Predicting Depression via Social Media

Munmun De Choudhury **Michael Gamon** **Scott Counts** **Eric Horvitz**

Microsoft Research, Redmond WA 98052
{munmund, mgamon, counts, horvitz}@microsoft.com

Abstract

Major depression constitutes a serious challenge in personal and public health. Tens of millions of people each year suffer from depression and only a fraction receives adequate treatment. We explore the potential to use social media to detect and diagnose major depressive disorder in individuals. We first employ crowdsourcing to compile a set of Twitter users who report being diagnosed with clinical depression, based on a standard psychometric instrument. Through their social media postings over a year preceding the onset of depression, we measure behavioral attributes relating to social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications. We leverage these behavioral cues, to build a statistical classifier that provides estimates of the risk of depression, *before* the reported onset. We find that social media contains useful signals for characterizing the onset of depression in individuals, as measured through decrease in social activity, raised negative affect, highly clustered egonetworks, heightened relational and medicinal concerns, and greater expression of religious involvement. We believe our findings and methods may be useful in developing tools for identifying the onset of major depression, for use by healthcare agencies; or on behalf of individuals, enabling those suffering from depression to be more proactive about their mental health.

Introduction

Mental illness is a leading cause of disability worldwide. It is estimated that nearly 300 million people suffer from depression (World Health Organization, 2001). Reports on lifetime prevalence show high variance, with 3% reported in Japan to 17% in the US. In North America, the probability of having a major depressive episode within a one year period of time is 3–5% for males and 8–10% for females (Andrade et al., 2003).

However, global provisions and services for identifying, supporting, and treating mental illness of this nature have been considered as insufficient (Detels, 2009). Although 87% of the world's governments offer some primary care health services to tackle mental illness, 30% do not have programs, and 28% have no budget specifically identified for mental health (Detels, 2009). In fact, there is no reliable laboratory test for diagnosing most forms of mental illness; typically, the diagnosis is based on the patient's self-reported experiences, behaviors reported by relatives or friends, and a mental status examination.

In the context of all of these challenges, we examine the potential of social media as a tool in detecting and predicting affective disorders in individuals. We focus on a common mental illness: Major Depressive Disorder or MDD[1]. MDD is characterized by episodes of all-encompassing low mood accompanied by low self-esteem, and loss of interest or pleasure in normally enjoyable activities. It is also well-established that people suffering from MDD tend to focus their attention on unhappy and unflattering information, to interpret ambiguous information negatively, and to harbor pervasively pessimistic beliefs (Kessler et al., 2003; Rude et al., 2004).

People are increasingly using social media platforms, such as Twitter and Facebook, to share their thoughts and opinions with their contacts. Postings on these sites are made in a naturalistic setting and in the course of daily activities and happenings. As such, social media provides a means for capturing behavioral attributes that are relevant to an individual's thinking, mood, communication, activities, and socialization. The emotion and language used in social media postings may indicate feelings of worthlessness, guilt, helplessness, and self-hatred that characterize major depression. Additionally, depression sufferers often withdraw from social situations and activities. Such changes in activity might be salient with changes in activity on social media. Also, social media might reflect changing social ties. We pursue the hypothesis that changes in language, activity, and social ties may be used jointly to construct statistical models to detect and even predict MDD in a fine-grained manner, including ways that can complement and extend traditional approaches to diagnosis.

Our main contributions in this paper are as follows:
(1) We use crowdsourcing to collect (gold standard) assessments from several hundred Twitter users who report that they have been diagnosed with clinical MDD, using the CES-D[2] (Center for Epidemiologic Studies Depression Scale) screening test.

[1] For the sake of simplicity, we would refer to MDD simply as "depression" throughout the paper.

(2) Based on the identified cohort, we introduce several measures and use them to quantify an individual's social media behavior for a year in advance of their reported onset of depression. These include measures of: user engagement and emotion, egocentric social graph, linguistic style, depressive language use, and mentions of antidepressant medications.

(3) We compare the behaviors of the depressed user class, and the standard user class through these measures. Our findings indicate, for instance, that individuals with depression show lowered social activity, greater negative emotion, high self-attentional focus, increased relational and medicinal concerns, and heightened expression of religious thoughts. Further, despite having smaller egonetworks, people in the depressed class appear to belong to tightly clustered close-knit networks, and are typically highly embedded with the contacts in their egonetwork.

(4) We leverage the multiple types of signals obtained thus to build an MDD classifier, that can predict, *ahead of MDD onset time*, whether an individual is vulnerable to depression. Our models show promise in predicting outcomes with an accuracy of 70% and precision of 0.74.

We believe that this research can enable new mechanisms to identify at-risk individuals, variables related to the exacerbation of major depression, and can frame directions on guiding valuable interventions.

Background Literature

Rich bodies of work on depression in psychiatry, psychology, medicine, and sociolinguistics describe efforts to identify and understand correlates of MDD in individuals. Cloninger et al., (2006) examined the role of personality traits in the vulnerability of individuals to a future episode of depression, through a longitudinal study. On the other hand, Rude et al., (2003) found support for the claim that negative processing biases, particularly (cognitive) biases in resolving ambiguous verbal information can predict subsequent depression. Robinson and Alloy, (2003) similarly observed that negative cognitive styles and stress-reactive rumination were predictive of the onset, number and duration of depressive episodes. Finally, Brown et al., (1986) found that lack of social support and lowered self-esteem are important factors linked to higher incidences of depression. Among a variety of somatic factors, reduced energy, disturbed sleep, eating disorders, and stress and tension have also been found to be correlates of depressive disorders (Abdel-Khalek, 2004).

In the field of sociolinguistics, Oxman et al., (1982) showed that linguistic analysis of speech could classify patients into groups suffering from depression and paranoia. Computerized analysis of written text through the LIWC program has also been found to reveal predictive cues about neurotic tendencies and psychiatric disorders (Rude, Gortner & Pennebaker, 2004).

