Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities

Stevie Chancellor
College of Computing
Georgia Tech
Atlanta, GA 30332
schancellor3@gatech.edu

Zhiyuan (Jerry) Lin
College of Computing
Georgia Tech
Atlanta, GA 30332
zlin48@gatech.edu

Erica L. Goodman
Department of Psychology
University of North Dakota
Grand Forks, ND 58202
erica.goodman@my.und.edu

Stephanie Zerwas
Department of Psychiatry
University of North Carolina
Chapel Hill, NC 27599
zerwas@med.unc.edu

Munmun De Choudhury
College of Computing
Georgia Tech
Atlanta, GA 30332
munmund@gatech.edu

ABSTRACT
Social media sites have struggled with the presence of emotional and physical self-injury content. Individuals who share such content are often challenged with severe mental illnesses like eating disorders. We present the first study quantifying levels of mental illness severity (MIS) in social media. We examine a set of users on Instagram who post content on pro-eating disorder tags (26M posts from 100K users). Our novel statistical methodology combines topic modeling and novice/clinician annotations to infer MIS in a user's content. Alarmingly, we find that proportion of users whose content expresses high MIS have been on the rise since 2012 (13%/year increase). Previous MIS in a user's content over seven months can predict future risk with ~81% accuracy. Our model can also forecast MIS levels up to eight months in the future with performance better than baseline. We discuss the health outcomes and design implications as well as ethical considerations of this line of research.

ACM Classification Keywords
H.4 Information Systems Applications: Miscellaneous

Author Keywords
Instagram; social media; mental health; mental illness; selfinjury; eating disorder

INTRODUCTION
Online communities can promote well-being and ailment management by improving perceived self-efficacy and mitigating psychological distress [33]. However, online platforms have been challenged by communities that encourage deliberate destruction of one's own body. In particular, social media platforms like Instagram have been scrutinized for the continued presence of communities that promote and share self-injurious[1], suicidal, and pro-eating disorder (pro-ED) content[2]. These communities promote negative actions as deliberate or impulsive choices rather than as symptoms to other mental disorders or threats to health [34].

The Diagnostic and Statistical Manual of Mental Disorders (DSM) [4] identifies specific behaviors and cognitions that promote self-injurious behavior, extreme weight control, and suicidal ideation. Such behaviors may be associated with specific mental illnesses like eating disorders. Based on the clinical psychology literature [49, 36], we map manifestations of such behaviors in social media content to markers of *mental illness severity* (MIS).

Vast epidemiological and psychiatric evidence of MIS exists in clinical research [21, 49, 36, 54]; however, studies of MIS on social media sites are limited. Most studies in CSCW or HCI have focused on identifying markers of a mental illness, e.g. depression [18, 47, 31, 15], not on MIS more broadly in any given mental illness-prone community. The increasing pervasiveness of mental illness-related content on social media (such as Instagram, Tumblr, Twitter, and Reddit)[1] now provides an opportunity for rigorous quantitative studies of markers of MIS on these platforms. The rich content on these sites may be used both to objectively quantify, measure, and characterize levels of MIS broadly and also to examine the distribution of such content over previously inaccessible timescales and population sizes.

In this paper, we study, estimate, and forecast MIS in users who share pro-ED content on Instagram. The pro-ED community glorifies eating disorders as alternative lifestyle choices rather than as psychosocial disorders [19]. Cognitions present in the pro-ED community include suicidal and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CSCW '16, February 27-March 02, 2016, San Francisco, CA, USA
© 2016 ACM. ISBN 978-1-4503-3592-8/16/02...$15.00
DOI: http://dx.doi.org/10.1145/2818048.2819973

[1] http://techcrunch.com/2013/06/20/over-a-year-after-new-content-policies-self-harm-social-media-still-thrives/
[2] http://www.huffingtonpost.com/laurenduca/thinspiration-banned-frominstagram_b_3829155.html

thin ideal ideation, and behavioral activities include direct self injury [36] and extreme weight control behaviors (which some have termed indirect self-injury [20, 25]). The pro-ED community is an ideal community to study because up to 70% of individuals with eating disorders report a history of some form of deliberate self-injury [12]. Because of this, the community presents the largest variety of markers of MIS not present in communities focused on other mental disorders (e.g., depression or obsessive-compulsive disorder) or specific activities (e.g., cutting or suicidal ideation).

We make the following contributions in this paper:

- We present a novel, scalable, and robust method to quantify and characterize MIS in pro-ED Instagram content. Our method employs topic modeling, specifically Latent Dirichlet Allocation, on the tag content of pro-ED posts. We present a clinically-grounded framework to obtain novice and expert annotations on low, medium, and high MIS of the extracted topics.
- We develop an algorithm to combine automatically extracted topics and their severity scores into inferences of MIS levels in user content over time.
- We present a supervised learning (regularized multinomial logistic regression) model to predict to what extent a user who shares pro-ED content on Instagram would share content with markers of low, medium, or high MIS in the future, based on MIS manifested in their historical content.

Our study uses a large dataset crawled from Instagram of more than 26M posts from 100K public users between 2010 and 2015. Our rating methodology that combines topic modeling with human annotations can measure MIS in posts with high accuracy, precision, and recall when compared against independent ratings from novices and experts. Second, MIS inferred from a user's posts over a seven month period is able to predict levels of MIS during a future month with over 81% accuracy. Our results show that despite Instagram-enforced efforts to curb dissemination of such vulnerable content[3], the proportion of pro-ED Instagram users sharing content with high levels of MIS has been on the rise (13% increase / year).

Our findings provide one of the first quantitative methodologies to quantify, measure, and forecast MIS on social media platforms that is automated and applicable at scale. Although our system is not intended to outright diagnose eating disorders or other mental illness without self-reported information on health status, we believe our findings can be used to complement mental wellness efforts towards identifying at-risk populations and to enable social media designers a gauge community health and well-being.

Ethics, Privacy, and Disclosure. This paper used publicly accessible Instagram data to conduct our analysis. No personally identifiable information was used in this study. Publicly accessible, blurred out images are used only for exemplary purposes and were not included in any of our algorithms. Because we did not interact with our subjects and the data is public, we did not seek institutional review board approval. Our work does not make diagnostic claims. Some of the images in this paper are graphic in nature.

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

BACKGROUND AND RELATED WORK

What is Mental Illness Severity (MIS)?

Clinical definitions of mental illness severity (MIS) are based on epidemiological issues and rely on the inclusion and operationalization of criteria such as diagnosis, disability, and duration [24]. MIS is typically associated with significant cognitive function and judgment impairment, emotional instability, and limitations in undertaking life activities [35]. The cognitive and behavioral markers of self-injurious behavior, extreme weight control measures, and suicidal ideation constitute some of the most crucial indicators of MIS. [50, 24].

Self Injurious Behavior and Suicidal Ideation. The DSM-5 [4] does not list self-injurious behavior as a unique disorder but rather as symptoms of several conditions. Self-injurious behavior involves deliberate injury or damage to one's own body. Examples include cutting, self-mutilation, burning, branding, scarring, scratching, picking at skin or reopening wounds, biting, and ingesting toxic chemicals [45]. One in 12 teenagers engage in such risky behavior, and around 10% continue to self-injure into young adulthood [43]. In some cases, self-injuirous behavior may lead to or is associated with serious thoughts of harming or killing oneself, also known as suicide ideation [58]. Self-injurious behavior is considered to be one of the strongest predictors of who will commit suicide in the near future [43]. Patients with eating disorders are known to engage in these behaviors extensively. Over 70% of patients with disordered eating habits report that they self-injure, [57] and eating disorders have some of the highest comorbidities (80%) with other mental disorders, such as anxiety, depression, dysthymia, and substance use disorders [32]. Mortality rates in anorexia nervosa are the highest of any psychiatric disorder – 12 times higher than the average mortality rate for females aged 15-24 [32].

Extreme Weight Loss Measures. Extreme weight control measures include abnormal behaviors related to eating and exercise, such as unusual food restrictions, fear of particular foods or kinds of food, binging, purging, or abuse of laxatives, and overexercising [54]. The DSM considers extreme weight control measures to be some of the defining symptoms of severe mental illness in eating disorder populations [4], in which anorexia nervosa and bulimia nervosa constitute the most well-known and diagnosed form of eating disorders [3].

Clinical and Psychiatric Work on MIS

Literature on these three forms of MIS largely comes from a psychiatric or medical perspective and focuses almost entirely on treatment. However, to the best of our knowledge, quantitative and rigorous studies of MIS online are limited. Since the majority of current studies have been conducted on individuals with a history of psychiatric treatment [27], little is empirically known about MIS among individuals who are not clinical inpatients, which might be the case with those frequenting today's social media platforms [56]. Moreover, many who suffer with these forms of MIS never seek the help of mental health professionals, often due to the stigmatized nature of these behaviors or thoughts [14, 52]. Obtaining at-scale self-identified sufferers of mental illness or of MIS in

any capacity may also be challenging, precluding large-scale investigations of forms of MIS in the larger population.

