Natural Language Processing of Social Media as Screening for Suicide Risk

Glen Coppersmith, Ryan Leary, Patrick Crutchley and Alex Fine

Qntfy, Boston, MA, USA.

Biomedical Informatics Insights
Volume 10: 1–11

DOI: 10.1177/1178222618792860

ABSTRACT: Suicide is among the 10 most common causes of death, as assessed by the World Health Organization. For every death by suicide, an estimated 138 people's lives are meaningfully affected, and almost any other statistic around suicide deaths is equally alarming. The pervasiveness of social media—and the near-ubiquity of mobile devices used to access social media networks—offers new types of data for understanding the behavior of those who (attempt to) take their own lives and suggests new possibilities for preventive intervention. We demonstrate the feasibility of using social media data to detect those at risk for suicide. Specifically, we use natural language processing and machine learning (specifically deep learning) techniques to detect quantifiable signals around suicide attempts, and describe designs for an automated system for estimating suicide risk, usable by those without specialized mental health training (eg, a primary care doctor). We also discuss the ethical use of such technology and examine privacy implications. Currently, this technology is only used for intervention for individuals who have "opted in" for the analysis and intervention, but the technology enables scalable screening for suicide risk, potentially identifying many people who are at risk preventively and prior to any engagement with a health care system. This raises a significant cultural question about the trade-off between privacy and prevention—we have potentially life-saving technology that is currently reaching only a fraction of the possible people at risk because of respect for their privacy. Is the current trade-off between privacy and prevention the right one?

KEYWORDS: Suicide, suicide screening, suicide prevention, social media, data science, natural language processing

RECEIVED: February 28, 2018. **ACCEPTED:** June 20, 2018.

TYPE: Proceedings from the Digital Mental Health Conference, London, 2017 - Review

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Qntfy is a for-profit company that designs analytic products related to mental health. Qntfy funded this research in the interest of sharing our discoveries with the scientific community.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Glen Coppersmith, Qntfy, Boston, MA 02115, USA.
Email: glen.coppersmith@qntfy.com

Introduction

An estimated 16 million suicide attempts occur each year. Of these, approximately 800 000 people will die from those attempts.[1] Suicide deaths have increased by 24% in the past 20 years, making suicide one of the top 10 causes of death in the United States,[2] a pattern that seems to be constant across geographic region within the country.[3] Not only is the magnitude of the problem large and worsening, there has been little progress made over the past 50 years in understanding suicide and improving outcomes in at-risk individuals.[4] The stubbornness of the problem reflects its complexity, and the densely interwoven causal factors underlying it. Here we focus on one piece of the puzzle: how can we identify those who are at risk of taking (or attempting to take) their own life, and how can this screening be used to foster effective interventions?

Assessing an individual's risk for suicidal behavior is difficult. Experienced and talented clinicians frequently struggle to correctly interpret signals in their patients' behavior that are indicative of suicide risk. Setting aside the profound difficulties associated with understanding an individual's personal history and its relationship to their capacity and motivations for self-harm, there are at least 2 practical reasons that assessing suicide risk is difficult: (1) the latency between the onset of acute risk for suicide and the suicide attempt itself may be too small for interventions requiring contact with health professionals, and (2) most existing methods for detecting high risk of suicide require that individuals disclose their wish to harm themselves to a health professional. In this article, we explore the possibility that *digital life data*—that is, the interactions that a person has with digital devices, through the daily course of their life—collected passively but with consent might at least partially address each of these difficulties.

Individuals come to be at risk for suicide at different temporal intervals relative to suicide attempts. For instance, the kind of social isolation that is frequently associated with suicide can gradually accumulate over the course of a person's life or may become acute in a very short period of time after a traumatic life event such as the loss of a loved one.

Moreover, once an individual is engaged with a health care professional, standard methods of suicide intervention require both that the clinician administer a standardized risk assessment (often in the form of a questionnaire) and that the patients *disclose* their intention to harm themselves. Each of these presents its own challenges. First, administering a suicide screening tool may place an unreasonable burden on the health care provider. The standard for suicide screening within the health care system is Beck's Scale for Suicide Ideation, a 5- or 19-item questionnaire examining the patient's active and passive desire for suicide, and any specific plans they might have.[5] Many patients who are at risk for suicide only interact with primary care physicians (PCPs) or emergency departments (EDs) rather than those with psychiatric specialties. Such health care providers may lack the time or the training to administer a specific questionnaire for suicide risk. Indeed, enabling PCPs and EDs to better screen for suicide risk has been posited as a method for reducing the suicide rate.[6,7] Second, patients cannot always be relied upon to disclose suicidal thoughts in the clinical setting.[8] These factors have the

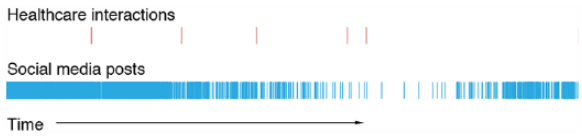

Figure 1. One example: person's interaction with the health care system (in red hashes) and with a social media platform (in blue hashes) over a period of 4 years (*x*-axis, left is earlier in time, right is later in time). Social media provides information in the "clinical whitespace"[15] between interactions with the health care system. Summarized from health record data presented in Padrez et al.[16]

potential for a large impact if missed screening opportunities can be capitalized upon[9]: 24.6% of patients attempting suicide visited a mental health professional in the 1-week period prior to their attempt, with 38.3% having visited a health professional of any kind in the same period of time.

Independent of the efficacy and specificity of these scales and instruments for screening for suicide risk, they are pragmatically limited in their application to times when a patient is interacting with the health care system, the health care professional they are interacting with deems administration of a suicide risk screening a worthy use of time, and the patient is willing to disclose suicidal thoughts, plans, or actions at that time.[10] Many at risk are not engaged with the health care system at all[11]: there are strong correlations of state-based suicide rates with indicators for lack of access to health care. Ahmedani et al[9] report that 26.7% of individuals attempting suicide had not seen a mental health care practitioner in the prior year, and 5.4% had not seen *any* health care practitioner in the prior year, with significant variation among racial and ethnic groups. Furthermore, the health care system seems to have a significant and systematic gap for helping individuals *after* their suicide attempt: Substance Abuse and Mental Health Services Administration (SAMHSA) survey data[12] from 2011 showed that 18.3% of drug-related suicide attempt ED admissions showed no evidence of follow-up treatment. Taken together, the existing infrastructure for suicide risk detection and intervention highlights the need for some way of screening individuals *outside* the context of their interactions with the health care system, and for detecting signals positively associated with suicidal behavior that are less overt than explicit disclosure.

