Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media

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

Emre Kiciman
Microsoft Research
Redmond WA 98052
emrek@microsoft.com

Mark Dredze
Johns Hopkins University
Baltimore MD 21218
mdredze@cs.jhu.edu

Glen Coppersmith
Qntfy.io
Crownsville MD, 21032
glen@qntfy.io

Mrinal Kumar
Georgia Tech
Atlanta GA 30332
mkumar73@gatech.edu

ABSTRACT
History of mental illness is a major factor behind suicide risk and ideation. However research efforts toward characterizing and forecasting this risk is limited due to the paucity of information regarding suicide ideation, exacerbated by the stigma of mental illness. This paper fills gaps in the literature by developing a statistical methodology to infer which individuals could undergo transitions from mental health discourse to suicidal ideation. We utilize semi-anonymous support communities on Reddit as unobtrusive data sources to infer the likelihood of these shifts. We develop language and interactional measures for this purpose, as well as a propensity score matching based statistical approach. Our approach allows us to derive distinct markers of shifts to suicidal ideation. These markers can be modeled in a prediction framework to identify individuals likely to engage in suicidal ideation in the future. We discuss societal and ethical implications of this research.

ACM Classification Keywords
H.4 Information Systems Applications: Miscellaneous

Author Keywords
social media; suicidal ideation; mental health; Reddit

INTRODUCTION
A central challenge in public health revolves around how to identify individuals who are at risk for taking their own lives [3, 8, 66]. One of the ten leading causes of death in the United States, suicide represents 1.4% of the total number of adult deaths[1]. Yet suicide prevention remains difficult. Suicidal acts are multifactorial events [65], and different categories of suicidal behavior have different pathogenesis, expression, and often an underlying mental illness [66]. Extending appropriate clinical and psychiatric care to suicidal patients relies heavily on identifying those at risk [29].

Suicidal ideation is defined as tendencies and cognitions related to ending one's life, ranging from the thought that life is not worth living, through concrete plans for killing oneself, to an intense delusional preoccupation with self-destruction [5]. Therefore, immense scientific and practical value lies in being able to understand the intensity, pervasiveness, and characteristics of the ideation, since this may predict later suicide risk or attempt [6].

Mental illness is a major risk factor of suicide — 80% of those who attempt or die by suicide are known to have had some form of mental illness [67]. However, the majority of those challenged by mental illness do not engage in suicidal ideation [3]. Hence, prior literature in cognitive and clinical psychology [26] has underscored the understanding specific "suicidogenic" elements in manifestations of mental illness.

Existing efforts toward discovering and recognizing suicidogenic elements have primarily been through the examination of psychological, psychiatric, and demographic variables of individuals [41, 4]. However these assessments face two significant methodological challenges: (1) In many studies, data is collected after the suicide attempt or completed suicide, providing "postdictors" rather than predictors of suicidal behavior and are therefore prone to include hindsight bias; and (2) the relatively rare occurrence of completed suicides and the stigma associated with suicide reporting in the general population has made studies challenging and expensive to conduct, additionally requiring extremely long follow-up intervals. Consequently, there is limited research on examining factors associated with the development of *future suicidal thoughts* among mental illness prone populations [5].

Contributions. This paper proposes social media as a way to characterize and predict shifts from discussion of mental health content to expression of suicidal ideation. We focus on a popular discussion-oriented social media site, Reddit, specifically several mental health and suicide support communities. Due to the semi-anonymous nature of these communities [51], the content shared by individuals allows us to

[1] http://www.cdc.gov/ViolencePrevention/pdf/Suicide_DataSheet-a.pdf

CHI'16, May 07-12, 2016, San Jose, CA, USA
 ISBN 978-1-4503-3592-7/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858207

obtain high quality, self-reported data around mental health concerns and suicidal ideation [31]. The central research question investigated in this paper involves: *Can we forecast whether an individual engaged in mental health discussions would, in the future, discuss suicidal ideation?* Towards this goal, we make the following two contributions:

(1) We characterize participants in Reddit's mental health communities who go on to post on the platform's suicide support forum using a number of linguistic and social interaction based measures that have been known to characterize an individual's behavioral and psychological state [65].

(2) We propose the novel application of propensity score matching to explore how users may share suicidal ideation content in the future, while controlling for the historical use of linguistic constructs of mental health. The challenges of interpreting correlational statistics from observational studies like ours is well-recognized [1, 72]. Through statistical analysis methods developed for causal inference, we isolate the effects of linguistic constructs from observed confounding factors, and are able to derive valuable insights into factors related to future suicide ideation.

Findings. From a population of individuals who post about mental health concerns on Reddit, we examine differences between those who proceed to discuss suicidal ideation in the future, from those who do not. We identify changes in linguistic structures, interpersonal awareness, social interaction and content between these two groups, some of which align with findings in the suicide literature [5]. Specifically, we observe transition to suicidal ideation to be associated with psychological states like heightened self-attentional focus, poor linguistic coherence and linguistic coordination with the community, reduced social engagement, and manifestation of hopelessness, anxiety, impulsiveness and loneliness. Finally, we examine whether we can automatically predict the tendency of individuals discussing mental health concerns to engage in these characteristic behaviors. For this purpose, we develop a logistic regression classifier that yields high accuracy. We situate our findings in the cognitive psychological integrative model of suicide [25] in order to derive qualitative interpretations, and discuss the implications of our work for HCI research, design and ethics as well as for developing timely interventions.

Privacy, Ethics and Disclosure. We use public data from Reddit. Personally identifiable information was removed and content was de-identified and paraphrased before being reported in the paper for exemplary purposes. This work has been approved by the appropriate Institutional Review Board (IRB). *Our work does not make any diagnostic claims related to mental illness or suicide.*

BACKGROUND AND PRIOR WORK

Mental Illness and Suicide

A number of mental health disorders, such as depression, tend to be closely related to suicide [65]. Reportedly, one in six patients who fall under the category of major depression as set forth in the Diagnostic and Statistical Manual of Mental Disorders (DSM) [2] dies as a result of suicide [11].

Literature in psychology has suggested the need to identify specific attributes of mental illnesses that relate to increased likelihood of suicidal thoughts [65, 52]. Nock and Kazdin [48] found that cognitive factors associated with depression are of greater importance than the affective dimension of depression in predicting suicide-related outcomes. Another strong correlation exists between affective disorders, attempted suicide, and borderline personality disorder [14]. Kashden et al. [37], who compared non-suicidal and suicidal psychiatric inpatients to community high school students, found suicidal inpatients to be characterized by impulsivity, hopelessness, and depression. Further, in a study by Lewinsohn et al. [42], the diagnoses with the strongest association with suicide attempts among young adults were combinations of depressive disorder with substance use, disruptive behavior, or anxiety (also see [63]).

Broadly, researchers have identified three stages leading to suicidal ideation among individuals with some form of mental illness [4, 66]: a) thinking, b) ambivalence, and c) decision making. Together, these stages define the *cognitive psychological integrative model of suicide* [55, 29], wherein the *thinking* stage may include thoughts of hopelessness, self-hatred, distress and anxiety; *ambivalence* relates to lowered self-esteem, regulation and reduced social cohesion; and *decision making* involves aggression and explicit plans of taking one's life [59]. Individuals may seek help, advice and support on mental health related social media forums during any of these stages, and thus these forums provide a non-reactive and non-intrusive way to measure risk factors of suicidal ideation among individuals vulnerable to different mental illnesses.

Mental Health and Suicide Studies on Social Media

In recent years, social media has been recognized to be a powerful "lens" that can provide insights into psychological states, health and well-being of individuals and populations [50, 19]. Linguistic attributes of shared content and social interactional patterns have been utilized to understand and infer risk to major depressive disorder [24, 49, 32, 16, 60, 70], postpartum depression [21, 22], addiction [47, 44], and other mental health concerns [35, 18, 17, 46]. Since social media is recorded in the present and preserved, it minimizes the hindsight bias sometimes induced by retrospective analyses. The rich repository of social media data also allows for the discovery, tracking, and perhaps forecasting of risk attributes longitudinally. Beyond observation and insight, social media may also provide mechanisms through which timely support may be extended to vulnerable communities.

