Prediction of depression and suicidality from social media activity using deep neural networks

Angelica Chen '17

Advisor: Professor H. Sebastian Seung

Submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Arts

Department of Computer Science

Princeton University

May 5, 2017

I pledge my honour that this paper represents my own work in accordance with University regulations.

Angelica Chen '17

Abstract

It's estimated that every 13 minutes, at least one American dies by suicide in the United States, making suicide the 2nd leading cause of death in the 15-34 age group (CDC, 2015). Suicide intervention is now one of the top priorities for public health agencies across the world, but unfortunately our screening efforts are still inadequate. In this study we propose a data-driven artificial intelligence approach towards more accurate and higher coverage screening for both suicidal ideation and depression. We trained deep neural networks on Facebook posts from 158 human subjects to distinguish depressed from non-depressed individuals and to distinguish suicidal from non-suicidal individuals, using their Beck Depression Inventory (BDI) and Columbia-Suicide Severity Rating Scale (C-SSRS) scores as the ground-truth labels. We achieved state-of-the-art accuracy, precision, and recall rates of 96%, 97%, and 94% respectively on the depression prediction task and accuracy, precision, and recall rates of 92%, 98%, and 83% respectively on the suicidality prediction task. In addition, the deep neural networks outperformed both a trained human psychiatrist as well as other standard machine learning-based classifiers on the same tasks. These results demonstrate that it is possible to achieve more accurate and timely screening for depression and suicidal ideation via deep learning, without the need to first seek out a physician or mental healthcare professional. This significantly lessens the time, effort, and money required to seek help, which is particularly important for low-income and marginalized communities with low access to adequate healthcare.

Acknowledgements

There are many people to whom I am indebted to for supporting me during both the writing of this thesis and my Princeton career as a whole. Words cannot fully express the extent of my gratitude to you, but I hope that this is an adequate start.

To Professor H. Sebastian Seung, thank you for giving up so much of your valuable time and energy to advise me on this thesis. I've learned so much from you, not just about deep learning but also about how to become a better researcher and scientist. You have taught me the importance of critically questioning all my assumptions and of constantly using failure to inform the research process. Thank you for always patiently answering my clueless questions, despite having many more years of experience than I!

To Professor Barbara Engelhardt, thank you for teaching me so much about natural language processing and for helping me brainstorm so many ideas for this thesis. Prior to this project I had absolutely no background in the field, and I have learned so much from your suggestions.

To Dr. Chin, Director of Princeton Counseling and Psychological Services (CPS), thank you for making time in your extremely busy schedule to contribute to this research and to teach me about the clinical side of suicidal intervention. Your consistently warm encouragement and helpful advice made a world of difference whenever I was feeling discouraged.

To the 233 incredibly kind strangers who participated in this study, thank you for being willing to donate your data and time towards answering 26 very awkward and uncomfortable questions. Without your generosity, this research would not have been possible - I hope you know what a difference you've made.

To my lovely roommates Mihaela Curmei, Catherine Niu, and Katie Awh, thank you for constantly being sources of brightness in my life. You've taught me how to balance work with life and how to always have fun, no matter the number of papers

or problem set deadlines approaching. For the nights spent learning German from strangers in Miami nightclubs, the days spent juggling schoolwork with trashy reality TV, and the moments spent doubled over in laughter - thank you.

To my dear friends Sonia Hashim, Jennifer Bu, Jonathan Tang, Albert Ho, Bo Moon, Kalina Petrova, Joyce Lee, Stan Palasek, Jonathan Liebman, Nicole Wang, Tony Lu, Zach Kendrick, Ben Eisner, Frank Jiang, Iris Rukshin - thank you for the never-ending encouragement and support you provided me with during the writing of this thesis. Whether it was sharing my thesis survey on social media, providing late-night snacks, sending me funny Snapchats, or running a half-marathon with me to deal with the stress (!), you were always there for me in my times of greatest need.

To my thesis fairy Shiye Su, thank you for providing chocolate and smiles during the times when they were needed the most, and for reminding a jaded senior of how to be passionate again about Princeton.

To the Princeton Office of the Dean of the College, thank you for providing the generous funding necessary for this research.

To the Princeton University Department of Computer Science and the Princeton Neuroscience Institute, thank you for providing the valuable teaching, education, and computational resources necessary for this research.

And last but certainly not least, to my family - Mom, Dad, Koert, and Medlyn - thank you for always supporting my dreams, whether they were to become a mathematician or a fashion designer, and for accepting me just as I am, both at my best and at my worst. You are the reason I am who I am today.

Contents

Chapter 1

Introduction

It's estimated that every 13 minutes, at least one American dies by suicide [11], causing over 42,000 deaths in the US in 2014 [13] and making it the 2nd leading cause of death in those between the ages of 10 and 34 [7]. This tragedy affects not just the individuals it kills, but also their loved ones and society as a whole. For every suicide, at least six close relatives or friends are bereaved and are exposed to a significantly higher risk of depression and suicide themselves [25, 27]. Furthermore, attempted and completed suicides account for an estimated $93.5 billion in lost wages and medical expenses each year in the US alone [119].

One of the main contributing factors towards the development of suicidal behaviors is untreated mental illness, in particular major depressive disorder (MDD). MDD is a mental illness characterized by persistent melancholy mood, abnormally low self-esteem, and disruptions in sleep and appetite [6]. Left untreated, it can ruin relationships and productivity at work, significantly diminishing one's quality of life. It's estimated that over 60% of suicide victims suffered from MDD prior to death [18, 55] and up to 15% of individuals suffering from clinical depression eventually die by suicide [5].

Hence, depression and suicide clearly exert an undue burden on society that de-

serves more of our efforts to combat. Yet despite growing public awareness of mental health issues and increased spending on mental health services [115], suicide rates are at an all-time high, increasing 24% between 1999 and 2014 [46]. A large contributor to this problem is insufficient mental health screening - although 64% of people who attempt suicide visit a doctor in the month before and 38% in the week prior [17] to the suicide attempt, research indicates that it is rare for primary care physicians to successfully detect suicidal ideation in patients, even in those who have been formally diagnosed with clinical depression [96].

To try to improve both the frequency and success rates of suicide screening, we aim to develop an automated deep learning-based screening tool capable of detecting moderate to severe depression and suicidal ideation from an individual's social media activity. We believe that such a tool can not only be effective but also avoid many of the drawbacks inherent to traditional clinical screening tools, such as the absence of timely and contextual data, lack of statistical verification of predictive value, and their dependence upon patient openness and honesty.

With this goal in mind, the rest of the paper is organized as follows: in Section 1.0.1 we discuss the current WHO-sanctioned public health model of suicide prevention and what has already been achieved in attempting to solve this problem; in Section 1.0.2 we discuss why the most significant oversight in the current suicide prevention pipeline is inadequate screening; in Section 1.0.3 we discuss why this problem may be better approached with the assistance of deep learning applied on social media; in Section 1.0.4 we summarize our concrete goals for this study; in Section 2 we summarize the related academic work on this topic; followed by our methodology (Section 3, results (Section 4), and discussion (Section 5).

1.0.1 The Public Health Model of Suicide Prevention

In order to better understand the problem at hand and to be able to offer an improved solution, it is important to be knowledgeable about the current state of suicide prevention and the pre-existing public health suicide intervention measures in place.

Despite the severity of untreated depression and suicidal ideation, suicide prevention remains one of the most dire and insufficiently addressed problems in public health systems. This is particularly surprising given the fact that significant advancements have been made in treating both conditions with psychotropic medications, talk therapy, and other extensively researched programs. However, such interventions require significant funding and infrastructure and are not always made a priority in developing countries with more meager resources. Current epidemiological surveys indicate that while suicide rates have decreased in certain well-off, developed countries like Finland and England, suicide rates have actually increased in developing countries such as Brazil, China and India [80]. For instance, between 1980 and 2000, Brazil's suicide rates increased by 21% [95], whereas the US's suicide rates decreased by 15% during the same time period [117].

Not only do these developing countries frequently have competing priorities such as alleviating poverty and sustaining steep population growth, but they are often prone to common misconceptions about depression and suicide that lead to grave consequences such as the downplaying or even illegalization of suicide. It is common for a 'reductionist' model of suicide to arise in the absence of widespread mental health education, which reduces the contributing factors of suicide to its most immediate causes, such as an argument with family or a recent career setback [80]. It fails to take into account more distal factors, such as underlying depression or other mental instability, and merely frames suicide as the impulsive result of a character flaw. As a result, it becomes all too easy for government officials and policymakers to lower the priority of establishing nationwide mental health programs, and even

provides rationale for criminalizing suicide. In fact, an estimated 25 countries currently have laws against attempted suicide [117]. However, criminalizing suicide often deters individuals from seeking help for mental health issues and/or reporting suicide attempts. Legalizing suicide, on the other hand, encourages open conversation about the mental health issues that contribute to suicide and has not been shown in any studies to increase the rate of suicide, to the best of our knowledge. In fact, suicide rates typically decline after legalization [75, 86, 66, 108].

Furthermore, research indicates that the elevated suicide rates in developing countries may be caused not just by a dearth of national suicide intervention programs and mental health education, but also by the strong relationship between socioeconomic status, level of education, and suicide risk. Studies of both individual-level [57, 50, 88, 87, 89] and population-level [15, 89, 120, 127] suicide correlates have shown that poor education, unemployment, and low income are all strongly correlated with increased depression and suicide risk. Even subtler factors associated with these socioeconomic attributes, such as if an individual does not have access to a car [87] or is the offspring of a mother with low academic aspirations for her children [88], are also directly linked to increased risk.

It is thus evident that developing countries possess an even more urgent need for national suicide prevention strategies than more developed countries, but they often lack the health infrastructure and investment to implement such policies. To aid these countries in developing their own suicide prevention strategies, the World Health Organization (WHO) has developed a formal framework for effective intervention, known as the public health model of suicide prevention. Below, we summarize each of these stages and briefly discuss what has already been accomplished in each stage.

Surveillance

This initial stage of the public health model involves developing surveillance systems for systematically recording vital outcome-specific data on both attempted and completed suicides in real-time, coupled with the prompt dissemination of the aforementioned data to local policymakers and public health agencies. Such programs are critical for informing communities on how to deal with the problem, particularly in a local context. Since both the contributing and the protective factors of depression and suicide vary widely across different cultures and demographics, local up-to-date data is necessary for the customization of suicide intervention approaches towards specific populations.

Furthermore, this data is also crucial from an a posteriori perspective for assessing the effectiveness of current interventions and determining whether or what modifications need to be made. For instance, some interventions may be more or less effective in specific areas or groups, but this difference would not observed or remedied without the collection of high-quality, granular data. Such data should include, at minimum, the means with which an individual attempted suicide, whether the attempt led to death, the individual's demographic and health background, any means of intervention experienced, and public health resources utilized.

However, successful suicide surveillance systems are both uncommon and difficult to implement. In fact, the WHO estimates that only 25 countries currently have any type of national suicide surveillance system in place [117]. Common difficulties include the stigmatization or criminalization of suicide in many countries (and the significant underreporting thereof), lack of sufficient infrastructure, and a dearth of both human and financial investment. Indeed, an adequate surveillance system requires numerous components, including data collection technologies, death registration/certification, and attempt reporting. Although the technology for efficient and granular data collection exists, it is often not utilized properly, and it may be used in inconsistent

ways across locales. Lack of or mis-communication between geographic regions may also result in different types of data being recorded, or ambiguities in the definitions of data classifications. (For instance, should all self-directed harm be recorded as suicide attempts? Does suicidal ideation without follow-through count?) Coroners across different regions may also classify deaths in different ways when investigating equivocal cases, and some regions might not even possess sufficient funds for thorough investigations by the medical examiner's office. Last but not least, suicide attempts are estimated to be under-reported by as much as 17.5% for females and 12% for males [122], indicating that many suicide surveillance systems do not report the true extent of the problem.

Identify risk & protective factors

One of the biggest difficulties in understanding suicidal behavior is identifying its causes. Usually, no single event or contributing factor can be held responsible, as suicide is more often the culmination of multiple causal pathways in conjunction with each other. By attempting to understand these many risk factors, we are better equipped to develop well-informed intervention and screening procedures. In addition, it is also important to identify factors that are protective against suicide in order to develop intervention strategies that not only minimize risk but also enhance resilience.

Risk factors are often categorized into three types: individual, socio-cultural, and situational [117]. Some examples of each of these categories include:

- *Individual*: History of mental illness, abuse, or trauma, chronic illness, and neurobiological imbalance
- *Socio-cultural*: Influences from religious or cultural beliefs about suicide, glamorization of suicide in the media, social contagion of suicide, and deterrents to accessing sufficient mental health care (such as high costs or stigma)

- *Situational*: Recent career setback, death of a loved one, social isolation, financial difficulties, ready access to means of suicide

On the other hand, protective factors might include:

- A strong social and moral support system
- Effective problem-solving and conflict resolution skills
- Adequate access to healthcare resources
- Cultural or religious beliefs that neither stigmatize nor glamorize suicidal behaviors
- Restricted access to lethal means

Both these lists are non-exhaustive; extensive research has already been conducted [104, 31, 109] into the various risk and protective factors of suicidal behaviors in a variety of different contexts, many of which are not listed here. What is more unclear is how to make use of this information - although we can recognize what characteristics make an individual particularly vulnerable to or resilient against suicidal behaviors, it is still very difficult to acquire an accurate picture of a specific individual's risk and protective factors and to determine whether or not they will actually develop suicidal behaviors, given their risk assessment.

Develop & evaluate interventions

The next part of the public health model of suicide intervention involves the actual design, implementation, and evaluation of suicide intervention programs. The WHO splits interventions into three types - universal, selective, and indicated.

Universal interventions target the entire population as a whole and focus on developing better mental healthcare systems that promote overall health and reduce impediments to care. In the past, most of these interventions have targeted either

increasing access to mental healthcare resources or restricting access to lethal means of suicide. Concerning the former, much work remains to be done. It is estimated that individuals who complete suicide are 2.25 times more likely to have interacted with a primary care provider than a mental health specialist prior to their death [92]. In particular, the vast majority of high-risk adolescents have received insufficient or no treatment whatsoever [79], and fewer than 20% of those who eventually complete suicide see a mental health provider in the few months preceding their deaths [47, 8].

However, instituting universal policies or infrastructure that either facilitate access to or increase the quality of mental health resources are generally quite effective when implemented. In particular, studies indicate that lower suicide rates are associated with greater local access to outpatient mental healthcare services, after controlling for socioeconomic and demographic factors [35, 105]. Even just living nearby more crisis intervention and case management health services is correlated with significantly reduced risk of suicidal behavior [40]. In addition to adding new mental healthcare resources, improving the quality of pre-existing programs can also significantly improve patients' quality of life and reduce adverse outcomes. Asarnow *et al.*[71] analyzed the effects of increasing primary care providers' training in evidence-based treatments for depression (such as applying cognitive behavioral therapy (CBT) and prescribing antidepressants) and found that their patients experienced significantly reduced depression symptoms and suicidal ideation. In another large review of suicide prevention strategies performed by experts from 15 countries, researchers found that providing professional education in suicide risk evaluation and depression screening/treatment at the primary care level was one of the most effective methods of reducing completed suicide [93].

Such universal interventions need not be geographically proximate in order to be effective - in the present Internet era of ubiquitous smartphones and online chatrooms, digital means of intervention can also be quite successful. For example, the National

Suicide Prevention Lifeline is a free and confidential hotline that provides emotional support and crisis counseling 24 hours a day, 7 days a week [1]. In evaluations conducted by Columbia University's Research Foundation for Mental Hygiene, it was found that approximately 50% of callers to the lifeline utilized mental healthcare resources recommended to them by the lifeline staff after their call [60]. Furthermore, after the Lifeline began to provide Applied Suicide Intervention Skills Training (ASIST) to their counselors, callers became significantly less likely to feel depressed or overwhelmed, and were generally much more hopeful after their calls to the Lifeline [58]. Although much work remains to be done in improving the effectiveness of remote interventions, other hotlines and textlines have reported similarly promising results. TextToday, the US's first text message crisis line, was found to significantly increase the likelihood that callers would seek professional help after the conversation, as compared to if they had never called at all [53]. In a study of California's network of suicide hotlines (a portion of which is part of the National Suicide Prevention Lifeline program), researchers reported that nearly half of callers experienced significant reductions in distress after the call, though the rest mostly experienced no change. Caller satisfaction was moderate, with an average rating of 3.4 on a 5-point scale, with 5 being the highest possible rating [107].