Although studies to date have improved our understanding of factors that are linked to mental disorders, a notable limitation of prior research is that it relies heavily on small, often homogeneous samples of individuals, who may not necessarily be representative of the larger population. Further, these studies typically are based on surveys, relying on retrospective self-reports about mood and observations about health: a method that limits temporal granularity. That is, such assessments are designed to collect high-level summaries about experiences over long periods of time. Collecting finer-grained longitudinal data has been difficult, given the resources and invasiveness required to observe individuals' behavior over months and years.

We leverage continuing streams of evidence from social media on posting activity that often reflects people's psyches and social milieus. We seek to use this data about people's social and psychological behavior to predict their vulnerabilities to depression in an unobtrusive and fine-grained manner.

Moving to research on social media, over the last few years, there has been growing interest in using social media as a tool for public health, ranging from identifying the spread of flu symptoms (Sadilek et al., 2012), to building insights about diseases based on postings on Twitter (Paul & Dredze, 2011). However, research on harnessing social media for understanding behavioral health disorders is still in its infancy. Kotikalapudi et al., (2012) analyzed patterns of web activity of college students that could signal depression. Similarly, Moreno et al., (2011) demonstrated that status updates on Facebook could reveal symptoms of major depressive episodes.

In the context of Twitter, Park et al., (2012) found initial evidence that people post about their depression and even their treatment on social media. In other related work, De Choudhury et al., (2013) examined linguistic and emotional correlates for postnatal changes of new mothers, and built a statistical model to predict extreme postnatal behavioral changes using only prenatal observations. The latter work highlights the potential of social media as a source of signals about likelihood of current or future episodes of depression. With the present work we: (1) expand the scope of social media-based mental health measures, describing the relationship between nearly 200 measures and the presence of depression; and (2) demonstrate that we can use those measures to predict, *ahead* of onset, depressive disorders in a cohort of individuals who are diagnosed with depression via a standard psychometric instrument.

Data

Ground Truth Data Collection

We employ crowdsourcing to collect labels we take as ground truth data on the presence of MDD. Crowdsourcing is an efficient mechanism to gain access to behavioral data from a diverse population, is less time consuming, and is

inexpensive (Snow et al., 2008). Using Amazon's Mechanical Turk interface, we designed human intelligence tasks (HITs) wherein crowdworkers were asked to take a standardized clinical depression survey, followed by several questions on their depression history and demographics. The crowdworkers could also opt in to share their Twitter usernames if they had a *public* Twitter profile, with an agreement that their data could be mined and analyzed anonymously using a computer program. We sought responses from crowdworkers who were located in the United States, and had an approval rating on Amazon Mechanical Turk (AMT) greater than or equal to 90%. Each crowdworker was restricted to take the HIT exactly once, and was paid 90 cents for completing the task.

Depression Screening Test

We used the CES-D (Center for Epidemiologic Studies Depression Scale)[2] questionnaire as the primary tool to determine the depression levels of the crowdworkers. The CES-D is a 20-item self-report scale that is designed to measure depressive symptoms in the general population (Radloff, 1977), and is one of the most common screening tests used by clinicians and psychiatrists for the purpose. It measures symptoms defined by the American Psychiatric Association Diagnostic and Statistical Manual (DSM-IV), and quantifies depressive feelings and behaviors during the past week. For example, the test seeks responses to questions such as: "*I thought my life was a failure*"; "*I felt lonely*"; "*I had crying spells*". Participants were asked to choose one of the following responses to each of the questions: (i) Rarely or none of the time (<1 day); (ii) Some or a little of the time (1-2 days); (iii) Occasionally or a moderate amount of the time (3-4 days); and (iv) Most or all of the time (5-7 days). A participant's minimum score could be zero and the maximum could be 60, where higher scores indicate the presence of more symptomatology.

Tackling Noisy Crowd Responses

While crowdsourcing has its benefits as a behavioral data collection paradigm, it may suffer from the problem of noisy responses to HITs. We control for this using two steps: (1) we discard data points where crowdworkers take insufficient time (less than two minutes) to complete the survey and (2) we deploy an auxiliary screening test in the same HIT, in addition to the CES-D questionnaire.

We used the Beck Depression Inventory (BDI) for this purpose (Beck et al., 1976). Like CES-D, BDI is also used commonly as a test for depression by healthcare professionals to measure depression. Our conjecture was that, for high quality responses to HITs, the scores in CES-D and BDI would correlate—i.e., individuals who are truly depressed (or not depressed) would score high (or low) in both the tests. This would help us to eliminate data, stemming from responses that may have been input without careful deliberation.

[2] http://www.bcbsm.com/pdf/Depression_CES-D.pdf

Note that in order to minimize bias, we refrained from indicating in our HITs that the two tests were depression screening questionnaires. Rather, we simply mentioned that they measure behavioral patterns. We also randomized the order of the CES-D and BDI questionnaires in the HITs to avoid biases stemming from the ordering of the surveys.

Self-reported Information

Finally, we collected information about the crowdworkers' depression history and experiences by assessing:

- *Whether or not they had been diagnosed with clinical depression in the past. If so, when.*
- *If they were clinically depressed, what was the estimated time of its onset.*
- *If they are currently depressed, or using any anti-depression medications.*

We also asked if crowdworkers could share their Twitter username for research analysis purposes, should they be owners of a *public* Twitter profile.

Statistics of Ground Truth Data

A total of 1,583 crowdworkers completed our HITs between September 15 and October 31, 2012. 637 participants (~40%) agreed to provide us with access to their Twitter feeds. Next we eliminated noisy respondents based on the two-step technique discussed earlier, which yielded a set of 554 users. Finally, we intended to focus on individuals with depression onset dates anytime in the last one year, but no later than three months prior to the day the survey was taken. This ensured that we could collect reasonably long historical social media data for each user prior to the onset (important for prediction). In this set, we further focused on users who reported to have suffered from at least two depressive episodes during the one-year period, so as to qualify for MDD (Posternak et al., 2006).

A set of 476 users out of the above cohort indicated in the self-report section of the HIT to have been diagnosed with depression with onset between September 2011 and June 2012. This comprised our final user set. The set contained 243 males and 233 females, with a median age of 25, with the two most frequent education levels being "Some college, no degree" and "Bachelor's degree," and the most reported income range of "$25,000-$50,000".