There exists some research examining suicide in social media [58, 43, 15, 9, 28, 38]. Authors in [71] focused on South Korean blogs to predict nationwide suicide rate data (also see [36] for a similar study in the US context). Studying linguistic features of suicidal ideation, authors in [68] surveyed a sample of Twitter users to examine the association between suicide-related tweets and suicidal behavior. However bulk of this prior work focused at macro-level trends (e.g., national suicide rates) or examined differences between suicide-related content and general content shared on social media. To the best of our knowledge, there has not been prior work forecasting likelihood of suicidal ideation in an individual based on mental health discussions on social media. It is

MHs (Mental Health subreddits)
I have been considering going for some formal therapy. Any suggestions?
Everyday I feel sad and lonely
Since past sometime I think I am having panic attacks. I really need help from you guys.
It has been so many years, I feel I still can't move on. I am noticing behavior what could be considered "triggers" now.
SW (SuicideWatch)
I know I was never meant to lead this life.
Don't want to hurt the people I care but I can't take this anymore.
Today I felt I have nothing left, why am I even living... I don't see a point.
I'd kill myself, but the other part of me tells me not to waste all the money my parents invested on me..

Table 1: Example titles of posts in the MHs and SW datasets; content has been carefully paraphrased to protect the privacy of the individuals.

important to identify and differentiate how and which social media markers of mental health concerns may relate to future suicidal ideation, given the important temporal link between history of mental illness and future suicide risk. Further, mental illness is associated with vulnerability, hence identifying markers that may indicate increased suicidal thoughts in the future may help the deployment of appropriate interventions.

Our paper builds on this emergent body of research by analyzing data shared on mental health communities in Reddit, to probe attributes of individuals contemplating suicide in the future. We additionally note that a major hurdle in studying suicide-related issues in the computing field has been the lack of appropriate ground truth data on individuals actually suffering from suicidal thoughts. The social stigma associated with such sensitive disclosure may further prevent individuals from self-reporting their condition on social media. In our work, we partially address this challenge by studying semi-anonymous communities on Reddit where vulnerable individuals voluntarily participate for seeking help and support. Finally, while existing work has often relied on lexicon matching or phrase identification techniques to identify correlational attributes of interest such as suicide intent or mental illness risk, in this paper we infer more robust insights by extending current methodology with causal inference techniques based on propensity score matching.

DATA

We begin with a brief description of the features of Reddit, which are important to understand the context of our research problem. Reddit has many of the characteristics of an online forum; users or "redditors" can submit content in the form of link posts or text posts. Posts are organized by areas of interest or sub-communities called "subreddits". Besides posting, redditors can also engage via "upvoting" or "downvoting" a post, or responding on a post through comments.

Data Collection. We obtained post and comment data from a number of mental health subreddits (henceforth MHs) and a suicide support subreddit "r/SuicideWatch" (henceforth SW). We focused on a set of 14 MHs that have been examined in prior work on mental health discourse [51, 39]. These subreddits included r/depression, r/mentalhealth, r/traumatoolbox, r/bipolarreddit, r/BPD, r/ptsd, r/psychoticreddit, r/EatingDisorders, r/StopSelfHarm, r/survivorsofabuse, r/rapecounseling, r/hardshipmates, r/panicparty, r/socialanxiety. While SW solely focuses on helping those contemplating suicide, the other MHs cover a variety of mental health concerns but not specifically suicidal ideation [31]. All of these subreddits host public content.

We used Reddit's official API to collect posts, comments, and associated metadata from the SW and MHs subreddits (http://www.reddit.com/dev/api). Our analysis in this paper is based on all content shared on MHs between February 11 and November 11 2014 (63,485 posts, 209,766 comments and 35,038 users). We refer to the data obtained from SW during the same time period (16,348 posts, 9,224 users) to identify those individuals in MHs who go on to post on SW over time.

MHs and SW Content Verification. Following our data collection, we focused on verifying whether MHs and SW subreddit content actually relate to discussion of mental health concerns and suicidal ideation. The MHs have been previously examined for understanding mental health discourse on Reddit [51, 39]. For SW, we consulted (1) a licensed clinical psychologist/suicide prevention expert and (2) two active moderators of SW to obtain qualitative grounding that the content in SW indeed related to expressions of suicidal ideation. Example (paraphrased) titles of posts from one of the MHs and SW are given in Table 1.

Constructing User Classes. We split our data into two sequential time periods (t_1 from Feb 11 2014 to Aug 11 2014, and t_2 from Aug 12 2014 to November 11 2014). Using these two time periods, we created two sets of users. Note that since Reddit does not enforce the real name rule of having exactly one account per person, our reference to "users" in this paper is equivalent to "user accounts". First, we identified those users that posted on MHs during t_1, but did not post on SW during t_1 or t_2 (i.e., users that discuss mental health topics but not on SW; hereafter "MH"). The second class included those who posted on MHs during t_1 and posted in SW during t_2 (i.e., users that discuss mental health topics, originally not related to suicide, but eventually transition to talk about suicide; hereafter "MH $\rightarrow$ SW"). Figure 1 shows a schematic description of our user class construction. Note that by focusing on users that initiate at least one post on SW or the MHs, as opposed to only commenting, we can focus on those frequenting the communities for support, disregarding those primarily providing help through commentary. This split yielded 440 MH $\rightarrow$ SW users; which is 1.52% of the total number of 28,831 accounts who posted in MHs but never on SW during either of the periods. To construct a MH cohort of equal size who did not post on SW in either period, we randomly sampled a set of 440 users from the 28,831 users. Note, although MH users did not post on SW during our timeframe of analysis, they may have done so outside the bounds of our analysis.

To support our goal of characterizing differences between the MH $\rightarrow$ SW and MH users, we obtained via Reddit's API the timeline of posts and comments authored by the 880 users (the API only provides the last 1000 public posts and comments for a user). For each post, we obtained their associated metadata (e.g., vote difference or score) and comments. Our final dataset contained 4,731 posts and 46,949 comments from the 440 MH $\rightarrow$ SW users, and 8,318 posts and 54,086 comments from the 440 MH users.

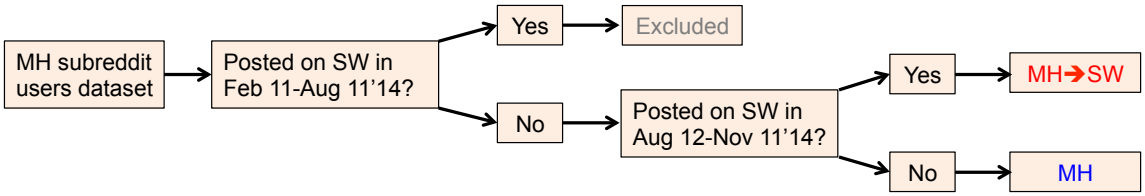

Figure 1: Schematic diagram of obtaining MH → SW and MH classes of users.

We note an important concern: individuals may post suicidal thoughts on MHs, never engaging on SW, and thus "corrupting" the MHs data with discussions of suicidal ideation. We argue against this possibility. (1) SW is a prominent suicide support forum, and the role of this community in suicide prevention and in acting as an inoculator of vulnerable thoughts is well-recognized [31]. (2) Most MHs (e.g., r/depression) clearly specify in their guidelines that suicidal thoughts should go to SW: "*It's usually better to post anything that specifically involves suicidal thoughts or intent in /r/SuicideWatch rather than here. If you're concerned about someone else who may be at risk for suicide, please check out their talking tips and risk assessment guide*." (3) Finally, discussions with the moderators of SW confirmed that active steps are taken to move all suicidal ideation related content to SW. Given these considerations, we expect that few suicidal ideation posts appear on subreddits outside of SW.

METHODS

We present our various measures and methods by which we identify differences between MH → SW and MH users. This will include both characterizations of differences as well as automated methods for differentiating between the groups.