We believe that social media platforms enable the study of MIS – particularly within pro-ED communities – that previously have been hard to study. Our work attempts to overcome these limitations by providing a methodology to detect MIS and allowing risk assessment over *those* populations who might be difficult to reach by clinical means.

Online Pro-ED Communities

Pro-ED communities are groups that promote eating disorders as an alternative lifestyle choice. These communities share content, provide advice, and glorify thinness and low body weight as ideal. These communities share restrictive dieting plans and advice, coaching, and sharing inspirational imagery of thin bodies, also known as "thinspiration" or "thinspo". Online platforms pose specific risks for those suffering with eating disorders that may dispose them to harm themselves through self-injurious behavior or extreme weight control measures. For instance, imagery and the visual nature of social networks, especially Tumblr and Instagram, may encourage disordered thoughts and actions, such as normative conceptions of ideal body shape [41]. Such behavior is supported by statistics that found that 69% of American girls five to 12-years old say pictures influence their concept of ideal body shape and 47% report that images make them want to lose weight. Moreover, due to the archival and social nature of these platforms, they can offer individuals opportunities to revisit their own or their peers' self-injury experiences or taking inspiration for continuing extreme weight control measures.

The body of research on pro-ED communities online is large and varied [8, 22, 38, 53]. However, the bulk of this work is qualitative in nature [6], aside from a few recent exceptions [60, 59, 16]. For instance, Yom-Tov et al compared the content associated with pro-anorexia and pro-recovery posters on Flickr [60]. However, this literature has not examined these communities from the angle of MIS, a significant aspect of and hurdle to recovery in these communities acknowledged in the DSM [4]. Our paper examines MIS in Instagram's pro-ED community to bring one of the first empirical insights into a problem on social media[1].

Health Behavior Inference from Social Media

Social computing research has shown that content and conversational patterns can be used to infer psychological states, well-being, and social support status. In an early work, De Choudhury et al. [17] analyzed how new mothers' risk to postpartum depression may be detected from content posted to Facebook and Twitter. Schwartz et al. [51] predicted life satisfaction score of counties with socioeconomic factors and Twitter language. Other research includes utilizing social media content and interactions to identify conditions and symptoms related to diseases [48], substance use [44, 40], mental health [18, 47, 31, 15, 2], and nutrition and health [1].

This body of work shows that valuable information is embedded in social media language, and that language may bear predictive power to detect health concerns. However, most prior research has focused on distinguishing between mental illness-prone and control populations; to the best of our knowledge, they have not focused on characterizing manifested levels of MIS. We also provide a systematic, carefully constructed, and clinically-grounded characterization of MIS in pro-ED content based on three types of markers: self-injurious behavior, suicidal ideation, and excessive weight control measures. Such cognitive and behavioral markers have not been discussed in prior social computing research. Our work offers what we believe is one of the first rigorous and sophisticated methodologies to measure and predict levels of MIS by using linguistic content on social media.

DATA

We gathered a dataset of *public* posts related to eating disorders on Instagram using the official Instagram API[4]. Note that Instagram does not have formalized community structures, like forums or private groups. Instead, communities form around more amorphous, public tags. In the case of the pro-ED community on Instagram, users cluster around tags relating to eating disorders (e.g., "anorexia", "proana").

Our data collection did not prioritize collecting content shared by tags directly associated with self-injury and suicide; those tags would bias the content and nature of our results. Observations indicated that specific self-injury tags on Instagram (e.g., "selfharmmmm") heavily favored cutting. Moreover, searching specific self-injury or suicide ideation related tags would generate only a partial sample because the set of all possible MIS tags on Instagram is unknown. Our crawl proceeded in two stages described below.

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 and users in our study.

Constructing Eating Disorder Post Set. First, we identified a set of nine "seed tags"[5] that have been found to be common pro-ED organizing tags and structures across social media platforms [16]. Then, two researchers searched for posts on each of these nine tags to ensure there was sufficient pro-ED content. With these tags, we conducted an initial month-long crawl using Instagram's official API; this gave us 434K posts with 234K unique tags. We identified 222 total tags that had at least a 1% co-occurrence rate in our dataset.

Next we expanded the initial seed set by collating a list of all tags that co-occurred with the seed tags in the initial 434K posts. From this list, we manually checked and removed tags that did not map directly to eating disorders. Specifically, we found three areas that these removals fell into: too generic, related to another disorder, or eating disorder recovery-related:

[4]http://instagram.com/developer/

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

(1) Tags that related to eating disorders but were broad enough to be used by the general population, e.g. "beautiful", "inspiration", or "fat".

(2) Tags that related to other mental disorders or distressing content, e.g. "nervous", "ocd", or "suicidal".

(3) Tags that were obviously related to the eating disorder recovery community, e.g. "anarecovery". We expect the eating disorder recovery community to not be engaging in self-injury behavior, excessive weight control measures, or suicide ideation, a finding supported by prior work [38].

This reduced the filtered co-occurrence tag list from 222 tags to 72 known to be related to eating disorders (see sample tags in Table 1). Then we conducted a second longer crawl of pro-ED content focusing on these 72 tags. This gave us over about 8 million posts dated between January 2011 and November 2014. We removed any posts that were cross-posted to any recovery tag as well as any that had three tags ("mia", "ana", and "ed") that did not also contain another tag from our list of 72. Qualitative observation indicated that these three tags are not associated with pro-ED when they are used in isolation; rather, they referred to first names or references to popular celebrities ("ed" for Ed Sheeran).

Our dataset at this phase had 6.5 million posts relating to pro-ED.

Gathering Pro-ED Users and Their Post Timelines. To construct our candidate set of pro-ED users and their posts, we obtained a random sample of 100K users from the authors of the 6.5 million posts collected above. Again, we used Instagram API to obtain the post timelines of these users (all public posts of the users). Our final dataset contains over 26M posts from 100K users, with post shared between October 2010 and March 2015.

As Figure 1 (a) and (b) show, the distribution of posts each user has and the number of tags used in each post are heavy-tailed. We also note that posts returned by the Instagram API do not distribute uniformly over time — most of the posts returned are after late 2012 as shown in Figure 1 (c). Distribution of volume of users with at least one post per month, referred to as "active" users per month, (Figure 1 (d)) also follows the same pattern.

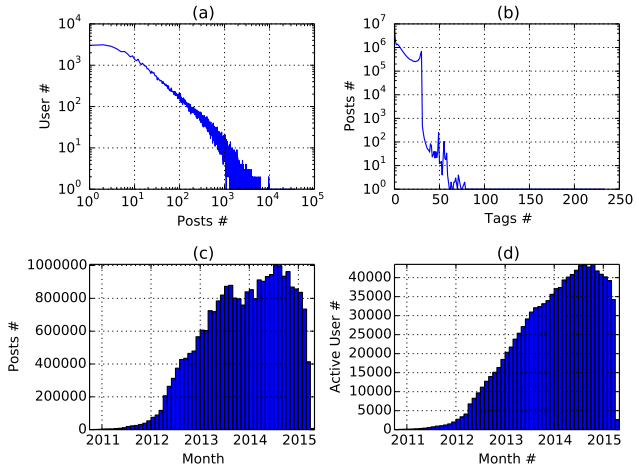


Figure 1: (a) User distribution over number of posts. (b) Distribution of tags used in posts. (c) Number of posts per month from October 2010 to April 2015. (d) Monthly volumes of active users from October 2010 to April 2015.

METHODS

Inferring Mental Illness Severity (MIS)

In this section we discuss our method of inferring and evaluating MIS in a user in a certain point in time. A significant challenge in quantitative studies of any health risk or behavior is the availability of labeled content (i.e., ground truth) on *which users* are susceptible. In the absence of self-reported information on the mental health status of individuals, tags may serve as a good indicator of whether a user's content expresses markers of high MIS. However, capturing the set of all tags related to MIS is difficult, as discussed earlier. Additionally, assessing in isolation a tag's MIS level may be difficult — e.g., "cutting" might be attributable to high MIS, however the tag "pain" is ambiguous. Furthermore, discrete human judgments on MIS may not be applicable to a user's individual posts, since a user may use their Instagram profile to share not only content with MIS but also on a variety of other topics.

