“This Post Will Just Get Taken Down”: Characterizing Removed Pro-Eating Disorder Social Media Content

Stevie Chancellor
Georgia Tech
Atlanta GA 30332
schancellor3@gatech.edu

Zhiyuan (Jerry) Lin
Georgia Tech
Atlanta GA 30332
zlin48@gatech.edu

Munmun De Choudhury
Georgia Tech
Atlanta GA 30332
munmund@gatech.edu

ABSTRACT

Social media sites like Facebook and Instagram remove content that is against community guidelines or is perceived to be deviant behavior. Users also delete their own content that they feel is not appropriate within personal or community norms. In this paper, we examine characteristics of over 30,000 pro-eating disorder (pro-ED) posts that were at one point public on Instagram but have since been removed. Our work shows that straightforward signals can be found in deleted content that distinguish them from other posts, and that the implications of such classification are immense. We build a classifier that compares public pro-ED posts with this removed content that achieves moderate accuracy of 69%. We also analyze the characteristics in content in each of these post categories and find that removed content reflects more dangerous actions, self-harm tendencies, and vulnerability than posts that remain public. Our work provides early insights into content removal in a sensitive community and addresses the future research implications of the findings.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

deviant behavior; content removal; social media; eating disorders; health

INTRODUCTION

Social media platforms like Twitter, Instagram, and Facebook are used to discuss a variety of topics, ranging from the mundane to sensitive. Content and user accounts are sometimes moderated to both sustain a healthy community and prevent behavior that violates established rules or social norms [41, 19, 25, 7]. Current moderation practices include platform-enforced policies (e.g., deleting content or banning users) or community-driven assessments (e.g., voting or rating mechanisms, block functionality, reporting infringing content).

The photo-sharing social network Instagram has several moderation approaches for posts that breach their Community Guidelines. Instagram removes sexual photos, illegal behaviors, and spam[1]. It also prohibits self-harm and pro-eating disorder (pro-ED) content — a behavior that promotes eating disorders as legitimate lifestyle choices instead of dangerous psychosocial disorders[2]. Users can also remove their own posts that they feel are not appropriate either for the platform or for the established norms of the community. Whatever the reason, intent, or authority behind removal, users often acknowledge in their posts that, “*this post will just get taken down*”. We consider these removed posts as a prime example of deviant behavior [3] because these posts do not conform to the personal or collective norms that the individual, the platform, or the community may have established in tandem.

When individuals experience vulnerability, they tend to reach out to others to “buffer” themselves against negative emotions and actions [16]. To the pro-ED community, social media sites like Instagram provide such an outlet to seek out these kind of “safety valves” to regulate their emotions [1]. Users in this community frequently discuss negative events, like intents to deliberately self-injure [12, 33]; we suspect that the perceived sensitivity expressed in pro-ED posts that removed from the platform is higher than normal – a factor that likely underlies the removal. There is an important opportunity to learn the characteristics of this deviant content for several reasons: learning more about motivations behind post removal, how this removal behavior may carry over to other online communities or technologies, and how other technologies may identify characteristics of deviant behavior on a platform.

Our Contributions and Ethics. This paper examines characteristics of pro-ED posts removed from Instagram. We build a supervised learning approach, a binary logistic regression classifier, to distinguish between the content of public pro-ED posts and removed posts. From here on, we refer to these removed posts as “deviant posts”. We build our classifier on a sample of over 30,000 deviant and an equal number of public pro-ED posts. We find that the two classes of posts can be distinguished with satisfactory performance (accuracy of 69% and an area-under-curve measure of 76%). Interpreting the variables with high predictive power, we find that deviant pro-ED posts indeed show heightened vulnerability compared to public posts, and these content markers provide important insights into deviant behavior. We also discuss the implications, future work, and power of this research.