Linguistic, Interpersonal, & Interaction Measures

Our first set of methods include developing three sets of measures spanning: *linguistic structure*, *interpersonal awareness* and *interaction*. The choice of these measures is motivated by literature that examines associations between the behavioral expression of individuals and their responses to crises, including vulnerability due to mental illness [13, 23]. Each of these measure categories consists of the following variables:

Linguistic Structure. For this measure, we compute the fraction of nouns, verbs[2], and adverbs in posts and comments; automated readability index, a measure to gauge the understandability of text [62]; and linguistic accommodation, a process by which individuals in a conversation adjust their language style according to that of others [20][3]. Together, these variables characterize the text shared by the user classes beyond their informational content. Per literature in psycholinguistics, such structure is known to relate to an individual's underlying psychological and cognitive state and can reveal cues about their social coordination [54].

Interpersonal Awareness. This measure category includes: proportion of first person singular (indicating pre-occupation with self), first person plural (indicating collective attention), second person and third person pronouns (indicating social interactivity and reference to people or objects in the environment). Literature has indicated that pronoun use can quantify an individual's self and social awareness and can reveal mental well-being, including that manifested in social media [21].

Interaction. Variables corresponding to this measure category include: volume of posts and comments authored, post length, length of comments authored, volume of comments received on shared posts, length of comments received, mean vote difference (difference between upvotes and downvotes on posts authored), and response velocity (in minutes), given by the time elapsed between the first comment and the time the corresponding post was shared.

Prediction Framework

We frame our prediction problem of identifying which MH user will go on to post on SW in the future as a supervised learning task. We first build a probability distribution over all unigrams and bigrams (referred to as tokens) in the posts and comments of both user classes. Thereafter, we construct several regularized logistic regression based binary classifiers, where the response variable is whether a user belongs to MH → SW or to MH.

We consider different sets of predictor variables for the classifiers based on our three measure categories and the tokens obtained from above. We consider five models: (1) `Linguistic Structure`; (2) `Interpersonal Awareness`; (3) `Interaction` where the predictor variables of each correspond to their respective variables; (4) `Content` consisting of the unigram and bigram tokens along with their relative frequencies given from above; and (5) `Full` comprising all variables from all measure categories and the tokens from the `Content` model.

Propensity Score Matching

We borrow methods from the causal analysis literature [1, 72], although we recognize that our data does not meet the strong assumptions that are required to infer true causality. First, all confounding variables are not included among the observed covariates. Secondly, the stable unit treatment value (SUTVA) assumption (where one individual's outcome must be independent of whether another individual takes a treatment) likely does not hold in these communities we study. Though this prevents us from making any causal claims, in practice we find that the results of our stratified analysis provides significant insight beyond simpler correlational techniques in this observational study.

To go beyond prediction and to better understand the possible causal factors involved in users' transitions from posting

[2]Nouns and verbs were detected using a standard POS tagger.
[3]We utilize a simplified version of Danescu-Niculescu-Mizil et al.'s [20] approach, where accommodation is measured in terms of change in function word use obtained from the psycholinguistic lexicon LIWC (http://www.liwc.net/).

in MH to posting in SW, we wish to isolate the "effects" of individual tokens (unigrams and bigrams) in posts and reduce bias due to confounding variables. Essentially, we wish to estimate the effect of a specific *treatment* (the use of a *target token* in an MH post) on a measured *outcome* (the likelihood of transitioning to post in SW) conditioned on confounding variables (all previously written tokens in MH posts). To do so, we work within the potential outcomes framework [57, 34] for causal modeling, applying stratified propensity score matching to estimate causal effects [56]. Stratified propensity score matching achieves this by subdividing the *treatment group* (individuals who used the target token) and the *control group* (individuals who did not use the target token) into comparable groups based on the individuals' *estimated* propensity to use the token. We learn a propensity estimating function based on covariates (past tokens from the histories of both the control and treatment group members). This balances the distributions of confounding factors within each strata, creating comparable treatment and control groups within each strata. The observed difference in the likelihood of SW transitions between the two groups then provides an estimate of the effect of the treatment, as the groups are otherwise comparable.

In our implementation, for a given target token, we estimate the propensity score using the averaged perceptron learning algorithm [27] and stratify the users into 10 strata. Estimation is conducted based on a binary vector representation of user posting history, $H = h_1, ..., h_n$, where h_i is 1 if the user posted the token i prior to posting the target token, and 0 otherwise. Per [10], we use trimming to limit our comparisons to strata with sufficient common support, and report the population average treatment effect over them, as well as the z-score and χ^2 tests of statistical significance. We perform this analysis for all target unigrams and bigrams (tokens) used by more than 10 individuals in MH (11278 tokens).

RESULTS

We present our results in three phases. We begin by describing differences characterizing the MH → SW and MH users through the linguistic structure, interpersonal awareness and interaction based measure categories we defined above. Next we discuss differences in content of posts and comments shared by the two user classes. Finally, we present the results of supervised classification of the users.

Linguistic, Interpersonal, & Interaction Differences

Table 2 shows differences between the MH → SW and MH users along the linguistic structure, interpersonal awareness and interaction based measure categories. We show each variable's mean value per measure type for both groups and the z-score of the difference, based on Wilcoxon signed rank tests.

Observation 1. *MH → SW users show poorer linguistic structure and accommodation including lowered readability.*

Per Table 2, MH → SW use more verbs ($z = 2.1$) and adverbs ($z = 4.8$) (which indicate discourse around actions), but less entities, e.g., nouns ($z = 6.5$). Together this reveals poor linguistic structure [69] and indicates lowered interest in objects and things [12]. Expressing more about actions is also known to be correlated with sensitive disclosure [33]. Further, we observe lower readability index of the posts shared by MH → SW users ($z = 5.5$); such language framing limitations are linked to decreased cognitive functioning and coherence [53]. Finally, we observe that MH → SW users exhibit lowered sense of linguistic accommodation to the general content on MHs ($z = 5.4$), compared to the MH users. This may indicate decreased involvement of the MH → SW users with the communities, as well as decreased ability or intent to adjust to their norms and conventions.

	MH	MH → SW	z	p
Linguistic Structure				
nouns	0.294	0.125	6.51	***
verbs	0.045	0.107	2.19	**
abverbs	0.048	0.099	4.87	***
readability index	0.609	0.232	5.51	***
accommodation	0.857	0.487	5.46	**
Interpersonal Awareness				
1st person singular	0.018	0.086	-10.6	***
1st person plural	0.093	0.078	4.53	*
2nd person	0.058	0.031	8.01	*
3rd person	0.087	0.042	6.32	***
Interaction				
posts authored	18.97	10.31	2.53	*
post length	215.62	443.73	-15.4	***
comments authored	122.42	106.22	0.95	-
comments received	19.862	13.414	1.05	*
comment length authored	63.417	87.116	-1.88	*
comment length received	42.323	26.362	5.44	**
response velocity (mins)	7.746	6.966	0.84	-
vote difference	28.788	7.681	7.18	***

Table 2: Differences between MH → SW and MH user classes based on linguistic structural, interpersonal awareness and interaction measures. Statistical significance is reported based on Wilcoxon signed rank tests at levels $p = .05/N; .01/N; .001/N$, ($N = 17$), following Bonferroni correction.

Observation 2. *MH → SW users show higher self-attentional focus and greater detachment from the social realm.*

MH → SW users also use greater number of first person singular pronouns ($z = -10.6$). This generally indicates that, MH → SW users convey more personal stories and may be high in self-preoccupation [7]. Lower use of second person pronouns ($z = 8$), first person plural pronouns ($z = 4.5$) and third person pronouns ($z = 6.3$) in posts from MH → SW users might imply less interactive users who are less socially bothered regarding the larger Reddit audience.

Observation 3. *MH → SW users show lowered social engagement and access to support and increased self-disclosure.*

Finally, MH → SW users tend to have longer ($z = -15.4$) but fewer ($z = 2.5$) posts. More verbosity in shared content has previously been shown to be a sign of increased self-disclosure and cognitive complexity; however less activity in community settings has also previously been shown to be indicative of social isolation [30]. MH → SW users receive fewer comments on their posts ($z = 5.4$) and had smaller differences in their voting scores ($z = 7.1$), which may be an indicator of lower engagement and lower access to social support from the community as compared to posts from the MH users, as also observed in [51].