The above suggests at least 2 paths to improve screening for suicide risk within the existing health care system: (1) providing evidence of risk without relying exclusively on self disclosure and (2) the pragmatic reduction of time needed with a health care worker to administer the screening tool (and thus reducing the resistance to administering the tool on the part of the health care professional). Furthermore, a model capable of identifying those at risk *outside* the health care system could be part of a system to funnel them toward appropriate care.

Interestingly, there has been significant progress in using data outside the health care system to assess and understand mental health and well-being in recent years (which some refer to as *digital phenotyping*, eg, Onnela and Rauch[13]). In particular, signals related to a person's mental health and well-being have been extracted from a person's *digital life*. Digital life generally covers all the interactions that a person has with digital devices, through the daily course of their life, including social media data (eg, Facebook, Instagram, Twitter, Reddit), data from wearable devices (eg, Fitbit, Jawbone), geolocation, actigriphy from phone sensors,[14] or interactions with smart devices (eg, Amazon's Alexa or other Internet of Things [IoT] devices). For one point of comparison, see Figure 1, which shows one person's interaction with the health care system (in red hashes) and posts on Facebook (in blue hashes) over the course of a few years.[16] A growing body of work suggests that various facts about an individual's mental health can be inferred from the text that a person generates, raising the intriguing possibility that social media data could be used as a screening and/or early detection tool. Social media data have been found to contain predictive signal for conditions, including major depressive disorder,[17,18] post-traumatic stress disorder,[19–25] schizophrenia,[26] eating disorders,[27,28] generalized anxiety disorder, bipolar disorder,[29] self-harm,[25] suicide,[30–38] borderline personality disorder, and others.[39,40]

Computational analysis of social media data may therefore fulfill the desiderata of a screening system that (1) captures an individual's behavior outside of their interactions with the health care system and (2) is amenable to the kind of automation and scalability that is often sorely lacking in underfunded and resource-constrained health care providers.

This article's primary contributions are as follows:

- The creation of an automated model for analysis and estimation of suicide risk from social media data.
- An examination of how this could be used to improve existing screening for suicide risk within the health care system.
- An exploration of the ethical and privacy concerns of creating a system for suicide risk screening not currently in care.

Data

The creation and evaluation of these machine learning algorithms depend on having social media data from people prior to a suicide attempt and a contrasting set of users who have not attempted suicide. To train the algorithms to differentiate between those who are at risk for suicide and those who are not, we also needed examples of users who are as close a match as possible to those who would attempt suicide, but did not attempt suicide (so far as we know). Social media posts from control users thus provide a baseline to which the data from those would go on to attempt suicide can be compared. We combine data from 2 sources to create this dataset—examining public self-stated data and using data donated through OurDataHelps.org.

I'm so glad I survived my suicide attempt to see the wedding today. I was so foolish when I was young, so many suicide attempts!
I have been out of touch since I was hospitalized after my suicide attempt last week. It's been half a year since I attempted suicide, and I wish I had succeeded
I'm going to go commit suicide now that the Broncos won... #lame It is going to be my financial suicide, but I NEEEEEEEEEED those shoes.

Figure 2. Fictitious example: posts of genuine statements of a suicide attempt (top), genuine statements of a suicide attempt indicating a time for the attempt (middle), and disingenuous statements of a suicide attempt (bottom). Only data from the middle are include in the analysis here (because we can ascertain a date and have a chance to obtain data prior to that date). In line with Benton et al,[41] we do not reproduce text directly from social media data and instead paraphrase to protect the user's identity.

OurDataHelps.org users

The first data source is from a set of users who have graciously donated their data to support research in this area through OurDataHelps.org. Users of this platform sign up and authorize access to data from their digital life—social media (eg, Facebook, Twitter, Instagram, Reddit, Tumblr), wearable (eg, Fitbit, Jawbone), and other technology (eg, Strava, Runkeeper). Users also fill out questionnaires asking for basic demographic data as well as for information about their history with various mental health conditions. Specifically relevant to this study, they note the number and dates of past suicide attempts. A handful of the users' data in OurDataHelps.org were provided by their loved ones, posthumously after their suicide. Through this authorized access, we examine the posts that the user made publicly and visible to their "friends and family"—this notably *excludes* private message data like Facebook Messenger or Twitter direct messages. From the users of OurDataHelps.org, we have 186 users who have attempted suicide. Many of the users who donated data at OurDataHelps.org did not attempt suicide or have any reported mental health diagnoses. For each user who attempted suicide from the OurDataHelps.org population, we find a user with the same gender and nearly the same age to serve as a control.

Public self-stated users

The second data source comes from users who publicly discuss past suicide attempts on social media, as originally described in Coppersmith et al[19] and adapted to examine suicide attempts in Coppersmith et al.[36] Here, we significantly increase the number of users in the dataset (4 times the size of the dataset in 2016), which increases the accuracy and generality of the machine learning algorithms. These users make posts that describe the date of a past suicide attempt, like those found in Figure 2. People may make statements like this (1) to explain past behavior, (2) as a method for fighting the stigma and discrimination associated with mental illness, or (3) as a way of offering support to those in a similar situation. From these statements, we can infer the date of the suicide attempt and examine public data prior to that attempt from this user. To find matched controls to the self-stated data, we examine a random sample of Twitter data, finding users whose posts were primarily in English, and were either human-annotated or estimated to be of the same gender and approximate age as a user who has attempted suicide. (Estimation performed via the methods described in Sap et al.[42])

Combined dataset

When the data from these 2 sources are combined, we have 547 users who have attempted suicide, 418 users of which we know the month of their suicide attempt (263 of which we know the exact date) and can access data prior to the attempt. There were 4 users present in both datasets (ie, someone who had both donated data and been found via the public self-stated methods), and these users are included only once in the combined dataset. For the analysis below, we restrict the data from these users to only the posts made in the 6 months prior to their suicide attempt. This results in a final dataset of 418 users who have attempted suicide for whom we have up to 6 months of posts prior to their suicide attempt (and an equal number of demographically matched controls). For comparison, Coppersmith et al[36] had 125 users compared with this article's 418. For each user, we have an average of 473 social media posts (and an equal number drawn from each matched control), for a total of 197 615 posts from those who would go on to attempt suicide in the next 6 months, and an additional 197 615 posts from their matched controls (for a total of 395 230 posts). The age and gender distribution of these data can be found in Figure 3. Most of this population is comprised of females aged 18 to 24, though there is a reasonable number of males from a similar age range as well. Less well represented, but not absent, are people older than 30 and those of a non-binary gender.

