Detecting and Characterizing Mental Health Related Self-Disclosure in Social Media

Sairam Balani
Georgia Institute of Technology
85 Fifth St NW
Atlanta, GA 30308 USA
s.balani@gatech.edu

Munmun De Choudhury
Georgia Institute of Technology
85 Fifth St NW
Atlanta, GA 30308 USA
munmund@gatech.edu

Abstract
Self-disclosure is an important element facilitating improved psychological wellbeing in individuals with mental illness. As social media is increasingly adopted in health related discourse, we examine how these new platforms might be allowing honest and candid expression of thoughts, experiences and beliefs. Specifically, we seek to detect levels of self-disclosure manifested in posts shared on different mental health forums on Reddit. We develop a classifier for the purpose based on content features. The classifier is able to characterize a Reddit post to be of high, low, or no self-disclosure with 78% accuracy. Applying this classifier to general mental health discourse on Reddit, we find that bulk of such discourse is characterized by high self-disclosure, and that the community responds distinctively to posts that disclose less or more. We conclude with the potential of harnessing our proposed self-disclosure detection algorithm in psychological therapy via social media. We also discuss design considerations for improved community moderation and support in these vulnerable self-disclosing communities.

CHI'15 Extended Abstracts, April 18–23, 2015, Seoul, Republic of Korea.
ACM 978-1-4503-3146-3/15/04.
http://dx.doi.org/10.1145/2702613.2732733

Author Keywords
Social media, self-disclosure, mental health, Reddit

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction
Over the past several years, online communities catering to health needs and challenges have shown considerable uptake [6]. These tools are known to act as a constantly available and conducive source of information, advice, and support. An important motivation behind the use of these online

bipolarReddit	depression
psychoticreddit	PTSD
mentalhealth	BPD
traumatoolbox	MMFB
SuicideWatch	panicparty
hardshipmates	feelgood
StopSelfHarm	Anger
survivorsofabuse	DPDR

Table 1. Sample of mental health related subreddits used to detect self-disclosure in this paper.

505Nerds	AskReddit
currentlyreading	cringe
philosophy	funny
friendship	news
MildlyInteresting	bestof

Table 2. Sample of control related subreddits used to detect self-disclosure in this paper.

I was bullied in the 5th class. this caused me to get depressed day and day again and ended with me getting depressions
today is the first day in at least a year that I have gotten this far w/o the feelings of self loathing, hopelessness and despair, overwhelming me to the point of feeling like ending it all
I haven't been able to sleep tonight (west coast) and I spent the last hour crying silently [..]. I'm 20 (m) years old, and for the last two years I haven't make any progress into adult life

Table 3. High self-disclosure posts.

resources for health concerns is that they support open and honest discourse [16]. Such discourse is often referred to as "self-disclosure" – self-disclosure is the telling of the previously unknown so that it becomes shared knowledge, the "process of making the self known to others" [12]. Literature indicates that self-disclosure can be an important therapeutic ingredient and is linked to improved physical and psychological well-being [5]. In the context of health conditions that are typically considered socially stigmatic, such as mental illness, self-disclosure has been noted to be a basic element in the attainment of improved health [14]. This is because self-disclosure results in disinhibition [16], which is known to play a positive role in psychological counseling.

Our motivation is rooted in this rich body of work examining the important role of self-disclosure in mental health. We present preliminary findings on automatic detection and characterization of self-disclosure in content shared on mental health online communities in Reddit. Automatic detection of self-disclosure levels in social media posts may help community moderators direct appropriate help and advice in a timely fashion to individuals with mental health challenges. It can also help build tailored recommender tools that, based on self-disclosure levels, match potential parties in the community to a post's author.

We focus on a large variety of mental health focused forums on Reddit and build a classifier to identify levels of self-disclosure in the content posted on these forums. Next we characterize attributes of mental health discourse in the forums via the lens of the levels of self-disclosure in them. We find noted differences in ways individuals socially engage on these forums, depending on their chosen level of self-disclosure. We discuss implications of our work in social system design, in the potential of harnessing social media for psychological therapy and counseling, and the ethical challenges in this line of research.

Related Work

Self-disclosure has been widely investigated both in the psychology and the computer mediated communication (CMC) literature. A rich body of this work has argued self-disclosure to be beneficial: having been linked to trust and group identity [12], as well as playing an important role in social interactions by reducing uncertainty [3]. Self-disclosures are usually made for a purpose [2], including expression, relationship development and social control [2].