Content Differences

Our next analysis focuses on the content (posts and comments) shared by the MH → SW and MH users. First we

establish differences between the two cohorts based on our propensity score matching methodology, and then present a qualitative interpretation of our quantitative observations.

Propensity Score Analysis

In Table 3 we report 70 tokens with the highest z scores that distinguished between MH → SW and MH users based on propensity score matching; specifically *increased* the likelihood of a MH account's posting in SW in the future based on a particular token used in the past. The tokens reported were all found to be significant at the $p = .001$ level. Corresponding to each token, we also report the absolute number of users in our dataset (out of a total of 880) who used the token (treatment count), the proportion of users in our data who fell into an unclipped strata of the token (population coverage), the percent increase in likelihood of posting in SW in the future based on use of the token in the past (average treatment effect), the z score of the token's likelihood of use between the two user classes, and associated χ^2 statistic of this difference.

We find that controlling for historical use of different tokens in MH content, use of tokens such as "depression" ($z = 8.04$), "useless" ($z = 7.05$), "suicide" ($z = 6.66$), "anxiety" ($z = 6.56$), "no_friends" ($z = 6$), "have_nothing" ($z = 5.98$), "kills" ($z = 5.9$) and "to_cry" ($z = 5.5$) significantly increases a user's likelihood to post in SW in the future. For "depression" this increase is 30%, for "suicide" it is 32%, for "no_friends" it is 51%, for "to_cry" it is 51%, while for "kills" it is 53%. In addition to the tokens in Table 3 we look at effects of pronoun usage ("I", "you", "he", "she", "we", "they") and corresponding possessive pronouns, and find that the use of "I" and "my" have a statistically significant large effect (I: effect=+37%, $z = 2.8$; my: effect=+28%, $z = 4.11$); the use of 3rd person female pronouns have some statistically significant effect (her: effect=+10%; $z = 3.01$; she: effect=+8%, $z = 2.09$); and all other pronouns have effect < 7% with lower or no statistical significance). Broadly use of these tokens indicate a negative attitude, experience of emotional distress and self-focus [7]; an observation that aligns with our observations made above related to the measure of interpersonal awareness.

Which are the tokens that *decrease* the likelihood of posting in SW in the future? We show tokens with the most negative treatment effects and high z-score values in Table 4. Use of tokens like "counseling" ($z = -4.09$), "relationship_that" ($z = -3.89$), "intimate" ($z = -3.73$), "hope_it" ($z = -4.28$), "i_agree" ($z = -4.54$) and "and_enjoy" ($z = -4.44$) result in reducing the likelihood of posting in SW in the future by 50-57%. This shows a tendency of the users of these tokens to maintain a positive outlook towards life, remain hopeful (perhaps of recovery), agreeableness and focus on valuing social ties, including discussion of treatment or therapy.

Note that the effects of using a token may not be homogeneous. Certain people may see no effect of using a token, while others see a large effect. Figure 2 explores this for the tokens *depression, suicide, anxiety, suicidal, and can_t*: the most significant target words with at least 100 people using the token. Within each plot, we show how the future likelihood to post in SW varies across strata for people using and not using the target token. In the case of "depression", for example, we see that for people with a very low estimated propensity to use the word "depression", use the word, it has a large effect on their likelihood to post on SW. While people who have the highest estimated propensity to use the word "depression" see no additional change from using the word. We see similar variances for "suicide", while the effect of using the word "can_t" is approximately constant across strata.

To investigate deeper into heterogeneous effects, we search across all treatment tokens to find those that, at different strata, *both increase and decrease* the likelihood of posting in SW. We find 161 such treatment tokens (62 unigrams and 99 bigrams) where there is at least one strata with a positive effect and one with a negative effect ($p < 0.05$). While their effects are significant within the strata, because they have contradictory effects at different strata, these tokens do not necessarily have large or significant *average* treatment effects. Table 5 shows a selection of these treatment tokens, along with the effect and the top 5 distinguishing tokens for those strata with the most significant positive and negative effects. Distinguishing tokens are ranked as the ratio of the frequency of occurrence within the strata to the ratio in the set as a whole. For example, the treatment token "stressed" increases SW posting likelihood in the 9th strata by 44%, but decreases SW posting likelihood by 33% in the 3rd strata. The former strata is distinguished by tokens as "i_do" and "i_hate", while the latter is distinguished by "there_and" and "deal_with". These results highlight the importance of context in interpreting the likely outcomes of many words used in MH posts.

We discuss the context of use of the different treatment tokens (e.g., "depression, "suicide") that are linked to increased likelihood of posting in SW. In Table 6 we present 20 tokens that were most predictive of the use of four treatment tokens. We observe that the predictive tokens are considerably different across the MH → SW and MH classes (based on Mann Whitney U tests; also ref. Kendall's τ for rank correlation), indicating that the context in which the treatment tokens are used are distinct across those who go on to post on SW in the future, versus individuals who do not.

Qualitative Interpretation

Are there meaningful themes that characterize the different treatment tokens (Table 3) linked to heightened likelihood of posting in SW in the future? Specifically, in what ways do these distinguishing tokens relate to known psychological attributes of suicide ideation examined in the literature?

Spectral Clustering of Treatment Tokens. To address these questions, we extract thematic clusters in an unsupervised manner from co-occurrence relationships between the tokens. That is, for each unique token pair, we compute their normalized frequency of appearing together in a post or comment of the 880 users in our dataset; we consider the top 100,000 most frequent co-occurring token pairs. Specifically, we use the normalized spectral clustering algorithm [64]. This algorithm accomplishes the partitioning by mapping the original space of pairwise co-occurrence relationships to an eigen space. We find that the clusters of tokens obtained through this method show significant differences among each other, based on the Kruskal-Wallis one way analysis of variance test ($p < .001$).

Extracting Themes from Clusters. Next, to examine the most dominant themes in the set of clusters obtained from