To overcome these challenges in MIS inference, we adopted a hybrid approach where we leveraged both automated natural language processing techniques and human annotations. We employed Latent Dirichlet Allocation (LDA) [7] on all posts from all users in our compiled dataset. LDA has been successfully employed in the inference of health phenomenon from social media data [48]. Our goal was to obtain a set of topics spanning the content (tags) of the posts, some of which we suspect to be aligned to increased MIS. This allowed us to go beyond simple tag-based MIS inference techniques — LDA would use all tags for topic inference including those that co-occur with known/unambiguous MIS tags. Moreover, LDA assumes each tag to be drawn from a mixture of topics instead of mapping each tag to a specific topic (MIS or otherwise). By using LDA, we were also able to obtain a posterior distribution of topics over posts of a user. This prevents assigning users to specific topics; instead, we could model their posts as a distribution over content with varying levels of MIS.

Our method proceeded in the following steps:

(1) Topic Inference. We built an LDA model on the posts of all 100K users. We first removed common English words[6] and converted tags from each post to bag-of-words format. We trained an online LDA model [29], which is expected to converge quickly given relatively stationary topics (in this case, pro-ED related topics) without much drift over time. Specifically, we updated the LDA model[7] for every chunk of 1 million posts for all 26M posts to obtain 100 topics, which was found to be appropriate based on initial experiments with our data.

[6]NLTK stop words, http://www.nltk.org/book/ch02.html
[7]We used Gensim library, https://radimrehurek.com/gensim/

(2) MIS Annotation. By reviewing the top 50 keywords of each topic provided by the LDA model, we then obtained human annotations of MIS on every topic. We defined MIS to span three levels — low (1), medium (2), or high (3). Our choice of the scale was motivated from prior work on modeling and inferring self-disclosure in social media content [5], on detecting suicide-risk behaviors in social media [30], and from the clinical psychology literature on MIS [50, 24].

Our annotators included four researchers. Two annotators were trained clinical psychologists with specific expertise in eating disorders and experience interacting with eating disorder in-patients, and the other two had considerable social computing research experience. The researchers created a set of rules to annotate each topic. The raters first manually browsed pro-ED posts on Instagram by searching over all of the nine seed tags used for crawling pro-ED data[4] — so the raters could familiarize themselves with MIS manifested in pro-ED posts. Then, the raters collated a set of rules which were used for the annotation task, heavily referencing the DSM-5 [4], clinical literature [50], experience of interacting with in-patients with eating disorders, and other related work [19]. We call this the MIS scoring or rating system:

- High MIS (score of 3): included extreme weight control behaviors and "thinspiration" (e.g., "purge", "thinspo", "starve", "donteat"), self-injurious behavior (e.g., "cutting", "blades", "slit"), and suicidal ideation (e.g., "killme". "suicidal"). This categorization is supported by the clinical psychology literature [24] showing ED sufferers to display significant suicidal thoughts, eating pathology, including increased body dissatisfaction, binging and purging, compulsive behaviors (e.g., hair pulling, nail biting, skin picking, self-biting) and impulsive self-injurious behaviors (e.g., cutting, burning, self-hitting, banging, scratching).
- Medium MIS (score of 2): included fat talk, self-deprecation, emotional instability, cognitive impairment, social isolation, and discussing eating disorders (e.g., "uglyandfat", "selfhate", "broken", "anorexia", "eatingdisorder", "mia"). In addition, manifestations of mental disorders but without any revealed vulnerability (e.g., "depression", "bpd", "anxiety") were also scored as medium MIS. Literature [35] suggests high comorbidity of eating disorders with other psychiatric conditions, such as major depression and anxiety disorders, which may or may not be severe. ED is also known to significantly interfere with judgment, social adjustment and interpersonal relationships, but these may not be associated with any emotionally or physically dangerous acts [36].
- Low MIS (score of 1): included tags not related to eating disorders or mental health (e.g., "nyc", "iphone", "biking", "cats", "fashion", "selfie").

Following the annotation task, the interrater reliability metric Fleiss' κ was found to be very high (.91); further, between the novice (non-clinician) and the expert (clinician) ratings there was high agreement. Discrepancies were resolved through mutual discussion. Example topics and assigned MIS scores are listed in Table 3. 5 topics out of the 100 given by the LDA model were high (3), 6 as medium (2) and 89 as low (1). We intend to release the LDA model, its associated topics, and MIS ratings of the topics for research use following acceptance.

Algorithm 1: Calculate post MIS rating

```
function getPostMISRating(tagList, ldaModel, topicMIS)
Input  : A list of tags used in a post tagList; LDA model
         ldaModel; human annotated topic MIS rating
         topicMIS
Output: Discretized post MIS rating, or 0 for empty tagList
bow ← bag of words(tagList);
MISScore ← 0;
if tagList is empty then return NULL;
for i = 1 to Number of Topics do
    topicScore ← ldaModel.getTopicProb(i) ∗ topicMIS(i);
    MISScore ← MISScore + topicScore;
end
MISScore ← round(MISScore);
return MISScore;
```

(3) Computing Users' MIS Rating. Given any post, we then use the topic probability distribution generated by the LDA model as weights and combine them with the annotated MIS levels of the topics, to obtain a weighted average MIS rating, as described in Algorithm 1. Table 2 lists example posts and the MIS level derived through our algorithm.

(4) Monthly MIS Rating Inference. To infer a user's MIS in a time slot (month), we obtained the discretized (rounded) mean MIS rating over all posts posted by the user on Instagram during the slot.

Evaluation of MIS Ratings

How effective is our LDA topic modeling approach combined with novice/expert annotations in identifying MIS in a user's content? To answer this question, we compare our algorithm-derived MIS rating in a sample of posts from a sample of users in our dataset with MIS ratings in the same sample obtained from four human raters – two novices (non-clinicians) and two experts (clinicians). We first randomly sampled a set of 150 posts with equal numbers for the three MIS ratings 1 (low), 2 (medium), and 3 (high). Thereafter for the sake of consistency, the same four researchers who labeled the LDA topics rated this sample of 150 posts for the three MIS ratings. The raters had high agreement (Fleiss' κ = .86) and resolved discrepancies mutually via discussion. Raters found 68 posts rated as low, 9 as medium, and 73 posts rated as high MIS.

Next, we compared the agreed upon set of human annotations with algorithm derived MIS ratings on the 150 post sample. This gave high accuracy, precision, recall, and F1 scores (mean accuracy >71%, mean precision > 68%, and mean recall > 70%). These scores were particularly high for MIS ratings 1 and 3; for these two classes respectively, precision was over 94% and recall above 64%. However we observe that the human annotations and algorithm differ significantly in the case of MIS rating 2. A closer look reveals that while recall is still very high for this class (66%), the low precision (37%) is responsible for the low F1 score. Manual inspec-

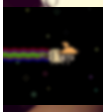
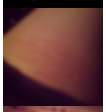
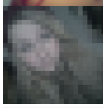
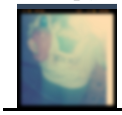
Post	Post tags	Raw	Rounded	Annotated
	fly, dragonfly, prints, fashion, sketch, illustration, ink, masonry, textiles	1.02	1	1
	rainbow, mexicancat, burrito, tablet, rainbowcat, rainbows, wallpaper, cat, livewallpapers, burritocat, livewallpaper, mexican, catwallpaper	1.15	1	1
	starve, bodydismorphia, scars, ana, selfharmmm, bulimia, dead, bones, suicide, triggerwarning, cuts, razorblade, depression, anorexia, razor, sue, cut, mia, tw, blade, fat, hungry, purge, ugly, cutting, deb, blood	2.87	3	3
	disappointment, pain, gayteen, selfharmmm, relapsed, selfinjury, selfhate, slit, cutting, fat, cuts, anorexic, depression, razor, gaykik, death, blood, depressed, ugly, blithe, imsorry, anxiety, triggering, cutter	2.77	3	3
	beautiful, suicidal, ana, bulimic, bulimia, fat, sad, tumblr, cut, blood, cuts, love, ew, hate, anorexic, depression, anorexia, funny, suicide, ugly, mia, helpme, depressed, cutting, killme, thin, skinny, happy	2.82	3	3
	tumblr, depressed, quote, scars, ehtilb, dark, selfharmmm, blithe, depression	2.01	2	3
	koolaid, lazy, girl, thinspoo, sugarfree, exercise	1.67	2	3

Table 2: Example posts with tags, raw MIS (Raw) generated using our method, discretized MIS rating (Rounded), and MIS per agreed upon annotations from four researchers (Annotated). All post images are blurred to avoid disclosure of user information. Most disagreements happen when MIS rating is medium, where it can be ambiguous.

tion of such misclassified posts reveals that the disagreement arises due to the inherent ambiguity in the posts of actual MIS rating 2 (ref. last two rows in Table 2). We note that the MIS ratings 1 and 3 are perhaps the most important and distinctively defined classes of practical importance (e.g., for designing interventions, see Discussion) and constitute over 93% of the content (see Table 4).