Methods and Results

For a screening method to be highly scalable, it must minimize the time taken by humans required to provide the assessment. Thus, we design methods based on automated and computerized assessment and analysis. The bulk of the technical design work necessary here is in the creation and evaluation of machine learning methods, which examine existing data to automatically

extract and determine the relevant patterns for who is at risk for suicide. The dataset described above supports this kind of analysis. Here, we describe the methods at work in some technical depth. However, the implications of the technology can be readily understood independent of the technical details in this section.

The classification given to each user by these models (ie, the model's best guess as to whether each user will go on to attempt suicide) is based on tens of thousands of small clues, too many to enumerate in any way that is accessible to human intuition. However, we can examine some text that is scored highly by the algorithm, shown in Figure 4. These illustrative examples give a feel for the validity of the technique. Similarly, Figure 5 provides a visualization of the scores of the model for each user over the last 200 tweets prior to their suicide attempt in blue (ie, all suicide attempts occur at 200 on the x-axis in this plot) and their matched controls in green. Higher is indicative of more risk for suicide. Although there is variability within each line, and portions of time when control users are above users at risk, they are generally separable—more blue on the top and more green on the bottom.

The key results are (1) that there are quantifiable signals present in the language used on social media that machine learning algorithms can use to separate users who would go on to attempt suicide from those who would not with relatively high precision and (2) that the machine learning algorithms depend on a wide swath of subtle clues, rather than a few indicative phrases. The algorithm's ability to distinguish users who would go on to attempt suicide—crucially, without input from a trained human—is good enough that it is worth considering how a tool incorporating the algorithm might fit into a clinical application. The implications of this are further discussed in the following section, and non-technical readers may safely skip to that section. The remainder of this section describes the machine learning techniques employed here in more detail.

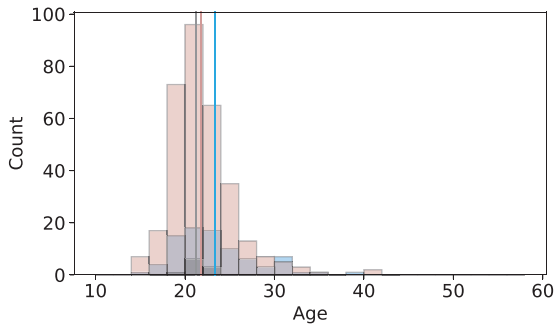

Figure 3. Histogram of age of users, separated by gender. Females in red, males in blue, non-binary in gray. The mean age of each gender is indicated by a vertical line of the same color.

Deep learning for analysis of language

Recent advancements in natural language processing have leveraged deep learning to improve the state of the art for language modeling[43] and text classification.[44,45] These techniques are loosely related to human neural architecture in that they are composed of networks of "neurons" capable of learning complex, non-linear relationships between input data (in this case, social media posts) and output data (whether or not the post was composed by someone who would go on to attempt suicide).

Although the dataset here is limited in the number of attempting users, these users together contribute hundreds of thousands of individual messages. Conceptually, we train a text classification model intended, from a single post, to predict whether the author is at risk for a suicide attempt. Rather than use these individual message-level scores, we use the aggregated scores from many posts from a single user to predict the individual's risk.

With the somewhat limited amount of training data available for the task (when compared with traditional "deep learning" tasks where tens or hundreds of millions of examples are used), we leverage both supervised and unsupervised learning methods to prevent overfitting the model to the training data. The model first uses a word embedding layer to project each word into a dense vector space. This low-dimensional vector space is crafted such that semantically similar words remain a short distance from each other in Euclidean or cosine space. Our model is initialized with pretrained GloVe embeddings[46] to reduce the risk of overfitting the model. The embeddings are trained by their original authors over a significantly larger dataset to learn to encode general language usage (like which words are semantically similar). This, in part, compensates for the relatively small sample of users and messages under consideration here. These embeddings are later fine-tuned during the

@user I'm sorry, this is terrible... I've cut and it won't stop bleeding.
I feel horrible sometimes. Hate how I look, how I sound, how I exist.
I guess there's not changing how much you mean to someone. Doesn't matter how worthless you feel.
I feel it'd be best if I weren't here. I couldn't annoy anyone that way.

i think it would be best if i weren't here . that way i couldn't annoy anyone . i've been cutting again

Figure 4. Fictitious rephrasing of posts scored highly by the suicide risk model. The bottom panel indicates what the attention mechanism within the neural architecture is highlighting as an important signal for a person at risk. Larger words (eg, "cutting" and "again") have more power in the current prediction than smaller words (eg, "be" or "been").

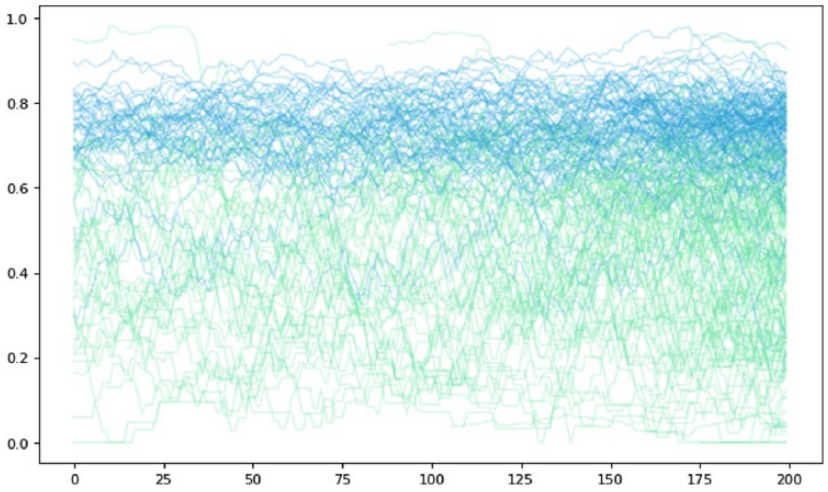

Figure 5. Scores for each user's 200 posts preceding the suicide attempt (and matched controls) are plotted with a moving window average; *y*-axis magnitude corresponds to risk for suicide (likelihood of messages belonging to a user from the positive class). Lines in blue are users who will go on to attempt suicide, whereas users in green are their matched controls.

model training process, shifting to better capture nuances of language related to mental health. Sequences of word vectors, one sequence per message, are processed via a bidirectional Long Short-Term Memory (LSTM) layer to capture contextual information between words. Next, the output of each layer is combined into a single vector using skip connections into a self-attention layer. The attention mechanism is used to apply weights to the timesteps of the sequence such that the most informative subsequences are more strongly considered in the final prediction.[47,48] Finally, a linear layer with *softmax* output predicts a posterior probability representing the likelihood that the text was written by an author at risk for suicide.