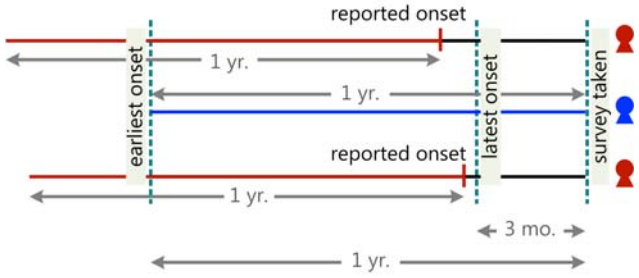

Figure 1: Twitter data collection method. For users (red) who scored positively for depression, we date back from the reported onset of depression up to one year, and collect all of their Twitter postings. For users (blue) scoring negatively for depression, we collect one year of their Twitter feeds dating back from the day the survey was taken.

In our final dataset of 476 users, we used the responses to the CES-D questionnaires as our primary depression level estimation tool. With CES-D, typically three groups of depression severity are calculated (Radloff, 1977): *low* (0-15), *mild to moderate* (16-22), and *high* range (23-60). The literature indicates that a cut-off of 16 is prone to yielding a high number of false-positives. Thus, several studies adopt higher cut-offs (e.g., 20, 22, or 27). In this work, we used 22 as our chosen threshold (also see Park et. al., (2012)), to minimize Type I and Type II errors in our subsequent prediction task.

We obtained 171 users (~36%) who scored positive for depression. We note that Park et. al., (2012) found a similar percentage of individuals in their study, scoring positively for depression. Two classes of users were constructed in this manner: an MDD positive class of 171 users scoring high for depression; and negative class of 305 users: ones with little or no signs of the condition.

Social Media Data

For behavioral exploration and prediction, we collected data from the Twitter feeds of all users. We used the Twitter Firehose made available to us via our organization's contract with Twitter. Depending on the depression onset date of each of the users in the positive class, we collected all of their Twitter posts in the past one year, dating back from the reported depression onset. For instance, for a user with depression onset date of July 17, 2012, we collected all of her posts made between July 17, 2011 and July 16, 2012. For users in the negative class, we collected all of their postings in the one year prior to the date they took our AMT survey. Figure 1 illustrates the data collection process. Note that we set an allowed time range for depression onset (max and min), in order to ensure that we captured users with a sufficiently long history of clinical MDD.

Total number of users	476
Total number of Twitter posts	2,157,992
Mean number of posts per user over the entire 1 year period	4,533.4
Variance of number of posts per user over the entire 1 year period	3,836
Mean number of posts per day per user	6.67
Variance of number of posts per day per user	12.42

Table 1: Statistics of Twitter data of study cohort.

Having a job again makes me happy. Less time to be depressed and eat all day while watching sad movies.
"Are you okay?" Yes…. I understand that I am upset and hopeless and nothing can help me… I'm okay… but I am not alright
"empty" feelings I WAS JUST TALKING ABOUT HOW I I HAVE EMOTION OH MY GOODNESS I FEEL AWFUL
I want someone to hold me and be there for me when I'm sad.
Reloading twitter till I pass out. *lonely* *anxious* *butthurt* *frustrated* *dead*

Table 2: Example posts from users in the depression class.

In Table 1 we list several statistics of the crawled dataset. A few samples of posts randomly selected from the data of users in the depression class are shown in Table 2.

Measuring Depressive Behavior

We first present a set of attributes that can be used to characterize the behavioral differences of the two classes of users—one of which consists of individuals exhibiting clinical depression, based on the year-long feed of their Twitter postings. Note these measures are defined to be dynamic measures, i.e., we calculate their values daily based on the activity of users over the entire year-long period preceding depression onset or the date survey was taken by them.

Engagement

We define five measures of engagement motivated from (De Choudhury et al., 2013): *Volume*, defined as the normalized number of posts per day made by the user; Proportion of *reply* posts (@-replies) from a user per day indicating her level of social interaction with other Twitter users; The fraction of *retweets* from a user per day, signaling how she participates in information sharing with her followers; The proportion of *links* (urls) shared by each user over a day; and fraction of *question-centric* posts from a user in a day, measuring the user's tendency to seek and derive informational benefits from the greater Twitter community.

Insomnia index: Our final engagement measure quantifies the pattern of posting made by a user during the course of a day (24 hour cycle). Literature on depression indicates that users showing depression signs tend to be active during the evening and night. Hence we define a "night" window as "9PM—6AM" (consequently the "day" window for the same user, in local time, would be "6:01AM-8:59PM"). For each user, we thus define the normalized difference in number of postings made between night window and day window to be the "*insomnia index*" on a given day.

Egocentric Social Graph

We define a number of egocentric network measures for users, based on both social graph structure, as well as the interactions with others on Twitter (through @-replies). These measures can be categorized into three types: (1) node properties (focusing on a particular user u); (2) dyadic properties (focusing on a user u and another user v with whom she interacts through an @-reply); and (3) network properties (focusing on a user u in the context of her entire egocentric network of @-reply exchanges). For the purposes of this paper, we consider the egocentric social graph of a user to be an undirected network of the set of nodes in u's two-hop neighborhood (neighbors of the neighbors of users in our dataset), where an edge between u and v implies that there has been at least one @-reply exchange each, from u to v, and from v to u on a given day.

(1) *Node properties.* We first define two measures that characterize the nature of a user's egocentric social network as in (De Choudhury et al., 2013). The first feature is the number of *followers* or inlinks of a user at a given day, while the second is the count of her *followees* or outlinks.

(2) *Dyadic properties.* In this category, we define a measure called *reciprocity*, which is measured as how many times a user u responds to another user v who had sent her @-reply messages. It is given by the mean of the ratio of the number of @-replies from u to any user v, to the number of @-replies from v to u. The second feature in this category is called the *prestige ratio*, and is given by the ratio of the number of @-replies that are targeted to u, to the number of @-replies targeted to a user v, where v is a user with whom u has a history of bi-directional @-replies.