CHI'16, May 07-12, 2016, San Jose, CA, USA
©2016 ACM. ISBN 978-1-4503-3362-7/16/05 $15.00
DOI: http://dx.doi.org/10.1145/2858036.2858248

[1] https://help.instagram.com/477434105621119/

[2] http://blog.instagram.com/post/21454597658/instagrams-new-guidelines-against-self-harm

We recognize that analysis of deleted or removed content is controversial territory in social computing. It has been studied in prior work [37, 4, 14], and we appreciate the authors of [32] for providing insight into the ethics of this research. However, we believe that the important health benefits of our work for those in need justifies the study of removed content in pro-ED communities. In that light, our work extends efforts in the social computing research community where large-scale data analytic approaches have been adopted to extend timely and tailored support to vulnerable communities online [18, 15].

Nonetheless, we took precautions to conduct this study as ethically as possible. The data used in this paper was collected through the official Instagram API when the posts were publicly accessible. There is no research interaction with the users, and the data is only analyzed by the algorithms we develop—i.e., we adopt an "eyes off" quantitative analysis approach (e.g., [11, 30]), since, ethically, users would need to consent to have their deleted content read by researchers. Thus, we did not seek institutional review board approval. To protect the privacy of our participants, we do not disclose any algorithmic output that contains personally identifying information, including usernames or personal tags.

RELATED WORK

Substantial research in HCI and CSCW has examined characteristics of deviant content and deviant behavior, including those associated with content removal or deletion.

Online Deviance. Rafferty et al. [35] qualitatively examined cyberbullying and aggression as forms of deviance (also [27, 44]), and Wang et al. [42] studied the nature of abusive deviant content on Twitter. Other work has examined the dynamics of and motivation behind sharing deviant content [23, 31, 25, 10], the role of online identity choices and accountability in such content [41, 19, 40], and management of design considerations and mechanisms around deviance [9, 38, 39, 5, 2]. Deviant behavior and content sharing has also been studied in specific contexts, such as online gaming [8], online news commentary [6, 14], feminist forums [26], grief expression [34], and peer production communities [43, 36, 28, 2].

Content Deletion on Social Media. Sleeper et al. [37] studied post deletion practices of individuals around regretful experiences on Twitter. Almuhimedi et. al [4] quantitatively examined deleted tweets on Twitter, and Cheng et. al [14] looked at comments removed by moderators that were deemed antisocial and toxic to three online news communities. Close to our work is Chancellor et al. [13] that examined behavior patterns following banning of tags in pro-ED communities; however, this study did not explore the characteristics of deleted and removed pro-ED posts. We build on this larger body of work to offer some of the first quantitative insights into deviant content from a sensitive and controversial online community [22].

DATA AND METHODS

Data Collection

To allow enough removal time for deviant pro-ED posts, our data collection occurred in three phases over ten months. In all phases, we used the tools in the official Instagram API (https://instagram.com/developer/).

skinny	thin	thinspo	bonespo
eatingdisorder	probulimia	anorexia	thighgap
proanorexia	mia	bulimia	promia
thinspiration	secretsociety	ana	proana
anorexianervosa			

Table 1. Example tags used for crawling pro-ED posts in our study.

Phase I: Obtaining Pro-ED Data. In September 2014, we created a large initial sample of *public* posts with pro-ED tags. We consider these posts to be shared by the "pro-ED community" on Instagram; although Instagram does not have defined community structures, users organize around tags to create and share a common identity [24]. To identify pro-ED tags, we first curated a set of nine "seed tags"[3] found to be common tags and structures illustrating pro-ED behaviors and attitudes across different social media platforms [17]. Two researchers then manually inspected posts on each tag to ensure there was sufficient pro-ED content. We crawled these tags for a month, which returned 434K posts and 234K unique tags. We selected 222 tags that had at least a 1% co-occurrence rate with other tags in this dataset.

From this set of 222, we manually filtered tags that did not map to eating disorders for three reasons: (1) Tags that were related to eating disorders but were generic enough to be used in many other contexts, e.g. "beautiful" and "inspiration". (2) Tags related to other mental conditions or disorders, e.g. "depression", "anxious", or "suicidal". Lastly (3) Tags that were tied to the eating disorder recovery community, e.g. "edrecovery". This reduced the filtered co-occurrence tag list from 222 tags to 72 verified to related to eating disorders (see sample tags in Table 1).

