups can be a considered feature in an exclusion detection algorithm. For example, Palla et al. develop an algorithm based on clique percolation [75] to discover relationships that characterize

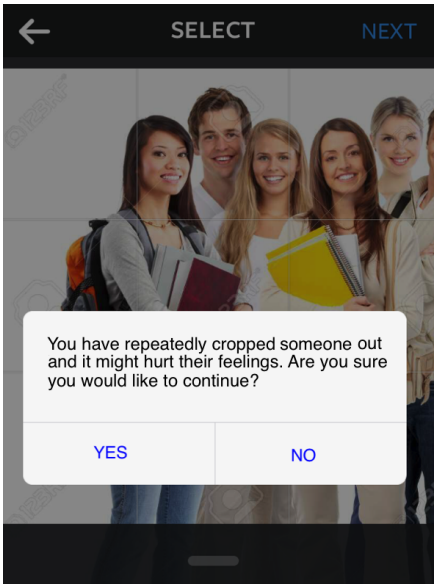


Figure 7.8: Prototype of Reflective Interface in "Exclusion Prevention" application

the evolution of communities [200]. Such methods can be applied to detect exclusion cyberbullying within mitigation tools.

7.5.2 Like Solicitation Exclusion

The studies in this dissertation reveal Like Solicitation exclusion across various social media platforms. On ASKfm, I discovered two types of discourse that exhibit qualities that allow "Like-for-Like" exclusion: *Like Solicitation and Rating Discourse* and *Listing All people You follow Discourse*. In *Like Solicitation and Rating Discourse*, the users ask that whoever "likes" the discourse will receive some sort of interaction on the website through "rating", "compliments", or reciprocated "likes". In *Listing All people you follow Discourse*, users ask a user to list everyone they follow on the site (via @username). This discourse type reveals "hidden" information as the site structure prevents users from seeing their followers list unless

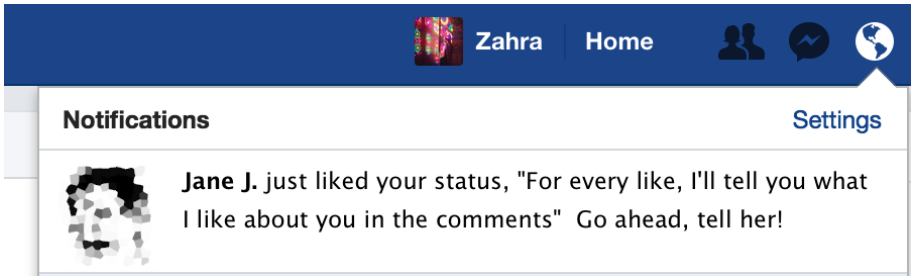


Figure 7.9: Prototype of notification interface to encourage Like-Solicitation exchange

they receive a "like" interaction or are tagged in a discourse type like this one. Both of these types of discourse exhibit qualities that allow exclusionary behavior.

Similarly, participants in the CTMM study (Chapter 6) described *Receiving Support* (Support in the form of text or PDA [120]) as a positive type of interaction that they experienced on Facebook. One participant noted, "The whole "like my status" and I'll say something nice about you." This type of discourse is similar to the *Like Solicitation and Rating Discourse* discovered in the askFM study.

While Like-Solitication may have been a part of the participants' (Chapter 6) positive weekly experiences, it was also included as a part of the weekly negative check-ins. Participants noted *Exclusionary Behavior* as the most negative part of their weekly experiences. In one instance, a user was upset that they were not invited to an event on Facebook event. In another instance, a participant who had promised to give compliments in exchange for "likes", did not. To prevent this type of exclusionary behavior, users can be reminded through the social media platform to deliver on their Like Solicitation statuses. Figure 7.9 shows a notification to encourage a user to deliver on the promise of commenting in exchange for the like

that has been redeemed.

7.6 Recommendations for Future Research

This study is the first of its kind to measure the influence of a cyberbullying mitigation tool on participants over an extended period of time. At the end of this study, I asked users to provide feedback on their thoughts about mode of delivery, aesthetic of delivery, timing, and the effect of the cyberbullying mitigation pings. Based on the results of this study, I make recommendations for running a study to evaluate *Tertiary Prevention*.

7.6.1 Logistical Challenges of Cyberbullying Mitigation Study

There are many ethical and logistical challenges of administering a study that measures the effectiveness of a cyberbullying mitigation tool over a period of time: 1) Recruitment 2) Preventing attrition 3) Linking Data 4) Identifying mode of delivery 5) Improving the User Experience of participants.

Assigning each individual a study ID is an important part of a cyberbullying mitigation study. Individuals were assigned an ID that links their social media data, their preliminary survey, their final exit survey, and their weekly check-ins. This means that individuals did not have to not have to re-enter information during every step of the data collection and linking different sets of data during different times of the study was quite straightforward.

7.6.1.1 Recruitment Challenges

A researcher who wants to study the effectiveness cyberbullying mitigation tools must first find a population that has been affected or is regularly affected by cyberbullying. In this study, I surveyed Freshman students at the University of Maryland about their past experiences about cyberbullying and online harassment. Those who were recruited in the study had experienced some version of online harassment as well as were active on Facebook, the platform through which the CTMM's would be created and curated. Studies say that 43% of undergraduate students have experienced some variation of online harassment [164]. For this reason, researchers at universities can use this population to study mitigation techniques.

7.6.1.2 Preventing Attrition

In order to prevent attrition, individual compensation increased dramatically if users completed the final survey. Since the final survey was necessary in order to compare well-being measures, as well as collecting feedback about perception of CTMM, compensation was maximized if an individual completed the final survey along with the check-ins. Below is a table of the break down for compensation for participants in this study. Participants should be given more incentive to complete both the pre-test and the post-test, so that measures could be accurately measured at the end of the study.

Table 7.1: Incentives and Compensation for Participants in Cyberbullying Mitigation Study

Degree of Completion	Incentive
Pre-test only	No Payment
Pre-test + 4 check-ins + Post-test	$10.00
Pre-test + 3 check-ins + Post-test	$8.00
Pre-test + 2 check-ins + Post-test	$7.00
Pre-test + 1 check-ins + Post-test	$6.00
Pre-test + 4 check-ins	$4.00
Pre-test + 3 check-ins	$3.00
Pre-test + 2 check-ins	$2.00
Pre-test + 1 check-in	$1.00

7.6.1.3 Linking Data

Check-ins were administered on a weekly basis. Emails were sent with a link including the Survey-Gizmo checkin. In order to keep track of the participant in the study, the original study ID participants were assigned was saved in the custom URL which was sent to the participant, so the participant did not have to enter personal information in the study.

Additionally, social media data was collected from participants' Facebook profiles. All participant data was identified by the unique ID assigned to participants when they first took the preliminary survey. This preliminary unique ID was vital in ensuring all the different aspects of data could later be linked and analyzed.

7.6.1.4 Identifying Mode of Delivery

The Cyberbullying Technology Mediated Memory Mitigation were administered through participant emails. This choice was made based on the fact that all participants would check their emails regularly and have access to it. However, the feedback in the study revealed that users likened some the emails to spam, since

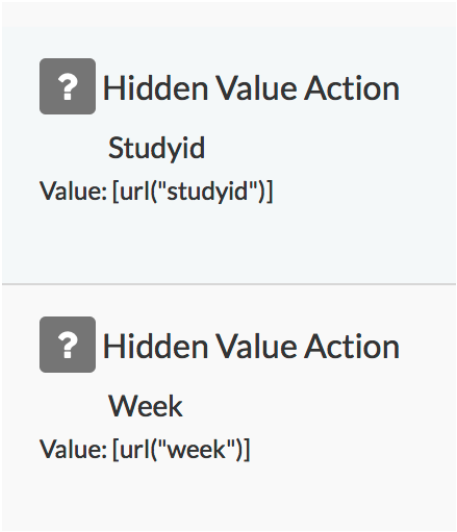


Figure 7.10: Survey Gizmo hidden URL variables for Study ID and Week number for weekly checkin.

one's email is a likely place to receive spam. Alternative modes of delivery that can be explored in future studies are application-specific deliveries, SMS, or third party mobile applications. The mode of delivery would likely influence how users interact with the mitigation and how likely or quickly they are to access it.

In the feedback exit survey, I asked users why they chose to interact with the links that were sent to them. Users commented that the location in which they received the pings affected whether they interacted with the links. Timing and location of a participant can further influence whether recipient of the mitigation pings described in this study would interact with the mitigations. In such an instance, third party applications can be used to identify a "home base" for a participant to ensure a higher participation rate.