(3) *Network properties.* In this final category, we define four measures. We define *graph density* to be the ratio of the count of edges to the count of nodes in u's egocentric social graph. The second feature is the *clustering coefficient* of u's ego network, which is a standard notion of local density, i.e. the average probability that two neighbors of u are neighbors of each other. The third feature, size of *two-hop neighborhood* is defined as the count of all of u's neighbors, plus all of the neighbors of u's neighbors. We define the next feature *embeddedness* of u with respect to her neighborhood as the mean of the ratio between the set of common neighbors between u and any neighbor v, and the set of all neighbors of u and v. The final feature in this category is the number of *ego components* in u's egonetwork, defined as the count of the number of connected components that remain when the focal node u and its incident edges are removed (De Choudhury et al., 2010).

Emotion

We consider four measures of the emotional state of users in our dataset: *positive affect* (PA), *negative affect* (NA), *activation,* and *dominance*. Daily measurements of PA and NA per user are computed using the psycholinguistic resource LIWC (http://www.liwc.net/), whose emotion categories have been scientifically validated to perform well for determining affect in Twitter (De Choudhury et al., 2013). We use the ANEW lexicon (Bradley & Lang, 1999) for computing activation and dominance. Activation refers to the degree of physical intensity in an emotion ("terrified" is higher in activation than "scared"), while dominance refers to the degree of control in an emotion ("anger" is dominant, while "fear" is submissive).

Linguistic Style

We also introduce measures to characterize linguistic styles in posts from users (Rude et al., 2004). We again use LIWC for determining 22 specific linguistic styles, e.g.: *articles, auxiliary verbs, conjunctions, adverbs, personal pronouns, prepositions, functional words, assent, negation, certainty* and *quantifiers.*

Depression Language

Finally we define two specialized features focused on characterizing the topical language of individuals detected positively with depression. While our previous measure focused on the linguistic style of depressive language, we are also interested in analyzing *what* people talk about.

(*a*) *Depression lexicon.* The first feature measures the usage of depression-related terms, defined broadly, in Twitter posts. For this purpose, we built a lexicon of terms that are likely to appear in postings from individuals discussing depression or its symptoms in online settings. We mined a 10% sample of a snapshot of the "Mental Health" category of Yahoo! Answers. In addition to already being categorized as relevant to depression, these posts are separated into questions and answers and are relatively short, making them well-aligned to the construction of a depression lexicon that can eventually be deployed on Twitter.

We extracted all questions and the best answer for each of question, resulting in 900,000 question/answer pairs. After tokenizing the question/answer texts, we calculated for each word in the corpus its association with the regex "depress*" using pointwise mutual information (PMI) and log likelihood ratio (LLR). We created the union of top 1% of terms in terms of LLR and PMI. To remove extremely frequent terms, we calculated the tf.idf for these terms in Wikipedia and used the top 1000 words with high tf.idf. Thereafter we deployed this lexicon to determine frequency of use of depression terms that appear in the Twitter posts of each user, on a given day.

(*b*) *Antidepressant usage.* The next feature measures the degree of use of names of antidepressants popular in the treatment of clinical depression (any possible overlap with the above lexicon was eliminated). Individuals with depression condition are likely to use these names in their posts, possibly to receive feedback on their effects during the course of treatment (Ramirez-Esparza et al., 2008). We used the Wikipedia page on "list of antidepressants" in order to construct a lexicon of drug names (http://en.wikipedia.org/wiki/List_of_antidepressants).

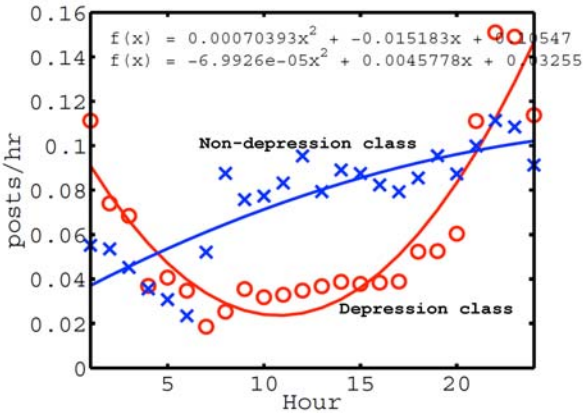

Figure 2: Diurnal trends (i.e. mean number of posts made hourly throughout a day) for the two classes. The line plots correspond to least squares fit of the trends.

Behavioral Attributes of Depression

In the light of the above measures, we present an exploration of behavior of depressed and non-depressed classes.

Diurnal Activity and Depression

Figure 2 shows the diurnal pattern of posting (in local time) from users of the two classes, measured as the mean number of posts made per hour, over the entire one year history of Twitter data of the users. We also show polynomial (of order 2) fits for both plots.

From the figure, we observe that for the non-depression class, most people are less active later in the night (i.e., post-midnight) and early in the morning, with activity generally increasing through the day. Evenings and early nights show peak, indicating that people are generally more likely to be using social media after the end of workday. On the other hand, the depression class trend shows peaks late in the night (post 8pm), with lower activity through the day (between 9am and 5pm). It is known from literature that for 8 out of 10 people suffering from depression, symptoms tend to worsen during the night (Lustberg & Reynolds, 2000). In fact, night time online activity is a known characteristic of these individuals, which may explain the increased levels of nighttime posting on Twitter.

Aggregated Behavior of Depression Sufferers

Next we discuss the patterns of some of the behavioral measures for both classes in Figure 3. We found marked differences across the two classes. At an aggregate level, for the depression class, we observe considerable decrease in user engagement measures, such as volume (38% lower; $p<.001$ based on a t-test) and replies (32% lower; $p<.001$). This indicates that these users are posting less, suggesting a possible loss of social connectedness. The same set of users exhibit higher expression of NA (28% higher; $p<.01$), possibly reflecting their mental instability and helplessness. Moreover, lower activation relative to the non-depression class (11% lower; $p<.01$) may indicate loneliness, restlessness, exhaustion, lack of energy, and sleep deprivation, all of which are known to be consistent depression symptoms (Rabkin & Struening, 1976; Posternak et al., 2006). Finally, we find that the presence of the first-person pronoun is considerably high (68% higher; $p<.0001$), while that of 3rd person pronouns is low in posts of the users in this class (71% lower; $p<.0001$), reflecting their high attention to self and psychological distancing from others (Rude et al., 2004).