In November 2014, we crawled all content across these 72 tags. This returned over 8 million posts between January 2011 and November 2014. We then removed any posts that were cross-posted to any recovery tags as well as any that had three specific tags ("mia", "ana", and "ed") that did not also contain another one of the initial list of 72. Qualitative observation indicated that these three tags used in isolation from pro-ED tags refers to first names (the name "Mia") or references to popular celebrities ("ed" for Ed Sheeran). Our dataset at the end of this phase had 6.5 million posts relating to pro-ED.

Phase II: Gathering Pro-ED Users and Post Timelines. In February 2015, we obtained a random sample of 100K active users from the authors of the 6.5 million posts above. We gathered the public timelines of each of the 100K users. This set contains over 26M posts from 100K users, with posts shared between October 2010 and March 2015.

Phase III: Gathering deviant pro-ED data. In August 2015, we used the Instagram API to check whether the posts from Phase II were still publicly accessible. We randomly sorted our posts and gathered the first 31K posts where the post was no longer available on the platform but the user's account still existed. Note that checking whether the user's account was still active was an important step to prevent confounding resulting from a post being deleted because the user removed their Instagram account altogether — we believe the characteristics of account deletion may be considerably distinct from

[3] Seed tags include: "ed", "eatingdisorder", "ednos", "ana", "anorexia", "anorexic","mia", "bulimia", and "bulimic".

those related to post removal. These 31K posts represent our deviant pro-ED posts for our classification task. To construct an equivalent still-public dataset of pro-ED posts, we gathered the first 31K random public posts, bringing the total number of posts in our dataset to 62K.

Characterization and Prediction Framework

Next, we develop several logistic regression models to learn the characteristics of deviant pro-ED posts as well as to automatically distinguish them from content that remains public. Regularization helps us control for collinearity (since we use the text of posts) and sparsity in our data. In our case, we used the model implementation given in the Python package statsmodels. Our response variable is the binary indicator of whether a post is deviant or still-public. For the predictor variables, we considered four different sets and build a model for each set. These variables capture straightforward linguistic constructs in the post's text. Note that we did not use the visual features of the photo or video included in the post, but it constitutes an important direction toward future work.

TagCt: uses the frequency of tags in a post as predictor variables; we consider all tags which occur 200 or more times, which gives 614 predictor variables.

TagCo: uses the 500 most frequent pairwise tag co-occurrences in posts[4], and the 614 tags with frequencies over 200 (from **TagCt**).

TagCtCo: uses 500 most frequent pairwise unigram co-occurrences in the captions of the posts, the 500 most frequent pairwise tag co-occurrences in posts, and the 614 tags with frequencies over 200 (from **TagCt**).

TagCtCoCap: uses 1,000 most frequent pairwise unigram co-occurrences in the captions of the posts, the 1,000 most frequent pairwise tag co-occurrences in posts, and the 614 tags with frequencies over 200 (from **TagCt**).

RESULTS

Our model used 80% of our data for training purposes and parameter tuning; the remaining 20% were heldout for testing.

On the training data, we first present the goodness of fit of our four models and how they fared against the Null model (Table 2). Compared to the Null model, all of our models provide considerable explanatory power (statistically significant based on Bonferroni correction) with significant reduction in deviances. Particularly, the TagCtCoCap model (that uses tag frequencies, tag co-occurrences and unigram and bigram tokens in post captions) yields the best fit. We find that the difference between the deviance of the Null model and the deviance of this model approximately follows a χ^2 distribution, with degrees of freedom equal to the number of additional variables in the latter model: $\chi^2(2613, N = 62K) = 107387 - 54487 = 5.29 \times 10^4, p < 10^{-10}$.

