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

Munmun De Choudhury  Emre Kiciman  Mark Dredze
Georgia Tech  Microsoft Research  Johns Hopkins University
Atlanta GA 30332  Redmond WA 98052  Baltimore MD 21218
munmund@gatech.edu  emrek@microsoft.com  mdredze@cs.jhu.edu

Glen Coppersmith  Mrinal Kumar
Qntfy.io  Georgia Tech
Crownsville MD, 21032  Atlanta GA 30332
glen@qntfy.io  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 deaths1. 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

1http://www.cdc.gov/ViolencePrevention/pdf/Suicide_DataSheet_a.pdf
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
Table 1: Example titles of posts in the MHs and SW datasets; content has been carefully paraphrased to protect the privacy of the individuals.

**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.

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, reditors 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/survivorsofsubabuse, 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 → 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 → 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 → 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 → SW users, and 8,318 posts and 54,086 comments from the 440 MH 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, verbs², 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...
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 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, \( H = h_1, \ldots, h_n \), where \( h_i = 1 \) 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 \( \chi^2 \) 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 defined above. 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.

<table>
<thead>
<tr>
<th>Linguistic Structure</th>
<th>MH Mean</th>
<th>MH → SW Mean</th>
<th>( z )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>0.294</td>
<td>0.125</td>
<td>6.51***</td>
<td></td>
</tr>
<tr>
<td>verbs</td>
<td>0.045</td>
<td>0.107</td>
<td>2.19**</td>
<td></td>
</tr>
<tr>
<td>abverbs</td>
<td>0.048</td>
<td>0.099</td>
<td>4.87***</td>
<td></td>
</tr>
<tr>
<td>readability index</td>
<td>0.609</td>
<td>0.232</td>
<td>5.51***</td>
<td></td>
</tr>
<tr>
<td>accommodation</td>
<td>0.857</td>
<td>0.467</td>
<td>5.46**</td>
<td></td>
</tr>
<tr>
<td>Interpersonal Awareness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st person singular</td>
<td>0.018</td>
<td>0.086</td>
<td>-10.8***</td>
<td></td>
</tr>
<tr>
<td>1st person plural</td>
<td>0.093</td>
<td>0.078</td>
<td>4.53*</td>
<td></td>
</tr>
<tr>
<td>2nd person</td>
<td>0.058</td>
<td>0.031</td>
<td>8.01*</td>
<td></td>
</tr>
<tr>
<td>3rd person</td>
<td>0.087</td>
<td>0.042</td>
<td>6.32***</td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>posts authored</td>
<td>18.97</td>
<td>10.31</td>
<td>2.53*</td>
<td></td>
</tr>
<tr>
<td>post length</td>
<td>215.62</td>
<td>443.73</td>
<td>-15.4***</td>
<td></td>
</tr>
<tr>
<td>comments authored</td>
<td>122.42</td>
<td>106.22</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>comments received</td>
<td>19.862</td>
<td>13.414</td>
<td>1.05</td>
<td></td>
</tr>
<tr>
<td>comment length authored</td>
<td>63.417</td>
<td>87.116</td>
<td>-1.88</td>
<td></td>
</tr>
<tr>
<td>response velocity (mins)</td>
<td>7.746</td>
<td>6.966</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>vote difference</td>
<td>28.788</td>
<td>7.681</td>
<td>7.18***</td>
<td></td>
</tr>
</tbody>
</table>

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 \( p = 0.05 / N; 0.01 / N; 0.001 / N, (N = 17) \), following Bonferroni correction.

Finally, 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.

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
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 $ightarrow$ SW and MH classes (based on Mann Whitney U tests; also ref. Kendall’s $\tau$ 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
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 spectral clustering, we analyze the clusters corresponding to the first six eigenvalues of the Laplacian matrix given by spectral clustering, we analyze the clusters corresponding to the first six eigenvalues of the Laplacian matrix given by spectral clustering, we analyze the clusters corresponding to the first six eigenvalues of the Laplacian matrix given by spectral clustering, we analyze the clusters corresponding to the first six eigenvalues of the Laplacian matrix given by
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 k=.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].