Finally, we found that the use of depression terms in postings from the positive class are significantly higher (89% higher; $p<.0001$). To delve more deeply into the content shared by users in this class, we report the unigrams from the depression lexicon, that were used the most in Table 3. We also validated these usage frequencies based on the β coefficients of the unigrams in a penalized logistic regression model—the model takes as input a predictive feature vector and predicts a binary response variable (depressed/not depressed), at the same time handling highly correlated and sparse inputs. In order to make better sense of the unigrams, we derived broad "themes" that would cluster them together, using responses from crowdworkers on Amazon's Mechanical Turk (inter-rater agreement Fleiss-Kappa: 0.66). These themes are motivated from prior work in (Ramirez-Esparza et al., 2008), where the language of depression forums was analyzed.

We observe that words about *Symptoms* dominate, indicating details about sleep, eating habits, and other forms of physical ailment—all of which are known to be associated with occurrence of a depressive episode (Posternak et al., 2006). The second theme shared by the depression class is *Disclosure*. It appears that sufferers may turn to social media platforms in order to share feelings with others, receive social support, or to express their emotional state—especially feelings of helplessness and insecurity. Users al-

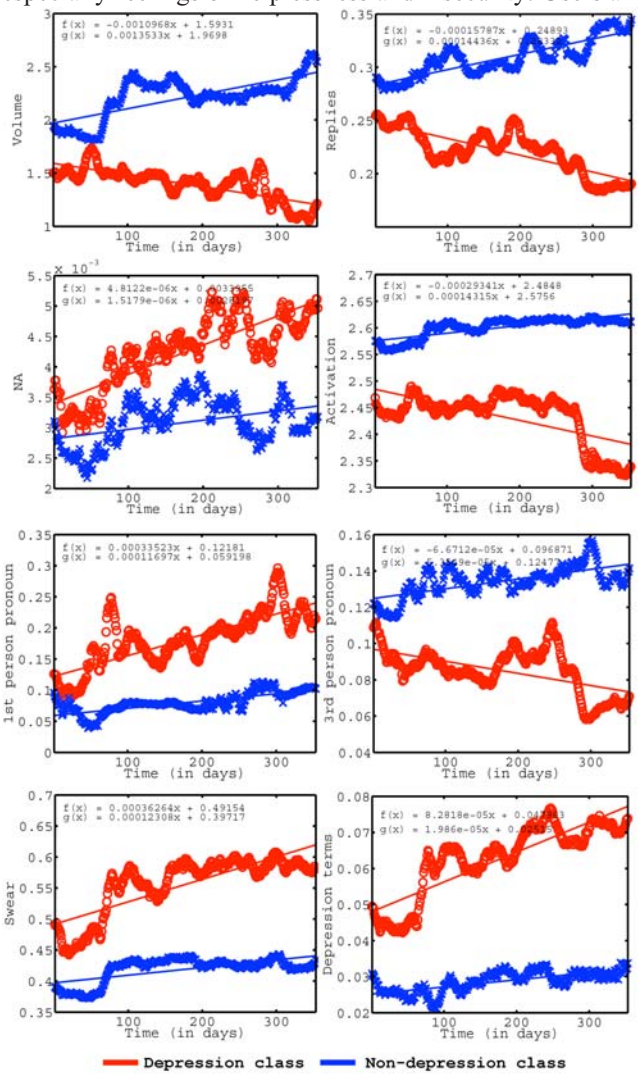

Figure 3. Trends for various features corresponding to the depression and non-depression classes. Line plots correspond to least squares fit.

so appear to discuss to some extent their therapy and treatment, even dosage levels of medication e.g., 150mg, 40mg (*Treatment* theme), as well as generally about concerns in life, work and relationships (*Relationships, Life* theme). In this last category we observe a noticeable volume of unigrams relating to religion or religious thoughts (*jesus, bible, church, lord*). On investigation of the literature, it appears that religious involvement is often found to be comforting to individuals experiencing psychological distress or sudden bereavement (McCullough et al., 1999).

Theme	Unigrams
Symptoms	anxiety, withdrawal, severe, delusions, adhd, weight, insomnia, drowsiness, suicidal, appetite, dizziness, nausea, episodes, attacks, sleep, seizures, addictive, weaned, swings, dysfunction, blurred, irritability, headache, fatigue, imbalance, nervousness, psychosis, drowsy
Disclosure	fun, play, helped, god, answer, wants, leave, beautiful, suffer, sorry, tolerance, agree, hate, helpful, haha, enjoy, social, talk, save, win, care, love, like, hold, cope, amazing, discuss
Treatment	medication, side-effects, doctor, doses, effective, prescribed, therapy, inhibitor, stimulant, antidepressant, patients, neurotransmitters, prescriptions, psychotherapy, diagnosis, clinical, pills, chemical, counteract, toxicity, hospitalization, sedative, 150mg, 40mg, drugs
Relationships, life	home, woman, she, him, girl, game, men, friends, sexual, boy, someone, movie, favorite, jesus, house, music, religion, her, songs, party, bible, relationship, hell, young, style, church, lord, father, season, heaven, dating

Table 3. Unigrams from the depression lexicon that appear with high frequency in the posts from the depression class. These terms had the largest standardized β coefficients based on penalized logistic regression.

In a similar manner, we further observe distinctively higher numbers of mentions of antidepressant medication among the depressed class, again based on a penalized logistic regression model: *serotonin* (β=.32); *amphetamine* (β=.28); *maprotiline* (β=.22); *nefazodone* (β=.13).