In the context of mental health, Ellis [5] reported that discourse on emotionally laden traumatic experiences can be a safe way of confronting mental illness. On similar lines seminal work by Pennebaker et al. [14] found that participants assigned to a trauma-writing condition (where they wrote about a traumatic and upsetting experience) showed immune system benefits. Disclosure has also been associated with reduced visits to medical centers and psychological benefits in the form of improved affect [15].

Additionally, prior research in CMC found that medical patients tend to report more symptoms and undesirable behaviors when interviewed by computer rather than face-to-face [8]. Ferriter [7] found that pre-clinical psychiatric interviews conducted using CMC yielded more honest, candid answers. In the UK, the Samaritans report that although only 20% of telephone callers report suicidal feelings, this number increases to around 50% of email contacts [11].

We situate our work within the findings from this body of literature. We build our observation from recent findings on the widespread adoption of online social platforms for health related discourse [6]. However note that literature examining levels of self-disclosure around *online* mental health discourse is limited. While there have been recent work on modeling self-disclosure manifested in social media content [10, 18], however they (a) focus primarily on Twitter, and (b) do not focus on health. Therefore the impact of online

Just now me and my mom went to get ice cream at Coldstone
Now I can focus on what I am doing and I just feel great. Does anyone else feel the same?
I'm pretty new to the area I just moved (been here 9 months now) and I've made a handful of friends through work.

Table 4. Low self-disclosure posts.

Can anyone help me figure out the name of this show? It was an animated show and was not that popular. If it helps
Believe in yourself. You can achieve your dream as long as you are willing to make it your number one priority and work your hardest to achieve it.
Just saw alcohol. A shot will probably dehydrate you. A beer will likely hydrate you a little.

Table 5. No self-disclosure posts.

	Acc (%)	**Prec**	**Rec**
Naïve Bayes	62.5 (±1.2)	.583 (±.08)	.965 (±.09)
k-NN (*k*=5)	60.7 (±.7)	.926 (±.07)	.258 (±.01)
Decision trees	58.8 (±2.2)	.862 (±.6)	.251 (±.04)
Perceptron	78.4 (±2.1)	.74 (±.02)	.869 (±.02)

Table 6. Performance of different self-disclosure classifiers based on accuracy, precision, and recall.

self-disclosure in individuals challenged with behavioral health concerns is under-examined; also little is understood about how self-disclosure manifests in other social platforms.

In our prior research [4], we examined how use of the social media Reddit for mental health purposes is sometimes characterized by disinhibiting behavior, along with high levels of social support from the larger community. Through this paper, we expand this research by examining how to detect and characterize levels of self-disclosure in health discourse on Reddit, focusing on mental health communities.

Self-Disclosure Data

Data Collection

We used Reddit's official API to collect posts, comments on posts, and associated metadata from several mental health focused subreddits. We build on the data collection methodology we used in [4]. In order to arrive at a comprehensive list of subreddits to focus on, we utilized Reddit's native subreddit search feature (*http://www.reddit.com/reddits*). We searched for subreddits on "mental health". Two researchers familiar with Reddit employed an initial filtering step on the search results returned, so that we "seed" on high precision subreddits discussing mental health concerns. Thereafter, we focused on a snowball approach to compile a second list of "related" or "similar" subreddits that are mentioned in the profile pages of the seed subreddits. Sample of subreddits (31 in all) we crawled are given in Table 1. Note all of these subreddits host public content.

For the purposes of self-disclosure detection, we also identified subreddits (sample listed in Table 2) as our *control group* (total of 12 subreddits) – meaning they are unrelated to mental health topics. For sanity check, we randomly sampled a set of 200 posts from the control subreddits, and two researchers familiar with Reddit manually checked their content for presence of any mental health content. We found that 97% of subreddit content in our sample were not about any mental health concern (Cohen's Kappa for inter-rater agreement was .84). In all, our dataset had 32,509 posts from 23,807 users in the mental health subreddits, and 15,383 posts from 13,216 users in the control forums.

For each of the unique users in the mental health forums, we further collected all of their Reddit post/comment histories (last 1000 posts/comments per Reddit API limits) if their number of posts and comments in our dataset was five or more – this gave us 7,248 users and 4.1M posts/comments.