Next we report our results on the 20% heldout dataset. For the sake of brevity, we only report expanded performance metrics on the model with the best performance (the TagCtCoCap model) in Table 3. In the confusion matrix, class 1 is deviant posts and class 0 is public posts. We find that the TagCtCoCap model gives satisfactory accuracy in classifying both deviant and public pro-ED posts, with a mean precision, recall, and F-1 score of .69 each. The accuracy of the model is 69%, an improvement of 19% over a chance model (50% baseline accuracy due to balanced class sizes). We report the receiver operating characteristic (ROC) curve in Figure 1; the area under curve (AUC) is 76%.

Model	Deviance	df	χ^2	p-value
Null	107387	0		
TagCt	61029	613	9.43e+03	$< 10^{-7}$
TagCo	58266	1113	6.67e+03	$< 10^{-8}$
TagCtCo	57632	1613	6.03e+03	$< 10^{-8}$
TagCtCoCap	54487	2613	2.89e+03	$< 10^{-10}$

Table 2. Summary of different model fits. Null is the intercept-only model. All comparisons with the Null model are statistically significant after Bonferroni correction for multiple testing ($\alpha = \frac{0.05}{4}$).

Actual/Predicted	Class 0	Class 1	Total
Class 0	4433	1913	6346
Class 1	2018	4328	6346
Accuracy	69.85%	68.2%	69.03% (mean)
Precision	.69	.69	.69 (mean)
Recall	.70	.68	.69 (mean)
F-1	.69	.69	.69 (mean)

Table 3. Performance of the TagCtCoCap model in distinguishing deviant and public posts.

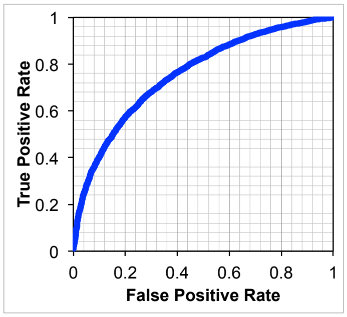

Figure 1. ROC curve for the tagCtCoCap model.

We also present 20 of the top 50 positive and negative β values of the TagCtCoCap model in Table 4. The positive and negative β values indicate increased likelihood of a post to be deviant or public, respectively. The variables that are most predictive of deviant posts are overwhelmingly associated with attitudes and behaviors that reinforce pro-ED lifestyles as well as self-injurious behaviors. These include “cutting”, “bodycheck” (where users invite others to suggest improvements for their body), and the desire to look or be skinny. There are also indicators of high vulnerability and threats to personal safety (“worthless”, “suicidal”, “razor”). In contrast, predictor variables that increase the likelihood of a post remaining public are closest to those reaching out to the eating disorder recovery community [45]. Public posts also emphasize a larger variety of emotions, cognitions, and confessions (“gourgeous”, “angry”, “misunderstood”).

DISCUSSION AND IMPLICATIONS FOR FUTURE WORK

Our findings show that characteristics in captions and tagging strategies of posts on Instagram can predict whether a post will be considered deviant on the platform and then removed. The beta weights displayed in Table 4 illustrates two important findings from our data. One, deviant posts have a higher chance of being related to self-injury, suicide, and the maintenance of eating disorders, confirming what has been observed in prior work [12]. This implies that, while these communities are allowed to exist on the platform, these particular deviant

[4] Two tags or unigrams co-occur if they are used together in a post.