Prediction of MIS

Our prediction task involves identifying low, medium, and high MIS in content of users in the future based on MIS in their previous Instagram posts. For this purpose, we employed regularized multinomial logistic regression. We considered using forms of autoregressive integrated moving average models (ARIMA) [42], as these models are fitted to time series data to predict future points in the series. However, since our response variable, i.e. MIS rating, is discrete (categorical) and not continuous, multinomial logistic regression was used as the suggested alternative to ARIMA [42].

As with other types of regression, for the regularized multinomial logit method, there is no need for predictor variables to be statistically independent from each other (unlike, for example, in a naïve Bayes classifier or an ordinary least squares model). Regularization helps us control for collinearity (i.e., excessive correlation between MIS ratings that are temporally close) and sparsity (i.e., users may not post in certain time slots, thus no MIS can be measured) in our data. Further regularization allows us to incorporate smoothness in our model — it is likely MIS changes smoothly across consecutive time slots for most users' content. In our case, we used the model implementation given in the Python package scikit-learn. The library implements regularized logistic regression using the liblinear library, newton-cg and lbfgs solvers. The newton-cg and lbfgs solvers support L2 regularization with primal formulation. Below are the components of the regression model:

Response Variable. The MIS ratings of users' content (1=low; 2=medium; 3=high) at time slot t. Here, the time slot is taken as a month.

Predictor Variables. We define a sliding window of size w, and consider the monthly MIS rating of users' content over all time slots between $t - w$ to $t - 1$ as w predictor variables of the model.

The class sizes (response variables) in our dataset were unbalanced (ref. Table 4), so we employed bagging and boosting to improve performance of the model [9]. We used 90% of our data as training data (90K users); the remaining 10% of users was set aside as the held-out test set on which we report our prediction results. Specifically, we first generated B bootstrap samples of the training data using random sampling with replacement – in these samples we selected users so that all three classes are balanced. We then trained our regularized multinomial logistic regression model on each of these bootstrap samples. Following training, in the boosting phase, we iteratively learned weak regression models using the Adaboost algorithm [13]. That is, we took the weighted sum of the coefficients of the model – this gave us a robust model that

Topic	Top Representative Tags	MIS Rating
2	nyc, newyork, london, la, losangeles, graffiti, streetart, california, frozen, ny, brooklyn, texas, newyorkcity, cali, miamiheat, wall, yup, hawaii, downtown, street, urban, stickers, miami, tokyo, dallas, printing, airport, philly, manhattan, jelsa , hollywood, houston, westcoast, statenisland, sign, photogrid, changedmylife, queens, mural, anna	1
3	chicago, fail, nashgrier, magcon, colorado, science, november, crystals, georgia, turkey, sober, gayboys, indigochild, scruff, flow, snowing, 1dedit, socal, southafrica, blizzard, fitblr, prettylittleliars, sobriety, photobooth, traffic, beachbum, eureka, pll, greece, eurekaca , hairy, stone, istanbul, bayarea, basic, minimalism, simplicity, blackboy, shine, citylife	1
22	tumblr, ugh, bye, idk, hi, tumblrgirl, tumblrpost, ok, weheartit, hello, textpost, tumblrquotes, adult, hey, theme, tumblrboy, lsd, justgirlythings, borderlinepersonalitydisorder, tumblrposts, notmine, post, fatty, cow, sfs, dm, idontcare, mysecretfamily, ifollowbackinstantly, tumblrtextpost , textposts, byee, idc, kbye, cuidandotusalud, bauchnabel, bauchfrei, tumblrquote, cantstop, hateschool	2
38	anorexia, ed, eatingdisorder, anarecovery, ana, eatingdisorderrecovery, edsoldier, prorecovery, beatana, foodisfuel, strongnotskinny, edfamily, edwarrior, edsoldiers, healthy, anawho, beated, anorexianervosa, edfam, edfighter, healthynotskinny, food , staystrong, togetherwecan, edwarriors, realrecovery, bulimiarecovery, fearfood, recoveryisworthit, eattolive, weightgain, anorexic, ednos, recover, recoveryisposMISle, anawarrior, foodporn, anasoldier, bulimia, realcovery, edarmy, recovering, fightana	2
32	anxiety, depression, mentalhealth, mentalillness, hospital, bipolar, bpd, recovery, support, ocd, addiction, insomnia, eatingdisorders, mental, borderline, therapy, ptsd, awareness, schizophrenia, panicattack, panic, kickanasass, stress, medicine, anxietyattack, schizophrenic, schizo, socialanxiety, hallucinations , selfharm, prosupps, addictions, mfp, abuse, bodyimage, mentaldisorder, paranoia, eatingdisorder, bingeeatingdisorder	3
52	ana, anorexia, mia, ed, bulimia, fat, anorexic, eatingdisorder, ednos, skinny, starve, bulimic, depression, depressed, binge, purge, thin, blithe, anamia, starving, donteat, ugly, size00, deb, fasting, fast, sad, suicidal, sue, bones , size0, anxiety, anatips, food, thinspoooo, cat, suicide, gross, diet, hungry	3
59	depression, depressed, suicide, suicidal, cutting, ana, blithe, worthless, anxiety, selfharmmm, cut, anorexia, sad, ugly, mia, fat, scars, cuts, anorexic, selfhate, alone, bulimia, selfharm, sue, deb, broken, killme, blood, death, cat, dead , lonely, cutter, secret_society123, blades, bulimic, die, pain, useless, hate	3
68	sad, depressed, depression, alone, pain, broken, hurt, lonely, lost, hate, help, crying, sadness, suicide, cry, suicidal, anxiety, love, tears, scared, dying, tired, life, done, death, sorry, blithe, heartbroken, helpme, society , die, girl, unhappy, worthless, dead, sadquotes, upset, quotes, empty, cutting	3
93	skinny, thin, thinspo, collarbones, hipbones, thighgap, bones, legs, perfect, beautiful, ana, body, diet, thinspiration, notme, want, weightloss, flatstomach, perfection, tiny, size0, pretty, thynspo, thinspoooo, ribs, girl, fat, stomach, fit, motivation , slim, perfectbody, thinstagram, size00, small, skinnylegs, thygap, fitspo, model, belly	3

Table 3: Examples of LDA derived topics and their associated human-annotated MIS. The top 40 most representative tags are displayed per topic. For MIS ratings 1 and 2, we have chosen to show two representative topics, whereas for MIS rating 3, we show all five topics which were identified by our annotators to be of MIS rating 3.

was not biased by class imbalance. The weighted sum of the coefficients may be interpreted as being equivalent to taking the majority vote from classifiers trained on bootstrap samples of the training data [23]. This ensemble regression model given by Adaboost was finally applied on the 10% heldout set to report performance on predicting MIS rating.

RESULTS

Description of High MIS Topics

We begin with a discussion of the high MIS topics derived from combining the LDA model and human annotations. Recall, out of the 100 topics generated by the LDA model, five were rated by the clinicians and researchers to have high MIS. From Table 3, we observe that tags like "depression", "anxiety", "suicide", "eatingdisorder", "ugly", "skinny", "self harm" consistently appear in almost all of the high MIS topics. However there are systematic differences in these topics despite the presence of these tags.

Statistical observations of these tags establishes that these topics are different. We performed a Kruskal-Wallis one-way analysis of variance test on the tag distributions (i.e., normalized frequency of tags) across all of the 100 topics; the test indicated statistically significant differences to exist in the content of the topics ($F = 8.94; p < 10^{-7}$). We then performed post-hoc analysis using Tukey's range test to identify whether the pairs of topics among the five high MIS topics are significantly different from each other. This indicated significance at the $p < .001$ level.

Qualitative differences are noticable across the high MIS topics. The first topic, topic 32, discusses expressions of a variety of different mental health disorders (e.g., "bpd", "ptsd", "social anxiety", "schizophrenic", "ocd", "panic"). The topic also contains mentions of stress and anxiety, some of the known concomitants of mental illnesses disorders like "addictions" and "insomnia", as well as some of the probable causes behind them (e.g., "abuse"). Additionally, the topic

revolves around calls for support, desire or need for recovery, and mentions of treatment efforts (e.g., "medicine"). Note that while eating disorders as a distinct illness in the DSM-5 [4], many other mental health disorders have high comorbidities — this explains the nature of content of this topic.

Next, topic 52 centers around commonly adopted harmful/dangerous habits of pro-ED lifestyles. For instance, we observe mentions of "binge", "purge", "starving", "blithe", "fast", "hungry". We also observe manifestation of self-loathing, e.g., "ugly", "gross", "fat". These tags capture attitudes and thoughts that reinforce these lifestyles ("donteat") and seem to share motivations towards continuing to do so ("anatips"). Literature identifies such manifestation of self-identification with pro-ED lifestyles to be an indirect form of self-injurious behavior [19].