Assessment and interpretation of results

To evaluate the efficacy of these deep learning techniques for estimating suicide risk from social media data, we evaluated the classification performance in 10-fold cross-validation across pairs of users (one user prior to their suicide attempt and their demographically matched control). Figure 6 shows the receiver operating characteristic (ROC) curve of performance for these models for separating users at risk for a suicide attempt from their matched controls. Each ROC curve shows the trade-offs between true positives and false alarms over the sensitivity of the model. For a single point of comparison, at 10% false alarm rate (0.1 on the *x*-axis), the models in Figure 6 range from 70% to 85% true positive rate (0.7-0.85 on the *y*-axis). This, to the best of our knowledge, is state of the art performance on suicide risk prediction from social media alone. Chance performance is denoted by the diagonal dashed line, and the further up and left of the dotted line a model's curve is, the better it generally is at the task. Area under the ROC curve (AUC) is often used as a singular scalar metric of performance (chance: 0.5; perfect discrimination: 1.0); this is given in the figure legend. Each line represents an amount of data used prior to the attempt to make

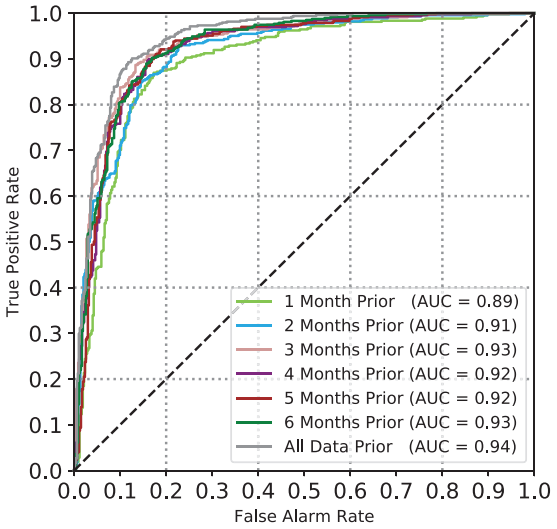

Figure 6. ROC curves for models separating users prior to a suicide attempt from their matched controls. The green line only uses data for the month prior to the suicide attempt to make the classification (30 to 0 days prior), the blue line uses data from 2 months prior (60 to 0 days prior), and so on. The black line indicates performance using all of the data available for that user prior to their attempt. ROC indicates receiver operating characteristic.

the prediction—the green line looks only at data from 30 to 0 days prior to the suicide attempt. Similarly, the blue line uses data from 60 to 0 days prior to the suicide attempt.

Excluded for brevity are similarly performant experiments examining data further out from the suicide attempt. Ultimately, we found that performance was roughly comparable if we examined data a few months prior (180 to 90 days prior) to the attempt and excluded data immediately preceding the attempt (90 to 0 days prior). Thus, that suggests that the model is capturing trait-type information (relevant to risk for suicide at some point in time) rather than state-type information (relevant to imminent risk of harm).

Although the majority of this dataset was female users (78%), the model seems to be sufficiently expressive to also capture information for males, with a slight loss of precision. There were not a sufficient number of users of a non-binary gender to fully assess performance, but anecdotally the model performance seemed to be on par with the other genders. This suggests that the model has learned information about suicide risk that seems to be relevant across genders. This does *not* necessarily mean that they are governed by a single theoretic model but simply that there are common language cues across the genders and the model built is sufficiently expressive to capture many of the differences.

To interpret the performance of this classifier as a screening tool, we can pick a few reasonable trade-offs between true positive and false alarm rates, and examine who would be deemed "at risk" by the algorithm. Let us assume a theoretical population of 1000 people who will get screened by this method. It is expected that 4% to 8% of them would go on to attempt suicide, so for the purposes of this illustration we use 6%.[49,50] That means that 60 people from this population would go on to attempt suicide, and thus 940 people would not. If we deployed this screener against this population, allowing 1% false alarm rate (thus 24% true positive rate), we would expect 25 people to be flagged as "at risk," 15 of which would go on to attempt suicide—that would mean that 60% of the population flagged at risk would go on to attempt suicide. Similarly, at 10% false alarms (84% true positive rate), 144 would be flagged, 35% of which would go on to attempt suicide. At 2% false alarms (35% true positive rate), 40 people would be flagged, 40% of which would go on to attempt suicide. A direct point of comparison for performance from clinicians is difficult to match exactly, but the results from Franklin et al[4] suggest that the equivalent expected proportion of users identified as "at risk" by clinicians who would go on to attempt suicide would be in the 4% to 6% range—significantly lower than the 40% to 60% here. Which of these operating points are used should be determined by what the next step taken after this screening is. If, for example, a positive flag means that their healthcare provider will be reminded of the importance of the clinical screen, perhaps 10% false alarms is a reasonable operating point. If the next step is, instead, a psychology consultation, perhaps something more restrictive (eg, 2% false alarms) is more appropriate, given the cost of scheduling and having such a consultation with an in-demand resource.

Discussion

The methods described here demonstrate that signals exist within social media data that are quantifiable and relevant to suicide risk. Concretely, we described algorithms that are able to identify people at risk for suicide from the analysis of the language of their social media posts, at levels of precision that suggest clinical utility, and at the period early enough to permit reasonably scalable and durable interventions (ie, months preceding crisis, rather than in the moment of crisis). The model, as configured here, was optimized toward detecting *trait*-level risk for suicide, or how at risk a person tends to be over a long period of time as opposed to *state*-level information that is more transient and related to a short period of risk. This was a deliberate choice as a screening tool to find users at risk well before the point of their attempt. Other parameterizations of this model and other methods of dividing the data for training may allow more state level and information about proximal risk of suicide to be discovered, but we leave that to future work. In our estimation, the ability for a system of care to respond in the moment of crisis is more costly, more dangerous (to the patient and potentially those intervening), and ultimately of less utility than intervening months or years prior to an attempt.