7.6.1.5 Improving the User Experience of Participants

In order to collect Facebook data, I was required to create a Facebook application that would improve the user experience of Facebook users. The requirement for each data point requires that it must provide value to people by analyzing the content collected or personalizes in-app content or experiences [4]. This means that one cannot simply create an application to gain access to user profiles and collect data. Upon completing the survey, users were linked to the Facebook application in which participants were prompted to sign in. The application I created asked users a series of social support questions [214] about their network and then created a photo collage of all of the friends from which participants indicated they felt some degree of social support. The login page for the application can be seen in Figure 7.11. The collage generated in the application "personalizes in-app content or experiences". In order to collect Facebook data, an application must be created that does more than merely collect data required for analyses, but also contributes to improving the user experience of the platform users.

7.6.2 Lessons Learned for Future Mitigation Studies

The study evaluating the effectiveness of CTMMs is the first of it's kind to evaluate a cyberbullying mitigation tool. For this reason, the results of this study not only inform us about the effectiveness of a cyberbullying mitigation tool, but the feedback collected in this study enables future researchers to better evaluate the effectiveness of cyberbullying mitigation tools by better designing their studies. In

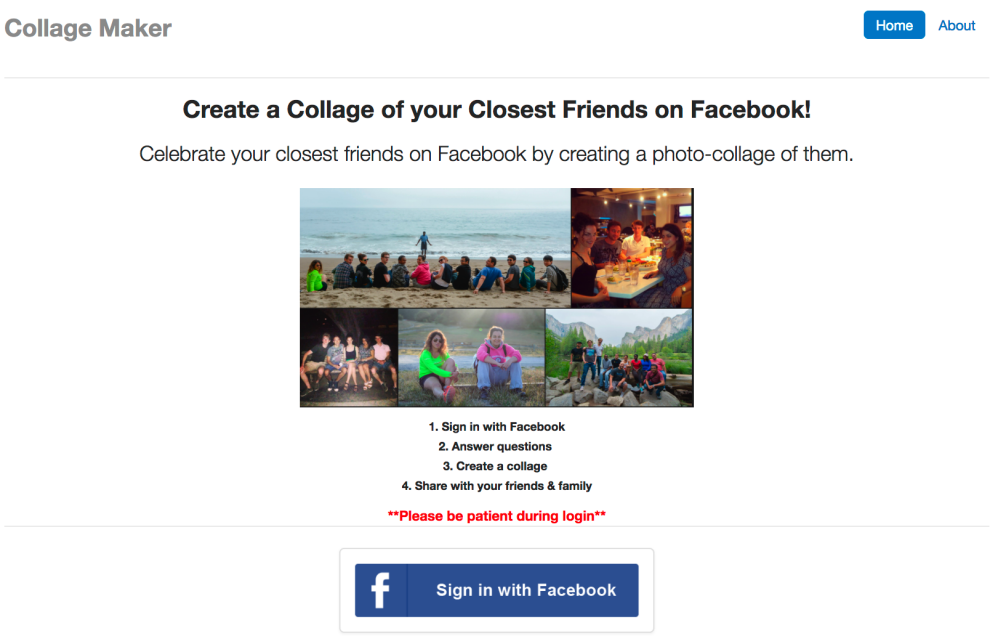


Figure 7.11: Login Page for the Collage Maker Application which was submitted to Facebook to personalize user experience.

this section, I provide insights into the lessons learned from conducting this study.

7.6.2.1 Include Control Group

The data collected during the administration and evaluation of the CTMMs occurred during the 2016 Presidential Elections. Beyond this specific event, historical events can occur may influence individuals' well-being over a period of time. For this reason, researchers who collect well-being measures over a period of time to measure the effectiveness of cyberbullying mitigation should include a control group who does not receive cyberbullying mitigation, to more accurately make conclusions about the effectiveness of the cyberbullying mitigation tool.

7.6.2.2 Appropriate Sample

Finding an appropriate sample on which to conduct a cyberbullying mitigation study is a challenge because it is difficult to predict whether the participants in the study will experience cyberbullying throughout the course of the study. In my evaluation of CTMMs, I invited college freshman who had experienced cyberbullying at some point in the course of their lives before the start of the study to participate in the longitudinal study. While I certainly found traces of cyberbullying in the sample of participants in the study, alternative heuristics can be considered when inviting participants to participate in a study that evaluates the effectiveness of a cyberbullying mitigation tool. For example, if a cyberbullying mitigation tool is being administered only on a particular social media platform, cyberbullying on that particular platform can be considered as a requirement for recruitment of the study. The proximity of the last cyberbullying incident experienced by participants should also be considered since individuals who have experienced cyberbullying more recently would be more likely to experience it again.

7.6.2.3 Consider Measures Collected

Since the final study in the dissertation was evaluating the effectiveness of a *Tertiary Prevention* mechanism for cyberbullying, well-being measures were considered for evaluating the effectiveness of the cyberbullying mitigation tool. However, effectiveness and success of a cyberbullying mitigation tool can be evaluated in different ways depending on the nature of the cyberbullying mitigation tool. In

this chapter, I introduced the notion of *Primary Prevention*, *Secondary Prevention*, and *Tertiary Prevention*. The effectiveness of *Primary Prevention* and *Secondary Prevention* may be measured differently. Beyond well-being measures, measures of social support can reflect the effectiveness of a cyberbullying mitigation tool.

Furthermore, the cyberbullying mitigation tool in this study was a *victim-centric* tool, evaluating how the mitigation effects the victim. A reflective interface aimed at influencing the decision-making process of a user who may send offensive content should be measured differently. The measure of effectiveness in such cases can be reflected in how a participant interacted with users when presented with a reflective interface and whether the boosting mechanism prevented from a negative interaction from occurring. Exit interviews with users who were presented with such cyberbullying mitigation tools would also provide more insight into the thought process and decision making process of individuals who are presented with reflective interfaces. This feedback could lead to an improvement of the design of reflective interfaces aimed at discouraging individuals from publishing malicious or incendiary content on social media platforms.

7.7 Other Ethical Considerations for Designing and Implementing Cyberbullying Mitigation

With increasing use and analysis of big data, ethical considerations have received significant attention [43], and the CSCW community has been at the forefront of discussions regarding ethical collection and analysis of user data [55]. Ensuring

the confidentiality of data and anonymity of participants is especially important when (1) analyzing adolescents and (2) analyzing sensitive events such as self-harm, bullying, and suicide. While standards vary across institutions and industries regarding what constitutes "human subjects data" researchers have a responsibility to take all necessary steps to protect the privacy and safety of individuals in a dataset.

Ethical considerations were built into the research design process to minimize any potential for harm. The researchers carefully weighed risks of data collection versus the potential benefits to the user population from this study. Scraping and data analytics in this study were only administered on public profiles. I do not present any information that could be used to re-identify an individual participant. In the pilot study, participants were presented with a number of resources upon completing the survey to consult if they had any questions about cyberbullying or wanted to discuss their experiences.

In line with work by Goode [107], I believe that deception is an integral part of testing the pipeline for administering support for cyberbullying victims. For example, in the study described in Chapter 4 evaluating *Cyberbullying Reversal Pings*,the impact of positive messages on a victim of cyberbullying in a natural (i.e., not lab) setting can only be effectively measured if recipients of positive messages are not explicitly told that the supportive messages are being sent through an automated system. One of the biggest challenges in designing this study was minimizing the possibility that the CTMM would have a negative effect on the recipient. Researchers working with automated response systems in any environment must consider all potential responses to content–no matter how banal in nature;

with young people experiencing cyberbullying, this becomes even more critical. Future researchers should carefully consider how the systems they design may affect the intended audience, both for the better and for the worse. As important as this technology is, the well-being of the people on the receiving side is always paramount. The biggest ethical issue with the study is participant risk (e.g., bringing cyberbullying up during mitigation make victims feel worse).

Per IRB protocol, participants must be asked to sign consent form in beginning of the study and should be informed that their profiles will be monitored throughout the study. Below, I explain the two main ethical challenges in this study: 1) Informed Consent and Scraping Profile Information and 2) Making participants potentially feel worse by referencing cyberbullying.

7.7.1 The Uncanny Valley, The Transparency Paradox and Informed Consent

By consenting to be involved in this study, users consent to have their Instagram data scraped. The consent of participants in these studies is particularly important. Zimmer et al. identifies the importance of ethical concerns before embarking on research on social networking sites including: anonymization of data prior to public release, respecting the expectations of privacy on a particular social networking site, and respecting the nature of consent [288]. Some participants in the CTMM study (Chapter 6) reported that they felt that weekly pings had too much information from their previous lives (information to which they had consented to being collected in the beginning of the study). While the Facebook application notified users of the data collected, the data was stored in a database on not sent

to users until much later in the study. Masahiro Mori, a robotics professor at the Tokyo Institute of Technology described the Uncanny Valley, the reaction to robots that resemble humans but are not quite human. Recently, the notion of the uncanny valley has gained popularity in other scientific circles and is used to describe phenomenon that seem human-like but are not and thus seem eery to humans [184].