Egonetwork Characteristics. Next we present differences across the two classes of users based on the egocentric network measures, as summarized in Table 4. We notice lower numbers of followers and followees for the depression class—possibly showing that these users exhibit reduced desire to socialize or tendency to consume external information and remain connected with others. They also show reduced reciprocity to others' communications, indicating decreased desire for social interaction. The lower value of the graph density of their egonetworks, and the smaller sizes of their 2-hop neighborhoods shows that the interactions per individual in their networks are limited, compared to the users in the other class. The prestige ratio, however, seems to be close to unity, compared to the other class, indicating that depressed individuals and their neighbors typically have similar numbers of neighbors. Near unity prestige ratio also makes us conjecture that the neighbors of users in the depressed class could be ones they trust and connect with on psychological issues, or through their experiences. In fact, we know from (Kawachi & Berkman, 2001) that depressed individuals are known to cluster together. However given the limited availability of data in our study, we cannot confirm this finding—however constitutes an interesting topic for future research.

In conjunction with the higher value of clustering coefficient, embeddedness, and number of ego components, we conjecture that these observations indicate that depression sufferers typically belong to high connectivity close-knit networks. This may be an indication that when depressed individuals turn to social media, they intend to leverage the tool to build a closed network of trusted people, with whom they are comfortable sharing their psychological experiences, seeking out social support, or gathering information regarding their treatment and medication.

Egonetwork measures	Depres. class	Non-depres. class
#followers/inlinks	26.9 (σ=78.3)	45.32 (σ=90.74)
#followees/outlinks	19.2 (σ=52.4)	40.06 (σ=63.25)
Reciprocity	0.77 (σ=0.09)	1.364 (σ=0.186)
Prestige ratio	0.98 (σ=0.13)	0.613 (σ=0.277)
Graph density	0.01 (σ=0.03)	0.019 (σ=0.051)
Clustering coefficient	0.02 (σ=0.05)	0.011 (σ=0.072)
2-hop neighborhood	104 (σ=82.42)	198.4 (σ=110.3)
Embeddedness	0.38 (σ=0.14)	0.226 (σ=0.192)
#ego components	15.3 (σ=3.25)	7.851 (σ=6.294)

Table 4. Average measures, along with std. dev. of the egocentric social graph, comparing the depression and non-depression classes over the year-long period of analysis.

Predisposition of Depression. In terms of the trends of each of the behavioral measures in Figure 3, we notice a general decrease over time in some measures, e.g., volume, replies, activation, 3rd person pronoun (note the negative slope in the trend lines), while a general increase over time for others like NA, 1st person pronoun usage, swear word use, and frequency of depression terms (positive slope in trend lines). We conjecture that this finding indicates individuals showing a shift in their behavior as they approach the onset of their depression—note that the year-long trends shown in the figure precede the reported onset of depression for the users. The clinical literature reports that a variety of predisposing/precipitating factors or states are associated with the onset of depression in people; these include mood disturbances, suicidal thoughts, cognitive impairments, or self-care, attention, judgment and communication (Rabkin & Struening, 1976). Through the general increase of NA, lowered activation or rise in use of depressive language over the period preceding depression onset, it seems that Twitter postings do indeed capture this.

Predicting Depressive Behavior

Given the two classes of users and their differences in behavior, how accurately can we forecast, prior to the reported onset of depression, whether or not a user is likely to be

in the depressed class? In the remainder of the paper, we propose and evaluate a model for the purpose.

Constructing Feature Vectors

For each set of behavioral measures, we obtained daily measurements per user, which helped us construct one time series per measure per user, over the entire one year of Twitter history. Next, we developed a series of numbers from each of these time series for a given user, to be used eventually in constructing feature vectors for the depression prediction framework. Note that these time series features take into account the aggregated value over the year-long period (given by mean), as well as its trend of change.

- *Mean frequency*: the average measure of the time series signal of a feature over the entire period of analysis.
- *Variance*: the variation in the time series signal over the entire time period. Given a time series $X_i(1)$, $X_i(2)$, …, $X_i(t)$, …, $X_i(N)$ on the i^{th} measure, it is given as: $(1/N)\sum_t(X_i(t)-\mu_i)^2$.
- *Mean momentum*: relative trend of a time series signal, compared to a fixed period before. Given the above time series, and a period length of M (=7) days, its mean momentum is: $(1/N)\sum_t(X_i(t)-(1/(t-M))\sum_{(M\le k\le t-1)}X_i(k))$.
- *Entropy*: the measure of uncertainty in a time series signal. For the above time series it is: $-\sum_t X_i(t)\log(X_i(t))$.

Besides these features, we also used the self-reported information on age, gender, education level, and income of the users as another set of features. This yields four numbers per measure for each user in our dataset; a total of 188 features (there are 43 dynamic features in all; 4 demographic features). We represent each user as such a 188-item feature vector, with the vector being standardized to zero mean and unit variance.

Prediction Framework

We now pursue the use of supervised learning to construct classifiers trained to predict depression in our two user classes. To avoid overfitting, we employ principal component analysis (PCA), although we report results for both all dimension-inclusive and dimension-reduced cases. We compare several different parametric and non-parametric binary classifiers to empirically determine the best suitable classification technique. The best performing classifier was found to be a Support Vector Machine classifier with a radial-basis function (RBF) kernel (Duda et al., 2000). For all of our analyses, we use 10-fold cross validation on the set of 476 users, over 100 randomized experimental runs.

Prediction Results

We now focus on prediction of future episodes of depression. We first present some results of statistical significance of the behavioral features, as measured through their mean, variance, momentum, and entropy values over the one year period of analysis (Table 5). We use independent sample *t*-tests, where *df*=474: the values of the *t*-statistic and the corresponding *p*-values are given in the table. Note that we have 188 feature variables; hence to counteract the problem of multiple comparisons, we adopt Bonferroni correction. We choose a significance level of α=0.05/188= 2.66e-4. In Table 5, we report the features for which we have at least one of mean, variance, momentum or entropy values to be statistically significant.