treat. token	count	coverage	treat. effect	z	χ^2	treat. token	count	coverage	treat. effect	z	χ^2
depression	318	0.901	0.3	8.04	7.78	money_i	35	0.801	0.52	5.89	3.96
useless	53	0.801	0.51	7.05	6.53	out_as	34	0.701	0.53	5.89	4.76
suicide	143	1	0.32	6.66	5.03	this_happened	35	0.901	0.51	5.89	3.72
anxiety	216	1	0.24	6.56	4.11	this_world	37	0.8	0.5	5.88	4.17
suicidal	111	0.9	0.34	6.56	5.37	over_i	35	0.901	0.51	5.86	3.58
i_almost	40	0.901	0.52	6.44	4.22	still_a	36	0.7	0.51	5.85	4.68
and_an	45	0.7	0.51	6.4	6.15	off_a	35	0.801	0.51	5.85	4.24
medicine	41	0.8	0.52	6.38	4.86	loneliness	37	0.8	0.5	5.84	3.99
unless_i	38	0.9	0.53	6.36	4.47	class_and	34	0.901	0.52	5.84	3.39
hug	42	0.8	0.52	6.36	4.9	alone_i	77	1	0.34	5.84	3.91
they_didn	42	0.801	0.51	6.33	4.72	am_the	31	0.8	0.54	5.82	3.77
take_me	40	0.9	0.52	6.32	4.33	care_i	34	0.701	0.52	5.79	4.59
and_give	42	0.8	0.51	6.23	4.61	giving_me	35	0.701	0.51	5.79	4.71
shirt	37	0.8	0.53	6.22	4.62	they_get	34	0.9	0.51	5.79	3.43
happy_i	37	1	0.52	6.21	3.59	capable	37	0.801	0.49	5.79	4.05
i_talk	41	0.8	0.51	6.2	4.81	keep_in	33	0.9	0.52	5.77	3.44
locked	39	0.8	0.51	6.17	4.64	the_amount	33	0.801	0.52	5.76	3.9
can_t	557	0.901	0.22	6.14	4.44	hate_it	38	0.7	0.48	5.76	4.37
people_on	40	0.801	0.5	6.12	4.53	socially	33	0.801	0.51	5.75	4.35
do_for	37	0.801	0.52	6.11	4.26	increase	34	0.901	0.51	5.75	3.36
problems_i	38	0.8	0.51	6.08	4.83	t_keep	33	0.901	0.52	5.75	3.56
anyone_i	37	0.701	0.51	6.07	5.36	just_in	34	0.9	0.51	5.73	3.28
thoughts_and	36	0.801	0.53	6.07	4.35	picked_up	35	0.801	0.5	5.73	3.95
ve_started	36	0.9	0.52	6.04	3.92	t_help	129	0.9	0.28	5.71	3.49
stuck_in	39	0.701	0.5	6	4.66	no_real	35	0.801	0.5	5.71	3.82
no_friends	37	0.9	0.51	6	3.85	alone	286	0.9	0.19	5.71	3.21
but_only	37	0.9	0.51	5.98	3.96	existing	36	0.8	0.49	5.71	3.84
have_nothing	36	0.901	0.51	5.98	3.41	an_idiot	34	0.7	0.51	5.71	4.47
require	36	0.9	0.52	5.97	3.94	just_trying	32	0.8	0.52	5.7	3.88
would_get	38	0.8	0.5	5.96	4.24	t_deserve	33	0.9	0.51	5.7	3.43
but_can	34	0.9	0.52	5.95	3.69	depressive	32	0.801	0.52	5.69	3.79
been_there	36	0.901	0.51	5.95	3.87	can_give	34	0.801	0.51	5.69	3.77
who_don	36	0.8	0.51	5.92	4.45	friends	502	1	0.17	5.69	2.83
world_of	35	0.901	0.52	5.92	3.67	end_i	33	0.9	0.51	5.68	3.52
kills	34	0.701	0.53	5.9	4.67	existence	35	0.801	0.5	5.68	3.89

Table 3: (Statistically significant) treatment tokens obtained via propensity score matching that contribute to *increased* change in likelihood of posting in SW.

treat. token	count	coverage	treat. effect	z	χ^2	treat. token	count	coverage	treat. effect	z	χ^2
captain	11	0.4	-0.6	-4	4.24	straight_up	12	0.601	-0.56	-3.82	2.38
differences	16	0.601	-0.57	-4.47	3.56	preferred	11	0.601	-0.56	-3.71	2.43
the_trip	11	0.601	-0.57	-3.76	3.2	awesome_i	11	0.501	-0.56	-3.68	2.86
intimate	11	0.501	-0.57	-3.73	2.93	s_at	21	0.801	-0.55	-4.83	3.33
to_in	20	0.701	-0.56	-4.92	4.1	stated	20	0.801	-0.55	-4.8	3.66
too_hard	16	0.601	-0.56	-4.4	3.56	slight	18	0.701	-0.55	-4.61	3.3
suspect	16	0.701	-0.56	-4.4	3.04	and_enjoy	17	0.601	-0.55	-4.44	3.48
always_a	14	0.601	-0.56	-4.15	3.29	gotten_to	16	0.7	-0.55	-4.35	2.77
be_working	14	0.601	-0.56	-4.12	2.73	it_work	15	0.501	-0.55	-4.22	4.17
keep_your	12	0.601	-0.56	-3.82	2.46	came_from	15	0.701	-0.55	-4.21	2.76

Table 4: (Statistically significant) treatment tokens obtained via propensity score matching that contribute to *decreased* change in likelihood of posting in SW.

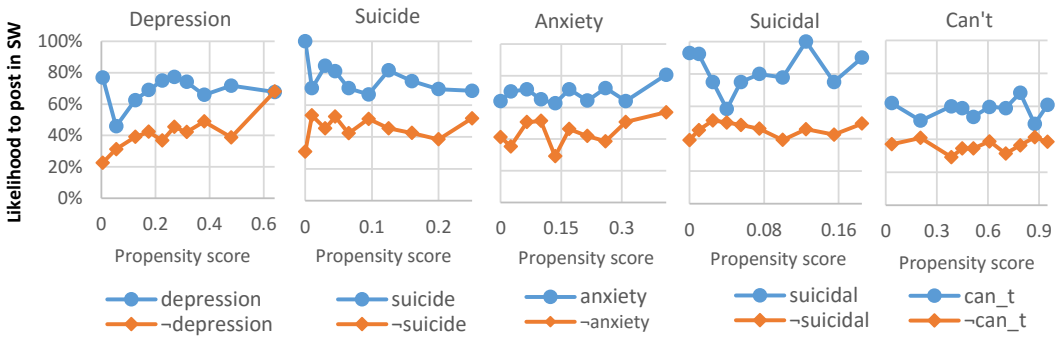

Figure 2: The likelihood of posting to SW for people using and not using target words varies over strata. Note that the strata divisions for each target word occur at different estimated propensity scores, baes.

spectral clustering, we analyze the clusters corresponding to the first six eigenvalues of the Laplacian matrix given by spectral clustering. Two researchers familiar with mental health content on social media inspected the set of tokens in

	Increase likelihood of MH → SW			*Decrease likelihood of MH → SW*		
treat. token	propensity (strata #)	effect	distinguishing historical words	propensity (strata #)	effect	distinguishing historical words
baby	0.03-0.04(5)	+44%	me_on, problem, mind, week, was_in	0.00-0.01(2)	-39%	wait, okay, thanks, re, her
have_sex	0.01-0.02(4)	+43%	instead_of, months, isn, well, a_girl	0.03-0.05(6)	-53%	months, dating, isn, such, there
medication	0.06-0.09(6)	+46%	seem_to, its, back, myself_and, seem	0.02-0.03(3)	-51%	but_then, not_to, d, i_never, everything
own	0.30-0.36(7)	+37%	all_i, can_be, we, of_this, this	0.45-0.59(9)	-32%	stop, my_own, yourself, you_can, needs
relationship_with	0.00-0.01(2)	+44%	spent, finally, my_mind, etc, think_it	0.13-0.17(9)	-36%	had_a, good, i_as, m_not, do_i
stressed	0.08-0.09(9)	+44%	i_do, as, i_hate, by, i_m	0.02-0.04(3)	-33%	there_and, but_they, deal_with, ve_had, took
upset	0.13-0.16(8)	+58%	but_i, some, a, haven_t, we	0.08-0.11(6)	-28%	living, is_this, with_the, i_started, sorry

Table 5: For selected tokens, strata with most significant increase and decrease in likelihoods of posting to SW. We show the top distinguishing tokens for each strata. All effects are significant at $p = 0.05$

treatment token	MH → SW	MH	τ
depression	make, to_the, m, around, times, bad, past, anxiety, call, sense, re, maybe, parents, world, yet, how_to, doing, at_my, my_own, point, well	make, to_the, m, around, bad, times, maybe, doing, past, yet, anxiety, how_to, world, re, parents, at_my, call, my_own, well, again, point	.36
useless	i, any, when_i, from, of_the, and_i, i_know, time, for_a, in_the, if_i, at_the, when, to_me, in, i_was, sure, because_i, i_love, but_i, about	if_i, when, i_m, in_my, day, of_my, the_time, its, is_that, so, place, different, than, but, up_i, too, haven_t, later, my_depression, my_life, anyone	.17
suicide	if_i, when, the_time, in_my, i_m, place, day, its, different, of_my, is_that, haven_t, up_i, so, my_parents, ve_never, thanks_for, later, but, parents, but_this	i, when_i, any, from, i_know, of_the, for_a, if_i, and_i, at_the, time, in_the, to_me, in, sure, i_love, it_but, when, i_was, because_i, enough_to	.18
anxiety	where, to_me, hard, i_have, recently, months, think, taking, i, let, weeks, least, issues, if_it, i_think, can_do, they_re, down, into, look, always	to_me, i, where, i_have, think, let, hard, least, why, recently, taking, i_think, always, into, months, can_do, re, if_it, weeks, not, they_re	.11

Table 6: Tokens predictive of high propensity (>0.5) of use of treatments "depression", "useless", "suicide", "anxiety", "no_friends" among MH → SW and MH users. Color intensity is proportional to frequency of the token. τ represents Kendall's τ for rank correlation.

these clusters for validation purposes. They used a semi-open coding approach to develop a codebook and extracted descriptive topical themes for the clusters (Cohen's κ=.74). During the codebook development, the two annotators referred to prior literature on the cognitive psychological integrative model of suicide [37, 29, 25].