One participant reported, *I was a little creeped out that I was being emailed compilations of photos.*

One potential recommendation as a result of this phenomenon is to be more transparent about how data is collected and how it is being used. While some participants reported that they felt uncomfortable with the amount of data the Facebook application had about them, they had consented to all of the data being collected at the start of the study. The question begets itself, would users feel better about the information being presented to them if there was more transparency on how it was being used?

For those who adhere to notice-and-consent policies, notice about how data is being collected needs to be simplified so that ordinary people can understand it. In the simplification process fine details are lost and thus full transparency is not achieved. If notice was delivered with all of the fine details, we know that it is unlikely to be understood by the average individual. An abbreviated policy of data collection is easy to read and understand but also filters away the fine details which are often the key to understanding the policy. This is defined as the transparency paradox, "transparency of textual meaning and transparency of practice conflict in all but rare instances" [193]. In the design recommendations for cyberbullying

mitigation tools, I draw from the results of this study on how to approach the transparency paradox when creating cyberbullying mitigation tools.

Participants in the study reported that the amount of information delivered in the CTMM's seemed "creepy" or "eerie". This feedback paves way for the "transparency paradox". The feedback received at the completion of the study revealed that users feel eerie with the amount of information that was sent to them in the emails and did not realize that the third party Facebook Application captured so much information about them. In this section, I make design recommendations for navigating the transparency paradox for a CTMM.

7.7.2 Referencing Cyberbullying

As mentioned previously, the ethical challenge unique to this study is the risk posed to an individual who might not want to be reminded of cyberbullying on their profile. The cyberbullying mitigation techniques used in this study only focus on sending positive messages and not referencing previous cyberbullying messages, so as to prevent the participants from experiencing further trauma.

7.8 Conclusion and Future Work: Measuring the Effectiveness of Cyberbullying Mitigation Solutions

While this study resulted in potential solutions for cyberbullying mitigation, much work lies ahead. I have proposed a number of potential mitigation solutions and the technologies required to implement these solutions. Future research should

implement and evaluate these solutions with users through longitudinal studies to evaluate the behavioral impact they have on bullies, victims, and bystanders. In addition, future work should leverage the existing technologies to implement the proposed solutions which are a result of co-design between researchers and adolescents.

My analysis and categorization of the different preventative types allows me to consider additional research questions, such as which preventative solution is most effective for cyberbullying prevention and how can I accurately measure this effectiveness. Until this point, technological cyberbullying prevention mechanisms have not been evaluated for effectiveness. The framework presented in this paper provides a straightforward way to begin to consider how one would compare different solutions. The ethical challenges of such a study are daunting, but would provide critical insights to preventing cyberbullying.

7.8.1 Domain Specific Detection

In the weekly checkins for the evaluation of CTMM, participants reported different types of negative experiences experienced during the week. Some users reported political debates escalating into corner cases of cyberbullying like sexism and racism [257]. While many algorithms automatically detect general harassment, In this section I make recommendations for the detection of contextual cyberbullying through the incorporation of semantic networks. These recommendations build on previous efforts to accurately detect traces of cyberbullying through Natural Language Processing methods [79, 80, 145, 279].

Based on the feedback received from the weekly checkins, racism and sexism were reoccurring themes in different instances of online harassment. In this section, I recommend using semantic networks to detect two types of context-specific cyberbullying: sexism racism. Majority of existing automated methods of cyberbullying have used expletives as features to detect cyberbullying. Kontostathis et al. use data from the social networking site formspring.me data and machine learning techniques to use off the shelf natural language processing tools to automatically detect cyberbullying content [145, 279].

7.8.1.1 Semantics and Cyberbullying

Dinakar et al. present an approach for cyberbullying detection on a limited range of subjects: appearance, intelligence, racial and ethnic slurs, social acceptance and rejection [79]. Dinakar et al. make a strong distinction between automatic spam detection and cyberbullying detection drawing the distinction that cyberbullying is specific to a person and contextual, while spam is sent to multiple people at once. The datasets from this project are from YouTube and Formspring (now spring.me). Dinakar et al. aim to detect sexuality insults and LGBT-related insults. Dinakar et al. make use of OMCS (Open Mind Common Sense) to better detect cyberbullying that was not picked up in their specific research [243]. OMCS is a project that has a broad collection of basic knowledge that is able to provide stereotypes and social constructions that are used to insult victims of bullying. For example, the comment, "put on a wig and lipstick be who you really are" might not be detected as cyberbullying by the previous classification method. However, using the OMCS

and different kinds of relations defined on ConceptNet, the authors use the list of assertions to conclude that indeed this post is cyberbullying.

7.8.1.2 ConceptNet and Open Mind Common Sense Project

The Open Mind Common Sense (OMCS) project is a knowledge base that gives access to basic knowledge so that can help applications understand the ways objects, people and entities interact with one another [243]. OMCS holds knowledge about typical gender roles. For example, it knows girls are capable of doing housework in the same way that boys are capable of wrestling. Dinakar et. al present ConceptNet as means to make the OMCS computationally useful [79]. ConceptNet is a semantic directed graph that shows the relationships between entities in the OMCS knowledge base.

ConceptNet is a Semantic Network of Common Sense knowledge. It is based on the Common Sense Project, a project aimed at crowdsourcing "common sense" relationships in the real world. An example of how the Common Sense Project works is through gamification. For example, users are asked questions like: "Hammers are used for:" and then asked to fill in a blank. Through such exercises "common sense" about the real world are constructed. While ConcepNet includes WordNet relationships like hyponym/hypernym (isA), it is capable of representing many more relations [95]. ConceptNet attempts to represent "common sense" assertions through these relations.

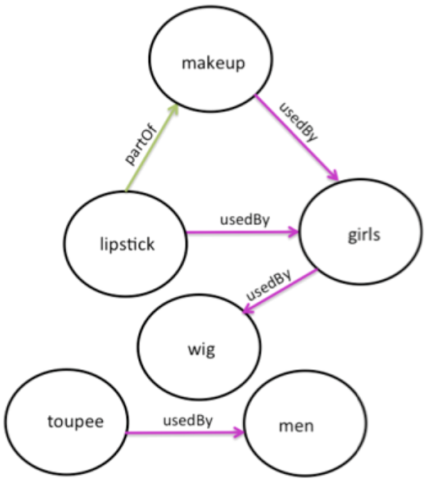


Figure 7.12: Semantic Graph of LGBT-related insult represented through ConceptNet relationships

7.8.1.3 Building SexismSpace and RacismSpace

In order to build SexismSpace, the sexist concepts and their relative assertions must be converted into a matrix. The matrix is represented as Concepts by features, which are relations and the object of the relation in the assertion. For example, let's consider the following sexist assertions: 1) "woman hasProperty delicate", "woman hasProperty inferior", and "man hasProperty logical". In the matrix, "woman" and "man"are represented as concepts, and "hasProperty-delicate','hasProperty-inferior", and "hasProperty-logical" are all represented as features for which the concepts "man" and "woman" can be true or false. Singular Value Decomposition must then be conducted on the Matrix [71].

ConceptNet includes the concepts "blackRace" and "whiteRace". However, similar to the domain of sexism, a world needs to be constructed based on racist

assertions. The assertions for the canonical concepts also must be encoded to accurately be able to compute a racism score. For example, in a Youtube comment about a girl in a Cheerios commercial who is the child of interracial parents and is caught "stealing" a cheerio. One comment says, "Just goes to show you they start stealing at an early age!!! First Cheerios next bikes then cars lol just like in real life" [53]. In order to use *Common Sense Reasoning* to detect racism in a comment like this, one must encode stereotypes about different races based on assertions in comments like these. Before I list the potential assertions one can extract from the statement above, I must reiterate that none of these stereotypes are true and I don't subscribe to them, but I am merely listing them in order to demonstrate how one would approach this problem using a semantic network like ConceptNet. The assertions for the statement "Just goes to show you they start stealing at an early age!!! First Cheerios next bikes then cars lol just like in real life" would include: "blackRace capableOf steal", "blackRace capableOf stealingCars", "blackRace capableOf stealingCheerios", "blackRace capableOf stealingBikes", "blackRace hasProperty thief", "thief isA blackRace", "bikes receiveAction steal", "cars receiveAction steal". Such a racist world would have to be built in order to detect that the comment above is racist. Furthermore, the canonical concepts in this domain would be the specific races for which you are trying to detect racism. In the case for this particular statement, the similarities between extracted concepts in a sentence and the canonical concepts blackRace and whiteRace can be used to measure the degree of sexism in a statement. As described in the related literature, majority of cyberbullying detection methods focus on expletive detections. While this method is useful is

some cases, bullying posts include a much wider range of language cues than simply expletives, and bullying messages may be highly context dependent.