Generating Ground Truth on Self-Disclosure

Automatic detection of self-disclosure levels in posts necessitates obtaining gold standard labels on self-disclosure, in essence, "ground truth". For the purpose, two raters familiar with Reddit and its mental health communities in particular, independently rated a small random sample (50 posts) with equal proportions from mental health and control subreddits, for three levels of self-disclosure – *no* self-disclosure, *low*, and *high* self-disclosure. These three classes of self-disclosure were chosen based on categorization by Bak et al in [10]. The raters mutually discussed their labels thereafter and thus came up with a set of rules for rating. The rules were further aligned with observations in prior work [9, 10, 11]. Per the rules:

a. Posts that either reveal personal information (e.g., age, location, gender etc.) or divulge sensitive or vulnerable thoughts, beliefs, or embarrassing/confessional experiences were to be considered to be indicative of high self-disclosure. Joinson [10] characterized sensitive disclosure in terms of the extent of "revealed vulnerability".
b. Posts about self but not disclosing any personal or emotionally vulnerable content was to be considered of low self-disclosure.

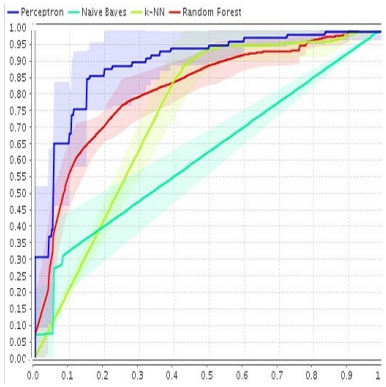

Figure 1. ROC curves showing the performance of various classifiers. Since our classification task involves three classes, ROC curves cannot be generated directly. Hence we perform training/testing for all pairwise cases (high/low, low/none, high/none). Each ROC curve is averaged over all three pairs of classification tasks.

I've lost	anymore	normal
want to kill	I hate myself	part of me
I've never	happy	live
because I	feel like	I can't
need to	Don't know	time
something wrong	difficult	even though
does anyone	did not realize	every time
need to quit	I'm trying	if I could

Table 7. Highly weighted features of the self-disclosure classifier.

c. No self-disclosure posts were those which were about people or things other than the posting author, and which divulged information unrelated to the self.

Following these mutually agreed upon rules, the previous two raters and an additional rater familiar with Reddit independently coded a larger sample of 800 posts to create a training set for the purposes of classification. The raters had good agreement in their ratings: Fleiss' Kappa was found to be .73. However, given the subjective nature of characterization of self-disclosure, we considered only those posts in the training set for which we had agreement across all three raters – this gave us 627 posts. Across the three categories, the coded set consisted of 38% posts of high self-disclosure, 35% posts of low self-disclosure, and 27% with no self-disclosure. Table 3 gives examples of mental health post excerpts with high self-disclosure, while Table 4 and 5 provide examples of low and no self-disclosure posts. Note that including non-mental health posts in the training set was essential so as to let the detector learn on posts of low and no self-disclosure and on those not mental health related.

Self-Disclosure Inference

Based on the training data thus created, we pursued the use of supervised learning to develop a classifier, which would indicate whether a post is of high, low or no self-disclosure. We tested a variety of different classification techniques (decision, trees, *k* Nearest Neighbor, naive Bayes). The best performing classifier was found to be a perceptron classifier, with adaptive boosting used to amplify performance [17], whose results will be used in the remainder of this paper.

We used the following feature generation rules: First we eliminated stopwords from each post based on standard list provided by Python's NLTK library. Next we performed stemming using Porter Stemmer. We extracted uni-, bi-, and tri-grams from each post, and considered those with five or more occurrences. We also computed two additional features – length of each post, and whether the author of the post is an exclusive poster on mental health forums, or is observed in our dataset, to post on other forums as well. Thus each post was characterized by 1070 features.

We used standard 10-fold cross validation (CV) to evaluate the classifier, and ran our model over 100 random 10-fold CV assignments for generalizability of the results. We report the average accuracy, precision, recall, F1, specificity as metrics of performance. We find that our classifier based on the perception model yields an average accuracy of 78.4% in detecting high or low self-disclosure, with .74 precision and .86 recall (see Table 4 for details). Other methods like *k*-NN (*k*=5) give higher precision, but at the expense of very low recall. Figure 1 gives the ROC (receiver operating characteristic) curves for all the models. Per the ROC curve corresponding to the perceptron model, we find it to yield the maximum area under curve (.81), hence best performance.