Importantly, these models depend on a wide variety of signals, most of which are not what a clinician would generally ascribe to association with suicide risk. This is similar in spirit to the findings of the Crisis Text Line (CTL), which found that the word "ibuprofen" was one of the most highly correlated words used by users at *imminent* risk of suicide, when talking to a peer providing private, anonymous, emotional support.[51] The findings here are complementary to the CTL findings in many ways, though the methods used to derive them are similar. Where their data are private, anonymous conversations, these results are based on data from public social media postings. Where their data are comprised of those at imminent risk, ours reflect users' behavior months prior to an attempt. Both cases, however, highlight the important role that language and linguistic analysis can play in understanding mental health.

These results suggest that an automated screening protocol using this sort of technology would be able to detect some users at risk. In many senses, that is the easiest part of meaningful impact on the suicide rates. This technology would, of course, represent a small portion of a larger system designed to address mental health issues. Most of that system is human in nature—friends, caregivers, clinicians, and the patients themselves. Although technology enabled by algorithms such as the one reported here may be able to identify people at risk, it will only be effective at reducing the suicide rates if (1) the technology facilitates the deployment of other, existing mental health resources, and (2) the system of care (inside or outside the formal health care system) functions to divert the person from risk.

Potential sources of bias and limitations

There are a few ways in which the findings reported here may fail to generalize to all segments of the population.

First, these data are predominantly derived from females aged 18 to 24. Although our results indicate that this approach works reasonably well for males of a similar age (and anecdotal evidence in support of those with non-binary gender), there is not sufficient evidence to assess the efficacy for people from other age groups. One could reasonably expect that their relationship to suicide is meaningfully different from this

demographic, and thus might not be as easily detected through approaches like this. To mitigate, this may require more specialized modeling for those demographics specifically, for example. Furthermore, we did not explicitly examine race or ethnicity in this work, though prior research indicates that mental illness presentation may be more reflective of deviations from cultural norms in functioning, as opposed to a consistent set of symptoms per disorder that holds cross-culturally.[52] Moreover, the strength of the stigma that surrounds discussion of suicide or other forms of psychological distress varies across communities,[53] which may affect the feasibility of using social media as a window onto mental health for certain types of users. It is also worth pointing out that, despite the seeming pervasiveness of social media, *not everyone* uses platforms like Twitter and Facebook. To the extent that there are systematic differences between those who use social media and those who do not, the findings described here simply cannot generalize to some segments of the population. Finally, almost all users in this analysis *survived* through their suicide attempt, at least long enough to post about it in social media. Although they may have died of a subsequent attempt, this may not be accurately reflective of the people who would go on to die by suicide or those who would die by suicide on their first attempt.

Ethical Implications

The technological results presented here demonstrate the feasibility of using automated machine learning techniques to identify people at risk for suicide, and with sufficient accuracy that it is worth examining how this might improve existing clinical systems of care. Essentially, the research presented here demonstrates that suicide screening at scale is possible, but we have not yet answered the question of how it should be used.

The technology described here raises a significant cultural question about the trade-off between privacy and prevention. These algorithms can provide early warnings in life-threatening situations—potentially a piece of technology that could enable saving lives. There are a few ways this might be integrated into systems of care that prevent loss of life:

1. In the health care system with existing patients, augmenting the capabilities of PCPs or other health care professionals to screen for suicide risk: Here, the patient agrees to be assessed for risk, authorizes the analysis of the data, and authorizes the health care provider to see the results.
2. Outside the health care system, empowering a person's support network (eg, friends or family members who are not health care professionals) to know when a person is in danger: Here, the patient authorizes the analysis of their data and authorizes their support networks to see results and alerts about their estimated risk.
3. As a screening tool to identify those at risk in the general population (and likely not receiving care): Here, public data can be screened and analyzed for risk, proactively identifying users outside the health care system.

All of these systems are technically feasible to build, but there are significant ethical considerations that should be discussed publicly prior to their implementation. Here, we consider some of the central concepts to the design of such screening systems, examine prior art in this space, examine analogous systems in other sectors, and highlight some points for consideration. We will not unilaterally claim an answer to these questions, but provide discussion of relevant analogous systems, and a framework for creating and assessing the viability of a path forward.

Opt-in versus opt-out

There are many considerations in how a system using this technology might be ethically implemented, but the most central theme is whether data collection is *opt-in* or *opt-out*. An *opt-in* system is where a person takes a conscious action to be part of the analysis and is generally in line with the concept of informed consent. Prior to any analysis, a person is informed about what the system does, what the likely outcomes are, and what actions (if any) might be taken as part of the process. This provides the person sufficient information to make an informed decision as to whether or not they wish to partake in the study or screening.

An opt-in system has the benefit that the users are willingly participating, and thus, any analysis or action is necessarily *not* a violation of a person's privacy. Similarly, this means that any opt-in system has a considerably limited reach. Any screening on an opt-in basis will necessarily only find the people at risk who *are already engaging with the system*. In scenario (1), this would be people who are engaging with the health care system for some sort of treatment. In scenario (2), this would constitute anyone who has found this system through any means (eg, web search). This does *not* cover scenario (3), because the whole population has not been made aware, informed, and agreed to analysis. Thus, opt-in systems are extremely unlikely to reach all of the users who are at risk.

In contrast, an opt-out system is one in which the user may not even know they are being analyzed or might be intervened with. Users in such systems may "opt out" by taking an explicit action (ie, asking to be removed). A screening system operating in an opt-out manner could examine all the public data from all the users in the world and identify those at risk, even if they have no connection to a health care system. Case (3) is an example of an opt-out system, where users have the ability to remove themselves from consideration from screening, but unless they have done so are being analyzed. For a screening system to reach and assess those not engaging in care or web searches related to the behavior in question, some sort of opt-out system is the only way to reach them.

Samaritans' radar

The most prominent parallel to the use of this technology for screening is the Samaritans' Radar App, briefly used in 2014.[54] The Samaritans are a well-known and well-respected suicide prevention organization, founded in 1952. They launched an app that purported to alert users when one of their friends exhibited signs of suicide risk. Unfortunately, a few design choices garnered significant public backlash, and the app was decommissioned shortly after its launch. We will describe the core functionality of the Samaritans' Radar and the design choices that, we suspect, were the root cause of public displeasure, then suggest methods for mitigating those concerns. Ultimately, we believe that any system architecture or process implementing automated screening tools of this kind should be guided and deeply informed by the advice of the lived experience community (those who have previously attempted suicide or have lost loved ones to suicide) as well as clinical experts involved in the daily provision of care to this population as well as to those who are at risk of suicide.