7.8.2 Evaluation of Primary and Secondary Prevention

This dissertation explores the design and evaluation of a type of *Tertiary prevention* of cyberbullying. In future work, additional modes of cyberbullying prevention (primary and secondary) should be evaluated with users. Beyond measures of well-being, open-ended instruments could capture the many nuanced reactions to the such prevention mechanisms that would ultimately lead to further design recommendations that would improve cyberbullying mitigation systems.

The studies in this dissertation demonstrate a mixed-method approach to studying cyberbullying mitigation. By developing new classification methods, extending existing cyberbullying intervention design themes, and implementing and evaluating cyberbullying solutions I have provided novel insights for the design of cyberbullying mitigation tools across various domains.

Bibliography

[1] The kanye west self confidence generator. `http://usatoday30.usatoday.com/exp/kanye/kanye.html`. Accessed: 2015-05-16.

[2] Erin gallagher, irish teen, commits suicide after battle with 'vicious' cyberbullying (photo), October 2012.

[3] Blogger caitlin seida: My photo went viral... and it led to cyber-bullying (video), October 2013.

[4] Facebook for developers, 2017.

[5] T. M. Achenbach, S. H. McConaughy, and C. T. Howell. Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological bulletin*, 101(2):213, 1987.

[6] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pages 665–674. ACM, 2008.

[7] L. Aksoy, A. van Riel, J. Kandampully, R. N. Bolton, A. Parasuraman, A. Hoefnagels, N. Migchels, S. Kabadayi, T. Gruber, Y. Komarova Loureiro, et al. Understanding generation y and their use of social media: a review and research agenda. *Journal of Service Management*, 24(3):245–267, 2013.

[8] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.

[9] L. Alvarez. Girls suicide points to rise in apps used by cyberbullies. *The New York Times*, 13, 2013.

[10] O. Amir and O. Lobel. Stumble, predict, nudge: How behavioral economics informs law and policy. *Columbia Law Review*, pages 2098–2137, 2008.

[11] N. Andalibi, O. L. Haimson, M. De Choudhury, and A. Forte. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3906–3918. ACM, 2016.

[12] D. M. Anderson and J. J. Sabia. Child access prevention laws, youth gun carrying, and school shootings. 2016.

[13] F. M. Andrews and S. B. Withey. Social indicators of well-being: The development and measurement of perceptual indicators. *New York: Plenum. doi*, 10:978–1, 1976.

[14] J. J. Arnett. Emerging adulthood: A theory of development from the late teens through the twenties. *American psychologist*, 55(5):469, 2000.

[15] J. J. Arnett. *Emerging adulthood: The winding road from the late teens through the twenties*. Oxford University Press, 2014.

[16] L. Arseneault, L. Bowes, and S. Shakoor. Bullying victimization in youths and mental health problems:much ado about nothing? *Psychological medicine*, 40(05):717–729, 2010.

[17] Z. Ashktorab, S. Kumar, S. De, and J. Golbeck. ianon: Leveraging social network big data to mitigate behavioral symptoms of cyberbullying. 2014.

[18] Z. Ashktorab and J. Vitak. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3895–3905. ACM, 2016.

[19] Ask.fm, 2013.

[20] H. U. Asuncion, A. U. Asuncion, and R. N. Taylor. Software traceability with topic modeling. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*, pages 95–104. ACM, 2010.

[21] I. Aviram and Y. Amichai-Hamburger. Online infidelity: Aspects of dyadic satisfaction, self-disclosure, and narcissism. *Journal of Computer-Mediated Communication*, 10(3):00–00, 2005.

[22] R. Bapna, A. Gupta, S. Rice, and A. Sundararajan. Trust and the strength of ties in online social networks: An exploratory field experiment. *MIS Quarterly*, 41(1):115–130, 2017.

[23] V. Barker. Older adolescents' motivations for social network site use: the influence of gender, group identity, and collective self-esteem. *CyberPsychology & Behavior*, 12(2):209–213, 2009.

[24] S. B. Barnes. A privacy paradox: Social networking in the united states. *First Monday*, 11(9), 2006.

[25] M. Bastian, M. Hayes, W. Vaughan, S. Shah, P. Skomoroch, H. Kim, S. Uryasev, and C. Lloyd. Linkedin skills: large-scale topic extraction and inference. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 1–8. ACM, 2014.

[26] N. N. Bazarova and Y. H. Choi. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication*, 64(4):635–657, 2014.

[27] A. V. Beale and K. R. Hall. Cyberbullying: What school administrators (and parents) can do. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 81(1):8–12, 2007.

[28] B. C. Becker and E. G. Ortiz. Evaluation of face recognition techniques for application to facebook. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[29] M. Bekker, J. Beusmans, D. Keyson, and P. Lloyd. Kidreporter: a user requirements gathering technique for designing with children. *Interacting with computers*, 15(2):187–202, 2003.

[30] T. Beran and Q. Li. The relationship between cyberbullying and school bullying. *The Journal of Student Wellbeing*, 1(2):16–33, 2008.

[31] M. S. Bernstein, E. Bakshy, M. Burke, and B. Karrer. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30. ACM, 2013.

[32] M. S. Bernstein, E. Bakshy, M. Burke, and B. Karrer. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30. ACM, 2013.

[33] M. S. Bernstein, A. Monroy-Hernández, D. Harry, P. André, K. Panovich, and G. G. Vargas. 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *ICWSM*, pages 50–57, 2011.

[34] J. L. Bevan, J. Pfyl, and B. Barclay. Negative emotional and cognitive responses to being unfriended on facebook: An exploratory study. *Computers in Human Behavior*, 28(4):1458–1464, 2012.

[35] J. C. Bezdek, J. Keller, R. Krisnapuram, and N. Pal. *Fuzzy models and algorithms for pattern recognition and image processing*, volume 4. Springer Science & Business Media, 2006.

[36] A. Binns. Facebooks ugly sisters: Anonymity and abuse on formspring and ask. fm. *Media Education Research Journal*, 2013.

[37] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[38] G. A. Bonanno, C. R. Brewin, K. Kaniasty, and A. M. L. Greca. Weighing the costs of disaster: Consequences, risks, and resilience in individuals, families, and communities. *Psychological Science in the Public Interest*, 11(1):1–49, 2010.

[39] L. Bowler, E. Mattern, and C. Knobel. Developing design interventions for cyberbullying: A narrative-based participatory approach. 2014.

[40] D. Boyd. Why youth (heart) social network sites: The role of networked publics in teenage social life. *Youth, identity, and digital media*, pages 119–142, 2009.

[41] D. Boyd. Social steganography: Learning to hide in plain sight, 2010.

[42] D. Boyd. *It's Complicated: the social lives of networked teens*. Yale University Press, 2014.

[43] D. Boyd and K. Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679, 2012.

[44] N. M. Bradburn. The structure of psychological well-being. 1969.

[45] P. B. Brandtzæg, M. Lüders, and J. H. Skjetne. Too many facebook friends? content sharing and sociability versus the need for privacy in social network sites. *Intl. Journal of Human–Computer Interaction*, 26(11-12):1006–1030, 2010.

[46] R. Broderick. 9 teenage suicides in the last year were linked to cyber-bullying on social network ask.fm, 2013.

[47] R. Broderick. Teenage suicides in the last year were linked to cyber-bullying on social network ask. fm. 2013.

[48] J. Brown. *Kanye West in the Studio: Beats Down! Money Up!(2000-2006)*. Amber Books Publishing, 2006.

[49] F. B. Bryant, C. M. Smart, and S. P. King. Using the past to enhance the present: Boosting happiness through positive reminiscence. *Journal of Happiness Studies*, 6(3):227–260, 2005.

[50] F. B. Bryant and J. Veroff. The structure of psychological well-being: A sociohistorical analysis. *Journal of Personality and Social Psychology*, 43(4):653, 1982.

[51] M. Burke, R. Kraut, and C. Marlow. Social capital on facebook: Differentiating uses and users. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 571–580. ACM, 2011.

[52] R. N. Butler. The life review: An interpretation of reminiscence in the aged. *Psychiatry*, 26(1):65–76, 1963.

[53] J. Campbell. Cheerios commercial: Racist slurs force youtube to close comments section for new ad, May 2013.

[54] J. V. Campo, J. Bridge, M. Ehmann, S. Altman, A. Lucas, B. Birmaher, C. Di Lorenzo, S. Iyengar, and D. A. Brent. Recurrent abdominal pain, anxiety, and depression in primary care. *Pediatrics*, 113(4):817–824, 2004.