We further identify, in Table 5, the *n*-grams (or features) with the highest weights given by the perceptron – it implies these features were the most significant in the classification task. We provide some brief qualitative examinations of these *n*-grams, in the light of prior psychology literature on self-disclosure and mental health [10, 11]. We find that the *n*-grams primarily are associated with vulnerable and self-loathing thoughts (e.g., thoughts of suicide), bear a negative tone, or depict confessional experiences. Based on prior research [10, 11, 12] and our own work on mental health discourse on Reddit [4], we find that these are the topical dimensions along which high self-disclosure and low/no self-disclosure posts vary. In essence, high self-disclosure posts share extensively their personal beliefs and fear, for instance, their vital constructs and private, sensitive informational attributes. The post excerpts below have been classified to be of high self-disclosure and through them, we demonstrate the use of some of the n-grams in Table 7:

	High SD	Low SD	No SD	*p*
Post length	326	152	84	***
upvotes	7.88	10.4	15	**
downvotes	8.76	7.93	7.5	*
comments	10.3	6.19	5.6	***
response time	192	247	439	**

Table 8. Comparison of high, low and no self-disclosure posts shared on mental health subreddits, (SD – self-disclosure). Statistical comparisons based on Kruskal-Wallis one way ANOVA are reported. Significance levels were subject to Bonferroni correction due to multiple comparisons. Here * is $p<.05/n$, ** is $p<.01/n$, and * is $p<.001/n$, where $n=5$, the number of comparisons.**

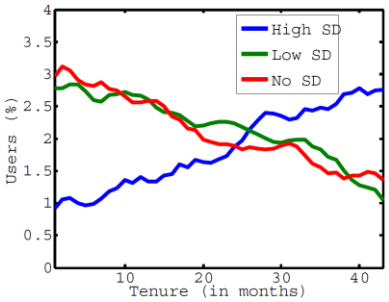

Figure 2. Showing how tenure varies for users with sharing high, low, and no self-disclosure posts.

"I don't want to kill myself, I haven't felt suicidal in a long time, but I just want to stop life for a while, you know?"

"My dad would beat the living shit out of me […]. I've been to the hospital so many times I've lost track"

*"I hate this. I hate myself. I don't want to f****** be this person anymore. I'm unmotivated, unfocused, immature."*

Characterizing Discourse via Self-Disclosure

Following the development of the self-disclosure classifier, we wanted to apply it on an unlabeled set of mental health posts in order to study how levels of self-disclosure characterize the nature of discourse and community response on these Reddit forums. For the purpose we used the unlabeled 4.1M mental health forum posts obtained by crawling the histories of those users in mental health forums who posted five or more times in our observed data.

On applying our trained self-disclosure classifier on these 4.1M posts, we find that 1.9M posts were classified as high self-disclosure posts, 1.3M posts were categorized to be of low self-disclosure, while .9M were assigned to be having no self-disclosure. This indicates that in general, mental health communities on Reddit are characterized by high self-disclosure in the shared content. Presumably, individuals find these forums to be an open platform that allows expressing views and thoughts about a topic often considered to be sensitive or unacceptable to the mainstream. The ability to be fairly anonymous on Reddit perhaps facilitates such high self-disclosure – unlike Facebook, a user does not need to specify any personal information in their profiles (e.g., location, age, name, gender etc.).

We further observe from Figure 2 that majority of the users who self-disclose more, tend to also have longer tenure on Reddit. By tenure, here we mean the total amount of time (in months) the user has been observed to be active on Reddit based on our dataset. Most low self-disclosure users have shorter tenure relatively, the same is true for the bulk of the no self-disclosure users ($F= 6.2$; $p<.01$ based on Kruskal-Wallis test). This shows that users find the ability to self-disclose more to be helpful, resulting in continued involvement and engagement on the social media site.

Finally, from Table 6 we observe that high self-disclosure posts receive fewer upvotes, greater downvotes, but higher number of comments as well as high response rate. Here response rate is calculated as the normalized time elapsed between post time and timestamp of first comment on the post. This shows that high self-disclosure on mental health subreddits, while may not be highly endorsed by the greater Reddit community due to their caustic and disinhibiting tone, however do receive social support through increased levels of engagement – commentary and rapidity of responses shows that the community is keen to help and advise these individuals with honest and disinhibiting self-disclosures. The greater conversational length as indicated by the high number of comments also aligns with prior literature, where it has been found that candid discourse results in longer verbal exchange [4, 10, 14].