Type / Content	β	Type / Content	β
TC/#different	2.22	TP/#edrecovery&#beated	-1.33
TC/#energy	1.95	TC/#eatittobeatit	-1.27
TC/#depressedteen	1.30	TC/#toned	-1.23
TP/#ana & #anorexia	1.10	TC/#nevergoodenough	-1.04
TC/#ptsd	1.08	TC/#angry	-0.97
TP/#anorexia & #anorexianervosa	1.00	TP/#ana & #edsoldiers	-0.92
TP/#cutting & #crying	0.98	TC/#misunderstood	-0.92
TP/#depression & #blade	0.94	CP/today & strong	-0.85
TC/#skinnyplease	0.90	CP/fat & depressed	-0.84
TP/#blade & #suicide	0.87	TC/#gourgeous	-0.80
CP/personal & account	0.86	TP/#anarecovery&#edfamily	-0.78
TP/#ana & #weightloss	0.85	TC/#edarmy	-0.77
TP/#sue&#anorexia	0.82	TP/#prorecovery &#anorexiarecovery	-0.71
TC/#harm	0.80	TC/#nourishnotpunish	-0.71
TP/#anxiety &#depressionquotes	0.80	TC/#eathealthy	-0.71
TC/#anagirl	0.79	TC/#ednosrecovery	-0.70
TC/#selfharmmm	0.79	CP/got & think	-0.70
TC/#bodycheck	0.79	CP/know & friend	-0.69
CP/want & look	0.77	TC/#prorecovery	-0.69

Table 4. Selected 20 out of the top 50 features' positive and negative beta weights in our `TagCtCoCap` model. There are 3 types of features: TC (tag with at least 200 occurrences), TP (co-occurring tag pair), and CP (co-occurring unigram pair in caption)

posts are either perceived negatively by the authors themselves later or are considered to violate community norms. Second, although further exploration is needed, we find that posts that remain public reach out to the pro-recovery community. This suggests a conscious effort by the community to tacitly endorse these posts, the users' ideas, and promoting a healthier lifestyle despite still being visible on pro-ED tags.

Implications

Because we can identify removed content with straightforward text characteristics, we foresee several exciting research opportunities as a result of this research.

Health and Just-In-Time Interventions. Some platforms have basic intervention systems to bring help to such vulnerable individuals. For example, Facebook's suicide intervention system prompts a user to contact a close friend or assistance hotline if another person reports their post to be suicidal[5]. Beyond these efforts, our predictors show that removed content overwhelmingly shows high vulnerability and threats to personal safety ("worthless", "suicidal", "razor"), making our classifier a good start to identifying moments for just-in-time intervention [29] to users who post this kind of content. Expansion of this approach to contain more cues would boost its accuracy and applicability to real-world systems. Potentials for interventions could include a prompt to direct users to contact a close friend or reach out to a specialist available on a hotline. This would be useful not only for pro-ED posts, but in other cases of mental illness.

Implications for Social Computing Research and HCI. Taking the meaning of intervention more broadly, our deviant post classifier can be used to facilitate better content moderation practices. Social media platforms could build tools leveraging our methodology to assist moderators in identifying deviant posts. Many platforms rely on the report/flag functionality to alert moderators of deviant content. Rather than waiting for reports to manually come in, our classifier could generate a list of posts that are potentially deviant, and then a human well-versed in the community guidelines of the platform can choose an outcome to the post. This way the system brings human judgment to an area where sensitivity and care are needed.

Broadly, through these investigations of deviant pro-ED content, social computing researchers can gain insights into the intent and motivation behind sharing of pro-ED content on a public social platform and the general goals of such sensitive self-disclosure. They can also understand how individuals are repurposing social media platforms in order to connect and bond with others with similar personal challenges relating to mental health, as well as to seek their help and support during moments of heightened physical or emotional vulnerability.

Extending beyond deviant health behavior, HCI researchers could incorporate our approach into analyzing deletion or removal practices behind other forms of user-generated deviant content, ranging from abuse, bullying and hate speech, including how different online communities shape their norms to tackle deviant content.

Limitations and Conclusion

Research has shown that communities like pro-ED can have detrimental effects on the health of the broader community, not just sufferers themselves [21]. Our algorithm could be implemented to help quickly remove deviant posts *before* the post author or the community perceives it to be deviant. However, great care must be taken when implementing any system that filters content. When posts are removed, what are the impacts on the users, their emotional state, and the broader goals of the community? At what point does automated removal shift from improving efficacy to chilling speech, removing important "safety valves" to dangerous behaviors [20], or potentially promote the discussion of more dangerous content [13]?