[55] T. P. A. B. M. G. J. H. W. L. Casey Fiesler, Alyson Young. Ethics for studying sociotechnical systems in a big data world. In *CSCW*. ACM, 2015.

[56] J. Charteris, S. Gregory, and Y. Masters. Snapchat selfies: The case of disappearing data. *eds.) Hegarty, B., McDonald, j., & Loke, S.. K., Rhetoric and Reality: Critical perspectives on educational technology*, pages 389–393, 2014.

[57] K. M. Christopherson. The positive and negative implications of anonymity in internet social interactions: "on the internet, nobody knows you're a dog". *Comput. Hum. Behav.*, 23(6):3038–3056, Nov. 2007.

[58] C. Chung and J. W. Pennebaker. The psychological functions of function words. *Social communication*, pages 343–359, 2007.

[59] D. K. Citron and M. A. Franks. Criminalizing revenge porn. 2014.

[60] E. A. Cook. Effects of reminiscence on life satisfaction of elderly female nursing home residents. *Health care for women international*, 19(2):109–118, 1998.

[61] H. Coolican. *Research methods and statistics in psychology.* Hodder & Stoughton Educational, 1990.

[62] T. Correa, A. W. Hinsley, and H. G. De Zuniga. Who interacts on the web?: The intersection of users personality and social media use. *Computers in Human Behavior*, 26(2):247–253, 2010.

[63] D. Cosley, K. Akey, B. Alson, J. Baxter, M. Broomfield, S. Lee, and C. Sarabu. Using technologies to support reminiscence. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, pages 480–484. British Computer Society, 2009.

[64] H. Cowie. Perspectives of teachers and pupils on the experience of peer support against bullying. *Educational Research and Evaluation*, 4(2):108–125, 1998.

[65] H. Cowie. Cyberbullying and its impact on young peoples emotional health and well-being. *The Psychiatrist*, 37(5):167–170, 2013.

[66] B. Cravey. With cyberbullying, threats go viral - and can turn deadly, September 2012.

[67] M. Csikszentmihalyi and R. Larson. Validity and reliability of the experience-sampling method. In *Flow and the foundations of positive psychology*, pages 35–54. Springer, 2014.

[68] C. E. Cunningham, L. J. Cunningham, V. Martorelli, A. Tran, J. Young, and R. Zacharias. The effects of primary division, student-mediated conflict resolution programs on playground aggression. *Journal of Child Psychology and Psychiatry*, 39(5):653–662, 1998.

[69] M. De Choudhury. Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd international workshop on Socially-aware multimedia*, pages 49–52. ACM, 2013.

[70] M. De Choudhury and S. De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*. Citeseer, 2014.

[71] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[72] M. de Sá, L. Carriço, L. Duarte, and T. Reis. A mixed-fidelity prototyping tool for mobile devices. In *Proceedings of the working conference on Advanced visual interfaces*, pages 225–232. ACM, 2008.

[73] D. Dearman and K. N. Truong. Why users of yahoo!: answers do not answer questions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 329–332. ACM, 2010.

[74] F. DeHue, C. Bolman, and T. Völlink. Cyberbullying: Youngsters' experiences and parental perception. *CyberPsychology & Behavior*, 11(2):217–223, 2008.

[75] I. Derényi, G. Palla, and T. Vicsek. Clique percolation in random networks. *Physical review letters*, 94(16):160202, 2005.

[76] F. g. Deters and M. R. Mehl. Does posting facebook status updates increase or decrease loneliness? an online social networking experiment. *Social psychological and personality science*, 4(5):579–586, 2013.

[77] T. Diamanduros, E. Downs, and S. J. Jenkins. The role of school psychologists in the assessment, prevention, and intervention of cyberbullying. *Psychology in the Schools*, 45(8):693–704, 2008.

[78] E. Diener and R. J. Larsen. Temporal stability and cross-situational consistency of affective, behavioral, and cognitive responses. *Journal of personality and social psychology*, 47(4):871, 1984.

[79] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18, 2012.

[80] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*, 2011.

[81] J. S. Donath et al. Identity and deception in the virtual community. *Communities in cyberspace*, 1996:29–59, 1999.

[82] A. Druin. A place called childhood. *Interactions*, 3(1):17–22, 1996.

[83] A. Druin. Cooperative inquiry: developing new technologies for children with children. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 592–599. ACM, 1999.

[84] A. Druin, B. Bederson, A. Boltman, A. Miura, D. Knotts-Callahan, and M. Platt. Children as our technology design partners. 1998.

[85] A. Druin, B. Bederson, J. Hourcade, L. Sherman, G. Revelle, M. Platner, and S. Weng. Designing a digital library for young children: An intergenerational partnership. chi 2001.

[86] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu. Community detection in large-scale social networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 16–25. ACM, 2007.

[87] M. Duggan and A. Smith. Social media update 2013. *Pew Internet and American Life Project*, 2013.

[88] J. Dye. Meet generation c: Creatively connecting through content-generation c is the" you" in youtube, the" my" in myspace, and the" i" in ipod. they're you (and me), and they're shaking up the way people. *EContent-Digital Content Strategies and Resources*, 30(4):38–43, 2007.

[89] A. H. Eagly, W. Wood, and S. Chaiken. Causal inferences about communicators and their effect on opinion change. *Journal of Personality and Social Psychology*, 36(4):424, 1978.

[90] N. Ellison, J. Vitak, R. Gray, and C. Lampe. Cultivating social resources: The relationship between bridging social capital and facebook use among adults. *Journal of Computer-Mediated Communication*, 2011.

[91] N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.

[92] N. B. Ellison, J. Vitak, R. Gray, and C. Lampe. Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. *Journal of Computer-Mediated Communication*, 19(4):855–870, 2014.

[93] D. L. Espelage and S. M. Swearer. *Bullying in American schools: A social-ecological perspective on prevention and intervention*. routledge, 2004.

[94] R. D. Fallot. The impact on mood of verbal reminiscing in later adulthood. *The International Journal of Aging and Human Development*, 10(4):385–400, 1980.

[95] C. Fellbaum. *WordNet*. Wiley Online Library, 1998.

[96] D. Fitton, B. Bell, J. C. Read, O. Iversen, L. Little, and M. Horton. Understanding teen ux: building a bridge to the future. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pages 79–82. ACM, 2014.

[97] E. B. Foa, L. Cashman, L. Jaycox, and K. Perry. The validation of a self-report measure of posttraumatic stress disorder: The posttraumatic diagnostic scale. *Psychological assessment*, 9(4):445, 1997.

[98] A. L. Forest and J. V. Wood. When social networking is not working individuals with low self-esteem recognize but do not reap the benefits of self-disclosure on facebook. *Psychological science*, page 0956797611429709, 2012.

[99] N. Friedkin. A test of structural features of granovetter's strength of weak ties theory. *Social networks*, 2(4):411–422, 1980.

[100] R. Galioto and J. H. Crowther. The effects of exposure to slender and muscular images on male body dissatisfaction. *Body image*, 10(4):566–573, 2013.

[101] G. Gigerenzer, P. M. Todd, t. ABC Research Group, et al. *Simple heuristics that make us smart*. Oxford University Press, 1999.

[102] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 211–220. ACM, 2009.

[103] T. Gilovich, D. Griffin, and D. Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002.

[104] K. Glasgow, C. Fink, and J. L. Boyd-Graber. " our grief is unspeakable": Automatically measuring the community impact of a tragedy. In *ICWSM*, 2014.

[105] A. L. Gonzales. Text-based communication influences self-esteem more than face-to-face or cellphone communication. *Computers in Human Behavior*, 39:197–203, 2014.

[106] A. Good, C. Wilson, C. Ancient, and A. Sambhanthan. A proposal to support wellbeing in people with borderline personality disorder: applying reminiscent theory in a mobile app. *arXiv preprint arXiv:1302.5200*, 2013.

[107] E. Goode. The ethics of deception in social research: A case study. *Qualitative sociology*, 19(1):11–33, 1996.

[108] L. A. Goodman, C. Corcoran, K. Turner, N. Yuan, and B. L. Green. Assessing traumatic event exposure: General issues and preliminary findings for the stressful life events screening questionnaire. *Journal of traumatic stress*, 11(3):521–542, 1998.

[109] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.

[110] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.

[111] M. S. Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973.

[112] R. Grenoble. Amanda todd: Bullied canadian teen commits suicide after prolonged battle online and in school, october 2012.

[113] E. Grossi, A. Compare, and S. F. PGWBI. Psychological general well-being index (pgwb). *Encyclopedia of Quality of Life and Well-Being Research. Netherlands: Springer*, pages 5152–6, 2014.