An alternative means of reducing suicide completion focuses instead on the practicalities of committing suicide. By restricting access to lethal means and frequently abused substances, policymakers are often able to successfully decrease the incidence rate of death by suicide, although such techniques do not solve the underlying problems. In particular, past legislation resulting in the restriction of firearms, pesticides, and barbiturates in China, Australia, India, Sweden, and Sri Lanka have resulted in observed decreases in suicide attempts [91, 30, 85, 121, 72, 100, 43, 97, 112, 37].

Increasingly, such universal interventions are shifting more towards mental health promotion via reducing societal stigma surrounding suicidal behavior and seeking

help; and educating the population about mental health, substance use disorders, and suicide. Although there exists less evidence directly linking mental health awareness campaigns to their concrete impact on health outcomes, such campaigns are generally thought to facilitate open discourse on the importance of promoting mental wellbeing and improving the efficacy of intervention programs.

A few key examples of evidence-based mental health promotion campaigns proven to directly reduce suicidal behaviors include the SEYLE (Saving and Empowering Young Lives in Europe) program and the NAD (Nuremberg Alliance against Depression) program. The SEYLE study was a multi-site longitudinal study that assessed the impacts of three types of educational interventions on the prevalence of suicidal behaviors in adolescents from 11 different European countries [130, 131]. The interventions consisted of training school teachers and faculty to recognize depressive and/or suicidal behaviors in at-risk pupils and to provide referrals whenever possible, a special mental health education program targeted at pupils, and a professional screening program also targeted at the pupils. The researchers observed a significant decrease in suicidal ideation, planning, and attempts after a 12-month followup and students even reported that they enjoyed the education portion of the program and felt less alone afterwards [129]. The NAD conducted a somewhat similar study focused on educating various members of the community in Nuremberg, Germany using a four-part program. The four parts consisted of: (1) training primary care providers to be more sensitive and on-the-lookout for suicidal behaviors, (2) to inform the local media about the Werther effect (suicide imitations after a single widely publicized or glamorized suicide), (3) training community pillars such as teachers, police, and religious leaders to be able to recognize warning signs and provide referrals when necessary, and (4) more directly supporting depressed individuals [65]. Not only was this program initially effective at reducing suicidal behaviors by 21.7% (as compared to a control group) during the year it was introduced, but its suicidality reduction

effects were observed and even magnified over a longer period of time, resulting in a 32.4% reduction the year after [63, 64].

The SEYLE and NAD campaigns are also examples of the second WHO category of suicide prevention interventions - selective interventions. Selective interventions focus on targeting specific high-risk groups that exhibit biological, psychological, socioeconomic, and/or environmental characteristics that increase their vulnerability to depression and suicidal behaviors, even though such individuals may not be in immediate danger. A key assumption is that developing custom-tailored intervention approaches for each of these groups will result in higher success than simply applying a more coarse-grained universal intervention approach. Specific groups of concern include individuals with an abusive or traumatic history, LGBTI (lesbian, gay, bisexual, transgender, and intersex) individuals, immigrants or refugees, the chronically ill, veterans or active members of the military, prisoners, and the homeless [117].

Selective interventions often involve specially training leaders in the community such as school staff, first responders, police officers, religious leaders, and social welfare workers. These individuals are known collectively as "gatekeepers" because they frequently act as confidants for members of the community and may be in a particularly apt position to identify individuals at high risk for depression or suicidal behaviors [93, 70]. Gatekeeper training programs focus on preparing the trainees to recognize at-risk individuals and to connect with them in a sympathetic, nonjudgmental way. Their primary goal is to get the individual to recognize the legitimacy of their mental health issues and to successfully convince him/her to see a mental healthcare professional [114]. Although more studies are needed in order to assess the long-term outcomes of gatekeeper training and its impact on the reduction of suicidal behaviors, initial studies of such training programs have shown promising improvements in gatekeeper knowledge, attitudes toward suicide and suicide prevention, and self-efficacy in school settings [126], workplaces [45], veteran groups [94],

and aboriginal communities [36].

Interventions can also be developed to be even more fine-grained, focusing instead on providing help at the individual level. Since everyone experiences suicidal ideation differently, it is important for individuals to be treated one-on-one by a qualified mental healthcare professional who can determine the unique mix of factors contributing to an individual's declining mental health. After this assessment is complete, the course of treatment must also be custom-tailored for the individual via trial-and-error. Typically, the most effective treatments are some combination of talk therapy and psychiatric medications such as selective serotonin reuptake inhibitors (SSRIs) or monoamine oxidase inhibitors (MAOIs) [101].

Implementation

Lastly, effective suicide interventions must be implementable at larger scales. Since this epidemic is a global problem and indiscriminately affects many different types of people all across the world, it is important that we be able to establish generalizable policies and interventions that are scalable in a cost- and time-effective manner. At the country level, this requires designing and committing to a national suicide intervention strategy, and allocating the necessary financial and human capital towards implementing such a strategy. Significant progress has already been made in this arena - an estimated 31% of the member countries of the International Association for Suicide Prevention (IASP) purportedly have some type of national strategy or action plan for suicide prevention [19].

Other campaigns that may be helpful in facilitating the scaling of suicide intervention programs include global public awareness campaigns for suicide prevention, such as the IASP-organized World Suicide Prevention Day (September 10 every year). This observance was instituted by the WHO as an opportunity to raise awareness internationally about the burden and preventability of suicide. It involves hosting events

and launching initiatives to train the public in various suicide prevention strategies, such as how to support a loved one affected by depression or suicidal ideation, and how to detect when an individual might be at high risk for suicidal behaviors [3].

1.0.2 The Leak in the Pipeline - Suicide Screening

Although many parts of the public health model of suicide prevention have already been well-researched and implemented, current progress has been significantly forestalled by the lack of effective detection of suicidal behaviors and ideation. Intervention will not be successful if we cannot accurately identify who needs our help the most, especially because depressed and/or suicidal individuals are generally less likely to openly discuss their feelings or reach out for help on their own. In fact, in a study conducted by Busch *et al.*[34], 80% of individuals who later completed suicide denied having suicidal thoughts in their last verbal communication.

By nature of the disease, depressives are often socially withdrawn, lethargic, and less motivated than their non-depressed peers [6]. Even if they recognize the need to seek help, they may have little desire to do so because they may feel ashamed or undeserving. In addition, there still exists significant societal stigma associated with seeking help for mental health issues, which negatively impacts many individuals' willingness to do so. A 2006 study conducted by the Centre for Mental Health Research at the Australian National University [21] discovered that 46% of people believed that others would think less of them for seeing a psychiatrist, psychologist, or counselor, and less than half of adults would seek help from a mental healthcare professional if they met the DSM-IV minimum criteria for major depression.

Yet other reasons exist for not seeking help for depression or suicidal ideation. Many individuals do not believe that anybody else could possibly help them, even those who are professionally trained to do so. Some also wish to handle their problems themselves and/or do not feel that their problems merit professional help. For

others, it is merely a logistical issue - getting the appropriate treatment would simply be too expensive or too inconvenient, and not worth the potential benefits. Another concerning reason is that some individuals are not aware that their distress is significant enough to constitute mental illness, or may even change their perception of mental illness in order to avoid seeking help [61, 125, 54].

But on the rare occasion when screening *is* successful, it is highly effective at reducing suicidal behavior and improving outcomes. According to a study conducted by Kapur *et al.*[76], if an individual with a history of self-harm or depression receives even a single mental health assessment in the ER, then they are significantly less likely to exhibit repeated suicide behavior - a risk which may have been as high as 40% in the short term. Another similar success story is that of the Air Force - ever since it began requiring annual suicide and mental health screens by specially trained psychiatrists, the suicide rate among Air Force Pilots dropped from 16 (per 100,000 individuals) to 9 [9]. Successful screening for depression can also be highly effective, as research indicates that treatment of depression is effective 60-80% of the time [10].

Yet despite its proven effectiveness, suicide screening is very rarely implemented and often not executed well when it is. The general consensus in the medical community is that depressed patients should be asked about suicidal ideation, but many general physicians have not received formal training in suicidal assessment [20]. Without such training, the physician may not be able to detect when a patient is downplaying their distress or to distinguish between varying levels of lethality of suicidal ideation. Research indicates that this is a surprisingly common problem - approximately 45% of Americans who die by suicide see a physician in the month preceding their suicide [16], but in the vast majority of these instances the physician does not detect the individual's suicidal thoughts [12].

Unfortunately, training physicians to successfully screen for suicidal ideation in patients is also an incredibly difficult problem. Efforts to facilitate and standardize

this task have been made via the invention of various clinical risk assessment scales such as the ReACT self-harm rule, suicide assessment scale (SUAS), and Karolinska interpersonal violence scale (KIVS), but very few of these scales were developed using empirical evidence [116]. There is also a dearth of research studies that statistically assess the predictive behavior of these tools. Among the few that exist, the studies' authors generally conclude that these scales have low predictive value (exhibiting positive predictive values no higher than 19% [116] and sometimes as low as $0.4-0.5\%$, in one particular study by Bolton [28]) and are oftentimes misused in clinical settings. Physicians often abridge the survey and neglect to ask all the required questions, significantly reducing the accuracy of the scale.

Given all of these challenges, it is clear that a depressed and/or suicidal individual encounters many barriers in the process of seeking help. Firstly, the individual must be self-aware enough to recognize the severity of their condition and the need to receive professional help. Then, he/she must find the motivation to seek help in spite of significant societal stigma and depression-induced lack of motivation. He/she must also possess either the necessary funds or health insurance to see a qualified health professional, as well as free time in their schedule. Once that individual has arrived in the doctor's office, he/she must be open and honest, and the doctor must be well-trained enough to accurately screen for and detect the level of severity of the patient's distress.

Taken together, this evidence suggests that our current healthcare system is severely underserving individuals suffering from depression or suicidal ideation. There exists a dire need for more effective screening tools with higher predictive value and lower barriers to use. In this study, we seek to solve this problem using a new and unconventional data-driven method that avoids many of the complications inherent to traditional clinical methods.

1.0.3 How Deep Learning and Social Media Data Can Help

Nowadays, we live our lives online. It's estimated that the average American spends nearly two hours per day on social media, totaling to an impressive 5 years and 4 months over the course of a lifetime. This is due in large part to the vast expansion of social media's role in our lives - what started merely as barebones online chatrooms in the AOL days of yore has now evolved into sophisticated news feeds, online photo sharing, public broadcasting, business advertising, and more. Its ubiquity has completely transformed the way we interact with our friends and family, increasing the frequency of casual remote communication and dramatically boosting the sizes of our social circles.

We invest not just our time into social media, but also large quantities of our personal data. There is perhaps no source of data about our personal everyday lives that is more abundant or rich. Online social networking sites such as Facebook and Twitter record everything from our demographic information and personal likes/dislikes to the exact amount of time that we spend looking at each individual post or advertisement. Not only are these data points much more dense than clinical data from doctors' visits, but they also encode contextual information about the subjects' interpersonal relationships, present wellbeing, and social network that may provide useful signals about mental health. In fact, a multitude of empirical evidence [22, 83, 42] suggests that both emotional state and mental health exhibit a social contagion effect, wherein more enduring moods such as depression and happiness can be spread through social networks via the creation and viewing of content of similar sentiment.

Plenitude and virulence aside, social media data is also useful for its ability to paint a picture of the direct link between mental illness and social media use. Although the use of social media generally decreases feelings of loneliness and enhances individuals' support networks [32], numerous studies have found that overuse (generally considered > 2 hours per day) is correlated with increased loneliness, social anxiety, and incidence

of depression [69, 26]. Whether social media overuse is a contributing factor to or simply a symptom of depression is unknown, but it is clear that analyzing social media usage patterns may be of assistance in screening for depression.

But in general extremely rich and unstructured data sets such as social media data can be as much of an inconvenience as they are a boon. Traditional machine learning classifiers often require extensive feature engineering - the manual identification and extraction of quantitative features known (or guessed) to be predictors of the desired label. Such feature engineering can be both time- and labor-intensive, and requires a much deeper intuitive understanding of the learning problem at hand. As we have already noted from our previous review of the current public health model of suicide intervention, identifying the risk factors of depression and suicidal ideation for use in accurately screening for these conditions is an enormously challenging task that modern psychiatric research is still working on. Deep learning is particularly helpful in such contexts because it replaces the task of feature engineering with architecture engineering. Rather than custom-designing features, the deep learning engineer focuses instead on experimenting with different numbers and types of hidden neural network layers, which is a much more generalizable task than feature engineering. It requires little to no specialty knowledge about the specific learning problem at hand and can be largely automated.

Furthermore, deep neural networks are capable of learning subtler features from the data that the human mind may not consciously perceive. In facial recognition, for example, the network may detect subtle wrinkles or feature ratios that the human eye may miss. Indeed, facial recognition is already a task that deep neural networks have surpassed human performance in [77]. The same might be true for predicting depression and suicidal ideation - deep neural networks may be able to detect subtler patterns in an individual's writing or social media usage that are not immediately obvious to humans.

Advanced feature extraction aside, deep neural networks are also significantly less prone to human biases than traditionally feature-engineered machine learning classifiers. Humans are the prisoners of their preconceptions, which may be reflected in the types of features used to train a statistical classifier. A neural network, on the other hand, learns features independently of human biases and may be able to exploit patterns that are counter-intuitive or unexpected.

This social media data-driven artificial intelligence approach towards developing screening tools for depression and suicidal ideation also has a number of other advantages over traditional screening methods. Typically, clinical screening is far more difficult in marginalized or "late-adopter" groups such as the unemployed, those involved with the courts and criminal justice system, and those who cannot afford regular doctors' visits [82]. However, making a Facebook account is easy and free, so a social media-based screening tool is far likelier to be used in marginalized communities than a clinical tool.

Also pertaining to coverage in different communities is the potential adaptability and generalizability of deep neural nets. Clinical tools are generally developed using research from a very specific and small group of people (usually middle-aged white men [82]), and are applied uniformly across all demographics. This is extremely problematic, given that suicidal behaviors vary widely across different countries and cultures. In fact, Leo *et al.*found that the average ratios between the lowest and highest suicide rates internationally were as extreme as 1 to 102.4 for men and 1 to 35.8 for women [49]. These differences may be attributed to a number of cultural factors - for example, both community attitudes about mental health and the availability of healthcare resources vary dramatically across the world. A comprehensive depression/suicide screening tool should attempt to take these cultural and demographic differences into account, changing its risk assessments based upon the local context. This is currently not implemented in standard clinical screening tools, but a neural

network can easily take this information into account if it is encoded in the input vectors.

In addition, behavioral research indicates that humans are often more open online than they are with their doctors, perhaps due to the impersonal and less confrontational nature of online communication. In a study conducted by Fein *et al.*, identification of adolescents with psychiatric problems increased by approximately 1.7 times when the adolescents were screened online instead of in-person by clinical staff using the same scale (the Behavioral Health Screening-Emergency Department (BHS-ED) system) [73]. Paperny *et al.*observed similar results when presenting patients with two versions of a questionnaire about high-risk psychosocial and health behaviors, one administered on a computer and the other by a physician [102]. The patients who took the digital version responded significantly more frequently about high-risk issues than the written questionnaire group, and 89% of all the participants in the study indicated that they preferred the digital version, even if the results were to be shared afterwards with their physician. Although few in number, the aforementioned studies that have assessed the effectiveness of online/digital screening in comparison with in-person clinical screening have overall concluded that digital screening is often just as, if not more, effective for identifying depression and suicidal ideation.

A further benefit of screening online is that such a screening program can directly link at-risk individuals to helpful online resources, such as the National Suicide Prevention Lifeline, Crisis Text Line, or other web applications for suicide prevention. These resources are not only extremely convenient and often free of charge, but have been shown in a large review of various suicide prevention web apps to be highly effective at reducing more immediately lethal thoughts related to self-harm and suicide [39]. This can be particularly helpful in the short term since it does not require the at-risk individual to have to find a nearby clinic or other mental healthcare facility and physically travel there, which requires far more energy than a depressed

individual might be willing to put in.

As such, a social media-based deep learning screening tool does not suffer many of the drawbacks that traditional clinical risk assessment scales do, such as low generalizability and vulnerability to patient dishonesty. It also offers a number of additional benefits, such as greater statistical power (due to a larger and richer dataset), ease of use, and immediate turnaround. To our knowledge, this is the first study to develop this type of depression/suicide screening tool and to statistically evaluate its predictive power.

1.0.4 Goals

With the aforementioned objectives and motivations in mind, the following are the specific goals of this study:

1. To gather a clinically-validated dataset consisting of individuals' social media data coupled with their psychiatrically assessed levels of depression and suicidality.