	Mean	Variance	Momentum	Entropy
volume	15.21***	14.88***	14.65***	17.57***
replies	22.88***	13.89	29.18***	19.48***
questions	8.205	7.14	23.06***	10.71
PA	14.64	10.94	13.25	17.74***
NA	16.03***	19.01***	17.54***	15.44***
activation	19.4 ***	17.56***	22.49***	17.84***
dominance	20.2 ***	18.33***	24.49***	12.92
#followees	28.05***	14.65	25.95***	16.85***
reciprocity	5.24 ***	5.35	7.93 ***	6.82 ***
clust. coeff.	12.33***	10.92	15.28***	11.91
#ego comp.	7.29	6.91 ***	9.04 ***	8.56
antidepress	8.68	10.13	10.17***	5.73
depr. terms	22.29***	16.28***	22.16***	18.64***
1st pp.	25.07***	15.26***	24.22***	19.77***
2nd pp.	13.03***	12.43	20.36***	11.49
3rd pp.	20.34***	14.60	21.47***	16.96***
article	9.75	14.41	16.68***	7.60
negate	8.42	6.33	16.7 ***	12.13
swear	12.91	6.12	20.8 ***	18.99***

*** $p \le \alpha$, after Bonferroni correction *df*=474

Table 5: Statistical significance (*t*-statistic values) of the mean, variance, momentum and entropy measures of selected dynamic features, comparing the depression and non-depression classes.

The results align with our findings described earlier. Across the feature types, certain stylistic, engagement, emotion measures, and use of depression terms and mentions of antidepressant medication bear distinctive markers across the two classes. In general, momentum seems to be a feature that shows statistical significance across a number of measures, demonstrating that not only is the absolute degree of behavioral change important (indicated by the mean), but the trend of its change over time bears useful markers of distinguishing depressive behavior.

Now we utilize our proposed classification framework to examine how well we can predict, whether or not an individual is vulnerable to depression, ahead of its onset. In order to understand the importance of various feature types, we trained a number of models.

We present the results of these prediction models in Table 6. The results indicate that the best performing model (dimension-reduced features) in our test set yields an average accuracy of ~70% and high precision of 0.74, corresponding to the depression class. Note that a baseline marginal model would yield accuracy of only 64%, i.e., when all data points are labeled per the majority class which is the non-depressed class. Good performance of this classifier is also evident from the receiver-operator characteristic (ROC) curves in Figure 4. The dimension-reduced feature

model gives slightly greater traction compared to the one that uses all features; demonstrating utility of reducing feature redundancy.

	precision	recall	acc. (+ve)	acc. (mean)
engagement	0.542	0.439	53.212%	55.328%
ego-network	0.627	0.495	58.375%	61.246%
emotion	0.642	0.523	61.249%	64.325%
linguist. style	0.683	0.576	65.124%	68.415%
dep. language	0.655	0.592	66.256%	69.244%
demographics	0.452	0.406	47.914%	51.323%
all features	0.705	0.614	68.247%	71.209%
dim. reduced	**0.742**	**0.629**	**70.351%**	**72.384%**

Table 6. Performance metrics in depression prediction in posts using various models. Third column shows the mean accuracy of predicting the positive class.

We also observe better performance of the model that uses the linguistic style features alone. Results in prior literature suggest that use of linguistic styles such as pronouns and articles provide information about how individuals respond to psychological triggers (Rude et al., 2004; Ramirez-Esparza et al., 2008). Next, we note that, one of the main characteristics of depression is disturbed cognitive processing of information as indexed by disturbed startle reflex modulation, as well as a reduced sense of interest or motivation in day-to-day activities (Billings et al., 1984; Oxman et al., 1982). Hence we observe better performance of depression language features in the prediction task. Finally the better performance of ego-network features shows that the network in which depressed individuals are embedded, serving as a proxy to their social and behavioral environment, bears key information in light of their condition. In essence, we conclude that social media activity provides useful signals that can be utilized to classify and predict whether an individual is likely to suffer from depression in the future.

Discussion

Implications. The ability to illustrate and model individual behavior using their social media data, that can predict depression *before* their estimated onset, shows promise in the design and deployment of next-generation wellness facilitating technologies. We envision privacy-preserving software applications and services that can serve as early warning systems providing personalized alerts and information to individuals. These tools perhaps can enable adjuvant diagnosis of depression and other mental illness, complementary to survey approaches (e.g., CES-D, BDI).

Beyond monitoring behavioral trends in real-time, social media-based measures, such as volume, NA, activation, use of depression language etc. can serve as a diary-type narrative resource logging "behavioral fingerprints" over extended periods of time. The application might even assign an "MDD risk score" to individuals based on predictions made about forthcoming extreme changes in their behavior and mood. In operation, if inferred likelihoods of forthcoming extreme changes surpass a threshold, they could be warned or engaged, and information might be provided about professional assistance and/or the value of social and emotional support from friends and family.

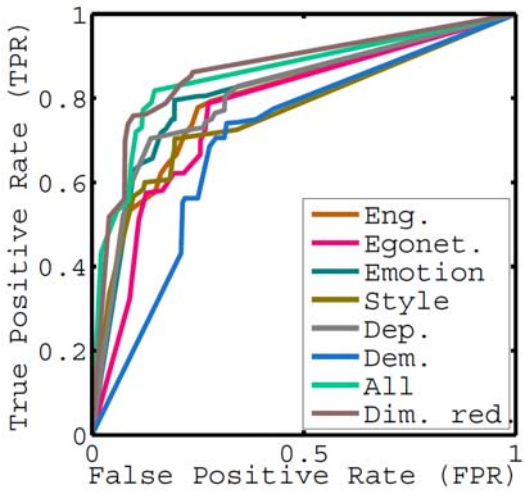

Figure 4. Receiver Operating Characteristic (ROC) curves in predicting labels of users. Each curve corresponds to a model trained on a particular feature type.

Privacy Considerations. Concerns regarding individual privacy, including certain ethical considerations, may arise with this form of analyses of social media as they ultimately leverage information that may be considered sensitive, given their focus on behavioral and emotional health. As we mentioned earlier, we collected data from crowdworkers on AMT whose Twitter profiles were public, and participants could opt out of sharing their data. For users who opted in, the AMT study obtained their consent that it would be okay to use their data anonymously in an automated setting, without active human intervention, in doing research analyses.