Future work will need to be delicately crafted to promote positive outcomes, as is the case with health communities. Collaborations between social media researchers, designers, ethicists, and psychologists will be essential in developing what actions are best for particular scenarios and when to deploy these interventions.

Given the constraints imposed by the Instagram API, we are not able to distinguish between post removal initiated by the individual and by the platform. Even though removed content provides a signal that such content is inappropriate and violates norms, some content may be removed due to mundane reasons such as a typo or accidental mis-post. This could potentially introduce noise into our model and analysis. Future work could expand this analysis to understanding deviant content in a deeper way, including how fast posts are removed from the platform as well as interviewing users on the motivation behind these practices.

This paper explored how timely, tailored support could be directed to sensitive online communities like pro-ED when content is removed from the platform. The moral and ethical grounds surrounding content moderation, banning, and removal are important conversations for the social computing community to consider going forward.

[5] http://www.washington.edu/news/2015/02/25/forefront-and-facebook-launch-suicide-prevention-effort/

REFERENCES

1. Jan Paul Acton. 1973. Evaluating public programs to save lives. (1973).

2. B Thomas Adler, Luca De Alfaro, Santiago M Mola-Velasco, Paolo Rosso, and Andrew G West. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Computational linguistics and intelligent text processing*. Springer, 277–288.

3. Ronald L Akers. 1977. Deviant behavior: A social learning approach. (1977).

4. Hazim Almuhimedi, Shomir Wilson, Bin Liu, Norman Sadeh, and Alessandro Acquisti. 2013. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 897–908.

5. Michael Bernstein, Michael Conover, Benjamin Mako Hill, Andres Monroy-Hernandez, Brian Keegan, Aaron Shaw, Sarita Yardi, R Stuart Geiger, and Amy Bruckman. 2012. Fail whaling: designing from deviance and failures in social computing. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1127–1130.

6. Amy Binns. 2012. DON'T FEED THE TROLLS! Managing troublemakers in magazines' online communities. *Journalism Practice* 6, 4 (2012), 547–562.

7. Jonathan Bishop. 2013. *Examining the Concepts, Issues, and Implications of Internet Trolling*. IGI Global.

8. Jeremy Blackburn and Haewoon Kwak. 2014. Stfu noob!: predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*. ACM, 877–888.

9. Amy Bruckman, Pavel Curtis, Cliff Figallo, and Brenda Laurel. 1994. Approaches to managing deviant behavior in virtual communities. In *Conference Companion on Human Factors in Computing Systems*. ACM, 183–184.

10. Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. 2014. Trolls just want to have fun. *Personality and individual Differences* 67 (2014), 97–102.

11. Moira Burke, Cameron Marlow, and Thomas Lento. 2010. Social network activity and social well-being. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1909–1912.

12. Stevie Chancellor, Zhiyuan (Jerry) Lin, Erica Goodman, Stephanie Zerwas, and Munmun De Choudhury. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*.

13. 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 *Proceedings of the 2016 Conference on Computer Supported Cooperative Work & Social Computing(CSCW)*. ACM. in press.

14. Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. In *International Conference on Weblogs and Social Media (ICWSM)*. AAAI.

15. Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *Proc. ICWSM*.

16. James C Coyne and Geraldine Downey. 1991. Social factors and psychopathology: Stress, social support, and coping processes. *Annual review of psychology* 42, 1 (1991), 401–425.

17. Munmun De Choudhury. 2015. Anorexia on Tumblr: A Characterization Study. In *Proc. Digital Health*.

18. Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*. ACM, 3267–3276.

19. Judith S Donath and others. 1999. Identity and deception in the virtual community. *Communities in cyberspace* 1996 (1999), 29–59.

20. T Emmens and A Phippen. 2010. Evaluating Online Safety Programs. *Harvard Berkman Center for Internet and Society.[23 July 2011]* (2010).

21. Christopher G Fairburn and Paul J Harrison. 2003. Eating disorders. *The Lancet* 361, 9355 (2003), 407–416.

22. David Giles. 2006. Constructing identities in cyberspace: The case of eating disorders. *British journal of social psychology* 45, 3 (2006), 463–477.