2. To train deep neural networks for two classification tasks: 1) to distinguish depressed and non-depressed individuals from their social media data (the depression prediction task), and 2) to distinguish individuals who have and have not experienced suicidal ideation sometime within the last six months from their social media data (the suicidality prediction task).

3. To compare the performance of these deep neural networks with that of other machine learning classifiers and a trained psychiatrist.

In achieving these goals, the present study is contributing the first dataset of this kind to the current literature and is also the first to attempt to apply deep learning towards developing such a classifier. To the best of our knowledge, no other studies

focused on developing depression or suicidality screening tools have applied deep learning to Facebook data previously.

Chapter 2

Related Work

Recently, significantly more attention has been paid to using social media data to develop public health tools but the research pertaining specifically to depression and suicidality are still limited.

2.0.1 Depression Prediction

To the best of our knowledge, there are only a handful of studies in the literature that have attempted to train machine learning classifiers on social media data to predict the presence or severity of depression - Choudhury *et al.*[38], Reece *et al.*[110], Reece and Danforth [111], Schwartz *et al.*[118], and Wang *et al.*[128]. Since none of these studies used deep learning, they all focused on feature engineering via the identification of linguistic, behavioral, and social predictors of depression.

The first major study, Choudhury *et al.*[38], used crowdsourcing via Amazon Mechanical Turk to recruit participants for their research. Each participant was asked to provide all Tweets going back one year and to take the Center for Epidemiologic Studies Depression Scale (CES-D) in order to provide predictive features and a ground truth label. The authors then extracted numerous depression-related features, including frequency of language about symptoms, disclosure of mental illness, treatment,

and religious involvement. They also computed emotional and linguistic features using the Linguistic Inquiry and Word Count (LIWC), a psychosocial linguistic tool developed by Tausczik *et al.*[124] for counting the number of words responding to any one of a number of psychologically significant dimensions, such as positive/negative affect, insight, assent, and many more. Another set of features consisted of the individual's local network graph properties, such as egocentricity, clustering coefficient, and embeddedness. Lastly, they also extracted behavioral features based off of Twitter metadata, such as the degree of reciprocity of the user's online interactions (such as @-mentions and re-Tweets) and diurnal activity levels. Rather than computing all these features per individual Tweet, the authors computed the aggregate statistics (mean, variance, mean momentum, and entropy) of each feature over all Tweets per user per day. These features were then inputted into several types of machine learning classifiers, the most predictive of which was a support vector machine (SVM) with a radial basis function (rbf) kernel. Its overall accuracy, precision, and recall were 70%, 74%, and 63% respectively.

Following up on this study was Reece *et al.*[110], who used the same data collection methodology (Tweet and CES-D collection via Amazon MTurk) but altered the set of features and used different units of observation. Like Choudhury *et al.*, the authors also used linguistic and emotional features computed by LIWC, but they also included unigram sentiment analysis features computed by labMT (a happiness metric computed specifically on social media text [51]) and ANEW (Affective Norms for English Words, a dictionary mapping words to various emotional ratings, developed in 1999 by the NIMH Center for the Study of Emotion and Attention [29]). Many of the labMT and ANEW features, such as labMT happiness, ANEW arousal, ANEW happiness, and ANEW dominance proved to be better predictors of depression than the LIWC features. The authors averaged these features over both daily and weekly time periods to produce two types of feature vectors, and trained random forest classi-

fiers on both datasets. The best performing classifier was trained on weekly data and resulted in an overall precision of 87% and recall of 52%. Hence this study improved upon the precision, but not the recall rate of Choudhury *et al.*[38].

Another similar study was Wang *et al.*[128], which differed mainly in that the authors trained their classifiers instead on Chinese posts from the Sina Micro-blog, the Chinese version of Twitter. They used a Chinese analog of LIWC, known as HowNet, to extract linguistic and emotional features. They also included features relating to Twitter behavior, such as the number of @-mentions, forwards, and comments, similar to Reece *et al.*[110] and Choudhury *et al.*[38]. However, their ground truth labels were more robust - each participant's depression status was confirmed via both psychiatric surveys and interviews, rather than surveys alone. After training Naive Bayes classifiers, decision trees, and decision tables, their best-performing classifier, a Naive Bayes classifier, distinguished between non-depressed and depressed individuals with 80% precision and 91% recall. These are particularly impressive results considering that research on Chinese sentiment analysis is still a very new and sparse field. It demonstrates that such social media analysis can also be effectively conducted in other languages, not just in English.

Schwartz *et al.*[118], on the other hand, took a very different approach and attempted to predict the degree of depression (denoted by the authors as $DDep$), rather than merely its presence. The authors operationalized $DDep$ as a continuous value computed from the depression characteristic scores defined by the International Personality Item Pool. They collected their data from a sample of Facebook users who completed a 100-item personality questionnaire and provided their entire history of Facebook status updates. Like in the previous studies, the authors extracted n-gram frequencies (with n ranging from 1 to 3), number of words per post, and LIWC features. Unlike previous studies, they also used latent Dirichlet allocation (LDA) to identify the top 2000 topics on Facebook and then to compute features relating to

the probability that the user would post about each topic. All features (computed across all posts in each user's entire Facebook history) were then inputted into a logistic regression model, which outputted test predictions with Pearson correlation coefficient $r = 0.386$, when compared to the true $DDep$ values.

Lastly, Reece and Danforth [111] conducted the only study to use images from social media, rather than text, to predict depression. Like their previous study that used Twitter data for the same purpose [110], they recruited human participants via Amazon MTurk and confirmed participants' depression statuses using the CES-D survey. They extracted features corresponding to image hue, image saturation, image brightness, number of comments, number of likes, number of posts/day, used filters, face presence, and face count across all posts per user per day. They then trained Bayesian logistic regression and random forest models to differentiate between the non-depressed and depressed individuals, of which the best-performing classifier was a random forest model with 60% precision and 70% recall. Although these metrics are somewhat lower than the metrics from the previous text-based studies, this research is a promising first step towards and proof-of-concept for predicting depression from image data.

2.0.2 Suicidality Prediction

The research on machine learning prediction of suicidality from social media data is similarly sparse, with all studies focusing solely on either Twitter [33, 68, 99] or Reddit data [48] and none focusing on Facebook data. In addition, the three Twitter studies used each Tweet as a single data point instead of all Tweets of a given user, in effect predicting whether a Tweet contained suicidal content rather than whether the Tweet's author was suicidal. This is subtly different from the classification problem we are attempting, since an individual who posts a Tweet containing suicidal content is not necessarily suicidal. For example, individuals in mental healthcare professions

or who have had similar experiences in the past may also post content related to suicide, but may not currently be suicidal themselves.

Homan *et al.*[68] was the earliest study to attempt this approach. The authors gathered a random sample of public Tweets from a 1-month period and pre-filtered them for suicidal content using LIWC and dictionaries of suicide-related search terms from Jashinsky *et al.*[74] and Crosby *et al.*[44]. They then further filtered the suicidal Tweets by recruiting both novices and experts to manually annotate the distress level of the Tweets as either happy (H), no distress (ND), low distress (LD), or high distress (HD). Similar to the depression prediction studies, the authors proceeded to convert each Tweet into a feature vector consisting of weighted n-gram frequencies (with n ranging from 1 to 3). The weights used were the n-grams' term frequency-inverse document frequency (TF-IDF) scores, a metric that is directly proportional to the n-gram's frequency and inversely proportional to the number of documents in which the n-gram appears. No other features were computed, resulting in no encoding of either behavioral or temporal information in the feature vectors. The authors trained a number of machine learning binary classifiers to distinguish between Tweets in the non-suicidal (the H and ND distress categories) and the suicidal (the LD and HD distress categories) classes, the best-performing of which was an SVM that used expert-annotated training data. It had an overall precision of 59% and recall of 71%.

O'Dea *et al.*[99] conducted a very similar study that differed only in its annotation methodology. The authors collected a Twitter dataset over a one-month period and manually annotated it, assigning labels of either "strongly concerning," "possibly concerning," or "safe to ignore" to each. Rather than using both novice and expert annotators, they recruited a team of three mental health researchers and two computer scientists that derived the labels together. Like Homan *et al.*, their feature vectors consisted only of linguistic features - in this case, TF-IDF-weighted unigram frequencies. They then trained logistic and SVM one-vs-all classifiers, achieving pre-

cision and recall rates of 80% and 53% respectively for the strongly concerning class, 76% and 91% for the possibly concerning class, and 75% and 53% for the safe to ignore class. This was a notable increase in accuracy over Homan *et al.*despite the similar methodology, perhaps due to the more refined classifications.

Burnap *et al.*[33] was the first study to increase the feature set beyond n-grams. Like the previous two studies, the authors used the Twitter API to collect a random sample of public Tweets and asked human annotators to classify the data as either suicidal or non-suicidal. They then extracted three sets of features - lexical characteristics such as part of speech and affective lexical domain, sentiment features as computed by LIWC, and the frequencies of informal words and phrases related to suicide. The authors proceeded to train Naive Bayes, decision tree, and SVM classifiers on the union of all three feature sets, the most accurate of which was an SVM that achieved an overall precision of 64% and recall of 74%.

Lastly, Choudhury *et al.*[48] was the only study to attempt to predict an individual's state of mind rather than the suicidal content of their social media posts. However, this was still achieved in an indirect manner - the researchers identified a random sample of users who regularly frequented mental health-related sub-Reddits such as r/depression, r/mentalhealth, and r/traumatoolbox and determined the subset of these users who had also frequented the suicide support sub-Reddit r/SuicideWatch. The users who had frequented the mental health sub-Reddits but not r/SuicideWatch were coined the MH group and the users who had frequented both were coined the $MH \rightarrow SW$ group. The authors then extracted four sets of features - linguistic structure (such as part of speech, readability, and accomodation), interpersonal awareness (such as use of the 1st person singular, 1st person plural, and 2nd person), interaction (such as the number of posts authored, post length, and number of comments), and content (relative unigram and bigram token frequencies). By training a logistic regression model on these features, the authors were able to distinguish between the

MH and $MH \rightarrow SW$ groups with 81% precision and 81% recall. Although this classification task is somewhat different from ours since the MH group is a specific subset of non-suicidal individuals who are interested in mental health issues, we consider this study to be the state-of-the-art in machine-learning prediction of suicidality from social media data.

2.0.3 Novelty

This study is novel in several ways, in particular for its dataset and deep learning methodology. To the best of our knowledge, none of the suicidality prediction studies mentioned here use a Facebook dataset, which may be more advantageous than a Twitter dataset due to its longer-form content and significantly larger user base [52]. In addition, ours is the first study to use psychiatrically-validated labels of individual suicidality. The other studies reviewed here either attempted to predict whether an individual Tweet contained suicidal content or labeled individuals according to their Reddit activity, rather than psychiatrically confirming whether the individuals had actually experienced suicidal ideation recently.

Regarding both the depression and suicidality prediction tasks, our study is also novel in its use of word embeddings and deep neural networks. To the best of our knowledge no other studies of this kind have attempted either technique, even though it may offer significant predictive advantages.

Chapter 3

Methodology

3.0.1 Data Collection

To compile training and testing data for the neural networks to learn on, we sought to collect text and image data from Facebook, Instagram, and Twitter, as well as psychiatric assessments of the presence and/or level of depression and suicidality in each participant to act as the classifier labels. We chose to assess depression using the Beck Depression Inventory (BDI), a 21-question multiple choice survey that seeks to measure an individual's level of depression by asking him/her to identify which statement they agree with most in groups of four statements, based upon their mental state of the last two weeks. Each group describes a particular characteristic of depression, with the individual statements ranging in intensity from minimal depression to severe depression. An example of one such question is shown below:

0. I do not feel sad.
1. I feel sad.
2. I am sad all the time and I can't snap out of it.
3. I am so sad and unhappy that I can't stand it.

As shown in this example, each choice is scored with an integer between 0 and

3, and the subject's overall BDI score is the sum of their individual question scores, resulting in a minimum score of 0 and a maximum of 63. Generally scores of 0-13 indicate minimal depression, 14-19 indicates mild depression, 20-28 indicates moderate depression, and 29-63 indicates severe depression [24]. A full copy of the BDI can be found in Appendix B. Although a number of other quantitative depression screening scales exist, we chose to use the BDI because it is very accurate when compared against psychiatric gold standards [14, 84, 113], it has high internal consistency [23, 113, 123], is not particularly sensitive to day-to-day mood fluctuations [23], and offers a finer-grained assessment of the degree of severity than other screening tools.

To assess suicidality, we chose to use the Columbia-Suicide Severity Rating Scale (C-SSRS), a 5-question yes/no survey that assesses both the severity and immediacy of an individual's risk of suicide [2]. (A full copy of the C-SSRS can be found in Appendix C.) In particular, the C-SSRS is one of the only clinical suicide screening tools to have been formally evaluated for predictive value. Like the BDI, the C-SSRS has been shown on multiple occasions to be highly accurate, sensitive, specific, and internally consistent [56, 106, 78]. We asked participants to answer each question twice - once with their response for the last 6 months, and once with their response for their entire lifetime.

Participants were recruited through several means - via public posts on Facebook, Instagram, and Twitter, emails sent to a random sample of Princeton students, and a website designed to advertise the study.

To collect the desired data from the participants, we used an online Qualtrics survey consisting of an informed consent page (a full copy of which can be found in the Appendix) followed by 29 questions. The first 21 questions corresponded to the questions from the BDI, the next 5 questions corresponded to the items on the C-SSRS, and the last three questions asked for the participant's Twitter, Instagram, and Facebook usernames respectively. The survey was automatically scored by the

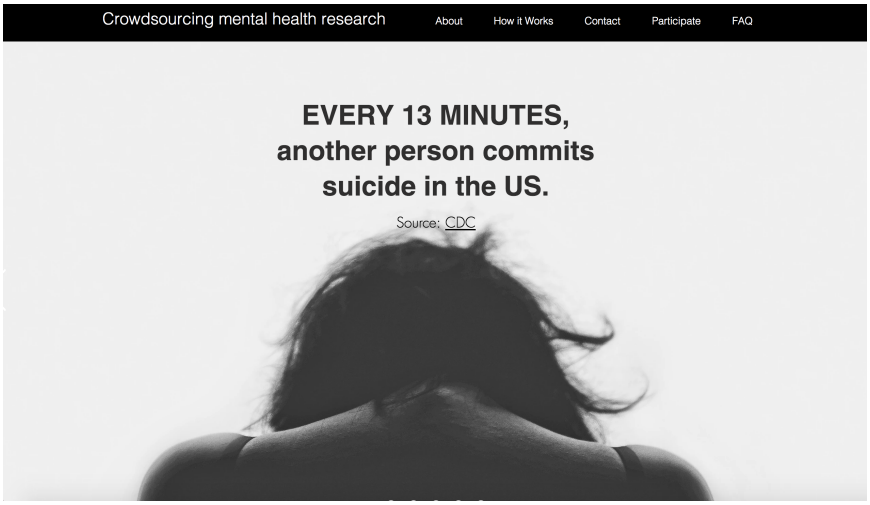

Figure 3.1: A screenshot of the website (located at `http://www.crowdsourcing-mental-health.com/`) designed to provide information about the study and to recruit human participants.

Qualtrics software - if a participant scored 20 or above on the BDI or had answered Yes to any of the questions in the C-SSRS corresponding to the last six months (indicating some level of suicidal ideation), they were automatically alerted at the end of the survey with a page listing the phone number, website, and other contact details of various mental health professionals and resources. Due to the anonymous nature of the data collection, no further contact could be made with these participants to encourage follow-through.

After completion of the survey, a Python script automatically sent friend/follow requests to each of the participant's social media accounts using the respective websites' developer APIs. Once the participants responded to the friend/follow requests, the Python package `BeautifulSoup` was used to gather data from the participants' Facebook feeds. We collected all text content from the user's Facebook stories going back six months, including statuses and captions of photos, links, or reshares; but not including comments.

Due to the intensely personal nature of the data collected in this study, confidentiality was of the utmost importance. All data collection, storage, and analysis proce-

dures were approved by the Princeton Institutional Review Board, Protocol #8049. As per HIPAA guidelines, we anonymized all data (after scraping the necessary social media feed) by removing the handle and replacing it with a randomly-generated unique numeric identifier. In addition, all data was stored on the University's H drive, which is automatically encrypted and can only be accessed via secure SSH connection. The University's Office of Information Technology has approved the H drive for storing and sharing restricted data such as the protected health information collected in this study. However, we still could not necessarily guarantee complete anonymity, since participants' identities could possibly be inferred from their social media posts. Consequently, access to the data was restricted solely to research personnel.