Topic 59 expresses hopelessness and dejection, including suicidal thoughts. Tags like "worthless", "alone", "broken", and "useless" reveal expression of lowered sense of self-esteem. Further, tags such as "selfharmmm", "cut", "scars", and "blood", "blades" capture explicit, graphic description of MIS. Such graphic description is known to allow normalization of behavior as the only way to deal with emotional distress [10]. The topic further illustrates the use of Instagram as a place to share markers of such behavior — "cutter" is a tag used to identify oneself as struggling with such challenges. Finally, we observe "killme", "death", "dead", "die" to be associated with extreme thoughts of hurting and killing oneself. Together, the tags in this topic capture a form of emotion expression that is known to align with perceptions of a conflicted identity [28].

Topic 68 expresses loneliness, pain, thoughts of death and desire to seek help and advice on these challenges. There are many tags capturing broad negative affect in this topic — "sad", "hurt", "lost", 'cry", "tears", "tired", "heartbroken", "unhappy". Regret and moroseness are visible through tags: "sorry", "done", empty", "sadquotes".

Finally, topic 93 is about manifestation of desire to be skinny and expression of normative perceptions of body image ("skinny", "perfect", "beautiful", "body", "tiny", "size00", "pretty", "slim", "perfect body", "model", "small"). The topic also includes descriptions of physical attributes of body ("collarbones", "hipbones", "skinnylegs", "thygap", "thighgap", "flatstomach", "stomach", "ribs", "bones", "legs", "belly"). Such description of potentially injurious attitudes and beliefs may indicate promoting co-construction of the pro-ED identity — "thinspiration", "motivation", "thinstagram", and "fitspo".

Together, we conclude that our hybrid model that combines LDA topics with human annotations on MIS captures a number of distinctive topics relating to MIS of pro-ED content sharing Instagram users.

Dynamics of MIS

Next we analyze levels of MIS in our data as well as their change over time. Table 4 gives the proportion of posts with low (1), medium (2), and high (3) MIS. The majority of posts are low MIS (88.8%) and the rest span medium and high MIS. While in general users who share pro-ED content are expected to show markers of high MIS through the sharing of content around physically or emotionally dangerous acts (Table 3), we find that a notable fraction of these users also use the Instagram platform for sharing non-mental illness related content, as indicated by the low level of MIS in majority of the posts.

MIS Rating	Post Count	Percentage
Low (1)	22,913,989	87.41%
Medium (2)	1,990,031	7.59%
High (3)	1,311,288	5.00%

Table 4: Proportion of posts with different MIS rating.

However, a deeper examination of the way the three levels of MIS change over time reveals that, while small in the proportion of posts, alarmingly, the *relative fraction* of users who share pro-ED content and show high MIS has been on the rise. Figure 2 presents the proportion of "active" users in our data with low, medium, and high MIS per month (55 months in all: Oct 2010 through Mar 2015). The figure indicates that from month 18 (Mar 2012) to month 48 (Oct 2014), both medium and high MIS rating user proportions show a steep increase, whereas low MIS rating shows decline during the same period. In fact at its peak, as many as 10% of users in our data are inferred to express high MIS rating. A Wilcoxon rank sum test shows statistically significant differences between the fraction of users with high MIS rating in month 48 and in month 18 ($z = 4.19; p < 10^{-5}$). This indicates that change of proportional volume of users with medium/high MIS rating over time is significant.

We also compute a rate of change metric to capture the trend of the proportion of medium and high MIS rating users over time, known as *momentum* — a measure that observes changes in time series data. It is given as the mean ratio of the difference between medium/high MIS rating users at time t and that at $t-1$ to the medium/high MIS rating users at time $t-1$. The momentum gives a positive value (13%/year), which indicates that overall proportion of pro-ED Instagram users with medium/high MIS rating in their content has been monotonically increasing over time.

MIS Rating Prediction

Fitting Models. In this final subsection, we start by reporting the measures of fitting the MIS rating prediction model on the bootstrap samples based on our regularized multinomial logistic regression framework. One of the aspects of this investigation was determining the appropriate sliding window size w for which we obtain the best model fit on the bootstrap samples as well as the one for which the prediction performance is optimal.

In Table 5 we present results on fitting a number of models to our bootstrap training samples with different values of the sliding window w. We report model performance in terms of *deviance*. Deviance measures the lack of fit to data, and lower values are better. It is calculated by comparing a model with the saturated model — a model with a theoretically perfect fit, or the intercept-only model – we refer to it

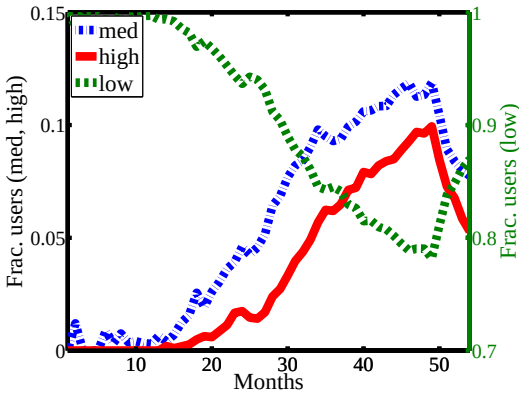


Figure 2: Fraction of users with low (1), medium (2), and high (3) MIS rating over time (in months). Here the fraction of users with a particular level of MIS in a certain month is given as the ratio between the number of "active" users with the particular MIS rating to the total number of "active" users during that month — "active" being defined as any user with at least one post during the month. Thus these fractions allow normalization and comparison of MIS ratings across months.

Model	Deviance	χ^2	p-value
Null model	915.523		
$w = 1$	633.028	282.495	$p < 10^{-4}$
$w = 2$	533.195	382.328	$p < 10^{-7}$
$w = 3$	489.055	426.468	$p < 10^{-8}$
$w = 5$	320.625	594.898	$p < 10^{-8}$
$w = 7$	279.801	635.722	$p < 10^{-10}$
$w = 10$	297.81	617.713	$p < 10^{-10}$
$w = 13$	319.259	596.264	$p < 10^{-10}$
$w = 15$	327.505	588.018	$p < 10^{-8}$
$w = 17$	417.77	497.753	$p < 10^{-7}$
$w = 20$	643.989	271.534	$p < 10^{-5}$

Table 5: Summary metrics of fitting the regularized multinomial logit model to our training data. Ten different models with different sliding window sizes w are reported, along with the null (intercept only) model.

as the "null model". Compared to the null model, all versions of our regularized multinomial logit models (with different values of w) provide considerable explanatory power with significant improvements in deviances. The difference between the deviance of the null model and the deviances of our models approximately follows a χ^2 distribution, with degrees of freedom equal to the number of additional variables in the more comprehensive model. As an example, comparing the deviance of $w = 5$ model with that of the null model, we see that the information provided by the MIS ratings over the past five months has significant explanatory power: $\chi^2(5, N = 90K) = 915.523 - 320.625 = 594.898, p < 10^{-8}$. This comparison with the null model is statistically significant after Bonferroni correction for multiple testing ($\alpha = 0.005$ since we consider 10 different models corresponding to the 10 values of sliding window w). The best model fit (in terms of lowest deviance) is given by the $w = 7$ model

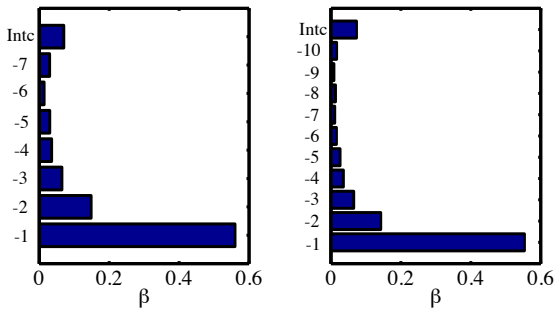


Figure 3: Coefficient weights (β) for model fits corresponding to the two best models: $w = 7$ and $w = 10$. "Intc" is intercept, and $-i$ is the coefficient weight corresponding to the i-th predictor variable, in other words it is the MIS rating at time $t - i$ when the prediction is for MIS rating at time t.

Model	Accuracy (%)	Precision	Recall	F1
$w = 1$	59.89	0.647	0.655	0.651
$w = 2$	62.76	0.674	0.706	0.690
$w = 3$	67.98	0.676	0.724	0.699
$w = 5$	69.32	0.723	0.797	0.758
$w = 7$	81.89	0.817	0.804	0.810
$w = 10$	73.99	0.808	0.790	0.799
$w = 13$	72.26	0.808	0.754	0.780
$w = 15$	67.42	0.774	0.728	0.751
$w = 17$	66.32	0.684	0.619	0.650
$w = 20$	61.50	0.638	0.617	0.627

Table 6: Predicting low, medium, high MIS ratings of users in heldout test set using the regularized multinomial logit model.

($\chi^2(7, N = 90K) = 915.523 - 279.801 = 635.722, p < 10^{-10}$), with best fits (low deviance) for models where w is closer to $w = 7$ and decreasing as w goes lower or higher.