Samaritans' Radar functioned roughly as follows: a user could install the app and provide the app access to their Twitter account. The app would then continuously analyze the posts of Twitter users the app user follows. If the app detected phrases that seemed to indicate suicide risk (eg, "I just want to kill myself") in the posts of any of the app user's followed accounts, it would alert the app user, so the app user could reach out and offer support. To fully understand the situation, one must also realize that any user can "follow" any other user on Twitter and thus have access to their posts (unless certain non-default privacy settings were activated to make their account "private").

The primary concern from the public was a variant of the following nightmare scenario: there was a user who downloaded and installed the app with nefarious intent. This user could follow arbitrary users (due to Twitter's following system as above) and get Samaritans' Radar to tell the nefarious app user when someone they followed was at risk. The nefarious app user could then take advantage of that situation and (perhaps) encourage them to take their life. Although this may be unlikely, it is worth taking seriously. Two underlying problems are (1) that the user being analyzed does not know that they are being analyzed (this is at best an *opt-out* scenario) and (2) anyone could have access to information indicating people at risk (without their consent or knowledge, given the previous point).

In contrast, the scenarios we described above, (1) and (2), provide some monitoring aspects similar to Samaritans' Radar but require the user being analyzed to *opt in*. Scenario (3) does not require an opt-in but could be restricted to providing this information only to a special class of users (eg, licensed health care staff or individuals certified in some manner). In all cases, the technology can be deployed in a way that defeats the nefarious actor above.

Analogues in advertising and marketing

Perhaps the closest analogue to the sorts of systems described by scenario (3) is found in the world of marketing and advertising. The companies in this space are able to infer latent attributes about users given their online behavior, and the companies' algorithms are designed to use these inferences to suggest products or services that the person may want. The types of personal data that the technology under consideration in this article depends on are a subset of those available to advertisers, and many of the analysis techniques are similar. Conceptually, wide-scale screening is similar in technological function to ad targeting; however, instead of steering a user's behavior toward making a purchase, the screening would be used to steer a user's behavior away from an adverse health event (in this case, a suicide attempt).

Although the first thought of how these interventions might be done is through a pop-up ad to a person indicating they should seek medical care, this is decidedly *not* what we are advocating. That is a strategy unlikely to be effective and has potential to induce stress in the viewer of the ad. If we assume, however, that effective interventions can be developed and deployed within the existing advertising framework (eg, "nudges" a la Leonard [2008][55]), then this may provide a reasonable framework for examining scenario (3) and also a reason parallel for framing the ethical and privacy discussions (eg, How is this different from sending an ad to buy something? How is it different from sending an ad for a prescription medication?).

Facebook's suicide prevention protocol

Among large technology companies, Facebook in particular has been making waves[51] with its announcement of an automated suicide prevention ("proactive detection") algorithm.[56] Facebook cites "pattern recognition" models trained from community flags of posts containing language indicative of suicidal ideation. Two points are of particular interest here: Facebook does not share information about the underlying model or human vetting process, and there is no way for Facebook users to opt-out of this program (without leaving the Facebook platform). Given that Facebook's process involves contacting first responders (police, paramedics) and directing them toward the user presumed to be in danger, one would expect a high degree of confidence in basing decisions off the model, solely or in part. Even with human "Community Operations Team" members in the loop, false positives from an automated model almost necessarily lead to some false positives passed along by humans. With no details from Facebook about the process behind or performance of this automated detection algorithm (or the human moderators, for that matter), it is impossible for a third party to evaluate its utility. These facts, interestingly, have led to Facebook being unable to deploy proactive detection for suicide in the European Union (EU), due to the EU's Data Protection Directive and General Data Protection Regulations.[57]

Privacy versus lives saved

At the crux of this dilemma is the trade-off between a person's right to privacy and the widely agreed-upon moral imperative to act on information that may save lives. For a simple and direct yet difficult example decision, consider that a hallmark feature of risk is being withdrawn, and a classic and effective way of intervention is to reach out. In many cases, the very act of reaching out can be perceived as an invasion of privacy.

Concretely, if one could intervene and help a person heading toward crisis by invading their privacy, ought one do it? This critically depends on 3 premises: first, the ability to detect an individual at risk is sufficiently precise; second, that this detection is at odds with popular conceptions of privacy; and third, there is capability to effectively intervene and reduce the (risk of) harm. The research presented here addresses the first premise, the second premise is a question that we hope is debated and discussed openly, but ultimately without also solving the third premise, no amount of careful crafting and ethical analysis will provide any measurable impact on lives saved.

Anecdotally, we have found that people tend to think differently about privacy when presented in this light, and seek to find mitigating factors that would make them feel comfortable about being analyzed in an opt-out manner (ie, a system-like scenario (3)), by some trained processionals solely for the purpose of preventing suicides.

To reiterate, we do not feel it is within the scope of this article—or within the bailiwick of the authors—to unilaterally suggest a specific framework for implementing algorithms of the kind described here. However, we do feel it is our responsibility to note that "the cat is out of the bag," so to speak. That is, the widespread availability of personal digital data (which is, under current paradigms, the price that consumers collectively pay for largely free Internet content), coupled with the rapidly increasing sophistication of classification algorithms of the kind described here, suggest that the question is not *if* technologies like this will be deployed, but *how*. The best-case scenario, in our opinion, is one in which government, academia, advocates from the mental health community, and clinicians work in concert to assure that the right individuals are benefiting from such technology (ie, those who are at risk of harming themselves), and that the risks are proactively identified, discussed, and mitigated in as thoughtful and transparent a way as possible. Perhaps a silver lining of the legally questionable actions of groups such as Cambridge Analytica during and leading up to the 2016 presidential election in the United States will be a more sophisticated and vigorous public discourse surrounding the use and protection of personal digital data.

Ethical path forward

We feel that the broad question at the crux of this discussion is beyond our capability to answer: "What is the right trade-off between privacy and prevention?" However, we do see a path toward making the best possible case to the general public to enable wide-scale suicide screening (as in Parker et al[58]). As to whether one ought or ought not ultimately construct a system of that sort (and the eventual implications thereof), we lay a path out for how one might credibly test the efficacy and public's opinion on it, as we go.