3.0.2 Language Processing

After completing data collection, we pre-processed all the text by lowercasing all alphabetical characters, shortening substrings of repeated characters of length longer than 3 characters to exactly 3 characters, and removing all English stopwords, which we obtained from the Python package `nltk` (Natural Language Toolkit [4]). We then tokenized the text, including URLs, emoticons, numbers, and hashtags as unique tokens, as listed in Table 3.1.

Ruby Regular Expression	Token
`/https?:\/\/\S +\b |www\.(\w +\.)+\S */`	`<URL>`
`[8:=;][’‘\-]?[)d]+|[)d]+[’‘\-]?[8:=;]i*`	`<SMILE>`
`/[8:=;][’‘\-]p+/i`	`<LOLFACE>`
`[8:=;][’‘\-]\(+|\)+[’‘\-][8:=;]*`	`<SADFACE>`
`/#{eyes}[’‘\-][\/|l*]/`	`<NEUTRALFACE>`
`/[-+]?[.\d ]*[\d ]+[:,.\d ]*/`	`<NUMBER>`
`/#\S +/`	`<HASHTAG>`
`/<3/`	`<HEART>`

Table 3.1: List of non-word tokens extracted from Facebook posts.

Next, we lexically normalized the tokens using (out-of-vocabulary, in-vocabulary)

pairs taken from Han *et al.*[62] and Liu *et al.*[90], both of which compiled their dictionaries from Twitter datasets.

To convert each participant's Facebook feed into a semantically meaningful numeric feature vector, we used GloVe embeddings, word vectors learned by an unsupervised learning algorithm designed by Pennington *et al.*[103]. The algorithm builds a word vocabulary, computes the co-occurrence matrix X of this vocabulary, and learns word vectors by minimizing the objective function

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \tag{3.1}$$

where V is the number of words in the vocabulary, w_i is the word vector for word i, $\tilde{w}_j$ is the word context vector for word j, b_i and $\tilde{b}_j$ are bias vectors, and f is a weighting function defined as

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases} \tag{3.2}$$

where $x_{\max}$ is a threshold specifying when a co-occurrence count is considered "large" and α is a non-negative constant. In practice, $x_{\max} = 100$ and $\alpha = 3/4$ [103]. The design of this objective function is motivated by the desire to find word vectors whose pairwise cosine similarities approximate the log of the corresponding co-occurrence counts. The log function acts as a smoother, and the weight $f(X_{ij})$ serves to ensure that co-occurrences that occur rarely have smaller weights than those that occur more frequently. GloVe has been shown to perform well on word analogy, similarity, and named entity recognition tasks [103], demonstrating that it adequately captures both semantic and syntactic information.

We used 25-dimensional GloVe vectors pre-trained on a Twitter corpus consisting of 2 billion Tweets with 27 billion tokens and a vocabulary of size 1.2 million obtained

from the original authors of the GloVe algorithm [103] to represent each Facebook post. Words not included in the GloVe vocabulary were encoded as 25-dimensional zero vectors, resulting in arrays of dimensions $(w, 25)$ for each Facebook story, where w is the word count of the story. Each story was then padded with 25-dimensional zero rows to the maximum post length, which was observed to be 142 words. For each participant, we vertically concatenated all Facebook posts from the last six months of their feed (resulting in each participant having a data array of dimensions $(\sum_{i=1}^{s} w_i, 25)$ where s is the number of stories and w_i is the number of words in story i), then padded the array with 25-dimensional zero rows to the maximum feed length, which we observed to be 298 posts.

To reduce dimensionality and memory usage, we then applied principal component analysis (PCA) to the data, transforming each user's data array by flattening it into a 1-dimensional vector and projecting it onto its top 50 components to form the final feature vector that was inputted into the neural networks.

3.0.3 Neural Network Training

For both classification tasks, we experimented with training five different types of neural network architectures, including fully connected neural networks (FCNNs), longer short-term memory networks (LSTMs), combinations of a convolutional neural network with an LSTM (LSTM-CNN), gated recurrent unit networks (GRUs), and combinations of a convolutional neural network with a GRU (GRU-CNN). For each of these types of neural networks, we trained 258, 42, 432, 42, and 432 different architecture designs respectively, resulting in a total of 1206 experiments per classification task.

For both prediction tasks, the input feature vector was the dimension-reduced GloVe representation of the participants' Facebook posts mentioned in the previous section. For the depression prediction task, an individual was labeled as depressed if

their BDI score was 20 or above. For the suicidality prediction task, an individual was labeled as suicidal if they had responded "Yes" to any question on the C-SSRS survey in the 6-month category.

All experiments were coded in Python using Google's `Tensorflow` package, and run on either an Nvidia Titan X GPU, Intel Xeon Phi processor, Nvidia Tesla GPU, or an Nvidia Fermi GPU.

Below are the five types of neural networks trained in our experiments:

Fully connected neural networks (FCNN)

This is the most basic type of neural network, consisting simply of a series of densely connected layers. Each dense layer can be represented by the equation

$$a(x) = f(Wx + b) \tag{3.3}$$

where a is the output (also known as the activation), $f(\cdot)$ is an activation function, W is the weight matrix, and b is a bias vector. Without applying f, this is equivalent to a linear regression model, so we use a nonlinear activation function in order to incorporate nonlinearity in the network. In our experiments we used the rectified linear unit (ReLU), defined as

$$f(x) = \max(0, x), \tag{3.4}$$

in the intermediate layers, and the softmax function, defined as

$$f(\mathbf{x}) = \frac{e^{x_j}}{\sum_{k=1}^{n} e^{x_k}}, \tag{3.5}$$

in the last layer to obtain an output between 0 and 1. In our experiments we trained 258 different FCNNs, consisting of 6 single-layer architectures, 36 2-layer architec-

tures, and 216 3-layer architectures. The parameters we experimented with are listed in Table 3.2.

Type of neural network	**Size of hidden layer 1**	**Size of hidden layer 2**	**Size of hidden layer 3**
FCNN with 1 hidden layer	32, 64, 128, 256, 512, 1024	–	–
FCNN with 2 hidden layers	32, 64, 128, 256, 512, 1024	32, 64, 128, 256, 512, 1024	–
FCNN with 3 hidden layers	32, 64, 128, 256, 512, 1024	32, 64, 128, 256, 512, 1024	32, 64, 128, 256, 512, 1024

Table 3.2: Architecture parameters of FCNN experiments. Each cell lists the values experimented with for that network architecture type.

Longer short-term memory networks (LSTM)

A longer short-term memory network is a type of recurrent neural network, named as such because it recursively applies the same set of weights over multiple time steps, providing the network with the capacity to memorize information and encode longer term dependencies [67]. An LSTM consists of a series of recurrent memory cells, defined as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{3.6}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{3.7}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{3.8}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{3.9}$$

$$h_t = o_t \circ \tanh(C_t) \tag{3.10}$$

where x_t is the input vector, f_t is the forget gate vector at time step t, i_t is the input gate vector at time step t, o_t is the output gate vector at time step t, C_t is the cell state at time t, h_t is the output vector, and $W_f, W_i, W_o, W_c, U_f, U_i, U_o, U_c, b_f, b_i, b_o, b_c$ are

weight matrices and bias vectors. Furthermore, $\circ$ represents the Hadamard (entry-wise) product and σ is the sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{3.11}$$

Intuitively, the forget gate decides whether to keep or forget the information from the last time step, the input gate decides what new information to store in the cell, and Equation (3.8) stores the update in the cell state. The output gate and output vector then determine what information from the cell state to output. In our experiments we applied ReLU activation to every intermediate layer and a softmax activation to the last recurrent layer to obtain an output between 0 and 1. We trained 42 different LSTMs, 6 with a single LSTM cell and 36 with two consecutive LSTM cells. The parameters we experimented with are listed in Table 3.3.

Type of neural network	**Size of recurrent layer 1**	**Size of recurrent layer 2**
LSTM with 1 recurrent layer	16, 32, 64, 128	–
LSTM with 2 recurrent layers	16, 32, 64, 128	16, 32, 64, 128

Table 3.3: Architecture parameters of LSTM experiments. Each cell lists the values experimented with for that network architecture type.

Combination of convolutional and longer short-term memory networks (LSTM-CNN)

To attempt to take into account the contextual nature of text data, we also trained hybrid combinations of convolutional neural networks (CNNs) and LSTMs consisting of a single 1-dimensional convolutional layer followed by one or more LSTM layers. A convolutional layer consists of convolving a filter (a small weight matrix) with various windows of the same size in the input vector/matrix, in effect creating a layer that is only locally connected to the previous layer, as opposed to the dense layers of FCNNs,

which are fully connected to all neurons of the previous layer. The stride of the layer determines how many units the filter shifts by between each convolution operation. In our experiments, we also inserted a max-pooling layer after the convolutional layer. A max-pooling layer performs downsampling by taking the maximum of each window in the input vector/matrix. We trained 432 different LSTM-CNN networks, each with a single convolutional layer followed by a max-pooling layer, followed by a single LSTM layer and a softmax activation layer for the output. The architecture parameters we experimented with are listed in Table 3.4.

Type of neural network	**Size of recurrent layer**	**Number of conv. features**	**Conv. filter length**	**Max pooling length**	**Max pooling stride**
LSTM-CNN	16, 32, 64, 128	16, 32, 64, 128	3, 4, 5	2, 3, 4	2, 3, 4

Table 3.4: Architecture parameters of LSTM-CNN experiments. Each cell lists the values experimented with for that network architecture type.

Gated recurrent unit networks (GRU)

A gated recurrent unit network (GRU) is a variation on the LSTM that replaces the forget gate with an update gate and lacks the internal memory that LSTMs possess. A GRU block is defined by the following equations:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{3.12}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{3.13}$$

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tanh(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h) \tag{3.14}$$

where x_t is the input vector, z_t is the update gate vector at time step t, r_t is the reset gate vector at time step t, h_t is the output vector at time step t, and $W_z, W_r, W_h, U_z, U_r, U_h, b_z, b_r, b_h$ are the weight matrices and bias vectors. Intuitively,

a GRU combines old information with new information, in effect updating its information rather than forgetting. Like with the LSTM network architectures, we applied ReLU activations to all intermediate GRU layers and a softmax activation to the last layer in order to produce an output between 0 and 1. We trained 42 different GRUs, 6 with a single GRU cell and 36 with two consecutive GRU cells. The parameters we experimented with are listed in Table 3.5.

Type of neural network	**Size of recurrent layer 1**	**Size of recurrent layer 2**
GRU with 1 recurrent layer	16, 32, 64, 128	–
GRU with 2 recurrent layers	16, 32, 64, 128	16, 32, 64, 128

Table 3.5: Architecture parameters of GRU experiments. Each cell lists the values experimented with for that network architecture type.

Combination of convolutional and gated recurrent unit networks (GRU-CNN)

Like with the LSTMs, we also trained hybrid combinations of convolutional neural networks (CNNs) and GRUs consisting of a single 1-dimensional convolutional layer followed by one or more GRU layers. We trained 432 different GRU-CNN networks, each with a single convolutional layer followed by a max-pooling layer, followed by a single GRU layer and a softmax activation layer for the output. The architecture parameters we experimented with are listed in Table 3.6.

Type of neural network	**Size of recurrent layer**	**Number of conv. features**	**Conv. filter length**	**Max pooling length**	**Max pooling stride**
GRU-CNN	16, 32, 64, 128	16, 32, 64, 128	3, 4, 5	2, 3, 4	2, 3, 4

Table 3.6: Architecture parameters of GRU-CNN experiments. Each cell lists the values experimented with for that network architecture type.

Regularization

We employed a number of regularization techniques in our experiments in order to prevent the models from overfitting. We applied dropout with a drop probability of 0.5 to every hidden layer of every network and l_2-regularization to all weights, biases, and activations. l_2-regularization is a method of penalizing weights with extreme values by modifying the cost function as follows:

$$J_{\text{reg}}(W) = J(W) + \lambda \|W\|_2 \tag{3.15}$$

where W is the weight matrix to be regularized, $J(W)$ is the original cost function, $\|\cdot\|_2$ is the l_2-norm (also known as the Euclidean norm), and λ is a constant regularization parameter specifying the severity of the penalty. In this study we used a λ value of 0.01 for the FCNN, LSTM, and GRU experiments, and a λ value of 0.001 for the LSTM-CNN and GRU-CNN experiments.

We also employed early stopping, a technique that halts the neural network training process as soon as the validation loss ceases to decrease. In our experiments we stopped training if the validation loss had not improved by more than 0.01 in three epochs.

Training and Cross-Validation

To train our neural networks we used the Adam optimizer [81] and clipped all parameters' gradients at a maximum norm of 1. As recommended by the original paper [81], we used a learning rate of 0.001, β_1 and β_2 values of 0.9 and 0.999 respectively, an ϵ value of 10^{-8}, and no learning rate decay. We used cross-entropy as our objective function, defined as

$$H(y, \hat{y}) = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{3.16}$$

where y is the observed distribution and $\hat{y}$ is the predicted distribution.

We also split the training data into 90% training, 10% validation. The validation loss was used to select the best-performing architecture design for each of the five types of neural network architectures and to determine when to stop training.

3.0.4 Evaluation Methodology

Due to the small size of the dataset, we estimated model performance using leave-one-out (LOO) cross-validation. Given a dataset with n samples (not including data used for hyperparameter validation), the neural network is trained n times; during each round, a new data point is selected as the test data point, and the neural network is trained on the remaining $n-1$ data points. All evaluation metrics are then computed as an average of the model's performance across all n LOO iterations.

To determine whether a specific neural network architecture type performed better than the other four in either classification task, we conducted pairwise independent t-tests comparing the accuracies of each pair of architecture types for both classification tasks.

We also trained a few other standard machine learning classifiers in order to determine whether a simpler model could be used to achieve comparable accuracy. These classifiers included:

- **Binary logistic regression**: A binary logistic regression model is simply a linear regression with a binary dependent variable. This is achieved by applying the sigmoid function to a linear model:

$$P(y=1|x) = \sigma(\theta^T x) \tag{3.17}$$

$$= \frac{1}{1+e^{-\theta^T x}} \tag{3.18}$$

$$P(y=0|x) = 1 - P(y=1|x) \tag{3.19}$$

$$= 1 - \frac{1}{1+e^{-\theta^T x}} \tag{3.20}$$

where θ is the weights vector and x is the input vector.

- **Nearest neighbor**: Nearest neighbor is a classifier that identifies the data point in the training set that is closest (by the Euclidean distance metric) to a given test data point and outputs the class of that data point:

$$\hat{y}(x) = y(\text{argmin}_{z \in T} \|z - x\|_2) \tag{3.21}$$

where $y(x)$ denotes the class of data point x, $\|\cdot\|_2$ is the Euclidean distance, T is the training set, and $\hat{y}(x)$ is the predicted class of x.

- **Decision tree**: A decision tree is a rules-based classifier that learns a function by recursively learning dichotomous rules for classifying data points. Given a set of training data S, a decision tree is created by selecting a feature k to "split" on, meaning that S is partitioned into two subsets/trees based upon the value of $x_i(k)$ for each data point x_i. The algorithm then chooses to split on the feature that offers the most information gain. For each subtree, another feature (possibly the same one as in the first split) is again selected to split the data on, and so on. There are a number of ways in which to compute information gain, but in our experiments we used Gini impurity, which is a measure of how much mis-classification would occur if we split on feature k and randomly assigned labels to the data points in each subtree according to the distribution of labels in that subtree:

$$I_G(f) = \sum_{i \neq k} f_i f_k \tag{3.22}$$

where f_i is the fraction of items labeled as class i in the subtree. In our experiments we stopped splitting a subtree whenever that subtree was either pure (meaning all samples were from the same class) or the subtree contained < 2 samples.

- **Support vector machine (SVM)**: A support vector machine is a type of classifier that seeks a hyperplane that maximally separates the classes [41]. It solves the following minimization problem:

$$\min_{w,b,\gamma} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \gamma_i \tag{3.23}$$

$$\text{such that } y_i(w^T \phi(x_i) + b) \geq 1 - \gamma_i, \tag{3.24}$$

$$\gamma_i \geq 0,\ i = 1, \cdots, n \tag{3.25}$$

  where C is a regularization parameter, x_i for $i = 1, \cdots, n$ are the training data vectors, y_i for $i = 1, \cdots, n$ are their labels, $\phi(\cdot)$ is the feature map of the kernel, and w, b, γ are the parameters we seek to learn. The decision function is

$$\hat{y}(x) = \text{sgn}\left(w^{*T} x + b^{*T}\right). \tag{3.26}$$

  In our experiments we used the radial basis kernel, defined as $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$.