Next, in Figure 3 we report the mean weights of the coefficients (β) in the models generated on the bootstrap training samples. We report on two models with best model fits in Table 5. We observe that the weights of the coefficients monotonically decrease backwards in time from the time when response (MIS) is estimated — that is, if the model is fit for MIS in month t_i, then the predictor variable (MIS) in month t_j has a larger coefficient compared to the predictor variable in t_k, if month t_j is closer to t_i relative to month t_k. Expectedly, more recent values of MIS in a user's content provide greater explanatory power towards future MIS values.

Prediction on Heldout Data. In the next part of our MIS prediction analysis, we present the performance of our approach when tested on the heldout test dataset of 10K users. We summarize performance metrics of this multi-class classification via Table 6. We report average accuracy, precision, recall, and F1 measures across the 10 different sliding window choices of our logit model. As also observed in the results on model fit, the highest accuracy (82%) and F1 (82%) are given by model $w = 7$. Specifically, we obtain high accuracy and F1 for the 1 (low) and 3 (high) MIS ratings (accuracy: 89% and

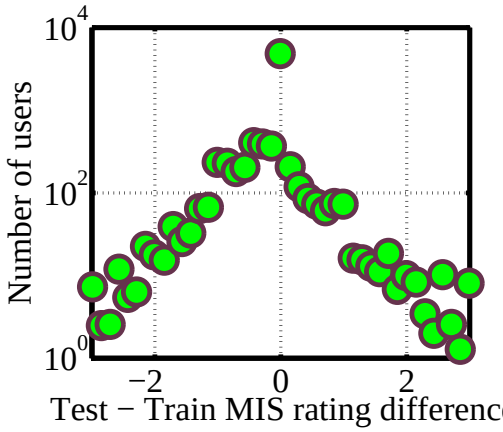


Figure 4: Distribution of differences between MIS ratings in test set and those in the training set for all correct predictions given by the $w = 7$ model (w is sliding window size). Positive difference (right side of the distribution) implies the model was able to predict that in the eighth month, the MIS rating measured from content of a user was *higher* than the mean MIS rating over the seven months preceding it.

87% respectively; and F1: 87% and 84% respectively). The performance is found to be relatively lower for MIS rating 2 (medium) (accuracy 70% and F1 71%), which we attribute to the ambiguity and subjectivity in such content. However all three classes perform above baseline models (majority vote models; since we use bagging and boosting, we are able to test against a baseline of 33% — equally split three classes), showing that past MIS rating of users measured from their Instagram postings is indeed able to forecast future MIS.

Predicting low/medium MIS to high MIS Transitions. Next, we wanted to investigate if our best performing model ($w = 7$) can predict increase in MIS in the future in cases where past MIS may be relatively lower. Being able to identify spike in MIS in a user's content where past MIS was lower can help direct appropriate help and support to those likely to be at a heightened risk in the future. We calculate the distribution of correctly classified users in our test set with respect to the difference between their predicted/test MIS rating and the mean MIS rating in historical/training data. Positive difference between predicted/test data MIS rating and that in historical/training data would indicate that our model predicted correctly the increase in MIS for the user's content (ref. distribution in Figure 4).

As reported earlier (Table 6), in the set of 10K users in our test dataset, 8,189 users are correctly classified i.e., we are able to infer MIS rating in their content in the eighth month correctly, based on ratings in the seven months before. We find that in the distribution given in Figure 4, for bulk of these correctly classified users predicted future MIS rating is approximately the same as past (notice the spike near zero difference). However, there is a notable fraction of the users ($\sim$ 11%) for whom our model correctly predicted an increase in MIS rating despite comparatively lower MIS rating values in the past (notice the right side of the distribution). Similarly, for $\sim$ 17% we were able to infer accurately a decrease in MIS in the future even though their MIS in the past was higher.

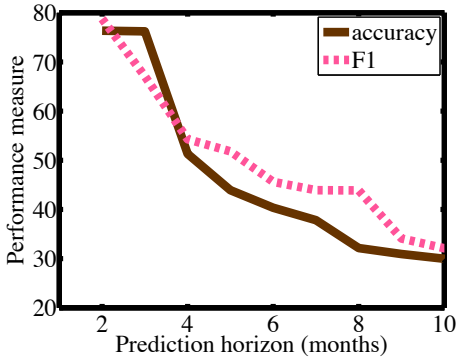


Figure 5: Change in accuracy and F1 score of the $w = 7$ MIS prediction model (w is sliding window size) for prediction horizons h between 2 months and 10 months. Here prediction horizon h implies the model is predicting MIS rating at month $t + h$ ($h > 1$) using MIS rating predictors from months $t - w$ through $t - 1$.

Broadly, this investigation shows that our model can not only infer future MIS based on historically manifested MIS of similar rating, but also is effective in capturing increases and decreases in MIS based on the content shared on Instagram.

Prediction Horizon. Finally, how far out into the future can we predict a user's MIS rating? In other words, so far we have shown that a model trained on MIS rating data between time $t - w$ and $t - 1$ (w is the sliding window size) can predict MIS rating at time t. Using the same model, could we predict MIS rating at time $t + h$, where $h > 1$? Here we define h as the "prediction horizon" and Figure 5 presents the accuracy and F1 measures of applying the $w = 7$ model (our best performing model from the previous paragraph) to predict MIS rating 2 months to 10 months out into the future ($h = 2, 3, ..., 10$).

Observations from Figure 5 indicate that, not surprisingly, performance as measured by accuracy and F1 steadily deteriorates as we attempt to predict MIS rating further out into the future. The best models, in other words, are found to be those where predictions are attempted to be made closer in time to the last observed time of MIS rating in the model (the $h = 2$ model gives an accuracy of 78% and F1 of 80%). However we do note that since bagging and boosting allows us to compare model performance at the baseline level (balanced class sizes, 33%), our predictions of MIS rating continue to be better than baseline through eight months into the future.

DISCUSSION

Summary of findings. Our results indicate that, despite the majority of posts in our data bearing low levels of mental illness severity (MIS), users that share pro-ED content on Instagram exhibit a trend of increasing MIS in their content over time. Examples of high MIS content spans from expression of negative self-perceptions to disordered thoughts about eating to graphic illustration of acts that could lead to physical and emotional harm or death. Notably, we have found that we can forecast a user's manifested MIS over time up to eight months into the future based on their MIS previously seen in their Instagram posts. We show that historical tagged content

and levels of MIS inferred from such content may predict the MIS level in a user's content in the future.

Implications of increase in MIS over time. A significant finding of our work is the steady increase in the proportion of medium and high MIS users who share pro-ED content over time. While pro-ED users would be expected to have a higher manifestation of MIS, we show that there is *a relative increase* in medium and high MIS expressed in content over time. Although causal relationships cannot be inferred without a more systematic investigation, examining *why* high MIS occurs opens up several avenues for future research. Are individual users showing markers of riskier behaviors because of contagion effects of the content they consume on Instagram? Could increases in MIS be indicative of a shift in social norms in the community?

It is important to note that, in response to media scrutiny in 2012 over the widespread prevalence of harmful content[8], Instagram banned some of the most common tags and started providing content advisories on certain others associated with self-injurious behavior, suicide, pro-ED, and related content. Per Instagram's policy, banning does not prevent use of a tag in a post but precludes these tags from returning in search results. Could high MIS tag usage reflect an adversarial response to these content moderation policies? Our conjecture springs from the observation that communities often adopt unprecedented and competing avenues to combat content regulation and moderation. For example, citizens of authoritarian regimes avoid censorship by embracing different forms of linguistic variation [39]. Similar adoption of unorthodox practices engender deviant communities engaging in cyberbullying and online harassment [55] as well as those involved in socially unacceptable or damaging activities (human trafficking, drug abuse, violence, or organized crime) [11]. Future research could explore the relationship between the observed increases in levels of MIS in connection with the larger moderation policy enforced by Instagram.

Design Implications

Limitations of Existing Efforts

Designing appropriate online interventions for populations at risk to mental illness has been of interest to the HCI and CSCW communities [37, 2]. Social media sites, in particular, have traditionally relied on active human interventions, such as reporting from peers, to identify users who may be sharing vulnerable or harmful content. In the context of pro-ED content with noticable levels of MIS, these efforts may be limiting. Often individuals sharing such content tend to remain socially cohesive, isolated from the larger online community, and exchange significant social support around the promotion and maintenance of harmful/dangerous behavior [46]. These factors make traditional methods of discovery and intervention challenging, if not impossible.