[114] T. Grüne-Yanoff and R. Hertwig. Nudge versus boost: how coherent are policy and theory? *Minds and Machines*, 26(1-2):149–183, 2016.

[115] M. L. Guha, A. Druin, G. Chipman, J. A. Fails, S. Simms, and A. Farber. Mixing ideas: a new technique for working with young children as design partners. In *Proceedings of the 2004 conference on Interaction design and children: building a community*, pages 35–42. ACM, 2004.

[116] T. Habermas and C. Paha. Souvenirs and other personal objects: Reminding of past events and significant others in the transition to university. *Critical advances in reminiscence work*, pages 123–138, 2002.

[117] M. Hauben and R. Hauben. Netizens: On the history and impact of usenet and the internet. *First Monday*, 3(7), 1998.

[118] D. M. Hausman and B. Welch. Debate: To nudge or not to nudge. *Journal of Political Philosophy*, 18(1):123–136, 2010.

[119] K. Hawton, K. E. Saunders, and R. C. O'Connor. Self-harm and suicide in adolescents. *The Lancet*, 379(9834):2373–2382, 2012.

[120] R. A. Hayes, C. T. Carr, and D. Y. Wohn. One click, many meanings: Interpreting paralinguistic digital affordances in social media. *Journal of Broadcasting & Electronic Media*, 60(1):171–187, 2016.

[121] W. Heirman and M. Walrave. Assessing concerns and issues about the mediation of technology in cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 2(2):1–12, 2008.

[122] J. Hicks, N. Ramanathan, D. Kim, M. Monibi, J. Selsky, M. Hansen, and D. Estrin. Andwellness: an open mobile system for activity and experience sampling. In *Wireless Health 2010*, pages 34–43. ACM, 2010.

[123] S. Hinduja and J. W. Patchin. Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, 14(3):206–221, 2010.

[124] S. Hinduja and J. W. Patchin. Social influences on cyberbullying behaviors among middle and high school students. *Journal of youth and adolescence*, 42(5):711–722, 2013.

[125] S. Hinduja and J. W. Patchin. *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Corwin Press, 2014.

[126] E. E. Hollenbaugh and M. K. Everett. The effects of anonymity on self-disclosure in blogs: An application of the online disinhibition effect. *Journal of Computer-Mediated Communication*, 18(3):283–302, 2013.

[127] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra. Analyzing negative user behavior in a semi-anonymous social network. *arXiv preprint arXiv:1404.3839*, 2014.

[128] H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and A. Ghasemianlangroodi. Towards understanding cyberbullying behavior in a semi-anonymous social network. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 244–252. IEEE, 2014.

[129] H. Hosseinmardi, S. Li, Z. Yang, Q. Lv, R. Ibn Rafiq, R. Han, and S. Mishra. A comparison of common users across instagram and ask. fm to better understand cyberbullying. In *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*, pages 355–362. IEEE, 2014.

[130] L. Humphreys. Tearoom trade. *Society*, 7(3):10–25, 1970.

[131] E. Isaacs, A. Konrad, A. Walendowski, T. Lennig, V. Hollis, and S. Whittaker. Echoes from the past: how technology mediated reflection improves well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1071–1080. ACM, 2013.

[132] E. M. Jaffe. Cyberbullies beware: Reconsidering vosburg v. putney in the internet age. *Charleston L. Rev.*, 5:379, 2010.

[133] A. N. Joinson. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, 31(2):177–192, 2001.

[134] J. J. Jones, J. E. Settle, R. M. Bond, C. J. Fariss, C. Marlow, and J. H. Fowler. Inferring tie strength from online directed behavior. *PloS one*, 8(1):e52168, 2013.

[135] D. B. Kandel, V. H. Raveis, and M. Davies. Suicidal ideation in adolescence: Depression, substance use, and other risk factors. *Journal of Youth and Adolescence*, 20(2):289–309, 1991.

[136] R. Kang, S. Brown, and S. Kiesler. Why do people seek anonymity on the internet?: informing policy and design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2657–2666. ACM, 2013.

[137] R. Kang, L. Dabbish, and K. Sutton. Strangers on your phone: Why people use anonymous communication applications. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 359–370. ACM, 2016.

[138] A. M. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[139] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business horizons*, 54(3):241–251, 2011.

[140] E. Killackey, A. L. Anda, M. Gibbs, M. Alvarez-Jimenez, A. Thompson, P. Sun, and G. N. Baksheev. Using internet enabled mobile devices and social networking technologies to promote exercise as an intervention for young first episode psychosis patients. *BMC psychiatry*, 11(1):80, 2011.

[141] Y. S. Kim and B. Leventhal. Bullying and suicide. a review. *International Journal of Adolescent Medicine and Health*, 20(2):133–154, 2008.

[142] Y. S. Kim, B. L. Leventhal, Y.-J. Koh, and W. T. Boyce. Bullying increased suicide risk: prospective study of korean adolescents. *Archives of suicide research*, 13(1):15–30, 2009.

[143] H. King. Facebook's 'secret conversations' mode deletes messages for extra security, October 2016.

[144] A. Konrad, E. Isaacs, and S. Whittaker. Technology-mediated memory: Is technology altering our memories and interfering with well-being? *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(4):23, 2016.

[145] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards. Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 195–204. ACM, 2013.

[146] M. Kosters and J. Van der Heijden. From mechanism to virtue: Evaluating nudge theory. *Evaluation*, 21(3):276–291, 2015.

[147] H. Krasnova, T. Widjaja, P. Buxmann, H. Wenninger, and I. Benbasat. Research notewhy following friends can hurt you: an exploratory investigation of the effects of envy on social networking sites among college-age users. *Information systems research*, 26(3):585–605, 2015.

[148] H. Kwak, H. Chun, and S. Moon. Fragile online relationship: a first look at unfollow dynamics in twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1091–1100. ACM, 2011.

[149] G. C. E. Kwan and M. M. Skoric. Facebook bullying: An extension of battles in school. *Computers in human behavior*, 29(1):16–25, 2013.

[150] N. Lang. Why teens are leaving facebook: Its meaningless, February 2014.

[151] A. Lanza, A. League, A. Spring, B. Streisand, B. Hot, F. S. Army, J. Lopez, J. D. R. Saavedra, J. Belcher, K. Annan, et al. Syrian civil war. *Timeline*, 2011.

[152] N. Lapidot-Lefler and A. Barak. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior*, 28(2):434–443, 2012.

[153] R. Larson and M. Csikszentmihalyi. The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, 1983.

[154] A. Leavitt. 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*, pages 317–327. ACM, 2015.

[155] W. Leggett. The politics of behaviour change: Nudge, neoliberalism and the state. *Policy & Politics*, 42(1):3–19, 2014.

[156] A. Lenhart. Cyberbullying. *Pew Internet & American Life Project*, 2007.

[157] A. Lenhart. Teens, social media, and technology overview 2015. *Pew Internet Project*, 2015.

[158] A. Lenhart, M. Madden, A. Smith, K. Purcell, K. Zickhur, and L. Rainie. Teens, kindness and cruelty on social network sites. *Pew Internet Project*, 2011.

[159] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr. Social media & mobile internet use among teens and young adults. millennials. *Pew Internet & American Life Project*, 2010.

[160] M. P. Levine and K. Harrison. Effects of media on eating disorders and body image. *Media effects: Advances in theory and research*, pages 490–516, 2009.

[161] T. Lewin. Teenage insults, scrawled on web, not on walls. *The New York Times*, page A1, 2010.

[162] C. N. Lewis. Reminiscing and self-concept in old age. *Journal of gerontology*, 1971.

[163] N. Lin, P. W. Dayton, and P. Greenwald. Analyzing the instrumental use of relations in the context of social structure. *Sociological Methods & Research*, 7(2):149–166, 1978.

[164] M. Lindsay and J. Krysik. Online harassment among college students: A replication incorporating new internet trends. *Information, Communication & Society*, 15(5):703–719, 2012.

[165] S. Livingstone and E. Helsper. Gradations in digital inclusion: children, young people and the digital divide. *New media & society*, 9(4):671–696, 2007.

[166] M. Madden, A. Lenhart, S. Cortesi, U. Gasser, M. Duggan, A. Smith, and M. Beaton. Teens, social media, and privacy. *Pew Internet & American Life Project*, 2013.

[167] C. K. Malecki and M. K. Demaray. The role of social support in the lives of bullies, victims, and bully-victims. *Bullying in American schools: A social-ecological perspective on prevention and intervention*, pages 211–225, 2004.

[168] P. V. Marsden. Core discussion networks of americans. *American sociological review*, pages 122–131, 1987.