Lastly, we also compared the performance of the deep neural networks against that of a trained human psychiatrist on the same tasks. We selected a random sample of 10 participants and asked the psychiatrist to predict whether each participant was depressed, suicidal, or both based upon their Facebook posts from the last six months. (Due to time constraints, we unfortunately could not generate psychiatric predictions for a larger sample.) To estimate the performance of the deep neural networks on the same task, we assigned the data corresponding to the same 10 participants to the test set, and retrained all the different architectures on 90% of the remaining data. The other 10% was used as cross-validation data to select the best architecture for each of the five model types. We then generated predictions on the test set using these five best-performing models and averaged their accuracies to obtain a final estimation of

the deep neural net performance.

Chapter 4

Results

4.0.1 Data Collection

A total of 233 individuals responded to the Qualtrics survey, but only 158 of these responses were considered complete - that is, the participant responded to all survey questions and accepted all friend and follow requests on their social media accounts. Histograms of participants' BDI scores and the number of "Yes" responses to questions on the C-SSRS are shown in Figures 4.1 and 4.2, using only the data from the completed responses. It is notable that 63 out of 158 participants (40%) scored in the moderately to severely depressed range and 78 out of 158 participants (49%) expressed that they had experienced suicidal ideation sometime within the past six months. We do not expect that these high numbers represent the true proportions of individuals in the general population who are depressed or suicidal, but is rather a result of the survey's voluntary response bias. Those who have had past experiences with depression or suicidal ideation are more likely to be concerned with and therefore interested in participating in mental health research.

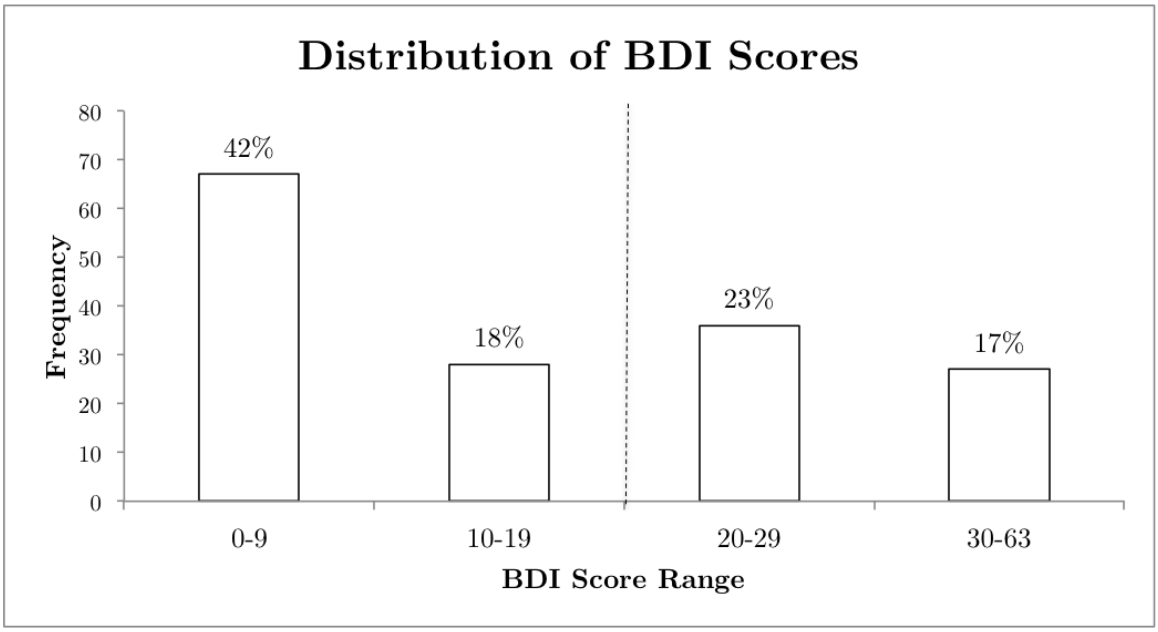

Figure 4.1: Histogram of participants' BDI scores. The dashed line indicates the class boundary between non-depressed (BDI scores < 20) and depressed (BDI scores ≥ 20).

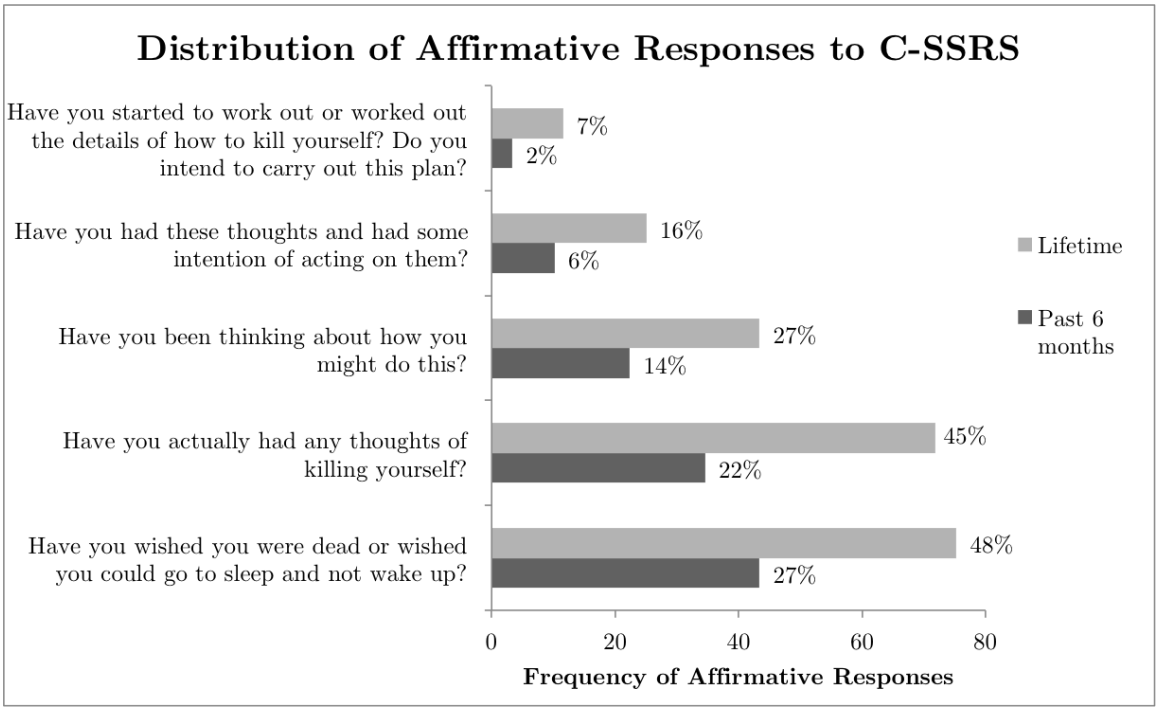

Figure 4.2: Number of "Yes" responses to questions on the C-SSRS.

4.0.2 Depression Prediction Task

For the depression prediction task, the accuracy, precision, and recall of the best-performing experiments for each deep neural network type are shown in Table 4.1.

Figure 4.3 illustrates the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) for each of the aforementioned experiments. The LSTM had the highest overall accuracy and recall rates, but the FCNN and LSTM-CNN had the highest precision rates. To determine whether the difference in performance between the deep neural network types was significantly different across all experiments, we conducted pairwise 1-tailed independent t-tests between the accuracies of each pair of neural network types. The results are shown in Table 4.2. Every test returned a statistically significant result for $\alpha = 0.05$, indicating that the mean accuracies are indeed significantly different between each pair of deep neural network types.

Neural Network Type	**Accuracy**	**Precision**	**Recall**
FCNN	0.9557	1.0000	0.8923
LSTM	0.9620	0.9683	0.9385
LSTM-CNN	0.9304	1.0000	0.8308
GRU	0.9494	0.9524	0.9231
GRU-CNN	0.9241	0.9818	0.8308

Table 4.1: Accuracy, precision, and recall of the best-performing experiment for each type of deep neural network on the depression prediction task.

	FCNN	LSTM	LSTM+CNN	GRU	GRU+CNN
FCNN	–	–	–	–	–
LSTM	1.06E-32	–	–	–	–
LSTM+CNN	8.34E-09	3.79E-41	–	–	–
GRU	2.80E-14	4.90E-06	5.77E-23	–	–
GRU+CNN	7.28E-20	1.77E-37	2.58E-15	2.84E-30	–

Table 4.2: Pair-wise independent 1-tailed t-tests between the accuracies (on the depression prediction task) of each type of deep neural network

Although it is difficult to compare the results of this study against that of other studies due to differences in datasets, definitions, and methodologies, we briefly summarize the precision and recall rates achieved by other studies (mentioned in Section 2) against ours in Figure 4.4. As indicated by the data point in the upper right, our

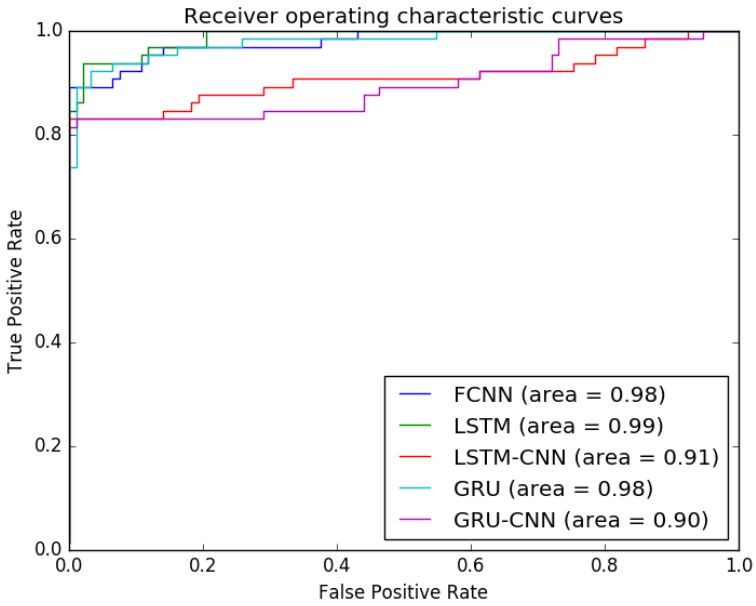

Figure 4.3: ROC curves and AUC of best-performing experiment for each deep neural network type on the depression prediction task.

study achieved both a higher precision and recall than other machine learning classifiers in the literature that attempted to predict depression from social media data. (However, we did not compare this study to Schwartz *et al.*[118] since the authors posed the problem as a regression rather than classification problem.)

4.0.3 Suicidality Prediction Task

For the suicidality prediction task, the accuracy, precision, and recall of the best-performing experiments for each deep neural network type are shown in Table 4.3. Figure 4.5 illustrates the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) for each of the aforementioned experiments. The GRU-CNN had the highest overall accuracy, but the FCNN had the highest precision and the GRU had the highest recall. To determine whether the difference in performance between the deep neural network types was significantly different across all experiments, we conducted pairwise 1-tailed independent t-tests between the ac-

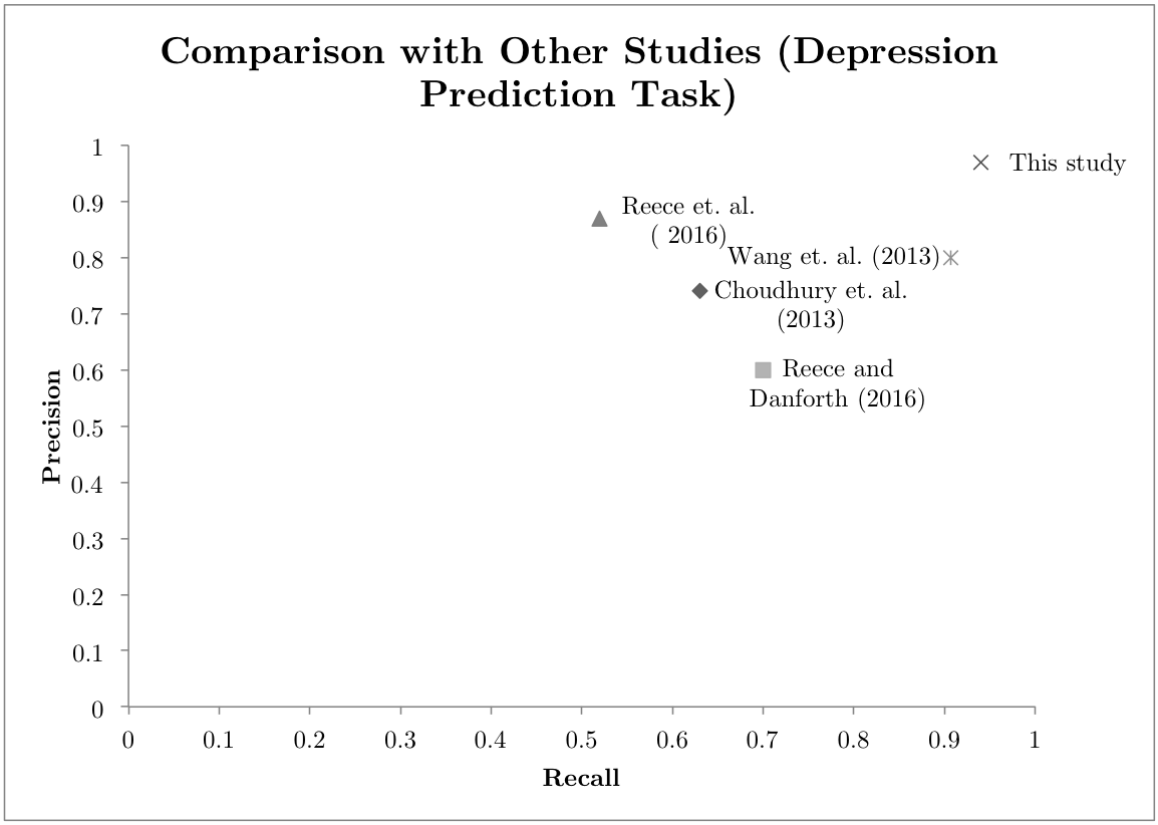

Figure 4.4: Comparison of the precision and recall of our study against that of other studies.

curacies of each pair of neural network types. The results are shown in Table 4.4. Every test returned a statistically significant result for $\alpha = 0.05$, indicating that the mean accuracies are indeed significantly different between each pair of deep neural network types.

Neural Network Type	Accuracy	Precision	Recall
FCNN	0.8734	0.9833	0.7564
LSTM	0.8987	0.9697	0.8205
LSTM-CNN	0.8608	0.9118	0.7949
GRU	0.9114	0.9324	0.8846
GRU-CNN	0.9240	0.9818	0.8308

Table 4.3: Accuracy, precision, and recall of the best-performing experiment for each type of deep neural network on the suicidality prediction task.

As with the depression prediction task, we also compared the results of this study

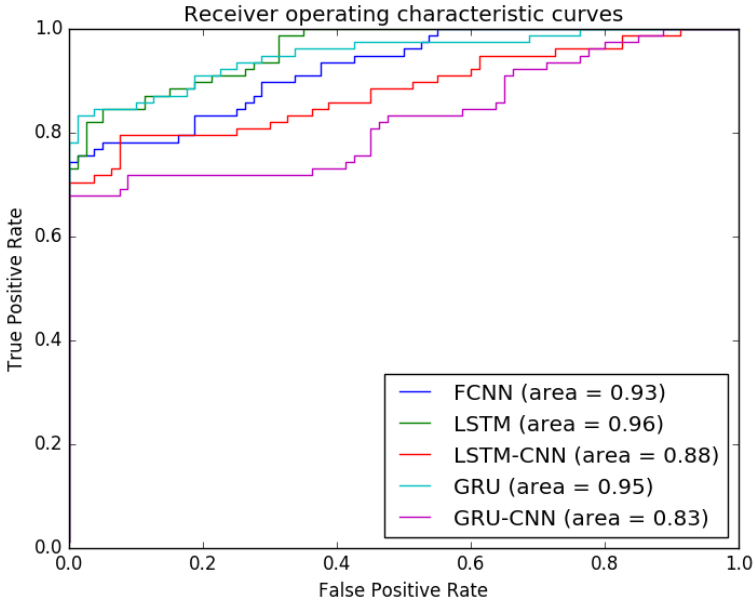

Figure 4.5: ROC curves and AUC of best-performing experiment for each deep neural network type on the suicidality prediction task.

	FCNN	LSTM	LSTM+CNN	GRU	GRU+CNN
FCNN	–	–	–	–	–
LSTM	2.54E-50	–	–	–	–
LSTM+CNN	1.41E-32	1.42E-21	–	–	–
GRU	4.85E-56	0.001	3.71E-31	–	–
GRU+CNN	5.54E-35	0.003	0.000212	3.75E-06	–

Table 4.4: Pair-wise independent 1-tailed t-tests between the accuracies (on the suicidality prediction task) of each type of deep neural network

against that of other studies in the literature that also attempted to predict suicidality from social media using machine learning techniques. The results can be seen in Figure 4.6. As indicated by the data point in the upper right, our study achieved both a higher precision and recall than other machine learning classifiers in the literature that attempted to predict suicidality from social media data.