Conclusion and Future Work

We have demonstrated the potential of using Twitter as a tool for measuring and predicting major depression in individuals. First we used crowdsourcing to collect gold standard labels on a cohort's depression, and proposed a variety of social media measures such as language, emotion, style, egonetwork, and user engagement to characterize depressive behavior. Our findings showed that individuals with depression show lowered social activity, greater negative emotion, high self-attentional focus, increased relational and medicinal concerns, and heightened expression of religious thoughts. They also appeared to belong to highly clustered close-knit networks, and were typically highly embedded with their audiences, in terms of the structure of their egonetworks. Finally, we leveraged these distinguishing attributes to build an SVM classifier that can predict, ahead of the reported onset of depression of an individual, his/her likelihood of depression. The classifier yielded promising results with 70% classification accuracy.

Among future directions, we hope to understand how analysis of social media behavior can lead to development of scalable methods for automated public health tracking at-scale. We are also interested in harnessing the potential of social media in tracking the diffusion of affective disorders in populations in a nuanced manner; for identifying the incidence and impact of trauma on individuals during crisis events, and for modeling of help-seeking behavior, health risk behaviors, and risk of suicide.

References

Abdel-Khalek, A. M. 2004. Can somatic symptoms predict depression? *Social Behavior and Personality: an international journal*, 32(7), 657-666.

Andrade L, Caraveo-A. 2003. Epidemiology of major depressive episodes: Results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys . *Int J Methods Psychiatr Res.*12(1):3–21.

Beck, A. T.; Steer, R. A.; & Brown, G. K. 1996. Manual for the Beck depression inventory-II. San Antonio, TX: *Psychological Corporation*, 1, 82.

Billings, A.; Moos, Rudolf H. 1984. Coping, stress, and social resources among adults with unipolar depression. *Journal of Personality & Social Psych.*, 46(4), 877-891.

Bradley, M.M.; & Lang, P.J. 1999. Affective norms for English words (ANEW). Gainesville, FL. *The NIMH Center for the Study of Emotion and Attention.*

Brown, G. W.; Andrews, B.; Harris, T.; Adler, Z.; & Bridge, L. 1986. Social support, self-esteem and depression. *Psychological medicine*, 16(4), 813-831.

Cloninger, C. R.; Svrakic, D. M.; & Przybeck, T. R. 2006. Can personality assessment predict future depression? A twelve-month follow-up of 631 subjects. *Journal of affective disorders*, 92, 35-44.

De Choudhury, M.; Mason, W. A.; Hofman, J. M.; & Watts, D. J. 2010. Inferring relevant social networks from interpersonal communication. In *Proc. WWW 2010*.

De Choudhury, M.; Counts, S.; and Gamon, M. 2012. Not All Moods are Created Equal! Exploring Human Emotional States in Social Media. In *Proc. ICWSM '12*.

De Choudhury, M.; Counts, S.; & Horvitz, E. 2013. Predicting Postpartum Changes in Behavior and Mood via Social Media. In *Proc. CHI 2013*, to appear.

Detels, R. 2009.The scope and concerns of public health. *Oxford University Press*.

Duda, Richard O.; Hart, Peter E.; & Stork, David G. 2000. *Pattern Classification*. 2nd Edition, Wiley.

Kawachi, I.; and Berkman, L. S. 2001. Social ties and mental health. *Journal of Urban Health*, 78(3), 458-467.

Kessler, R.C.; Berglund, P.; Demler, O. et al. 2003. The Epidemiology of Major Depressive Disorder: Results from the National Comorbidity Survey Replication. *Journal of the American Medical Association 289 (23)*: 3095-3105.

Kotikalapudi, R.; Chellappan, S.; Montgomery, F.; Wunsch, D.; & Lutzen, K. 2012. Associating depressive symptoms in college students with internet usage using real Internet data. *IEEE Technology and Society Magazine*.

Lustberg L; & Reynolds CF 2000. Depression and insomnia: questions of cause and effect. *Sleep Medicine Reviews* 4 (3): 253–262.

McCullough, Michael E.; & Larson, David B. 1999. Religion and Depression: A Review of the Literature. *Twin Research*, 2(2), 126-136.

Moreno, M.; Jelenchick, L.; Egan, K.; Cox, E. et al. 2011. Feeling bad on Facebook: depression disclosures by college students on a social networking site. *Depression and Anxiety* 28(6):447–455.

Oxman T.E.; Rosenberg S.D.; & Tucker G.J. 1982. The language of paranoia. *American J. Psychiatry* 139:275–82.

Park, M.; Cha, C.; & Cha, M. 2012. Depressive Moods of Users Captured in Twitter. In *Proc. ACM SIGKDD Workshop on Healthcare Informatics* (HI-KDD).

Paul, M., J.; & Dredze, M. 2011. You are What You Tweet: Analyzing Twitter for Public Health. In *Proc. ICWSM '11*.

Posternak MA; Solomon DA; Leon AC. 2006. The naturalistic course of unipolar major depression in the absence of somatic therapy. *J. Nerv & Mental Disease* 194(5):324–29.

Rabkin, J. G.; & Struening, E. L. 1976. Life events, stress, and illness. *Science*, 194(4268), 1013-1020.

Radloff, L.S. 1977. The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement* 1: 385-401.

Ramirez-Esparza, N.; Chung, C. K.; Kacewicz, E.; & Pennebaker, J. W. 2008. The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. In *Proc. ICWSM 2008*.

Robinson, M. S.; & Alloy, L. B. 2003. Negative cognitive styles and stress-reactive rumination interact to predict depression: A prospective study. *Cognitive Therapy and Research*, 27(3), 275-291.

Rude, S. S.; Valdez, C. R.; Odom, S.; & Ebrahimi, A. 2003. Negative cognitive biases predict subsequent depression. *Cognitive Therapy and Research*, 27(4), 415-429.

Rude, S.; Gortner, E.; & Pennebaker, J. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 1121-1133.

Sadilek, A.; Kautz, H.; & Silenzio, V. 2012. Modeling Spread of Disease from Social Interactions. In *Proc. ICSWM '11*.

Snow, R.; O'Connor, B.; Jurafsky, D.; & Ng, A. Y. 2008. Cheap and fast—But is it good?: Evaluating non-expert annotations for natural language tasks. In *Proc. EMNLP'08*.

World Health Organization. 2001. The world health report 200—Mental Health: New Understanding, New Hope.