Some social media platforms have been making alternative efforts to contain the dissemination of such risky and vulnerable content. As mentioned earlier, Instagram already has a moderation policy in place for pro-self injury, pro-suicide, and pro-ED content. Pinterest outright bans such content, YouTube censors it, and Tumblr has taken a slightly different stance by providing public service announcements when a user searches on such tags for the first time. Certainly, there are other platforms which do not take specific action on such content – examples include Twitter, Flickr and Reddit.

Interventions

Our analytical methods may be used to ameliorate current efforts to help discover vulnerable groups faster with greater coverage and in a more robust, data-driven manner. This is particularly valuable since MIS-prone individuals are known to be 100 times more likely to commit suicide compared to the general population [43]. In fact, we believe our models can be used not only to predict individual MIS risk, but also to identify at-risk communities outside of the pro-ED community. Importantly, being able to trigger appropriate interventions to help vulnerable communities may also help dampen the effect of social transmission of such content and behaviors and even lead to dispersion of such clustered and cohesive groups. We propose the following design considerations:

(1) Social and psychiatric support. Searches on content with high MIS ratings by our method may automatically be directed to links hosting helpful and research-supported resources, highlighting the health risks of such activities. Similarly, our methods could be used to detect whether a user is attempting to post triggering content (e.g., pictures of cutting). At that point the system could interject with a private message that provides link to an appropriate psychological disorder helpline.

(2) Self-monitoring. Individuals may volunteer to self-monitor their MIS estimated through our methods from their social media content. Those keen on recovery may generate and share abstracted trends of their MIS with a trusted friend, family member, or therapist. These logs of risk over time may provide more temporally nuanced assessments on MIS than is possible through surveys, interviews, or other self-reported information. This information can complement existing forms of psychological therapy, help establish rapport between therapists and clients, and help overcome difficulties encountered in these settings due to clients' reluctance in sharing sensitive information about their mental health. However, disclosures through social media also raise ethical issues for clinicians who need additional information in order to make a determination of their patient's level of risk to his or herself. Use of our rating system would need to be negotiated on a personal basis to ensure that each patient receives the best treatment or intervention.

(3) Early-warning systems. Online communities and social media platforms struggling with the contentious issue of pro-ED, eating disorders, and mental illness more broadly may leverage our method and findings to design and deploy personalized early-warning systems. In cases when inferred vulnerability is forecasted to surpass safety levels, support communities may be engaged who can help and provide encouragement and psychosocial support. In fact, many online communities also harbor a thriving "recovery" community to discourage the adoption and manifestation of such at-risk behav-

[8] http://www.telegraph.co.uk/technology/social-media/9775559/Concerns-raised-over-Instagram-after-app-allows-users-to-see-photos-promoting-anorexia.html

iors. For instance, "StopSelfHarm" is a subreddit on Reddit; tags "stopselfharm", "edrecovery", and "dontgiveup" are often used on Instagram and Tumblr to raise awareness about the harmful effects of eating disorders and self-injurious behavior. For individuals whose social media content bears markers of high mental health severity, such content may be promoted or recommended in their respective social media profiles, so that it increases exposure to alternative perspectives on the issue.

Ethical Considerations

We note that intervention design in this space needs to honor the privacy of the sensitive populations we study. It is also crucial to critically consider ethical issues arising with algorithmically-driven interventions, such as confidentiality, risk of false alarms, and potential loss of control caused by surveillance, forecasting, and regulation. Although we consider public data in this work, social media posts are personal information, and therefore appropriate ways need to be devised that allow secure transmission and storage of content used or generated by the interventions.

Mental illness is also a controversial topic. Is manifestation of self-injurious behavior or suicidal ideation tendencies in a public social media platform a "bad" thing? Who decides what is "good" and what is "bad"? How can interventions be implemented without infringing on the right for individuals to express their ideas? We also need to think carefully about the consequences of any kind of intervention. Communities around mental illness have been known to prefer remaining hidden, and any intervention made through social media platforms like Instagram needs to ensure that it does not drive the community in the fringes where they would be difficult to discover and extend help to. Platforms that wish to design and implement interventions will need to consult with relevant stakeholders, such as clinicians, designers, and researchers, to balance the desire to help individuals who may need help, medically compliant/effective diagnosis, and the rights of users on these platforms.

Limitations and Future Work

Risk Factors of MIS Versus Diagnosis. We note that our prediction is built for population-level analysis. Care should be taken to extending these methods and findings to specific individuals without direct clinical advice. Although our work is corroborated by clinicians familiar with eating disorders, our model accounts for *risk factors to MIS only*. Our research does not make any claims to the Instagram users we study; it is not clear to what extent individual users meet DSM criteria for having an eating or other psychological disorder. We note that we base our estimates of MIS solely on the content posted on Instagram. We cannot be certain if the shared content is an actual reflection of one's mental health status, or there could be under or over-reporting of the risk factors of MIS. Thus our model does not claim to diagnose any user with a psychological disorder or that a specific user will indeed act out any of these dangerous behaviors. However, we do note the value of our contributions in providing a complementary source of data which can reveal markers of or levels of MIS above and beyond self-reported or other clinical diagnosis information.

Offline Behavior and Online Presence. We also recognize that individuals, depending on their perception of the social audience on Instagram, may carefully craft and maneuver how they report MIS through their content. Our method solely depends on the nature of content (tags) in the Instagram posts, and hence our findings cannot account for inter-individual differences due to personality attributes, self-presentation, identity manifestation, or self-censorship [26]. However, since we measure changes in MIS *per user* over time, we measure *relative increase or decrease in MIS* and we believe that our method is not affected by these concerns.

Beyond Eating Disorders. We presented a study focusing on one kind of community (pro-ED) that shares content with markers of high MIS. While the nature of MIS this community portrays on Instagram is incredibly varied, it is possible they do not include certain forms of MIS more prevalent in other communities. We suggest caution in generalizing our findings to arbitrary mental disorder communities or arbitrary social media platforms; however, it remains an exciting direction for future research.

Future Directions. Finally, although we found that past low, medium, or high MIS may be predictive of future MIS, we need deeper understanding of the drivers and design characteristics of the corresponding social media platform where this kind of dangerous and vulnerable content is shared. Additionally, future research could collaborate with clinicians and their patients to examine the kinds of content patients and those in recovery share on social media.

CONCLUSION

In this paper, we provided one of the first empirical insights into quantifying and characterizing levels of mental illness severity (MIS) in social media content. Specifically, we proposed a new method to predict future MIS in users who share pro-ED content on Instagram. Our method combines LDA topic modeling with clinicially-grounded human annotations on MIS. Our results on a large dataset of 26M posts and 100K users showed that prior content inferred to be of low, medium or high MIS can be used to predict future MIS rating with 81.5% accuracy. This method was also found to perform better than baseline up to eight months into the future. Importantly, we found that the proportional volume of Instagram users who share pro-ED content and are inferred to express heightened MIS has increased since 2012. We hope our findings to be valuable in the design of interventions that can bring help and support to this proliferating vulnerable community on social media platforms.

REFERENCES

1. Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You Tweet What You Eat: Studying Food Consumption Through Twitter. In *Proc. CHI*.
2. Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2015. Depression-related Imagery on Instagram. In *Proc. CSCW'15 Companion*. 231–234.
3. Jon Arcelus, Alex J Mitchell, Jackie Wales, and Sren Nielsen. 2011. Mortality rates in patients with anorexia nervosa and other eating disorders: a meta-analysis of 36 studies. *Archv. Gen. Psychiatry* 68, 7 (2011), 724–731.