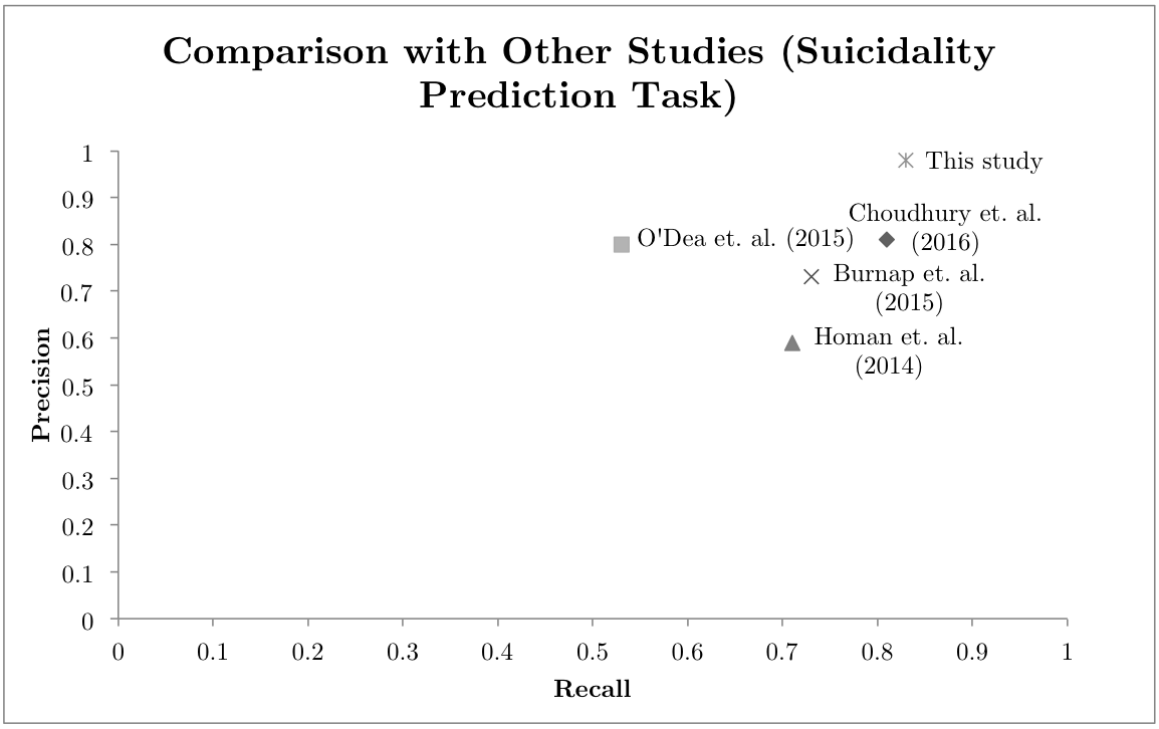

Figure 4.6: Comparison of the precision and recall of our study against that of other studies.

4.0.4 Comparison against other machine learning classifiers

To confirm that a deep neural network can indeed perform better on both tasks than a simpler model, we also compared the accuracies of other standard machine learning classifiers such as a logistic regression model, a nearest neighbor classifier, a decision tree, and a support vector machine against that of the deep neural networks. A comparison of their accuracies can be seen in Figure 4.7. In both tasks, the deep neural network has the highest accuracy. Excepting the case of the decision tree in the depression prediction task (due to high error), this difference is statistically significant.

4.0.5 Comparison against human psychiatrist performance

The results of the trained psychiatrist's performance in comparison with that of the deep neural network on both tasks can be seen in Figure 4.8. On both tasks, the

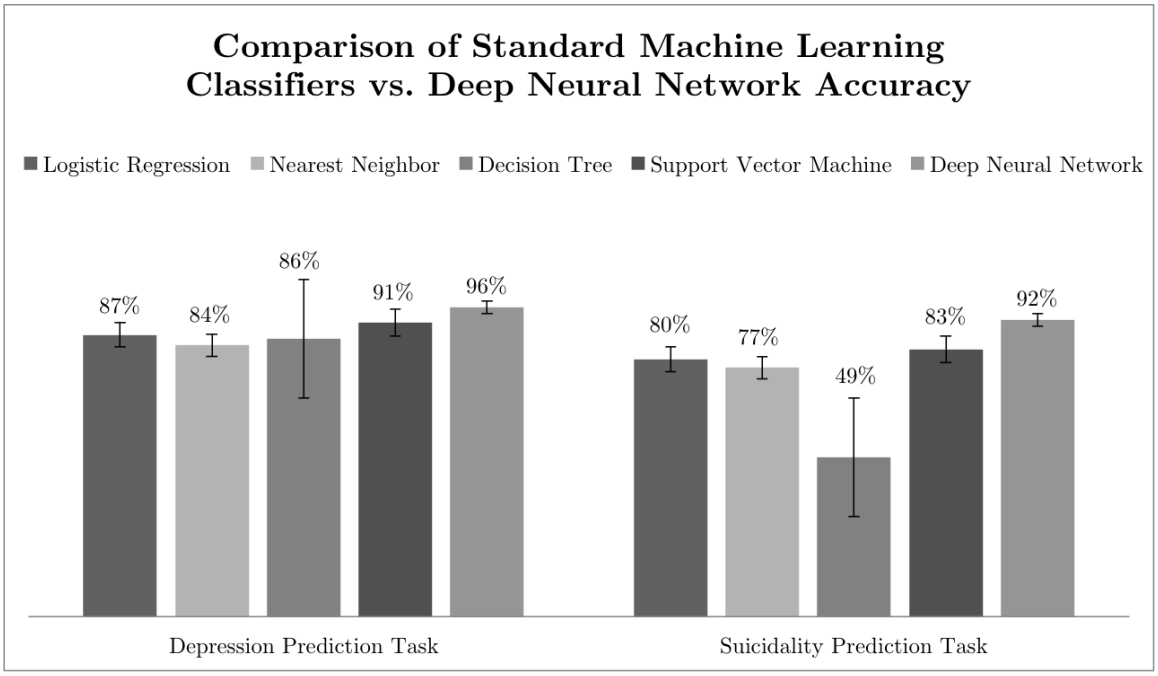

Figure 4.7: Comparison of accuracy of the best-performing deep neural network against accuracies of a logistic regression model, nearest neighbor classifier, decision tree, and support vector machine.

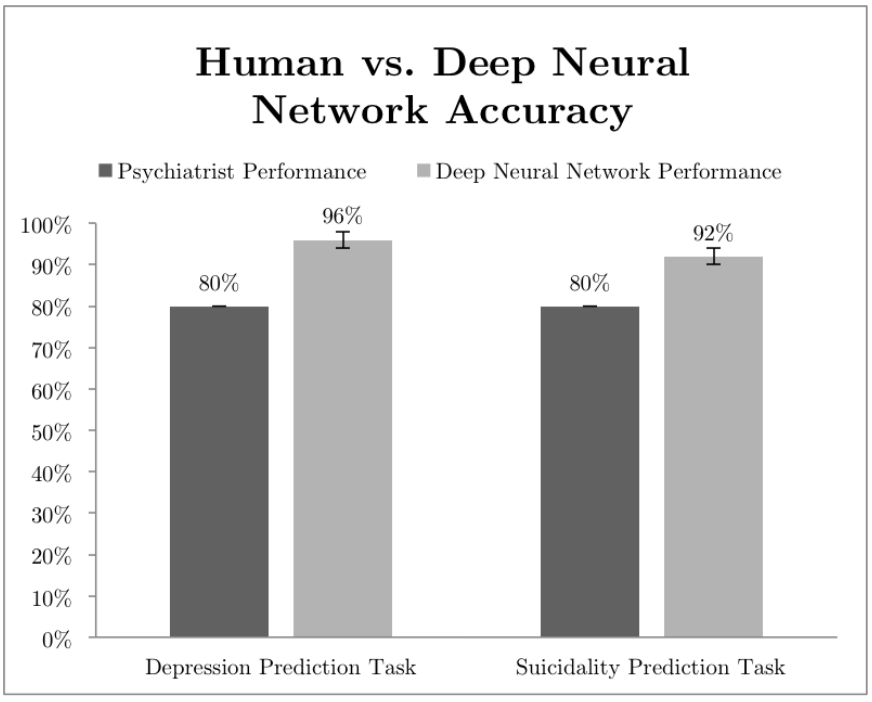

Figure 4.8: Comparison of average accuracy of the best-performing deep neural networks against the human psychiatrist's accuracy.

deep neural network performed significantly better than the psychiatrist. However, we conducted this analysis on a very small sample of points due to time and energy constraints - it would be more informative to compare the performances of the psychiatrist versus the deep neural network on a much larger scale.

Chapter 5

Conclusion

As seen in our results, a deep neural network is capable of predicting both depression and suicidality from social media data with far higher accuracy, precision, and recall than other standard machine learning classifiers and a human psychiatrist. Although much more investigation is required in order to determine whether these results are generalizable and to understand why certain neural network architectures perform better than others on each task, we believe that this research offers an adequate proof-of-concept for the use of artificial intelligence in depression screening and suicide intervention.

However, this classifier is by no means a panacea for either of these issues and must be applied in conjunction with other tools in order to be effective at reducing under-treatment. For example, this model can be incorporated into an online tool on a social network such as Facebook or Twitter and used to passively monitor a user's mental health status. Upon detecting that the user may be at high risk for severe depression and/or suicidal ideation, the tool may take a number of preventive actions, including but not limited to:

- sending the user a list of free online resources and hotline numbers. According to research by Gould *et al.*[59], suicide hotlines significantly decrease callers'

levels of suicidality.

- reaching out to close friends and family members to ask them to offer their support to the at-risk individual.
- letting the individual know that they are not alone, and connecting them to an online support community.
- identifying local mental healthcare providers and clinics where the individual may go for further help and treatment.
- directly connecting the user to online treatment programs. Research indicates that many online interventions are often just as effective at preventing suicide as in-person clinical treatments [39].

Due to its low-effort and low-cost implementation, a social media-based screening tool for depression and suicidality may offer significantly higher coverage than traditional clinical risk assessment scales. As previously discussed, traditional scales require an at-risk individual to take many steps in order to get screened - they must first recognize the need to seek help, find an appropriate medical professional or counselor, and possess the financial means and/or health insurance to pay for their appointment. In many low-income communities or marginalized populations where significant stigma against mental health exists, this process is far too arduous and the perceived importance far too low to be worth it. As such, we believe that implementing this online AI-based screening tool would be particularly beneficial for individuals in such communities.

Furthermore, this type of social media data-driven artificial intelligence approach towards health monitoring has potential applications beyond depression and suicidality - in the future it may also be applied towards a number of other conditions, such as eating disorders, PTSD, and anxiety. Such e-health interventions will be a boon not

just for the individuals affected, but also for those in medical professions. Having the aid of artificial intelligence tools will not only increase the accuracy of their diagnoses but also lessen their manual workloads, freeing up more time and energy for other pursuits, such as medical research.

5.0.1 Limitations

Our study is limited by a number of factors, including but not limited to a small and sparse dataset, selection bias, lack of introspective modeling, and limited generalization.

Deep neural networks perform best when trained on large datasets, and are quite prone to overfitting otherwise. Although we tried our best in the present study to impose regularization to prevent this, we cannot be sure that our results will generalize when applied to larger populations and new datasets. Furthermore, some of our data points may have been quite sparse, consisting of Facebook feeds with very few or short posts.

In addition, it is likely that the majority of our sample consisted of members of the Princeton community, since the survey was first circulated within the University email system before being more widely publicized. It is unclear what kinds of effects this bias may have had on the results, but the Princeton population is certainly not representative of the general population. Also, as aforementioned, the data likely suffers from significant voluntary response bias. Due to the intensely personal nature of the data collected and the lack of a monetary incentive, individuals may be far more likely to participate in the study if they have special interest in mental health research and/or have had past experiences with depression and suicidal ideation. Lastly, as is the case with most survey research studies, the data may also suffer from response bias as a result of participants misrepresenting their true responses. Survey respondents may want to provide responses that are deemed more socially acceptable, even when

their responses are anonymous. Indeed, in a number of post-survey communications between participants and the researcher, some participants expressed a desire for the researcher to not worry about them, even though the researcher did not have direct access to their survey responses since the data was anonymized.

Data limitations aside, this study is also limited in the types of qualitative insights it can provide, since neural networks are not introspective models, so to speak. Although deep neural networks are often quite effective at learning to complete a number of different classification tasks, they typically act as black boxes, learning a complex set of features that we do not intuitively understand the meaning of. It is typically unclear how these features are related to the prediction of the label, which prevents us from making useful psychiatric inferences.

Lastly, our suicidality prediction model may not be as generalizable as desired, since we focused on predicting suicidal ideation rather than actual suicide attempts. In fact, research indicates that only a minority of individuals who consider committing suicide actually proceed to do so [98]. Although both groups would benefit from receiving psychiatric care, our model does not offer a way of triaging between those who are at moderate versus more urgent risk of attempting suicide.

5.0.2 Future Work

As such, it is evident that there are a number of ways in which our research is limited and may be further extended in the near future. Firstly, as aforementioned in the discussion of the limitations of this study, it would be helpful to develop a model to predict not just a binary classification of suicidal versus non-suicidal, but also the degree of severity of suicidality. This can be helpful not just in the context of developing mental health monitoring tools for use on social media websites, but also for congested hotlines and crisis centers. Having the capacity to accurately triage incoming callers is crucial for effective resource allocation and can help ensure that

those who have the most urgent crises will receive help first.

Furthermore, our model may also be improved by gathering a larger dataset and taking better advantage of the richness of the data. In our study we also collected participants' Instagram and Twitter handles but did not take advantage of this data because not enough participants submitted this information. Training the classifier on this information in addition to the data from Facebook may increase its accuracy, as well as reveal interesting insights about behavioral differences across disparate social networks. In doing so we might also encode additional information about participants' social media activity beyond the textual content of their posts, such as timestamps, number of likes, number of reshares, and more. Effectively incorporating such multifarious sources of information into a single feature vector or matrix is a non-trivial task, but is worth conducting further research into.

Appendix A

Informed Consent Form

ADULT CONSENT FORM
PRINCETON UNIVERSITY

TITLE OF RESEARCH: *Deep-Learning Detection of Depression and Suicidal Ideation from Social Media Activity*

PRINCIPAL INVESTIGATOR: *Angelica Chen*

PRINCIPAL INVESTIGATOR'S DEPARTMENT: *Department of Computer Science*

You are being invited to take part in a research study. Before you decide to participate in this study, it is important that you understand why the research is being done and what it will involve. Please take the time to read the following information carefully. Please ask the researcher if there is anything that is not clear or if you need more information.

Purpose of the research:

The purpose of this research is to develop a machine-learning model for detecting the presence of (or lack thereof) depression and/or suicidal ideation from an individual's social media activity. You are being asked to participate in this study to provide a source of labeled data for the model to be trained and evaluated on.

Study Procedures:

All participants in this study must be 18 years old or older.

In this study, you will be asked to answer 26 multiple-choice questions about your current mental health status. You will also be asked to provide access to the public and friend-viewable content on your social media pages (Facebook, Twitter, or Instagram) from the last year. "Friend-viewable content" is any content you have posted, including but not limited to statuses, Tweets, photos, and comments, for which the privacy settings are set such that any of your friends or followers on the same site may view the content. No private data, such as Facebook posts set to the privacy setting "Only Me," will be collected.

Research personnel will be analyzing both the participants' responses to the survey questions and their social media data. This data will be anonymized and used to train various machine-learning models to predict level of depression and/or suicidal ideation. It will also be used in the evaluation of the models.

Your total expected time commitment for this study is: 20 minutes

In the case that your survey responses indicate that you may be moderately to severely depressed or experiencing suicidal ideation, you will be automatically notified upon completion of the survey. We advise that you contact Princeton Counseling & Psychological Services at (609) 258-3141 or (609) 258-3285, the National Suicide Prevention Lifeline at (800) 273-8255, or your local hospital emergency room. In addition, the following websites offer online tools for locating mental healthcare professionals in your local area:

- National Institute of Mental Health (NIMH): www.nimh.nih.gov (1-866-615-6464)

- Substance Abuse and Mental Health Services Administration (SAMHSA): https://findtreatment.samhsa.gov (1-800-662-4357)

Benefits and Risks:

Participants will not benefit directly from the study, but will be contributing to the advancement of mental health and applied machine learning research.