Using opt-in analysis of risk, with the technology detailed here, is a step that we have heard little disagreement for both in our personal discussions and in any of the literature around this sort of work (eg, Mikal et al[59]). Building a system based on opt-in principles would allow the efficacy of these methods to be tested in the real world. Building this to better empower the health care system (case 1, above) would allow for such an empirical evaluation, and convincing evidence of the technology's efficacy. Demonstrating this efficacy and sharing it widely would allow for an informed discussion among clinical researchers and healthy policy stakeholders. If, in turn, they were convinced that it is of sufficient utility for a more widespread or opt-out-style monitoring approach, they would then have the data in hand to demonstrate efficacy and the human systems to put in place to adequately handle the increase in demand that such a system would create.

Conclusion

We have demonstrated state-of-the-art results for the detection of people at risk for suicide through the automatic examination of the language posted on social media. These results from what we consider to be a foundational piece of a new kind of screening system, often discussed in the crisis prevention community, but not yet implemented. These machine learning algorithms are of sufficiently high accuracy to be fruitfully used in an envisioned screening system, but the remaining parts of the system are not yet ready for implementation. We examined the ethical and privacy concerns around the use of these algorithms for screening and monitoring, concluding that there are novel ways to consider using information from these algorithms to aid intervention, but the general public has voiced opposition to related approaches. Although the design of an intervention system powered by algorithmic screening is technically possible, the cultural implications of implementation are far from settled. It is our hope that this serves as a forcing function to have the discourse about the ramifications on culture and society.

Acknowledgements

The authors cannot express sufficient gratitude for the people who donated their data at OurDataHelps.org and thus powered this study. The authors would also like to thank Dr Becky Inkster for her organization of the Digital Innovation in Mental Health workshop and this accompanying Special Issue. Finally, the authors would like to acknowledge the role that the Suicide Prevention and Social Media (#SPSM) community played in the refinement of these ideas.

Author Contributions

Conceived and designed the experiments: GC, RL, PC, and AF. Analyzed the data: GC, RL, PC, and AF. Wrote the paper: GC, RL, PC, and AF.

REFERENCES

1. World Health Organization. *Mental health action plan 2013–2020*. Geneva, Switzerland: World Health Organization; 2013.
2. Curtin SC, Warner M, Hedegaard H. Increase in suicide in the United States, 1999-2014. *NCHS Data Brief*. 2016;241:1–8.
3. Sullivan E, Annest JL, Luo F, Simon T, Dahlberg L. Suicide among adults aged 35–64 years—United States, 1999–2010. *Center for Disease Control and Prevention (Morbidity and Mortality Weekly Report)*, 2013. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6217a1.htm.
4. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull*. 2016;143:187–232.
5. Beck AT, Kovacs M, Weissman A. Assessment of suicidal intention: the scale for suicide ideation. *J Consult Clin Psychol*. 1979;47:343–352.
6. Schulberg HC, Bruce ML, Lee PW, Williams JW, Dietrich AJ. Preventing suicide in primary care patients: the primary care physician's role. *Gen Hosp Psychiat*. 2004;26:337–345.
7. Boudreaux ED, Jaques ML, Brady KM, Matson A, Allen MH. The patient safety screener: validation of a brief suicide risk screener for emergency department settings. *Arch Suicide Res*. 2015;19:151–160.
8. Gnambs T, Kaspar K. Disclosure of sensitive behaviors across self-administered survey modes: a meta-analysis. *Behav Res Methods*. 2015;47:1237–1259.
9. Ahmedani BK, Stewart C, Simon GE, et al. Racial/ethnic differences in health care visits made before suicide attempt across the United States. *Med Care*. 2015;53:430–435.
10. Gaynes BN, West SL, Ford CA, Frame P, Klein J, Lohr KN. Screening for suicide risk in adults: a summary of the evidence for the U.S. Preventive Services Task Force. *Ann Intern Med*. 2004;140:822–835.
11. Tondo L, Albert MJ, Baldessarini RJ. Suicide rates in relation to health care access in the United States: an ecological study. *J Clin Psychiat*. 2006;67:517–523.
12. Drug Abuse Warning Network. 2011 selected tables of national estimates of drug-related emergency department visits. Technical report, Center for Behavioral Health Statistics and Quality, SAMHSA, 2013.
13. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*. 2016;41:1691–1696.
14. Wang R, Chen F, Chen Z, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. Paper presented at: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 13-17, 2014; Seattle, WA:3–14. New York, NY: ACM.
15. Coppersmith G, Hilland C, Frieder O, Leary R. Scalable mental health analysis in the clinical whitespace via natural language processing. Paper presented at: 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); February 16-19, 2017; Orlando, FL:393–396. New York, NY: IEEE.
16. Padrez KA, Ungar L, Schwartz HA, et al. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department [published online ahead of print October 13, 2015]. *BMJ Qual Safety*. doi:10.1136/bmjqs-2015-004489.
17. Chung C, Pennebaker J. The psychological functions of function words. In: Fiedler K, ed. *Frontiers of Social Psychology: Social Communication*. New York, NY: Psychology Press, 2007:343–359.
18. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. Paper presented at: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM); July 8–11, 2013; Cambridge, MA.
19. Coppersmith G, Harman C, Dredze M. Measuring post traumatic stress disorder in Twitter. Paper presented at: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM); June 1–4, 2014; Ann Arbor, MI.
20. Cohan A, Young S, Goharian N. Triaging mental health forum posts. Paper presented at: Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (Vol. 16); June 16, 2016; San Diego, CA.
21. Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M. CLPsych 2015 shared task: depression and PTSD on Twitter. Paper presented at: Proceedings of the Shared Task for the NAACL Workshop on Computational Linguistics and Clinical Psychology; June 5, 2015; Denver, CO.
22. Resnik P, Armstrong W, Claudino L, Nguyen T, Nguyen VA, Boyd-Graber J. The University of Maryland CLPsych 2015 shared task system. Paper presented at: 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); June 5, 2015; Denver, CO.
23. Preotiuc-Pietro D, Sap M, Schwartz HAS, Ungar L. Mental illness detection at the World Well-Being Project for the CLPsych 2015 shared task. Paper presented at: 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); June 5, 2015; Denver, CO.
24. Pedersen T. Screening Twitter users for depression and PTSD with lexical decision lists. Paper presented at: 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); June 5, 2015; Denver, CO.
25. Yates A, Cohan A, Goharian N. Depression and self-harm risk assessment in online forums. https://aclweb.org/anthology/D17-1322, 2017.
26. Mitchell M, Hollingshead K, Coppersmith G. Quantifying the language of schizophrenia in social media. Paper presented at: 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); June 5, 2015; Denver, CO.
27. Walker M, Thornton L, De Choudhury M, et al. Facebook use and disordered eating in college-aged women. *J Adolesc Health*. 2015;57:157–163.
28. Chancellor S, Mitra T, De Choudhury M. Recovery amid pro-anorexia: analysis of recovery in social media. In: Kaye J, ed. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY: ACM; 2016:2111–2123.
29. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. Paper presented at: Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology; June 27, 2014; Baltimore, MD.
30. Coppersmith G, Leary R, Whyne E, Wood T. Quantifying suicidal ideation via language usage on social media. Paper presented at: Joint Statistics Meetings Proceedings, Statistical Computing Section (JSM); August 2015; Seattle, WA.
31. Kumar M, Dredze M, Coppersmith G, De Choudhury M. Detecting changes in suicide content manifested in social media following celebrity suicides. Paper presented at: Proceedings of the 26th ACM conference on Hypertext & Social Media; September 1–4, 2015; Guzelyurt, Northern Cyprus. New York, NY: ACM.
32. O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting suicidality on Twitter. *Internet Interven*. 2015;2:183–188.
33. Wood A, Shiffman J, Leary R, Coppersmith G. Discovering shifts to suicidal ideation from mental health content in social media. Paper presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; May 7–12, 2016; San Jose, CA. New York, NY: ACM.
34. Kiciman E, Kumar M, Coppersmith G, Dredze M, De Choudhury M. Discovering shifts to suicidal ideation from mental health content in social media. Paper presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; May 7–12, 2016; San Jose, CA. New York, NY: ACM.
35. Bryan CJ, Butner JE, Sinclair S, Bryan ABO, Hesse CM, Rose AE. Predictors of emerging suicide death among military personnel on social media networks [published online ahead of print July 28, 2017]. *Suicide Life Threat Behav*. doi:10.1111/sltb.12370.
36. Coppersmith G, Ngo K, Leary R, Wood T. Exploratory data analysis of social media prior to a suicide attempt. Paper presented at: 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); June 16, 2016; San Diego, CA.
37. Dinakar K, Chen J, Lieberman H, Picard R, Filbin R. Mixed-initiative real-time topic modeling & visualization for crisis counseling. Paper presented at: Proceedings of the 20th International Conference on Intelligent User Interfaces; March 29-April 1, 2015; Atlanta, GA:417–426. New York, NY: ACM.
38. Pestian JP, Sorter M, Connolly B, et al. A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial. *Suicide Life Threat Behav*. 2017;47:112–121.
39. Coppersmith G, Dredze M, Harman C, Hollingshead K. From ADHD to SAD: analyzing the language of mental health on Twitter through self-reported diagnoses. Paper presented at: 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); June 5, 2015; Denver, CO.
40. Loveys K, Crutchley P, Wyatt E, Coppersmith G. Small but mighty: affective micropatterns for quantifying mental health from social media language. Paper presented at: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality; August 3, 2017; Vancouver, BC, Canada:85–95. New York, NY: ACL.