[169] A. E. Marwick and d. boyd. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1):114–133, 2011.

[170] A. E. Marwick et al. The drama! teen conflict, gossip, and bullying in networked publics. 2011.

[171] D. P. McAdams and C. A. Constantian. Intimacy and affiliation motives in daily living: An experience sampling analysis. *Journal of personality and social psychology*, 45(4):851–861, 1983.

[172] R. McCormick. Instagram's anti-abuse comment filter is rolling out now, August 2016.

[173] M. L. McCreary, L. A. Slavin, and E. J. Berry. Predicting problem behavior and self-esteem among african american adolescents. *Journal of Adolescent Research*, 11(2):216–234, 1996.

[174] M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[175] P. E. McKnight and J. Najab. Mann-whitney u test. *Corsini Encyclopedia of Psychology*, 2010.

[176] J. H. McLaughlin. Crime and punishment: Teen sexting in context. *Penn St. L. Rev.*, 115:135, 2010.

[177] E. Menesini, A. Nocentini, B. E. Palladino, A. Frisén, S. Berne, R. Ortega-Ruiz, J. Calmaestra, H. Scheithauer, A. Schultze-Krumbholz, P. Luik, et al. Cyberbullying definition among adolescents: A comparison across six european countries. *Cyberpsychology, Behavior, and Social Networking*, 15(9):455–463, 2012.

[178] K. W. Merrell, B. A. Gueldner, S. W. Ross, and D. M. Isava. How effective are school bullying intervention programs? a meta-analysis of intervention research. *School Psychology Quarterly*, 23(1):26, 2008.

[179] G. S. Mesch. Parental mediation, online activities, and cyberbullying. *CyberPsychology & Behavior*, 12(4):387–393, 2009.

[180] E. Milvy. Eavesdropping on the seventh grade instagram show, 2015.

[181] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[182] T. R. Mitchell, L. Thompson, E. Peterson, and R. Cronk. Temporal adjustments in the evaluation of events: The rosy view. *Journal of Experimental Social Psychology*, 33(4):421–448, 1997.

[183] M. J. Moore, T. Nakano, A. Enomoto, and T. Suda. Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior*, 28(3):861–867, 2012.

[184] M. Mori, K. F. MacDorman, and N. Kageki. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012.

[185] A. Muise, E. Christofides, and S. Desmarais. More information than you ever wanted: Does facebook bring out the green-eyed monster of jealousy? *CyberPsychology & behavior*, 12(4):441–444, 2009.

[186] M. J. Muller, D. M. Wildman, and E. A. White. equal opportunity pd using pictive. *Communications of the ACM*, 36(6):64, 1993.

[187] T. R. Nansel, M. Overpeck, R. S. Pilla, W. J. Ruan, B. Simons-Morton, and P. Scheidt. Bullying behaviors among us youth: Prevalence and association with psychosocial adjustment. *Jama*, 285(16):2094–2100, 2001.

[188] S. M. Nasehi, J. Sillito, F. Maurer, and C. Burns. What makes a good code example?: A study of programming q&a in stackoverflow. In *Software Maintenance (ICSM), 2012 28th IEEE International Conference on*, pages 25–34. IEEE, 2012.

[189] P. Naylor and H. Cowie. The effectiveness of peer support systems in challenging school bullying: the perspectives and experiences of teachers and pupils. *Journal of adolescence*, 22(4):467–479, 1999.

[190] C. Newton. Twitter begins filtering abusive tweets out of your replies, February 2017.

[191] D. Nicolalde and P. Brennan. Involving young adults in the design of health interventions. In *CHI 14 Workshop on Understanding Teen UX: Building a Bridge to the Future*, 2014.

[192] H. Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

[193] H. Nissenbaum. A contextual approach to privacy online. *Daedalus*, 140(4):32–48, 2011.

[194] F. H. Norris, M. J. Friedman, and P. J. Watson. 60,000 disaster victims speak: Part ii. summary and implications of the disaster mental health research. *Psychiatry: Interpersonal and biological processes*, 65(3):240–260, 2002.

[195] J. Nott et al. Video gaming on the feminist frontlines. 2014.

[196] S. O'Hear. Ask.fm claims its overtaken formspring. *http://techcrunch.com/2012/06/27/ask-fm-claims-its-overtaken-qagiant-formspring-whats-going-on-here*, 2012.

[197] G. S. O'Keeffe, K. Clarke-Pearson, et al. The impact of social media on children, adolescents, and families. *Pediatrics*, 127(4):800–804, 2011.

[198] D. Olweus. *Bullying at school*. Springer, 1994.

[199] J. Palfrey and U. Gasser. *Born digital: Understanding the first generation of digital natives*. Basic Books, 2013.

[200] G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[201] N. Pappas, G. Katsimpras, and E. Stamatatos. Distinguishing the popularity between topics: a system for up-to-date opinion retrieval and mining in the web. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 197–209. Springer, 2013.

[202] N. Park, K. F. Kee, and S. Valenzuela. Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes. *CyberPsychology & Behavior*, 12(6):729–733, 2009.

[203] M. Pasupathi and L. L. Carstensen. Age and emotional experience during mutual reminiscing. *Psychology and aging*, 18(3):430, 2003.

[204] M. Q. Patton. *Qualitative research*. Wiley Online Library, 2005.

[205] S. A. Paul, L. Hong, and E. H. Chi. Is twitter a good place for asking questions? a characterization study. In *ICWSM*, 2011.

[206] S. T. Peesapati, V. Schwanda, J. Schultz, M. Lepage, S.-y. Jeong, and D. Cosley. Pensieve: supporting everyday reminiscence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2027–2036. ACM, 2010.

[207] A. D. Pellegrini and J. D. Long. A longitudinal study of bullying, dominance, and victimization during the transition from primary school through secondary school. *British journal of developmental psychology*, 20(2):259–280, 2002.

[208] R. M. Perloff. Social media effects on young womens body image concerns: Theoretical perspectives and an agenda for research. *Sex Roles*, 71(11-12):363–377, 2014.

[209] E. Poole. Hey girls, did you know? slut-shaming on the internet needs to stop. *USFL Rev.*, 48:221–221, 2013.

[210] J. J. Prochaska, M. W. Rodgers, and J. F. Sallis. Association of parent and peer support with adolescent physical activity. *Research quarterly for exercise and sport*, 73(2):206–210, 2002.

[211] P. PROTECTION. Childrens online privacy protection act. 2002.

[212] R. Putnam. Social capital: Measurement and consequences. *Canadian Journal of Policy Research*, 2(1):41–51, 2001.

[213] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. *ICWSM*, 10:1–1, 2010.

[214] N. Rascle, M. Bruchon-Schweitzer, and I. G. Sarason. Short form of sarason's social support questionnaire: French adaptation and validation. *Psychological Reports*, 97(1):195–202, 2005.

[215] K. Raynes-Goldie. Aliases, creeping, and wall cleaning: Understanding privacy in the age of facebook. *First Monday*, 15(1), 2010.

[216] K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 241–244. IEEE, 2011.

[217] H. Rheingold. *The virtual community: Homesteading on the electronic frontier*. MIT Press, 2000.

[218] I. Rivers and P. K. Smith. Types of bullying behaviour and their correlates. *aggressive Behavior*, 20(5):359–368, 1994.

[219] S. Robson and L. Warren. 'can you kill yourself already?' the vile online messages from internet trolls 'that led girl, 16, to hang herself', December 2012.

[220] J. B. Rollman, K. Krug, and F. Parente. The chat room phenomenon: Reciprocal communication in cyberspace. *CyberPsychology and Behavior*, 3(2):161–166, 2000.

[221] M. Rosenberg. *Society and the adolescent self-image (rev.* 1989.

[222] M. Rosenberg, C. Schooler, and C. Schoenbach. Self-esteem and adolescent problems: Modeling reciprocal effects. *American Sociological Review*, pages 1004–1018, 1989.

[223] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 20(1):23–38, 1998.

[224] M. C. Ruedy. Repercussions of a myspace teen suicide: Should anti-cyberbullying laws be created. *NCJL & Tech.*, 9:323, 2007.

[225] D. Russell, L. A. Peplau, and M. L. Ferguson. Developing a measure of loneliness. *Journal of personality assessment*, 42(3):290–294, 1978.

[226] C. D. Ryff. Happiness is everything, or is it? explorations on the meaning of psychological well-being. *Journal of personality and social psychology*, 57(6):1069, 1989.

[227] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 771–780. ACM, 2006.

[228] I. G. Sarason, H. M. Levine, R. B. Basham, and B. R. Sarason. Assessing social support: the social support questionnaire. *Journal of personality and social psychology*, 44(1):127, 1983.