We now present a qualitative analysis on the context in which the different tokens in each of the six clusters are used in posts in our dataset. We frame our discussion using the cognitive psychological integrative model of suicide [37, 29, 25].

Hopelessness: Tokens in the first theme cluster ("have_nothing", "no_real", "kill_myself", "abandoned", "die") were found to relate to signals of hopelessness among individuals. We note that the cognitive psychological integrative model of suicide [25] has identified hopelessness as an important mediating variable between mental illness and suicidal ideation and there is ample evidence of the decisive role of hopelessness as an indicator both of current suicide intent and as a predictor of future suicidal behavior [37, 29]:

> But I want to *die*. I feel so *abandoned*. I must be *an idiot*. I hope for some random event to *kill me* so that nobody has to be guilty. My loved ones would mourn me but they would move on. At least easier than if I actively killed myself.

Anxiety: The second theme cluster with the highest eigenvalue related to signs of anxiousness ("anxiety", "panic", "to_cry"). The cognitive psychological model of suicide has also attached great importance to individuals' feelings of anxiety despair as a predictor of future suicide [52]:

> There are times when my brain seems to shut off and I calm down. But then I *panic*... about anything. I think I have *anxiety*. I feel like nothing is mine. I don't, I genuinely don't, remember a time where I felt normal.

Impulsiveness: We observed manifestation of impulsive tones in tokens of the third theme cluster. The cognitive suicide model also suggests that impulsivity resulting from cognitive deficits (e.g., cognitive rigidity, dichotomous thinking, and inability to generate or act on alternative solutions) are prominent markers of suicide ideation [4, 37]:

> Theres a terrible feeling through my whole body every waking moment I have and theres only 2 ways to *ending it*. It hasnt been getting better only worse, I am *freaking out*. The only thing stopping me is I dont know about/have access to anything that would make it quick and clean

Self-Esteem: The cognitive suicide model has further found lowered self-esteem and self-efficacy to be important attributes among those who are prone to suicide ideation [61]. Feelings of social isolation and loneliness, conceptualized as a part of the cognitive vulnerability, have consistently been shown to be related to suicidal ideation, attempts, and completions [8]. We find that tokens of the fourth cluster appear in posts bearing a tone of decreased self-esteem, including that of guilt, self-loathing and regret:

> I am too ugly to even make friends. I *hate it*. People do not want to be associated with me because of my image. I have tried talking to girls and they've all told me to go away and to just give up. So here I am, *giving up* and ending everything.

Loneliness: The suicide model also situates loneliness as a risk that exacerbates the frequency of thoughts of suicide [4]. Our fifth theme cluster includes tokens that indicate

Model	Deviance	df	χ^2	p-value
Null	9190.6	0		
Linguistic Structure	5083.7	5	4106.9	$< 10^{-6}$
Interpers. Awareness	7949.6	4	1241	$< 10^{-9}$
Interaction	4429.2	8	4761.4	$< 10^{-6}$
Content	2793.5	15000	6397.1	$< 10^{-10}$
Full	1864.4	15017	7326.2	$< 10^{-10}$

Table 7: Summary of different model fits. Null is the intercept-only model. All comparisons with the Null models are statistically significant after Bonferroni correction for multiple testing ($\alpha = \frac{0.05}{5}$).

expressions of social isolation and detachment from the social realm, consisting of friend and family ties:

> I honestly just don't think *my friends* I've been helping out so much and supporting so much really even care about me. I've been living with *my parents* since i graduated and they aren't happy about it. today i had to deal with being yelled at by each of them. i have *no one* and i never felt such pain, i am hurt and i am *alone*.

Severe or Stigmatized Illness: Per the cognitive suicide model, experience of stigmatized and/or terminal illness (e.g., cancer) is linked to bereavement, marginalization and perceived lack of social support [54]. Tokens like "depression", "disorder", "psychosis" indicate expression of such distress:

> *Depression* and *psychosis* suck. I've been battling these for many desperate years, turns out, I can't win, no matter how hard I try. I'm *tired* of pretending. I'm *tired* of making others happy to forget about my own issues.

Classification Results

In this final subsection, we examine to what extent the linguistic structural, interpersonal awareness, interaction and content variables may be able to predict and classify MH → SW users from MH users. For this supervised learning task, we set aside 20% of our user set (total 880 users) as our held out validation set. We performed k-fold cross validation on the rest 80% users ($k = 10$) for tuning parameters of all of the five regularized logistic regression models discussed in section 4. To evaluate the goodness of fits of the regularized logistic regression models we use *deviance*. Due to the randomness introduced by cross-validation, we ran our models $k = 10$ times and here we report the results corresponding to the lowest deviances that we obtained in any of the runs.

Compared to the `Null` model, we observe that all of our models provide considerable explanatory power with significant reduction in deviances (Table 7). Particularly, the `Full` model, that uses all variables yields the best fit. We find that the difference between the deviance of the `Null` model and the deviance of the `Full` model approximately follows a χ^2 distribution: $\chi^2(15017, N = 4769) = 9190.6 - 1864.4 = 7326.2, p < 10^{-9}$.

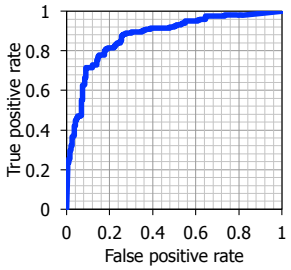

Figure 3: ROC curve in classification of MH → SW and MH users.

Finally, we summarize the performance of the `Full` regularized logistic regression model on the heldout dataset of 176 users (88 MH → SW and 88 MH users). From Table 8 we find that our model gives high accuracy in classifying the two classes, with a precision, recall and F-1 score of .8 each. Table 8 also gives the confusion matrix corresponding to the binary classification, from where we observe that MH users are marginally better classified (higher accuracy of 83.5%) compared to the MH → SW users (accuracy: 77.5%). We report the receiver operating characteristic (ROC) curve in Figure 3; the area under curve (AUC) is found to be .87.

Actual/Predicted	Class 0	Class 1	Total
Class 0	73	15	88
Class 1	20	68	88
Accuracy	83.5%	77.5%	80% (mean)
Precision	.79	.82	.81 (mean)
Recall	.83	.78	.81 (mean)
F-1	.81	.8	.80 (mean)

Table 8: Classifier performance distinguishing MH → SW and MH.

DISCUSSION

Clinical and Societal Relevance

Through this paper, we have provided a methodology to help identify individuals engaging in mental health discussions who are at a greater likelihood of transitioning to suicidal ideation discussion. An important contribution of our propensity score matching approach, in particular, has been the ability to identify linguistic constructs prior to any overt posts linked to suicidal ideation, which indicate ripe areas for further study involving causal inference. Thus we believe our methods can pave the way for longitudinal analysis of mental health content. This can help create provisions for early diagnosis of predisposition to suicidal thoughts, including therapeutic arrangements for suicide prevention. Furthermore, our work indicates linguistic constructs that should be further investigated for their ability to forecast risk, contrary to existing post-hoc approaches identifying the behavioral and cognitive markers of suicidal ideation. Broadly, our work opens up some promising opportunities of employing an unobtrusive data source like social media to understand and infer macroscale rates of suicidal ideation among those challenged by mental health concerns.

However our approach does not act as a standalone mechanism for estimating risk to suicidal ideation among those involved in mental health discussions. We caution against employing the social media predictor variables and the linguistic tokens our methods extracted as blanket filtering approaches to judge possible suicidal ideation. Such decisions are not only a controversial territory[4], but also can have drastic implications for one's health, well-being and self-esteem. Our methods and findings can be best leveraged as a *complementary screening tool* and used in conjunction with clinical, validated and conventional forms of well-being assessment.