Design Implications and Future Directions

Our findings, though preliminary, bear several implications in design. Psychology literature indicates mental illness challenged people's apprehensions about experiencing and disclosing negative emotions is the greatest barrier to seeking counseling [14, 15]. Hence the fact that these mental health subreddits are allowing individuals to be more expressive or engage in greater self-disclosure shows promise for web based psychotherapy. High self-disclosing redditors could be provided with services that can draw their attention to relevant recent conversations that have happened in the past and the community's response to it, so that they perceive a sense of support when they visit these mental health communities. These users could also be

flagged in the interfaces of the community moderators so that they can pay special attention to their requests for support, advice, or help, especially because some of them seem to be sharing emotionally vulnerable content.

Nevertheless, future research in online health intervention design needs to consider ways in which the identities of these vulnerable communities may be protected against revelation to unintended audiences. Conversations on the ethical dimensions of this line of research has been on the rise, and we hope to engage with the ethics and clinician communities to ensure that the algorithms of self-disclosure detection proposed here are utilized in the interest of the affected communities in ethically appropriate ways.

In subsequent research, we intend to investigate how individuals tailor their self-disclosure on these forums over time, based on the feedback from the community. Further it will be interesting to study how the extent of social stigma typical of mental illness or the online identity of an individual relates to levels of self-disclosure on social media.

Conclusion

We presented preliminary investigation of data-driven methods to detect levels of mental health related self-disclosure in social media content, particularly focusing on Reddit. Since greater self-disclosure facilitates improved mental health, we hope our research suggests ways to leverage shared online content to measure self-disclosure. Thus it may enable us design better health interventions and support systems that can cater to the needs of these emotionally vulnerable communities better.

References

1. Bak, J., Lin, C. Y., & Oh, A. Self-disclosure topic model for classifying and analyzing Twitter conversations. In Proc. EMNLP 2014, to appear.
2. Chelune, G. J. (1979). Self-disclosure: Origins, patterns, and implications of openness in interpersonal relationships. San Francisco: Jossey-Bass.
3. Cozby, P. Self-disclosure: a literature review. Psychological bulletin, 79(2):73, 1973.
4. De Choudhury, M., and De, S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In Proc. ICWSM, AAAI (2014).
5. Ellis, D., & Cromby, J. Emotional inhibition: A discourse analysis of disclosure. Psych & health 27, 5 (2012), 515–532.
6. Fox, S. The Social Life of Health Information, Pew Internet, 2011.
7. Ferriter, M. Computer aided interviewing and the psychiatric social history. Social Work & Social Sc Rev, 1993.
8. Greist, J., Klein, M., and Van Cura, L. A computer interview for psychiatric patient target symptoms. Archives of General Psychiatry, 29(2): 247, 1973.
9. Houghton, D. & Joinson, A. Linguistic markers of secrets and sensitive self-disclosure in twitter. In Proc. HICSS 2012.
10. Jin Yeong Bak, Suin Kim, and Alice Oh. 2012. Self-disclosure and relationship strength in Twitter conversations. In *Proc.* ACL '12, 60-64.
11. Joinson, A. N. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. European J. Social Psych 31, 2 (2001), 177–192.
12. Joinson, A., and Paine, C. Self-disclosure, privacy and the internet. The Oxford handbook of Internet psychology, page 2374252, 2007.
13. Jourard, S. M. Healthy personality and self disclosure. Mental Hygiene. New York (1959).
14. Pennebaker, J. W., and Chung, C. K. Expressive writing, emotional upheavals, and health. Foundations of health psychology (2007), 263–284.
15. Smyth, J. M. Written emotional expression: Effect sizes, outcome types, and moderating variables. Journal of consulting and clinical psychology 66, 1 (1998), 174.
16. Suler, J. The online disinhibition effect. Cyberpsychology & behavior 7, 3 (2004), 321–326
17. Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
18. Walton, S. and Rice, R. 2013. Mediated disclosure on twitter: The roles of gender and identity in boundary impermeability, valence, disclosure, and stage. Computers in Human Behavior, 29(4): 1465–1474