[229] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

[230] C. Sas and S. Whittaker. Design for forgetting: disposing of digital possessions after a breakup. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1823–1832. ACM, 2013.

[231] M. Scaife and Y. Rogers. Kids as informants: Telling us what we didnt know or confirming what we knew already. *The design of childrens technology*, pages 27–50, 1999.

[232] C. Schaefer, J. C. Coyne, and R. S. Lazarus. The health-related functions of social support. *Journal of behavioral medicine*, 4(4):381–406, 1981.

[233] H. J. Schau and M. C. Gilly. We are what we post? self-presentation in personal web space. *Journal of consumer research*, 30(3):385–404, 2003.

[234] A. M. Schenk and W. J. Fremouw. Prevalence, psychological impact, and coping of cyberbully victims among college students. *Journal of School Violence*, 11(1):21–37, 2012.

[235] S. K. Schneider, L. O'Donnell, A. Stueve, and R. W. Coulter. Cyberbullying, school bullying, and psychological distress: A regional census of high school students. *American Journal of Public Health*, 102(1):171–177, 2012.

[236] S. Y. Schoenebeck. The secret life of online moms: Anonymity and disinhibition on youbemom. com. In *ICWSM*. Citeseer, 2013.

[237] D. A. Schön. *The reflective practitioner: How professionals think in action*, volume 5126. Basic books, 1983.

[238] V. Schwanda Sosik, X. Zhao, and D. Cosley. See friendship, sort of: How conversation and digital traces might support reflection on friendships. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1145–1154. ACM, 2012.

[239] F. Scogin and L. McElreath. Efficacy of psychosocial treatments for geriatric depression: a quantitative review. *Journal of consulting and clinical psychology*, 62(1):69, 1994.

[240] A. Ševčíková and D. Šmahel. Online harassment and cyberbullying in the czech republic. *Zeitschrift für Psychologie/Journal of Psychology*, 2015.

[241] B. Shaul. Honestly looks to combat cyberbullying on ios, android. 2015.

[242] J. Shute. Cyberbullying suicides: What will it take to have ask.fm shut down?, August 2013.

[243] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237. Springer, 2002.

[244] R. Slonje and P. K. Smith. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology*, 49(2):147–154, 2008.

[245] P. Smith, J. Mahdavi, M. Carvalho, and N. Tippett. An investigation into cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying. *Research Brief No. RBX03-06. London: DfES*, 2006.

[246] P. K. Smith and S. Sharp. *School bullying: Insights and perspectives.* ERIC, 1994.

[247] R. H. Smith and S. H. Kim. Comprehending envy. *Psychological bulletin*, 133(1):46, 2007.

[248] L. Smith-Spark. Hanna smith suicide fuels calls for action on ask. fm cyberbullying, cnn. 2013.

[249] A. Sourander, A. B. Klomek, M. Ikonen, J. Lindroos, T. Luntamo, M. Koskelainen, T. Ristkari, and H. Helenius. Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Archives of general psychiatry*, 67(7):720–728, 2010.

[250] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315. ACM, 2004.

[251] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*, pages 177–184. IEEE, 2010.

[252] J. Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.

[253] L. K. Suzuki and J. P. Calzo. The search for peer advice in cyberspace: An examination of online teen bulletin boards about health and sexuality. *Journal of applied developmental psychology*, 25(6):685–698, 2004.

[254] R. H. Thaler and C. R. Sunstein. Nudge: Improving decisions about health, wealth, and happiness.

[255] R. H. Thaler and C. R. Sunstein. Libertarian paternalism. *The American Economic Review*, 93(2):175–179, 2003.

[256] D. R. Thomas. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246, 2006.

[257] A. Thompson. Teens report onslaught of bullying during divisive election, January 2017.

[258] R. S. Tokunaga. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior*, 26(3):277–287, 2010.

[259] C. L. Toma and J. T. Hancock. Self-affirmation underlies facebook use. *Personality and Social Psychology Bulletin*, 39(3):321–331, 2013.

[260] J. W. Treem and P. M. Leonardi. Social media use in organizations. *Communication Yearbook 36*, 36:143–189, 2012.

[261] J. W. Turner, J. A. Grube, and J. Meyers. Developing an optimal match within online communities: An exploration of cmc support communities and traditional support. *Journal of Communication*, 51(2):231–251, 2001.

[262] P. M. Valkenburg and J. Peter. Social consequences of the internet for adolescents a decade of research. *Current Directions in Psychological Science*, 18(1):1–5, 2009.

[263] P. M. Valkenburg, J. Peter, and A. P. Schouten. Friend networking sites and their relationship to adolescents' well-being and social self-esteem. *CyberPsychology & Behavior*, 9(5):584–590, 2006.

[264] B. A. Van der Kolk and A. C. McFarlane. *Traumatic stress: The effects of overwhelming experience on mind, body, and society*. Guilford Press, 2012.

[265] H. Vandebosch and K. Van Cleemput. Cyberbullying among youngsters: Profiles of bullies and victims. *New media & society*, 11(8):1349–1371, 2009.

[266] J. Vitak. The impact of context collapse and privacy on social network site disclosures. *Journal of Broadcasting & Electronic Media*, 56(4):451–470, 2012.

[267] J. Vitak, K. Chadha, L. Steiner, and Z. Ashktorab. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1231–1245. ACM, 2017.

[268] J. Vitak and J. Kim. You can't block people offline: examining how facebook's affordances shape the disclosure process. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 461–474. ACM, 2014.

[269] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. I regretted the minute i pressed share: A qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, page 10. ACM, 2011.

[270] J. D. Webster and M. E. McCall. Reminiscence functions across adulthood: A replication and extension. *Journal of Adult development*, 6(1):73–85, 1999.

[271] K. R. Wentzel. Social relationships and motivation in middle school: The role of parents, teachers, and peers. *Journal of educational psychology*, 90(2):202, 1998.

[272] M. White. The challenges of building a compassionate robot, November 2014.

[273] J. L. Whitlock, J. L. Powers, and J. Eckenrode. The virtual cutting edge: the internet and adolescent self-injury. *Developmental psychology*, 42(3):407, 2006.

[274] T. M. Wilkinson. Nudging and manipulation. *Political Studies*, 61(2):341–355, 2013.

[275] N. Willard. Educators guide to cyberbullying and cyberthreats. *Center for safe and responsible use of the Internet*, 2007.

[276] T. A. Wills, J. A. Resko, M. G. Ainette, and D. Mendoza. Role of parent support and peer support in adolescent substance use: a test of mediated effects. *Psychology of Addictive Behaviors*, 18(2):122, 2004.

[277] D. A. Wolfe and P. G. Jaffe. Emerging strategies in the prevention of domestic violence. *The future of children*, pages 133–144, 1999.

[278] B. Xu, P. Chang, C. L. Welker, N. N. Bazarova, and D. Cosley. Automatic archiving versus default deletion: What snapchat tells us about ephemerality in design. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1662–1675. ACM, 2016.

[279] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666. Association for Computational Linguistics, 2012.

[280] K. Ybarra, boyd d and O. J. Defining and measuring cyberbullying within the larger context of bullying victimization. *Journal of Adolescent Health*, 51(1):53–58, 2012.

[281] M. L. Ybarra, M. Diener-West, and P. J. Leaf. Examining the overlap in internet harassment and school bullying: Implications for school intervention. *Journal of Adolescent Health*, 41(6):S42–S50, 2007.

[282] J. Yip, T. Clegg, E. Bonsignore, H. Gelderblom, E. Rhodes, and A. Druin. Brownies or bags-of-stuff?: domain expertise in cooperative inquiry with children. In *Proceedings of the 12th International Conference on Interaction Design and Children*, pages 201–210. ACM, 2013.

[283] D. Young. Now you see it, now you dont... or do you?: Snapchats deceptive promotion of vanishing messages violates federal trade commission regulations, 30 j. marshall j. info. tech. & privacy l. 827 (2014). *The John Marshall Journal of Information Technology & Privacy Law*, 30(4):6, 2014.

[284] J. F. Young, K. Berenson, P. Cohen, and J. Garcia. The role of parent and peer support in predicting adolescent depression: A longitudinal community study. *Journal of Research on Adolescence*, 15(4):407–423, 2005.

[285] H. Yun. *The creation and validation of a perceived anonymity scale based on the social information processing model and its nomological network test in an online social support community*. ProQuest, 2006.

[286] J. M. Zelenski and R. J. Larsen. The distribution of basic emotions in everyday life: A state and trait perspective from experience sampling data. *Journal of Research in Personality*, 34(2):178–197, 2000.

[287] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.

[288] M. Zimmer. but the data is already public: on the ethics of research in facebook. *Ethics and information technology*, 12(4):313–325, 2010.