**Hoplessness:** 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, 25].

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]:

<table>
<thead>
<tr>
<th>treatment token</th>
<th>propensity (strata #)</th>
<th>effect distinguishing historical words</th>
<th>propensity (strata #)</th>
<th>effect distinguishing historical words</th>
</tr>
</thead>
<tbody>
<tr>
<td>baby</td>
<td>0.03-0.04(5)</td>
<td>+44% me_on, problem, mind, week, was_in</td>
<td>0.00-0.01(2)</td>
<td>-39% wait, okay, thanks, re, her</td>
</tr>
<tr>
<td>have_sex</td>
<td>0.01-0.02(4)</td>
<td>+43% instead_of, months, isn, well, a_girl</td>
<td>0.03-0.05(6)</td>
<td>-53% months, dating, isn, such, there</td>
</tr>
<tr>
<td>medication</td>
<td>0.06-0.09(6)</td>
<td>+46% seem_to, its, back, myself_and, seem</td>
<td>0.02-0.03(5)</td>
<td>-51% but_then, not_to, d, i_never, every</td>
</tr>
<tr>
<td>own</td>
<td>0.30-0.36(7)</td>
<td>+37% all_i, can_be, we, of_this, this</td>
<td>0.45-0.59(9)</td>
<td>-32% stop, my_own, yourself, you_can, needs</td>
</tr>
<tr>
<td>relationship_with_stressed</td>
<td>0.00-0.01(2)</td>
<td>+44% spent, finally, my_mind, etc, think_it</td>
<td>0.13-0.17(9)</td>
<td>-36% had_a_good, i_as, m_not, do, there_and, but_they, deal_with, ve_had_took</td>
</tr>
<tr>
<td>upset</td>
<td>0.13-0.16(8)</td>
<td>+58% but_j, some_a, a_haven_t, we</td>
<td>0.08-0.11(6)</td>
<td>-28% living, is_this, with_the, i_started, sorry</td>
</tr>
</tbody>
</table>

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

Table 6: Tokens predictive of high propensity (p > 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. r represents Kendall’s τ for rank correlation.

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 time 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...
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 χ² distribution:

\[ \chi^2(15017, N = 4769) = 9190.6 - 1864.4 = 7326.2, p < 10^{-9}. \]

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.

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 macro-scale 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, 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

---

Table 7: Summary of different model fits. All comparisons with the Null models are statistically significant after Bonferroni correction for multiple testing (α = $\frac{0.05}{10}$).

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>df</th>
<th>χ²</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>9190.6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linguistic Structure</td>
<td>5083.7</td>
<td>5</td>
<td>4106.9</td>
<td>&lt; 10^{-6}</td>
</tr>
<tr>
<td>Interpers. Awareness</td>
<td>7949.6</td>
<td>4</td>
<td>1241</td>
<td>&lt; 10^{-6}</td>
</tr>
<tr>
<td>Interaction</td>
<td>4429.2</td>
<td>8</td>
<td>4761.4</td>
<td>&lt; 10^{-6}</td>
</tr>
<tr>
<td>Content</td>
<td>2793.5</td>
<td>15000</td>
<td>6397.1</td>
<td>&lt; 10^{-10}</td>
</tr>
<tr>
<td>Full</td>
<td>1864.4</td>
<td>15017</td>
<td>7326.2</td>
<td>&lt; 10^{-10}</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Actual/Predicted</th>
<th>Class 0</th>
<th>Class 1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 0</td>
<td>73</td>
<td>15</td>
<td>88</td>
</tr>
<tr>
<td>Class 1</td>
<td>20</td>
<td>68</td>
<td>88</td>
</tr>
<tr>
<td>Accuracy</td>
<td>83.5%</td>
<td>77.5%</td>
<td>80% (mean)</td>
</tr>
<tr>
<td>Precision</td>
<td>.79</td>
<td>.82</td>
<td>.81 (mean)</td>
</tr>
<tr>
<td>Recall</td>
<td>.83</td>
<td>.78</td>
<td>.81 (mean)</td>
</tr>
<tr>
<td>F-1</td>
<td>.81</td>
<td>.8</td>
<td>.80 (mean)</td>
</tr>
</tbody>
</table>

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

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\(^3\). 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.

**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 we 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 that 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