23. Lynne Hall and Carlisle E George. 1999. Law and Punishment in Virtual Communities. *Proceedings of Cybersociety* (1999).

24. Kyungsik Han, Jin Yea Jang, and Dongwon Lee. 2015. Exploring tag-based like networks. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1941–1946.

25. Claire Hardaker. 2010. Trolling in asynchronous computer-mediated communication: from user discussions to theoretical concepts. *Journal of Politeness Research* 6, 2 (2010), 215–242.

26. Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing" trolling" in a feminist forum. *The Information Society* 18, 5 (2002), 371–384.

27. Sameer Hinduja and Justin W Patchin. 2014. *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Corwin Press.

28. Sara Javanmardi, David W McDonald, and Cristina V Lopes. 2011. Vandalism detection in Wikipedia: a high-performing, feature-rich model and its reduction through Lasso. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. ACM, 82–90.

29. Predrag Klasnja and Wanda Pratt. 2012. Healthcare in the pocket: mapping the space of mobile-phone health interventions. *Journal of biomedical informatics* 45, 1 (2012), 184–198.

30. Cliff Lampe, Rebecca Gray, Andrew T Fiore, and Nicole Ellison. 2014. Help is on the way: Patterns of responses to resource requests on Facebook. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 3–15.

31. Danielle Lawson. 2008. Negotiating social and moral order in internet relay chat. (2008).

32. Jim Maddock, Robert Mason, and Kate Starbird. 2015. Using Historical Twitter Data for Research: Ethical Challenges of Tweet Deletions. In *CSCW 2015 Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World*. ACM.

33. Jessica A Pater, Oliver L Haimston, Nazanin Andalibi, and Elizabeth D Mynatt. 2016. "Hunger Hurts but Starving Works:" Characterizing the Presentation of Eating Disorders Online. In *Proceedings of the 19th ACM conference on Computer Supported Cooperative Work & Social Computing (CSCW)*. ACM. in press.

34. Whitney Phillips. 2011. LOLing at tragedy: Facebook trolls, memorial pages and resistance to grief online. *First Monday* 16, 12 (2011).

35. Rebecca Rafferty and Thomas Vander Ven. 2014. âĂIJI Hate Everything About YouâĂİ: A Qualitative Examination of Cyberbullying and On-Line Aggression in a College Sample. *Deviant behavior* 35, 5 (2014), 364–377.

36. Pnina Shachaf and Noriko Hara. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science* 36, 3 (2010), 357–370.

37. Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. 2013. I read my Twitter the next morning and was astonished: A conversational perspective on Twitter regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3277–3286.

38. Janet Sternberg. 2000. Virtual misbehavior: breaking rules of conduct in online environments. In *Proceedings of the Media Ecology Association*, Vol. 1. 53–60.

39. Janet Sternberg. 2012. *Misbehavior in cyber places: The regulation of online conduct in virtual communities on the Internet*. Rowman & Littlefield.

40. John Suler. 2004. The online disinhibition effect. *Cyberpsychology & behavior* 7, 3 (2004), 321–326.

41. John R Suler and Wende L Phillips. 1998. The bad boys of cyberspace: Deviant behavior in a multimedia chat community. *CyberPsychology & Behavior* 1, 3 (1998), 275–294.

42. Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2014. Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 415–425.

43. William Yang Wang and Kathleen R McKeown. 2010. Got you!: automatic vandalism detection in Wikipedia with web-based shallow syntactic-semantic modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 1146–1154.

44. Jun-Ming Xu, Benjamin Burchfiel, Xiaojin Zhu, and Amy Bellmore. 2013. An Examination of Regret in Bullying Tweets.. In *HLT-NAACL*. 697–702.

45. Elad Yom-Tov, Luis Fernandez-Luque, Ingmar Weber, and Steven P Crain. 2012. Pro-anorexia and pro-recovery photo sharing: A tale of two warring tribes. *Journal of medical Internet research* 14, 6 (2012).