Computational Methods to Understand and Design for Deviant Mental Health Communities

Stevie Chancellor

Areas of Relevance: HCI, Data Science

Introduction and Motivations: Social media has changed how individuals cope with health challenges in good and bad ways. Especially for severely stigmatized mental health conditions like depression, online groups provide guidance, advice, and a sense of community [1]. However, individuals frequently turn to online apps and communities to promote deliberate self-injury, disordered eating habits, and suicidal ideation. These behaviors can have latent, "contagion" effects on others [2] as well as on social networks that struggle with managing such dangerous content [3].

I call these behaviors "deviant" – actions that violate the social norms and behaviors of a particular community. Deviant behaviors that promote self-injury violate platform policies as well as social expectations that individuals do not harm themselves. To precisely identify and manage these pernicious issues is a challenge by itself as is designing sensitive and appropriate interventions. Identifying and reacting to such issues becomes even more difficult when peer-to-peer technologies are appropriated for sharing such information in private.

My doctoral research investigates online deviant mental health communities with computational social science techniques at scale. Using large-scale social media datasets and techniques like machine learning and statistical modeling, I analyze and understand patterns of deviant behavior, these community's connections to mental health and support, and platform ethics in dealing with this behavior. Drawing on my prior and current work, I want to work with Snapchat to design data-driven interventions to facilitate intimate disclosure and support about mental health challenges.

Prior work: My body of work explores a specific deviant mental wellness community -- *the pro-eating disorder* (pro-ED) community, a clandestine group that glorifies excessive calorie restriction, binging/purging, and excessive exercise. One project examined the impacts of content moderation policies on Instagram pro-ED tags and how the community responded to these bans. Using a dataset of 8 million Instagram posts, I found that these communities develop elaborate lexical variants to avoid the bans, changing tags from "thighgap" to "thyghgappp,'" and alarmingly discuss more toxic, vulnerable, and self-harm content [3]. Using machine learning, two projects additionally have looked at posts removed or deleted from Instagram in the pro-ED community [4]. I've also built hybrid human and deep learning models to identify content that violates Tumblr's Community Guidelines [5].

Current and Future Work: Given that individuals use social networks for these deviant behaviors, how can we use at-scale, computational techniques to design new strategies to promote healthier behaviors? In the projects I outline, I discuss the ways that will answer these questions - blending data-driven analysis with design to better support individuals dealing with deviant mental health behaviors.

Part of my dissertation considers the perspectives of human moderators who manage deviant mental health content. I am partnering with a Reddit crisis support community that assists people at risk of committing suicide. Moderators in this community are overwhelmed with the emotional severity and volume of posts, and need a way to quickly assess who needs

what kind of assistance. I will build a mixed-methods labeling system to identify the riskiest posts in this crisis community. The first portion of the system will teach Mechanical Turk workers without any experience to identify risky behaviors. In prior work, I found that novice and clinical annotators had nearly perfect inter-rater reliability at identifying dangerous content [6]. I extend this finding to design methods to teach Turkers how to label posts to improve the quality of machine learning later on. Using these ratings to build a supervised machine learning algorithm, I will provide an urgency rating that moderators can use to better direct their efforts.

Future Work with Snapchat: I see myself working with Snapchat to promote supportive interactions for mental wellness self-disclosure with friends as well as trained counselors. My focus on deviant behavior in online communities is a lens for understanding both bad and good behaviors to improve online platforms [7]. How does my work on public communities extend to the unique context of Snapchat, where peer-to-peer communication dominates?

One future project to that end is designing for better support when friends disclose mental health challenges over Snapchat. Intimate disclosure of mental health issues to friends and family helps alleviate negative emotions [8]; however, many individuals are not equipped to empathetically handle these disclosures, especially when they are troubling or "deviant"[. I would use a data-driven approach to understand how disclosure works in a peer-to-peer setting – how do individuals disclose on Snapchat, if at all, and how do friends respond? This would be a mixed methods study that combines large-scale analysis of disclosures and responses as well as an interview study. Using the results of this as a guide, I will help design interaction techniques that would facilitate better mental health support for disclosers and friends who offer support. These data-driven design strategies complement Snapchat's goals of being an authentic, intimate platform for people to share their lives, from the mundane to the playful to the serious.

Second, Snapchat's platform might also facilitate collaborations with crisis support providers through thoughtfully designed emotional support lines. Snapchat is primarily used by teenagers and young adults, who have some of the highest rates of mental wellness challenges [9]. Snapchat is also an excellent platform that combines audio/visual short messages that can convey emotions more deeply than just text-based platforms. I envision a partnership with current online providers for mental health crises, like 7 Cups [10], to deliver emotional support for people struggling. For individuals who want support, these trained talk counselors could help individuals work through their feelings or empower them to seek formal assistance. This kind of work would advance the field in novel methods of online talk therapy as well as making Snapchat a positive influencer in the social media space to provide meaningful support around mental wellness.

Finally, Snapchat will need to tackle complex questions about false alarms or false positives, confidentiality, and the social obligations of intervention when dealing with sensitive and deviant mental health disclosure. There are no right answers in this space, and through my work, I hope to bring awareness to these ethical questions and potential ways that platforms can begin to address them.

Overall, my research on deviant mental health communities online impacts not only social networks, but also designing for better peer support. I'm excited to work with Snapchat to help support those with mental wellness challenges, and I'm grateful for being considered for the Fellowship.

Works Cited

[1] Jina Huh and Mark S Ackerman. 2012. Collaborative help in chronic disease management: supporting individualized problems. In CSCW.

[2] Dina LG Borzekowski, Summer Schenk, Jenny L Wilson, and Rebecka Peebles. 2010. e-Ana and e-Mia: A Content Analysis of Pro–Eating Disorder Web Sites. American journal of public health 100, 8 (2010), 1526.

[3] Stevie Chancellor, Jessica Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In CSCW.

[4] Stevie Chancellor, Zhiyuan Jerry Lin, and Munmun De Choudhury. 2016. "This Post Will Just Get Taken Down": Characterizing Removed Pro-Eating Disorder Social Media Content. In CHI.

[5] Stevie Chancellor, Yannis Kalantidis, Jessica A Pater, Munmun De Choudhury, and David A. Shamma. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In CHI.

[6] Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In CSCW.

[7] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA (2012).

[8] Adam N Joinson and Carina B Paine. 2007. Self-disclosure, privacy and the Internet. The Oxford handbook of Internet psychology (2007).

[9] Diane M Quinn and Stephenie R Chaudoir. 2009. Living with a concealable stigmatized identity: the impact of anticipated stigma, centrality, salience, and cultural stigma on psychological distress and health. Journal of personality and social psychology 97, 4 (2009), 634

[10] 7 Cups. https://www.7cups.com/.