Implications for HCI Research and Design

Provisions for Support and Interventions

Social media platforms, although do not have any legal obligation, have recently been stepping up to the cause of extending help to those who are perceived to be vulnerable. For instance, recently, Facebook, in partnership with the National

[4]Samaritans pulls 'suicide watch' Radar app (Nov 2014): http://www.bbc.com/news/technology-29962199

Suicide Prevention Lifeline, added a new feature of suicide prevention, through which an individual whose post is perceived to be distressful by a Facebook contact, could receive a support related intervention[5]. An important consideration of most of these efforts is that they either rely on people to report concerning posts or users, or apply rudimentary blanket policies around specific keywords/phrases. Both of these approaches are prone to missing vulnerable posts (those not reported), or misjudge flippant references as relating to dangerous behaviors. Our methods and findings may be utilized to expand these efforts, for instance, towards designing (semi)-automated personalized and adaptive interventions toward curbing suicidal tendencies, at the same time to improve access to appropriate peer and expert social and emotional support. We outline the following two design directions:

(1) Moderation Efforts. Individuals whose content contain phrases and other linguistic constructs relating to suicidal ideation, as revealed by our methods, may be flagged in the interfaces of moderators and other clinical experts for help and support. Community moderators may also be allowed to maintain a "risk list" in their interfaces that would include individuals forecasted by our methods to exhibit signs of suicidal ideation in the future. This would allow improved preparedness to bring timely and appropriate help to those in need. Further, on being informed that an individual in the community could be prone to suicidal thoughts in the future, moderators and experts may make provisions to connect them with appropriate mental health resources (e.g.. a hotline or a community like 7CupsofTea), encouraging peers or trusted friends and family, or field private messages with relevant information on help seeking or therapy.

(2) Self-Reflection. Interventions may also be designed that promote self-reflection of one's activity and behavior on these mental health support-seeking social media platforms. Our methods may be employed for automated (self)-assessment of behavior, cognition and affect, including serving as an early warning mechanism to individuals struggling with mental health concerns. Building off our methods, reflective interventions could also be designed to reveal longitudinal trends relating to specific markers of suicide ideation; for instance, to identify time periods of anomalous patterns, which are known to be otherwise difficult for individuals to keep track of [45]. Logging of these longitudinal trends can also serve as a diary-style data source to aid care-givers or other trained professionals and clinicians gain a deeper understanding of an individuals risk to dangerous behaviors in the future.

Ethical Considerations

Attempts to extend support to vulnerable populations, like those examined here, need a careful consideration of the risks and ethical challenges. Most importantly, at the time of design of the above suggested interventions, acceptability to social media users needs to be thoroughly investigated. In general, any intervention built out of automated algorithms like the one we proposed here, needs to honor the privacy of the individuals and those who volunteer to provide help and support. Further, beyond the design suggestions outlined above, actual modes of intervening and offering support (when, where, how) to individuals forecasted to express suicidal ideation in the future is a research and ethical question of its own. For instance, one point of intervention design is how to lead to positive behavior change, instead of counter-helpful outcomes. An unhelpful outcome could include chilling effects in participation in the community, or suicide ideation moving on to fringe or peripheral platforms where such populations might be difficult to extend help to. Finally, caution also needs to be adopted to ensure that alongside the interventions and analysis of the behavior of vulnerable communities to allow extending help and advice, ecosystems like Reddit continue to be perceived as a safe place for seeking support, and for therapeutic self-disclosure.

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

Limitations and Future Directions

There are some limitations to our work. We presume *self-selection biases* in who posts on MH and SW. Reddit allows the use of throwaway accounts or semi-anonymous identities [40], including having multiple accounts. Hence, given the sensitive and stigmatized perception of suicide, shift to SW from MH communities may happen via such an account. Such shifts would not be captured by our data collection process. Users with history of mental illness may also directly post on SW without ever posting on any mental health subreddits. Despite these limitations, we believe our work allows us to focus on a high precision dataset where do see these transitory thoughts and cognitions and our statistical method allows us characterize and predict these shifts. We also acknowledge that although our dataset is relatively larger than what has been studied in the psychology literature [11], the Reddit communities are help-seeking forums. We cannot guarantee *generalizability* with respect to the larger population.

Additionally, our findings do not provide insights into *why* an individual posting on a mental health community might decide to make the transition to SW. We also acknowledge that our algorithm for predicting which MH user will go on to post on SW in the future, is not 100% accurate, hence caution is suggested in interpreting cases of false positives or false negatives. Importantly, our inferences do not directly imply the individuals are at risk of suicide, or acted on their thoughts. Further, we note thar there are latent factors, beyond what can be observed on Reddit, that may be driving the patterns of suicidal ideation we observed.

CONCLUSION

In this paper, we presented a statistical methodology to identify whether an individual engaged in mental health discourse on social media is likely to transition to that around suicidal ideation in the future. We leveraged a large dataset from a number of mental health and suicide support communities on Reddit to address our research problem. We discovered a number of distinct markers characterizing these shifts: heightened self-attentional focus, poor linguistic coherence and coordination with the community, reduced social engagement and manifestation of hopelessness, anxiety, impulsiveness and loneliness in shared content. Through a logistic regression framework, we were also able to distinguish between individuals likely to undergo these shifts versus others who do not with high accuracy. Our findings indicate the potential of developing new kinds of technological provisions for social support and interventions catering to vulnerable populations.

REFERENCES

1. John Aldrich. 1995. Correlations genuine and spurious in Pearson and Yule. *Statistical science* (1995), 364–376.
2. APA American Psychiatric Association, American Psychiatric Association, and others. 1980. Diagnostic and statistical manual of mental disorders. (1980).
3. Roy F Baumeister. 1990. Suicide as escape from self. *Psychological review* 97, 1 (1990), 90.
4. Aaron T Beck. 1979. *Cognitive therapy of depression*. Guilford press.
5. Aaron T Beck, Roy Beck, and Maria Kovacs. 1975. Classification of suicidal behaviors: I. Quantifying intent and medical lethality. *The American journal of psychiatry* (1975).
6. Aaron T Beck, Maria Kovacs, and Arlene Weissman. 1979. Assessment of suicidal intention: the Scale for Suicide Ideation. *Journal of consulting and clinical psychology* 47, 2 (1979), 343.
7. Adriel Boals and Kitty Klein. 2005. Word use in emotional narratives about failed romantic relationships and subsequent mental health. *Journal of Language and Social Psychology* 24, 3 (2005), 252–268.
8. Ronald L Bonner and Alexander Rich. 1988. Negative life stress, social problem-solving self-appraisal, and hopelessness: Implications for suicide research. *Cognitive Therapy and Research* 12, 6 (1988), 549–556.
9. Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 75–84.
10. Marco Caliendo and Sabine Kopeinig. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* 22, 1 (2008), 31–72.
11. Jonathan TO Cavanagh, Alan J Carson, Michael Sharpe, and Stephen M Lawrie. 2003. Psychological autopsy studies of suicide: a systematic review. *Psychological medicine* 33, 03 (2003), 395–405.
12. David B Centerbar, Simone Schnall, Gerald L Clore, and Erika D Garvin. 2008. Affective incoherence: when affective concepts and embodied reactions clash. *Journal of personality and social psychology* 94, 4 (2008), 560.
13. Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication* (2007), 343–359.
14. John F Clarkin, RC Friedman, SW Hurt, R Corn, and M Aronoff. 1984. Affective and character pathology of suicidal adolescent and young adult inpatients. *Journal of Clinical Psychiatry* (1984).
15. Gualtiero B Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. 2015. Analysing the connectivity and communication of suicidal users on twitter. *Computer Communications* (2015).
16. Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *ACL Workshop on Computational Linguistics and Clinical Psychology*.
17. Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Chapter of the Association for Computational Linguistics, Denver, Colorado, USA.
18. Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *International Conference on Weblogs and Social Media (ICWSM)*.
19. Aron Culotta. 2014. Estimating county health statistics with Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1335–1344.
20. Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*. ACM, 745–754.
21. Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*. ACM, 3267–3276.
22. Munmun De Choudhury, Scott Counts, Eric Horvitz, and Aaron Hoff. 2014. Characterizing and Predicting Postpartum Depression from Facebook Data. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM.
23. Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-disclosure, Social Support, and Anonymity. In *Proc. ICWSM*. AAAI.
24. Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *AAAI Conference on Weblogs and Social Media*.
25. Gudrun Dieserud, Espen Røysamb, Øivind Ekeberg, and Pl Kraft. 2001. Toward an integrative model of suicide attempt: A cognitive psychological approach. *Suicide and Life-Threatening Behavior* 31, 2 (2001), 153–168.
26. Thomas E Ellis and Katharine G Ratliff. 1986. Cognitive characteristics of suicidal and nonsuicidal psychiatric inpatients. *Cognitive therapy and research* 10, 6 (1986), 625–634.
27. Yoav Freund and Robert E Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine learning* 37, 3 (1999), 277–296.