Although all possible efforts will be made to securely store and anonymize the data, there is still some risk that a participant's identity may be inferred from their social media data. However, all research personnel will keep this information confidential and will not release or publish any individual-level results.

Confidentiality:

All records from this study will be anonymized and kept confidential. Your responses will be kept private, and we will not include any information that will make it possible to identify you in any report we might publish. Research records will be stored securely on encrypted, password-protected servers. Furthermore, only a qualified psychiatrist (Dr. Chin, director of CPS) will have access to the human-readable form of your data. The other research personnel (Angelica Chen and Professor Sebastian Seung) will only have access to numerically encoded data that is not interpretable by the human eye.

Compensation:

Participants are asked to volunteer their time and data – no compensation will be provided.

Who to contact with questions/concerns:

1. PRINCIPAL INVESTIGATOR:

   Angelica Chen (ac17@princeton.edu)

2. If you have questions regarding your rights as a research subject, or if problems arise which you do not feel you can discuss with the Investigator, please contact the Institutional Review Board at:

   Assistant Director, Research Integrity and Assurance
   Phone: (609) 258-8543
   Email: irb@princeton.edu

3. For mental health concerns:

   Princeton Counseling & Psychological Services
   Phone: (609) 258-3141 or (609) 258-3285
   Website: https://uhs.princeton.edu/counseling-psychological-services

   National Suicide Prevention Lifeline
   Phone: (800) 273-8255

   National Institute of Mental Health (NIMH)
   Phone: (866) 615-6464

Website: www.nimh.nih.gov

Substance Abuse and Mental Health Services Administration (SAMHSA)
National helpline: (800) 662-4357
Website: https://findtreatment.samhsa.gov

---

3. I understand the information that was presented and that:

   A. My participation is voluntary, and I may withdraw my consent and discontinue participation in the project at any time. My refusal to participate will not result in any penalty.

   B. I do not waive any legal rights or release Princeton University, its agents, or you from liability for negligence.

□ By clicking this checkbox, I hereby give my consent to be the subject of your research.

Appendix B

Beck Depression Inventory (BDI)

Beck Depression Inventory.
For each question, please select the choice you most agree with.

1.
0 I do not feel sad.
1 I feel sad
2 I am sad all the time and I can't snap out of it.
3 I am so sad and unhappy that I can't stand it.

2.
0 I am not particularly discouraged about the future.
1 I feel discouraged about the future.
2 I feel I have nothing to look forward to.
3 I feel the future is hopeless and that things cannot improve.

3.
0 I do not feel like a failure.
1 I feel I have failed more than the average person.
2 As I look back on my life, all I can see is a lot of failures.
3 I feel I am a complete failure as a person.

4.
0 I get as much satisfaction out of things as I used to.
1 I don't enjoy things the way I used to.
2 I don't get real satisfaction out of anything anymore.
3 I am dissatisfied or bored with everything.

5.
0 I don't feel particularly guilty
1 I feel guilty a good part of the time.
2 I feel quite guilty most of the time.
3 I feel guilty all of the time.

6.
0 I don't feel I am being punished.
1 I feel I may be punished.
2 I expect to be punished.
3 I feel I am being punished.

7.
0 I don't feel disappointed in myself.
1 I am disappointed in myself.
2 I am disgusted with myself.
3 I hate myself.

8.

0 I don't feel I am any worse than anybody else.
1 I am critical of myself for my weaknesses or mistakes.
2 I blame myself all the time for my faults.
3 I blame myself for everything bad that happens.

9.
0 I don't have any thoughts of killing myself.
1 I have thoughts of killing myself, but I would not carry them out.
2 I would like to kill myself.
3 I would kill myself if I had the chance.

10.
0 I don't cry any more than usual.
1 I cry more now than I used to.
2 I cry all the time now.
3 I used to be able to cry, but now I can't cry even though I want to.

11.
0 I am no more irritated by things than I ever was.
1 I am slightly more irritated now than usual.
2 I am quite annoyed or irritated a good deal of the time.
3 I feel irritated all the time.

12.
0 I have not lost interest in other people.
1 I am less interested in other people than I used to be.
2 I have lost most of my interest in other people.
3 I have lost all of my interest in other people.

13.
0 I make decisions about as well as I ever could.
1 I put off making decisions more than I used to.
2 I have greater difficulty in making decisions more than I used to.
3 I can't make decisions at all anymore.

14.
0 I don't feel that I look any worse than I used to.
1 I am worried that I am looking old or unattractive.
2 I feel there are permanent changes in my appearance that make me look unattractive
3 I believe that I look ugly.

15.
0 I can work about as well as before.
1 It takes an extra effort to get started at doing something.
2 I have to push myself very hard to do anything.

3 I can't do any work at all.

16.
0 I can sleep as well as usual.
1 I don't sleep as well as I used to.
2 I wake up 1-2 hours earlier than usual and find it hard to get back to sleep.
3 I wake up several hours earlier than I used to and cannot get back to sleep.

17.
0 I don't get more tired than usual.
1 I get tired more easily than I used to.
2 I get tired from doing almost anything.
3 I am too tired to do anything.

18.
0 My appetite is no worse than usual.
1 My appetite is not as good as it used to be.
2 My appetite is much worse now.
3 I have no appetite at all anymore.

19.
0 I haven't lost much weight, if any, lately.
1 I have lost more than five pounds.
2 I have lost more than ten pounds.
3 I have lost more than fifteen pounds.

20.
0 I am no more worried about my health than usual.
1 I am worried about physical problems like aches, pains, upset stomach, or constipation.
2 I am very worried about physical problems and it's hard to think of much else.
3 I am so worried about my physical problems that I cannot think of anything else.

21.
0 I have not noticed any recent change in my interest in sex.
1 I am less interested in sex than I used to be.
2 I have almost no interest in sex.
3 I have lost interest in sex completely.

Appendix C

Columbia-Suicide Severity Rating Scale (C-SSRS)

Columbia-Suicide Severity Rating Scale
For each question select Yes or No under both "Lifetime" or "Past 6 months"

1. Have you wished you were dead or wished you could go to sleep and not wake up?

Lifetime:
Yes No
☐ ☐
Past 6 months:
Yes No
☐ ☐

2. Have you actually had any thoughts of killing yourself?

Lifetime:
Yes No
☐ ☐
Past 6 months:
Yes No
☐ ☐

3. Have you been thinking about how you might do this?

Lifetime:
Yes No
☐ ☐
Past 6 months:
Yes No
☐ ☐

4. Have you had these thoughts and had some intention of acting on them?

Lifetime:
Yes No
☐ ☐
Past 6 months:
Yes No
☐ ☐

5. Have you started to work out or worked out the details of how to kill yourself? Do you intend to carry out this plan?

Lifetime:
Yes No
☐ ☐
Past 6 months:
Yes No
☐ ☐

Bibliography

[1] About.

[2] About the C-SSRS Scale.

[3] IASP - World Suicide Prevention Day - September 10, 2017 - International Association for Suicide Prevention.

[4] Natural Language Toolkit NLTK 3.0 documentation.

[5] Suicide and Depression.

[6] Symptoms of Depression.

[7] Ten Leading Causes of Death and Injury.

[8] Psychiatric risk factors for adolescent suicide: A case-control study. *Journal of the American Academy of Child & Adolescent Psychiatry*, 32(3):521 – 529, 1993.

[9] Suicide Prevention Among Active Duty Air Force Personnel – United States, 1990-1999. Technical Report 46, Centers for Disease Control and Intervention, November 1999.

[10] Depression and Suicide Risk. Technical report, American Association of Suicidology, 2014.

[11] Suicide Facts at a Glance. Technical report, Centers for Disease Control and Intervention, 2015.

[12] Detecting and treating suicide ideation in all settings. Technical Report 56, The Joint Commission, February 2016.

[13] Suicide and Self-Inflicted Injury, October 2016.

[14] Anna-Mari Aalto, Marko Elovainio, Mika Kivimki, Antti Uutela, and Sami Pirkola. The beck depression inventory and general health questionnaire as measures of depression in the general population: A validation study using the composite international diagnostic interview as the gold standard. *Psychiatry Research*, 197(12):163 – 171, 2012.

[15] Esben Agerbo, Merete Nordentoft, and Preben Bo Mortensen. Familial, psychiatric, and socioeconomic risk factors for suicide in young people: nested case-control study. *BMJ*, 325(7355):74, 2002.

[16] Brian K Ahmedani, Gregory E Simon, Christine Stewart, Arne Beck, Beth E Waitzfelder, Rebecca Rossom, Frances Lynch, Ashli Owen-Smith, Enid M Hunkeler, Ursula Whiteside, et al. Health care contacts in the year before suicide death. *Journal of general internal medicine*, 29(6):870–877, 2014.

[17] Brian K Ahmedani, Christine Stewart, Gregory E Simon, Frances Lynch, Christine Y Lu, Beth E Waitzfelder, Leif I Solberg, Ashli A Owen-Smith, Arne Beck, Laurel A Copeland, et al. Racial/ethnic differences in healthcare visits made prior to suicide attempt across the united states. *Medical care*, 53(5):430, 2015.

[18] J. Angst, F. Angst, and H. H. Stassen. Suicide risk in patients with major depressive disorder. *The Journal of Clinical Psychiatry*, 60 Suppl 2:57–62; discussion 75–76, 113–116, 1999.

[19] Ella Arensman. Suicide prevention in an international context, 2017.

[20] Priya Bajaj, Elena Borreani, Pradip Ghosh, Caroline Methuen, Melissa Patel, and Michael Joseph. Screening for suicidal thoughts in primary care: the views of patients and general practitioners. *Mental Health in Family Medicine*, 5(4):229, 2008.

[21] Lisa J. Barney, Kathleen M. Griffiths, Anthony F. Jorm, and Helen Christensen. Stigma about depression and its impact on help-seeking intentions. *Australian & New Zealand Journal of Psychiatry*, 40(1):51–54, 2006. PMID: 16403038.

[22] Tarun Bastiampillai, Stephen Allison, and Sherry Chan. Is depression contagious? the importance of social networks and the implications of contagion theory. *Australian & New Zealand Journal of Psychiatry*, 47(4):299–303, 2013. PMID: 23568155.

[23] Aaron T. Beck, Robert A. Steer, and Margery G. Carbin. Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, 8(1):77 – 100, 1988.

[24] Aaron T Beck, Calvin H Ward, Mock Mendelson, Jeremiah Mock, and John ERBAUGH. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571, 1961.

[25] Alan L Berman. Estimating the population of survivors of suicide: Seeking an evidence base. *Suicide and Life-Threatening Behavior*, 41(1):110–116, 2011.

[26] Paul Best, Roger Manktelow, and Brian Taylor. Online communication, social media and adolescent wellbeing: A systematic narrative review. *Children and Youth Services Review*, 41:27 – 36, 2014.

[27] James M Bolton, Wendy Au, William D Leslie, Patricia J Martens, Murray W Enns, Leslie L Roos, Laurence Y Katz, Holly C Wilcox, Annette Erlangsen, Dan Chateau, et al. Parents bereaved by offspring suicide: a population-based longitudinal case-control study. *JAMA psychiatry*, 70(2):158–167, 2013.

[28] James M Bolton, David Gunnell, and Gustavo Turecki. Suicide risk assessment and intervention in people with mental illness. *BMJ*, 351:h4978, 2015.

[29] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.

[30] F. Stephen Bridges and Julie C. Kunselman. Gun availability and use of guns for suicide, homicide, and murder in Canada. *Perceptual and Motor Skills*, 98(2):594–598, April 2004.

[31] Gregory K Brown, Aaron T Beck, Robert A Steer, and Jessica R Grisham. Risk factors for suicide in psychiatric outpatients: a 20-year prospective study. *Journal of consulting and clinical psychology*, 68(3):371, 2000.

[32] Moira Burke, Cameron Marlow, and Thomas Lento. Social network activity and social well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1909–1912, New York, NY, USA, 2010. ACM.

[33] Pete Burnap, Walter Colombo, and Jonathan Scourfield. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 75–84. ACM, 2015.

[34] Katie A Busch, Jan Fawcett, and Douglas G Jacobs. Clinical correlates of inpatient suicide. *Journal of Clinical Psychiatry*, 64(1):14–19, 2003.

[35] John V Campo. Youth suicide prevention: does access to care matter? *Current opinion in pediatrics*, 21(5):628–634, 2009.

[36] Kim Capp, Frank P. Deane, and Gordon Lambert. Suicide prevention in aboriginal communities: application of community gatekeeper training. *Australian and New Zealand Journal of Public Health*, 25(4):315–321, 2001.

[37] A. Carlsten, P. Allebeck, and L. Brandt. Are suicide rates in sweden associated with changes in the prescribing of medicines? *Acta Psychiatrica Scandinavica*, 94(2):94–100, 1996.

[38] Munmun De Choudhury, Michael Camon, and Scott Counts. Predicting depression via social media.

[39] Helen Christensen, Philip J Batterham, and Bridianne O'Dea. E-health interventions for suicide prevention. *International journal of environmental research and public health*, 11(8):8193–8212, 2014.

[40] Sara L. Cooper, Dennis Lezotte, Jillian Jacobellis, and Carolyn DiGuiseppi. Does availability of mental health resources prevent recurrent suicidal behavior? an ecological analysis. *Suicide and Life-Threatening Behavior*, 36(4):409–417, 2006.

[41] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[42] Lorenzo Coviello, Yunkyu Sohn, Adam D. I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, and James H. Fowler. Detecting emotional contagion in massive social networks. *PLOS ONE*, 9(3):1–6, 03 2014.

[43] P Crome. The toxicity of drugs used for suicide. *Acta Psychiatrica Scandinavica*, 87(S371):33–37, 1993.

[44] AE Crosby, L Ortega, and C Melanson. Self-directed violence surveillance: uniform definitions and recommended data elements. atlanta, georgia: Cdc; 2011, 2016.

[45] Wendi Cross, Monica M. Matthieu, Julie Cerel, and Kerry L. Knox. Proximate outcomes of gatekeeper training for suicide prevention in the workplace. *Suicide and Life-Threatening Behavior*, 37(6):659–670, 2007.

[46] Sally C. Curtin and Margaret Warner. Increase in suicide in the United States, 1999-2014. Technical Report 241, National Center for Health Statistics, 2016.

[47] Shaffer D, Gould MS, Fisher P, and et al. Psychiatric diagnosis in child and adolescent suicide. *Archives of General Psychiatry*, 53(4):339–348, 1996.

[48] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2098–2110. ACM, 2016.

[49] DIEGO DE LEO. Why are we not getting any closer to preventing suicide? *The British Journal of Psychiatry*, 181(5):372–374, 2002.

[50] Justin T. Denney, Richard G. Rogers, Patrick M. Krueger, and Tim Wadsworth. Adult suicide mortality in the united states: Marital status, family size, socioeconomic status, and differences by sex*. *Social Science Quarterly*, 90(5):1167–1185, 2009.

[51] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLOS ONE*, 6(12):e26752–, 12 2011.

[52] Maeve Duggan and Dana Page. Social Media Update 2016. Technical report, Pew Research Center, November 2016.

[53] William P. Evans, Laura Davidson, and Lorie Sicafuse. Someone to listen: Increasing youth help-seeking behavior through a text-based crisis line for youth. *Journal of Community Psychology*, 41(4):471–487, 2013.

[54] B.A. Ezra Golberstein, Ph.D. Daniel Eisenberg, and B.A. Sarah E. Gollust. Perceived stigma and mental health care seeking. *Psychiatric Services*, 59(4):392–399, 2008. PMID: 18378838.

[55] Bradley N. Gaynes, Suzanne L. West, Carol A. Ford, Paul Frame, Jonathan Klein, Kathleen N. Lohr, and U.S. Preventive Services Task Force. Screening for suicide risk in adults: a summary of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 140(10):822–835, May 2004.

[56] Polly Y Gipson, Prachi Agarwala, Kiel J Opperman, Adam Horwitz, and Cheryl A King. Columbia-suicide severity rating scale: Predictive validity with adolescent psychiatric emergency patients. *Pediatric emergency care*, 31(2):88–94, 02 2015.

[57] E Goodman. The role of socioeconomic status gradients in explaining differences in us adolescents' health. *American Journal of Public Health*, 89(10):1522–1528, 2017/04/11 1999.

[58] Madelyn S. Gould, Wendi Cross, Anthony R. Pisani, Jimmie Lou Munfakh, and Marjorie Kleinman. Impact of applied suicide intervention skills training on the national suicide prevention lifeline. *Suicide and Life-Threatening Behavior*, 43(6):676–691, 2013.