4. American Psychiatric Association and others. 2013. *Diagnostic and statistical manual of mental disorders, (DSM-5)*. American Psychiatric Pub.
5. Jin Yeong Bak, Suin Kim, and Alice Oh. 2012. Self-disclosure and relationship strength in twitter conversations. In *Proc. ACL*. 60–64.
6. Anna M Bardone-Cone and Kamila M Cass. 2006. Investigating the impact of pro-anorexia websites: A pilot study. *European Eating Disorders Review* 14, 4 (2006), 256–262.
7. David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)* 3 (2003), 993–1022.
8. 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.
9. Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
10. Jay Callahan. 1996. A specific therapeutic approach to suicide risk in borderline clients. *Clinical Social Work Journal* 24, 4 (1996), 443–459.
11. Richard F Catalano and J David Hawkins. 1996. A Theory of Antisocial Behavior. *Delinquency and crime: Current theories* (1996), 149.
12. Laurence Claes, Koen Luyckx, Patricia Bijttebier, Brianna Turner, Amarendra Ghandi, Jos Smets, Jan Norre, Leen Van Assche, Els Verheyen, Yvienne Goris, and others. 2015. Non-Suicidal Self-Injury in Patients with Eating Disorder: Associations with Identity Formation Above and Beyond Anxiety and Depression. *European Eating Disorders Review* 23, 2 (2015), 119–125.
13. Michael Collins, Robert E Schapire, and Yoram Singer. 2002. Logistic regression, AdaBoost and Bregman distances. *Machine Learning* 48, 1-3 (2002), 253–285.
14. Karen Conterio, W Lader, and JK Bloom. 1998. Bodily harm: The breakthrough treatmentprogramfor self-injurers. (1998).
15. Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proc. ACL CLCP Workshop*.
16. Munmun De Choudhury. 2015. Anorexia on Tumblr: A Characterization Study. In *Proc. Digital Health*.
17. Munmun De Choudhury, Scott Counts, Eric Horvitz, and Aaron Hoff. 2014. Characterizing and Predicting Postpartum Depression from Facebook Data. In *Proc. CSCW*.
18. Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proc. ICWSM*.
19. Sharon K Farber, Craig C Jackson, Johanna K Tabin, and Eytan Bachar. 2007. Death and annihilation anxieties in anorexia nervosa, bulimia, and self-mutilation. *Psychoanalytic Psych* 24, 2 (2007), 289.
20. Angela Favaro and Paolo Santonastaso. 1998. Impulsive and compulsive self-injurious behavior in bulimia nervosa: Prevalence and psychological correlates. *The Journal of Nervous and Mental Disease* 186, 3 (1998), 157–165.
21. Armando R Favazza, Lori DeRosear, and K Conterio. 1989. Self-Mutilation and Eating Disorders. *Suicide & Life-Threatening Behavior* 19, 4 (1989), 352–361.
22. Rachel A Fleming-May and Laura E Miller. 2010. I'm scared to look. But I'm dying to know: Information seeking and sharing on Pro-Ana weblogs. *Proc. ASIST* 47, 1 (2010), 1–9.
23. Yoav Freund, Robert E Schapire, and others. 1996. Experiments with a new boosting algorithm. In *ICML*, Vol. 96. 148–156.
24. W Gaebel, H Zäske, and AE Baumann. 2006. The relationship between mental illness severity and stigma. *Acta Psychiatrica Scandinavica* 113, s429 (2006), 41–45.
25. Sarah A St Germain and Jill M Hooley. 2012. Direct and indirect forms of non-suicidal self-injury: Evidence for a distinction. *Psychiatry research* 197, 1 (2012), 78–84.
26. Erving Goffman and others. 1959. The presentation of self in everyday life. (1959).
27. Kim L Gratz, Sheree Dukes Conrad, and Lizabeth Roemer. 2002. Risk factors for deliberate self-harm among college students. *American Journal of Orthopsychiatry* 72, 1 (2002), 128.
28. P Hasking, R Momeni, S Swannell, and S Chia. 2008. The nature and extent of non-suicidal self-injury in a non-clinical sample of young adults. *Archives of Suicide Research* 12, 3 (2008), 208–218.
29. Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *Proc. NIPS*.
30. Christopher M Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Alm. 2014a. Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale. *ACL 2014* (2014), 107.
31. Christopher M Homan, Naiji Lu, Xin Tu, Megan C Lytle, and Vincent Silenzio. 2014b. Social structure and depression in TrevorSpace. In *Computer-Supported Cooperative Work and Social Computing (CSCW)*.
32. James I Hudson, Eva Hiripi, Harrison G Pope, and Ronald C Kessler. 2007. The prevalence and correlates of eating disorders in the National Comorbidity Survey Replication. *Biological psychiatry* 61, 3 (2007), 348–358.
33. Grace J Johnson and Paul J Ambrose. 2006. Neo-tribes: The power and potential of online communities in health care. *Commun. ACM* 49, 1 (2006), 107–113.
34. Adrienne S Juarascio, Amber Shoaib, and C Alix Timko. 2010. Pro-eating disorder communities on social networking sites: a content analysis. *Eating disorders* 18, 5 (2010), 393–407.

35. Kelly L Klump, Cynthia M Bulik, Walter H Kaye, Janet Treasure, and Edward Tyson. 2009. Academy for eating disorders position paper: eating disorders are serious mental illnesses. *Int J Eat Disord* 42, 2 (2009), 97–103.
36. Katrina Kostro, Jessica B Lerman, and Evelyn Attia. 2014. The current status of suicide and self-injury in eating disorders: a narrative review. *Journal of Eating Disorders* 2, 1 (2014), 19.
37. Reeva Lederman, Greg Wadley, John Gleeson, Sarah Bendall, and Mario Álvarez-Jiménez. 2014. Moderated online social therapy: Designing and evaluating technology for mental health. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 1 (2014), 5.
38. Stephen P Lewis and Alexis E Arbuthnott. 2012. Searching for thinspiration: the nature of internet searches for pro-eating disorder websites. *Cyberpsychology, Behavior, and Social Networking* 15, 4 (2012), 200–204.
39. Rebecca MacKinnon. 2008. Flatter world and thicker walls? Blogs, censorship and civic discourse in China. *Public Choice* 134, 1-2 (2008), 31–46.
40. Diana MacLean, Sonal Gupta, Anna Lembke, Christopher Manning, and Jeffrey Heer. 2015. Forum77: An Analysis of an Online Health Forum Dedicated to Addiction Recovery. In *Proc. CSCW*.
41. Jeanne B Martin. 2010. The development of ideal body image perceptions in the United States. *Nutrition Today* 45, 3 (2010), 98–110.
42. Terence C Mills. 1991. *Time series techniques for economists*. Cambridge University Press.
43. Paul Moran, Carolyn Coffey, Helena Romaniuk, Craig Olsson, Rohan Borschmann, John B Carlin, and George C Patton. 2012. The natural history of self-harm from adolescence to young adulthood: a population-based cohort study. *The Lancet* 379, 9812 (2012), 236–243.
44. Elizabeth L Murnane and Scott Counts. 2014. Unraveling abstinence and relapse: smoking cessation reflected in social media. In *Proc. CHI*.
45. Matthew K Nock and Mitchell J Prinstein. 2004. A functional approach to the assessment of self-mutilative behavior. *Journal of consulting and clinical psychology* 72, 5 (2004), 885.
46. Mark L Norris, Katherine M Boydell, Leora Pinhas, and Debra K Katzman. 2006. Ana and the Internet: A review of pro-anorexia websites. *International Journal of Eating Disorders* 39, 6 (2006), 443–447.
47. Minsu Park, David W McDonald, and Meeyoung Cha. 2013. Perception Differences between the Depressed and Non-depressed Users in Twitter. In *Proc. ICWSM*.
48. Michael J Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proc. ICWSM*.
49. Rebecka Peebles, Jenny L Wilson, and James D Lock. 2011. Self-injury in adolescents with eating disorders: Correlates and provider bias. *Journal of Adolescent Health* 48, 3 (2011), 310–313.
50. Mirella Ruggeri, Morven Leese, Graham Thornicroft, Giulia Bisoffi, and Michele Tansella. 2000. Definition and prevalence of severe and persistent mental illness. *The British Journal of Psychiatry* 177, 2 (2000), 149–155.
51. Hansen Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, and Lukasz et al. Dziurzynski. 2013. Characterizing Geographic Variation in Well-Being Using Tweets. In *ICWSM*.
52. Yukari Seko. 2013. Picturesque Wounds: A Multimodal Analysis of Self-Injury Photographs on Flickr. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, Vol. 14.
53. Leslie Regan Shade. 2003. Weborexics: The Ethical Issues Surrounding Pro-Ana Websites. *SIGCAS Comput. Soc.* 33, 4 (Dec. 2003).
54. Frédérique RE Smink, Daphne van Hoeken, and Hans W Hoek. 2012. Epidemiology of eating disorders: incidence, prevalence and mortality rates. *Current psychiatry reports* 14, 4 (2012), 406–414.
55. 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.
56. Karen L Suyemoto. 1998. The functions of self-mutilation. *Clinical psych rev* 18, 5 (1998), 531–554.
57. Brianna J Turner, Angelina Yiu, Brianne K Layden, Laurence Claes, Shannon Zaitsoff, and Alexander L. Chapman. 2015. Temporal associations between disordered eating and nonsuicidal self-injury: examining symptom overlap over 1 year. *Behavior Therapy* 46, 1 (2015), 125–138.
58. Robert F Valois, Keith J Zullig, and Amy A Hunter. 2013. Association between adolescent suicide ideation, suicide attempts and emotional self-efficacy. *Journal of Child and Family Studies* 24, 2 (2013), 237–248.
59. Markus Wolf, Florian Theis, and Hans Kordy. 2013. Language Use in Eating Disorder Blogs Psychological Implications of Social Online Activity. *Journal of Language and Social Psychology* 32, 2 (2013), 212–226.
60. Elad Yom-Tov, Luis Fernandez-Luque, Ingmar Weber, and P Steven Crain. 2012. Pro-Anorexia and Pro-Recovery Photo Sharing: A Tale of Two Warring Tribes. *J Med Internet Res* (2012).