28. King-wa Fu, Qijin Cheng, Paul WC Wong, and Paul SF Yip. 2015. Responses to a self-presented suicide attempt in social media. *Crisis* (2015).
29. Lawrence M Glanz, Gretchen L Haas, and John A Sweeney. 1995. Assessment of hopelessness in suicidal patients. *Clinical Psychology Review* 15, 1 (1995), 49–64.
30. Jeffrey T Hancock, Christopher Landrigan, and Courtney Silver. 2007. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 929–932.
31. Amanda Hass. "Please Do Not Downvote Anyone Whos Asked for Help". http://www.slate.com/articles/technology/users/2015/03/reddit_and_suicide_intervention_how_social_media_is_changing_the_cry_for.single.html. (????). Accessed: 2015-09-10.
32. Christopher M Homan, Naiji Lu, Xin Tu, Megan C Lytle, and Vincent Silenzio. 2014. Social structure and depression in TrevorSpace. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 615–625.
33. David J Houghton and Adam N Joinson. 2012. Linguistic markers of secrets and sensitive self-disclosure in Twitter. In *System Science (HICSS), 2012 45th Hawaii International Conference on*. IEEE, 3480–3489.
34. Guido W Imbens and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
35. Sue Jamison-Powell, Conor Linehan, Laura Daley, Andrew Garbett, and Shaun Lawson. 2012. I can't get no sleep: discussing# insomnia on twitter. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, 1501–1510.
36. Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2015. Tracking suicide risk factors through Twitter in the US. *Crisis* (2015).
37. Jody Kashden, William J Fremouw, Ty S Callahan, and Michael D Franzen. 1993. Impulsivity in suicidal and nonsuicidal adolescents. *Journal of abnormal child psychology* 21, 3 (1993), 339–353.
38. Ashiqur R. KhudaBukhsh, Paul N. Bennett, and Ryen W. White. 2015. Building Effective Query Classifiers: A Case Study in Self-harm Intent Detection. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*.
39. Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 85–94.
40. Alex Leavitt. 2015. This is a Throwaway Account: Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 317–327.
41. David Lester. 1974. Demographic versus clinical prediction of suicidal behaviors: A look at some issues. (1974).
42. Peter M Lewinsohn, Ian H Gotlib, and John R Seeley. 1995. Adolescent psychopathology: IV. Specificity of psychosocial risk factors for depression and substance abuse in older adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry* 34, 9 (1995), 1221–1229.
43. David D Luxton, Jennifer D June, and Jonathan M Fairall. 2012. Social media and suicide: A public health perspective. *American Journal of Public Health* 102, S2 (2012), S195–S200.
44. 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 *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1511–1526.
45. Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: an intelligent system for emotional memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 849–858.
46. Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the Language of Schizophrenia in Social Media. In *Proceedings of the NAACL Workshop on Computational Linguistics and Clinical Psychology*.
47. Elizabeth L Murnane and Scott Counts. 2014. Unraveling abstinence and relapse: smoking cessation reflected in social media. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 1345–1354.
48. Matthew K Nock and Alan E Kazdin. 2002. Examination of affective, cognitive, and behavioral factors and suicide-related outcomes in children and young adolescents. *Journal of clinical child and adolescent psychology* 31, 1 (2002), 48–58.
49. Minsu Park, David W McDonald, and Meeyoung Cha. 2013. Perception Differences between the Depressed and Non-depressed Users in Twitter. In *Proceedings of ICWSM*.
50. Michael J Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health.. In *ICWSM*.
51. Umashanthi Pavalanathan and Munmun De Choudhury. 2015. Identity Management and Mental Health Discourse in Social Media. In *Proceedings of the 24th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 315–321.

52. C Pawlak, T Pascual-Sanchez, P Rae, W Fischer, and F Ladame. 1999. Anxiety disorders, comorbidity, and suicide attempts in adolescence: A preliminary investigation. *European Psychiatry* 14, 3 (1999), 132–136.
53. Keith Petrie and Richard Brook. 1992. Sense of coherence, self-esteem, depression and hopelessness as correlates of reattempting suicide. *British Journal of Clinical Psychology* 31, 3 (1992), 293–300.
54. Daniel W Prezant and Robert A Neimeyer. 1988. Cognitive predictors of depression and suicide ideation. *Suicide and Life-Threatening Behavior* 18, 3 (1988), 259–264.
55. Michael J Priester and George A Clum. 1993. Perceived problem-solving ability as a predictor of depression, hopelessness, and suicide ideation in a college population. *Journal of Counseling Psychology* 40, 1 (1993), 79.
56. Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
57. Donald B Rubin. 2011. Causal inference using potential outcomes. *J. Amer. Statist. Assoc.* (2011).
58. Thomas D Ruder, Gary M Hatch, Garyfalia Ampanozi, Michael J Thali, and Nadja Fischer. 2011. Suicide announcement on Facebook. *Crisis: The Journal of Crisis Intervention and Suicide Prevention* 32, 5 (2011), 280–282.
59. Anita Saltz and Stanley Marsh. 1990. Relationship between hopelessness and ultimate suicide: a replication with psychiatric outpatients. *American Journal of Psychiatry* 147 (1990), 190–195.
60. H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 118–125.
61. Ralf Schwarzer and Reinhard Fuchs. 1995. Changing risk behaviors and adopting health behaviors: The role of self-efficacy beliefs. *Self-efficacy in changing societies* (1995), 259–288.
62. RJ Senter and EA Smith. 1967. *Automated readability index*. Technical Report. DTIC Document.
63. David Shaffer. 1988. The epidemiology of teen suicide: an examination of risk factors. *Journal of Clinical Psychiatry* (1988).
64. Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22, 8 (2000), 888–905.
65. Edwin S Shneidman. 1993. Commentary: Suicide as psychache. *The Journal of nervous and mental disease* 181, 3 (1993), 145–147.
66. Morton M Silverman and Ronald W Maris. 1995. The prevention of suicidal behaviors: An overview. *Suicide and Life-Threatening Behavior* 25, 1 (1995), 10–21.
67. Steven John Stack. 2014. Mental Illness and Suicide. *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society* (2014).
68. Hajime Sueki. 2015. The association of suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young internet users in Japan. *Journal of affective disorders* 170 (2015), 155–160.
69. Katherine M Thomas and Marshall Duke. 2007. Depressed writing: Cognitive distortions in the works of depressed and nondepressed poets and writers. *Psychology of Aesthetics, Creativity, and the Arts* 1, 4 (2007), 204.
70. Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing Depression from Twitter Activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3187–3196.
71. Hong-Hee Won, Woojae Myung, Gil-Young Song, Won-Hee Lee, Jong-Won Kim, Bernard J Carroll, and Doh Kwan Kim. 2013. Predicting national suicide numbers with social media data. *PloS one* 8, 4 (2013), e61809.
72. G Udny Yule. 1897. On the theory of correlation. *Journal of the Royal Statistical Society* 60, 4 (1897), 812–854.