41. Benton A, Coppersmith G, Dredze M. Ethical research protocols for social media health research. Paper presented at: Proceedings of the First Workshop on Ethics in Natural Language Processing; April 4, 2017; Valencia, Spain:94–102. New York, NY: ACL.
42. Sap M, Park G, Eichstaedt JC, et al. Developing age and gender predictive lexica over social media. Paper presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25–29, 2014; Doha, Qatar:1146–1151. New York, NY: ACL.
43. Kim Y, Jernite Y, Sontag D, Rush AM. Character-aware neural language models, 2015. http://arxiv.org/abs/1508.06615.
44. Kim Y. Convolutional neural networks for sentence classification, 2014. http://arxiv.org/abs/1408.5882.
45. Yang Z, Yang D, Dyer C, He X, Smola AJ, Hovy EH. Hierarchical attention networks for document classification. Paper presented at: HLT-NAACL; June 12–17, 2016; San Diego, CA.
46. Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. Paper presented at: Empirical Methods in Natural Language Processing (EMNLP); October 25–29, 2014; Doha, Qatar:1532–1543. New York, NY: ACL.
47. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate, 2014. https://arxiv.org/abs/1409.0473.
48. Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. Paper presented at: International Conference on Machine Learning; July 6-11, 2015; Lille, France:2048–2057.
49. Nock MK, Borges G, Bromet EJ, et al. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *Br J Psychiat*. 2008;192:98–105.
50. Kann L, Kinchen S, Shanklin SL, et al. Youth risk behavior surveillance—United States, 2013. *MMWR Suppl*. 2014;63:1–168.
51. Reardon S. AI algorithms to prevent suicide gain traction, 2017. http://www.nature.com/articles/d41586-017-08307-0.
52. Chentsova-Dutton YE, Chu JP, Tsai JL, Rottenberg J, Gross JJ, Gotlib IH. Depression and emotional reactivity: variation among Asian Americans of East Asian descent and European Americans. *J Abnorm Psychol*. 2007;116:776–785.
53. De Choudhury M, Sharma SS, Logar T, Eekhout W, Nielsen RC. Gender and cross-cultural differences in social media disclosures of mental illness. Paper presented at: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17); February 25-March 1, 2017; Portland, OR:353–369. New York, NY: ACM.
54. Hsin H, Torous J, Roberts L. An adjuvant role for mobile health in psychiatry. *JAMA Psychiatry*. 2016;73:103–104.
55. Thaler TH, Sunstein CR. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. London, England: Penguin Books; 2008.
56. Facebook. Getting our community help in real time (press release), 2017. https://newsroom.fb.com/news/2017/11/getting-our-community-help-in-real-time/.
57. Murphy M. EU data laws block Facebook's suicide prevention tool. *The Telegraph*, 2017. https://www.telegraph.co.uk/technology/2017/11/28/eu-data-laws-block-facebooks-suicide-prevention-tool/.
58. Parker J, Wei Y, Yates A, Frieder O, Goharian N. A framework for detecting public health trends with Twitter. Paper presented at: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '13); August 25–28, 2013; Niagara, ON, Canada:556–563. New York, NY: ACM.
59. Mikal J, Hurst S, Conway M. Ethical issues in using twitter for population-level depression monitoring: a qualitative study. *BMC Med Ethics*. 2016;17:22.