[59] Madelyn S. Gould, John Kalafat, Jimmie Lou HarrisMunfakh, and Marjorie

Kleinman. An Evaluation of Crisis Hotline Outcomes Part 2: Suicidal Callers. *Suicide and Life-Threatening Behavior*, 37(3):338–352, June 2007.

[60] Madelyn S. Gould, Jimmie L. H. Munfakh, Marjorie Kleinman, and Alison M. Lake. National suicide prevention lifeline: Enhancing mental health care for suicidal individuals and other people in crisis. *Suicide and Life-Threatening Behavior*, 42(1):22–35, 2012.

[61] Amelia Gulliver, Kathleen M. Griffiths, and Helen Christensen. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. *BMC Psychiatry*, 10(1):113, 2010.

[62] Bo Han, Paul Cook, and Timothy Baldwin. Lexical normalization for social media text. *ACM Trans. Intell. Syst. Technol.*, 4(1):5:1–5:27, February 2013.

[63] ULRICH HEGERL, DAVID ALTHAUS, ARMIN SCHMIDTKE, and GUENTER NIKLEWSKI. The alliance against depression: 2-year evaluation of a community-based intervention to reduce suicidality. *Psychological Medicine*, 36(9):12251233, 2006.

[64] Ulrich Hegerl, Roland Mergl, Inga Havers, Armin Schmidtke, Hartmut Lehfeld, Günter Niklewski, and David Althaus. Sustainable effects on suicidality were found for the nuremberg alliance against depression. *European Archives of Psychiatry and Clinical Neuroscience*, 260(5):401–406, 2010.

[65] Ulrich Hegerl, Christine Rummel-Kluge, Airi Vrnik, Ella Arensman, and Nicole Koburger. Alliances against depression a community based approach to target depression and to prevent suicidal behaviour. *Neuroscience & Biobehavioral Reviews*, 37(10, Part 1):2404 – 2409, 2013. Discovery research in Neuropsychiatry - anxiety, depression and schizophrenia in focus.

[66] Heidi Hjelmeland, Joseph Osafo, Charity S. Akotia, and Birthe L. Knizek. The Law Criminalizing Attempted Suicide in Ghana: The Views of Clinical Psychologists, Emergency Ward Nurses, and Police Officers. *Crisis*, 35(2):132–136, March 2014.

[67] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[68] Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilla Ovesdo er Alm. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. *ACL 2014*, page 107, 2014.

[69] Xiaomeng Hu, Andrew Kim, Nicholas Siwek, and David Wilder. The facebook paradox: Effects of facebooking on individuals social relationships and psychological well-being. *Frontiers in Psychology*, 8:87, 2017.

[70] Michael Isaac, Brenda Elias, Laurence Y Katz, Shay-Lee Belik, Frank P Deane, Murray W Enns, and Jitender Sareen. Gatekeeper training as a preventative intervention for suicide: a systematic review. *The Canadian Journal of Psychiatry*, 54(4):260–268, 2009.

[71] Asarnow J, Jaycox LH, Duan N, and et al. Effectiveness of a quality improvement intervention for adolescent depression in primary care clinics: A randomized controlled trial. *JAMA*, 293(3):311–319, 2005.

[72] Ludwig J and Cook PJ. Homicide and suicide rates associated with implementation of the brady handgun violence prevention act. *JAMA*, 284(5):585–591, 2000.

[73] Fein JA, Pailler ME, Barg FK, and et al. Feasibility and effects of a web-based adolescent psychiatric assessment administered by clinical staff in the

pediatric emergency department. *Archives of Pediatrics & Adolescent Medicine*, 164(12):1112–1117, 12 2010.

[74] Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. Tracking suicide risk factors through twitter in the us. *Crisis*, 2014.

[75] Deborah L Kahn and David Lester. Efforts to decriminalize suicide in ghana, india and singapore. *Suicidology Online*, 4:96–104, 2013.

[76] Nav Kapur, Sarah Steeg, Roger Webb, Matthew Haigh, Helen Bergen, Keith Hawton, Jennifer Ness, Keith Waters, and Jayne Cooper. Does clinical management improve outcomes following self-harm? results from the multicentre study of self-harm in england. *PloS one*, 8(8):e70434, 2013.

[77] Michal Kawulok, M. Emre Celebi, and Bogdan Smolka, editors. *Advances in Face Detection and Facial Image Analysis*. Springer International Publishing, Cham, 2016. DOI: 10.1007/978-3-319-25958-1.

[78] David C. R. Kerr, Brandon Gibson, Leslie D. Leve, and David S. DeGarmo. Young adult follow-up of adolescent girls in juvenile justice using the columbia suicide severity rating scale. *Suicide and Life-Threatening Behavior*, 44(2):113–129, 2014.

[79] Ronald C. Kessler, Patricia Berglund, Guilherme Borges, Matthew Nock, and Philip S. Wang. Trends in suicide ideation, plans, gestures, and attempts in the United States, 1990-1992 to 2001-2003. *JAMA*, 293(20):2487–2495, May 2005.

[80] M. M Khan. Suicide prevention and developing countries. *Journal of the Royal Society of Medicine*, 98(10):459–463, October 2005.

[81] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[82] Kerry L. Knox, Yeates Conwell, and Eric D. Caine. If suicide is a public health problem, what are we doing to prevent it? *American Journal of Public Health*, 94(1):37–45, 2017/04/20 2004.

[83] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[84] L Lasa, J.L Ayuso-Mateos, J.L Vázquez-Barquero, F.J Díez-Manrique, and C.F Dowrick. The use of the beck depression inventory to screen for depression in the general population: a preliminary analysis. *Journal of Affective Disorders*, 57(13):261 – 265, 2000.

[85] D. Lester and A. Leenaars. Suicide rates in Canada before and after tightening firearm control laws. *Psychological Reports*, 72(3 Pt 1):787–790, June 1993.

[86] David Lester. Decriminalization of suicide in seven nations and suicide rates. *Psychological reports*, 91(3):898–898, 2002.

[87] Glyn Lewis and Andy Sloggett. Suicide, deprivation, and unemployment: record linkage study. *BMJ*, 317(7168):1283–1286, 1998.

[88] Selma A. Lewis, Jim Johnson, Patricia Cohen, Marc Garcia, and Carmen Noemi Velez. Attempted suicide in youth: Its relationship to school achievement, educational goals, and socioeconomic status. *Journal of Abnormal Child Psychology*, 16(4):459–471, 1988.

[89] Zhuoyang Li, Andrew Page, Graham Martin, and Richard Taylor. Attributable risk of psychiatric and socio-economic factors for suicide from individual-level,

population-based studies: A systematic review. *Social Science & Medicine*, 72(4):608–616, 2011.

[90] Fei Liu, Fuliang Weng, and Xiao Jiang. A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 1035–1044. Association for Computational Linguistics, 2012.

[91] Colin Loftin, David McDowall, Brian Wiersema, and Talbert J. Cottey. Effects of restrictive licensing of handguns on homicide and suicide in the district of columbia. *New England Journal of Medicine*, 325(23):1615–1620, 1991. PMID: 1669841.

[92] Jason B. Luoma, Catherine E. Martin, and Jane L. Pearson. Contact with mental health and primary care providers before suicide: A review of the evidence. *American Journal of Psychiatry*, 159(6):909–916, 2002. PMID: 12042175.

[93] J. John Mann, Alan Apter, Jose Bertolote, Annette Beautrais, Dianne Currier, Ann Haas, Ulrich Hegerl, Jouko Lonnqvist, Kevin Malone, Andrej Marusic, Lars Mehlum, George Patton, Michael Phillips, Wolfgang Rutz, Zoltan Rihmer, Armin Schmidtke, David Shaffer, Morton Silverman, Yoshitomo Takahashi, Airi Varnik, Danuta Wasserman, Paul Yip, and Herbert Hendin. Suicide Prevention Strategies: A Systematic Review. *JAMA*, 294(16):2064, October 2005.

[94] Monica M. Matthieu, Wendi Cross, Alfonso R. Batres, Charles M. Flora, and Kerry L. Knox. Evaluation of gatekeeper training for suicide prevention in veterans. *Archives of Suicide Research*, 12(2):148–154, 02 2008.

[95] Carolina de Mello-Santos, Jos Manuel Bertolote, and Yuan-Pang Wang. Epidemiology of suicide in Brazil (1980-2000): characterization of age and gender

rates of suicide. *Revista Brasileira De Psiquiatria (Sao Paulo, Brazil: 1999)*, 27(2):131–134, June 2005.

[96] J. Milton, B. Ferguson, and T. Mills. Risk assessment and suicide prevention in primary care. *Crisis*, 20(4):171–177, 1999.

[97] AS Nielsen and B Nielsen. Pattern of choice in preparation of attempted suicide by poisoning–with particular reference to changes in the pattern of prescriptions. *Ugeskrift for laeger*, 154(28):1972–1976, 1992.

[98] Matthew K Nock, Guilherme Borges, Evelyn J Bromet, Jordi Alonso, Matthias Angermeyer, Annette Beautrais, Ronny Bruffaerts, Wai Tat Chiu, Giovanni De Girolamo, Semyon Gluzman, et al. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry*, 192(2):98–105, 2008.

[99] Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188, 2015.

[100] A. Ohberg, J. Lonnqvist, S. Sarna, E. Vuori, and A. Penttila. Trends and availability of suicide methods in Finland. Proposals for restrictive measures. *The British Journal of Psychiatry: The Journal of Mental Science*, 166(1):35–43, January 1995.

[101] World Health Organization et al. Public health action for the prevention of suicide: a framework. 2012.

[102] David M Paperny, June Y Aono, Robert M Lehman, Sherrel L Hammar, and Joseph Risser. Computer-assisted detection and intervention in adolescent high-risk health behaviors. *The Journal of pediatrics*, 116(3):456–462, 1990.

[103] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[104] Michael R Phillips, Gonghuan Yang, Yanping Zhang, Lijun Wang, Huiyu Ji, and Maigeng Zhou. Risk factors for suicide in china: a national case-control psychological autopsy study. *The Lancet*, 360(9347):1728–1736, 11 2002.

[105] Sami Pirkola, Reijo Sund, Eila Sailas, and Kristian Wahlbeck. Community mental-health services and suicide rate in finland: a nationwide small-area analysis. *The Lancet*, 373(9658):147 – 153, 2009.

[106] Kelly Posner, Gregory K. Brown, Barbara Stanley, David A. Brent, Kseniya V. Yershova, Maria A. Oquendo, Glenn W. Currier, Glenn A. Melvin, Laurence Greenhill, Sa Shen, and J. John Mann. The columbia–suicide severity rating scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry*, 168(12):1266–1277, 2017/04/16 2011.

[107] Rajeev Ramchand, Lisa Jaycox, Pat Ebener, Mary Lou Gilbert, Dionne Barnes-Proby, and Prodyumna Goutam. Characteristics and proximal outcomes of calls made to suicide crisis hotlines in california. *Crisis*, 38(1):26–35, 2017. PMID: 27338290.

[108] Rajeev Ranjan, Rajesh Sagar, Anju Dhawan, Saurabh Kumar, and RamanDeep Pattanayak. (De-) criminalization of attempted suicide in India: A review. *Industrial Psychiatry Journal*, 23(1):4, 2014.

[109] Kessler RC, Borges G, and Walters EE. Prevalence of and risk factors for lifetime suicide attempts in the national comorbidity survey. *Archives of General Psychiatry*, 56(7):617–626, 07 1999.

[110] A. G. Reece, A. J. Reagan, K. L. M. Lix, P. Sheridan Dodds, C. M. Danforth, and E. J. Langer. Forecasting the onset and course of mental illness with Twitter data. *ArXiv e-prints*, August 2016.

[111] Andrew G. Reece and Christopher M. Danforth. Instagram photos reveal predictive markers of depression. *CoRR*, abs/1608.03282, 2016.

[112] Professor N. Retterstl. Norwegian data on death due to overdose of antidepressants. *Acta Psychiatrica Scandinavica*, 80(S354):61–68, 1989.

[113] P. Richter, J. Werner, A. Heerlein, A. Kraus, and H. Sauer. On the validity of the beck depression inventory. *Psychopathology*, 31(3):160–168, 1998.

[114] Philip Rodgers. Review of the Applied Suicide Intervention Skills Training Program (ASIST). Technical report, LivingWorks Education Inc., April 2010.

[115] C. Roehrig. Mental Disorders Top The List Of The Most Costly Conditions In The United States: $201 Billion. *Health Affairs*, 35(6):1130–1135, June 2016.

[116] Leslie Roos, Jitender Sareen, and James M Bolton. Suicide risk assessment tools, predictive validity findings and utility today: time for a revamp? *Neuropsychiatry*, 3(5):483–495, 2013.

[117] Shekhar Saxena, Etienne G. Krug, Oleg Chestnov, and World Health Organization, editors. *Preventing suicide: a global imperative*. World Health Organization, Geneva, 2014.

[118] H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, 2014.

[119] Donald S. Shepard, Deborah Gurewich, Aung K. Lwin, Gerald A. Reed, and Morton M. Silverman. Suicide and suicidal attempts in the united states: Costs and policy implications. *Suicide and Life-Threatening Behavior*, 46(3):352–362, 2016.

[120] Ahmad Shojaei, Saadolah Moradi, Farshid Alaeddini, Mahmood Khodadoost, Abdolrazagh Barzegar, and Ali Khademi. Association between suicide method, and gender, age, and education level in iran over 20062010. *Asia-Pacific Psychiatry*, 6(1):18–22, 2014.

[121] J. Snowdon and L. Harris. Firearms suicides in Australia. *The Medical Journal of Australia*, 156(2):79–83, January 1992.

[122] M Speechley and KM Stavraky. The adequacy of suicide statistics for use in epidemiology and public health. *Canadian journal of public health = Revue canadienne de sante publique*, 82(1):3842, 1991.

[123] Robert A Steer, Thomas A Cavalieri, Douglas M Leonard, and Aaron T Beck. Use of the beck depression inventory for primary care to screen for major depression disorders. *General Hospital Psychiatry*, 21(2):106 – 111, 1999.

[124] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. 29(1):24–54.

[125] M. ten Have, R. de Graaf, J. Ormel, G. Vilagut, V. Kovess, and J. Alonso. Are attitudes towards mental health help-seeking associated with service use? results from the european study of epidemiology of mental disorders. *Social Psychiatry and Psychiatric Epidemiology*, 45(2):153–163, 2010.

[126] Tanya L. Tompkins, Jody Witt, and Nadia Abraibesh. Does a gatekeeper suicide prevention program work in a school setting? evaluating training outcome and

moderators of effectiveness. *Suicide and Life-Threatening Behavior*, 40(5):506–515, 2010.

[127] Lakshmi Vijayakumar, K. Nagaraj, Jane Pirkis, and Harvey Whiteford. Suicide in developing countries (1). *Crisis*, 26(3):104–111, 2005. PMID: 16276752.

[128] Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. A depression detection model based on sentiment analysis in micro-blog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 201–213. Springer, 2013.

[129] Camilla Wasserman, Christina W. Hoven, Danuta Wasserman, Vladimir Carli, Marco Sarchiapone, Susana Al-Halabí, Alan Apter, Judit Balazs, Julio Bobes, Doina Cosman, Luca Farkas, Dana Feldman, Gloria Fischer, Nadja Graber, Christian Haring, Dana Cristina Herta, Miriam Iosue, Jean-Pierre Kahn, Helen Keeley, Katja Klug, Jacklyn McCarthy, Alexandra Tubiana-Potiez, Airi Varnik, Peeter Varnik, Janina Žiberna, and Vita Poštuvan. Suicide prevention for youth - a mental health awareness program: lessons learned from the saving and empowering young lives in europe (seyle) intervention study. *BMC Public Health*, 12(1):776, 2012.

[130] Danuta Wasserman, Vladimir Carli, Camilla Wasserman, Alan Apter, Judit Balazs, Julia Bobes, Renata Bracale, Romuald Brunner, Cendrine Bursztein-Lipsicas, Paul Corcoran, Doina Cosman, Tony Durkee, Dana Feldman, Julia Gadoros, Francis Guillemin, Christian Haring, Jean-Pierre Kahn, Michael Kaess, Helen Keeley, Dragan Marusic, Bogdan Nemes, Vita Postuvan, Stella Reiter-Theil, Franz Resch, Pilar Sáiz, Marco Sarchiapone, Merike Sisask, Airi Varnik, and Christina W. Hoven. Saving and empowering young lives in europe (seyle): a randomized controlled trial. *BMC Public Health*, 10(1):192, 2010.

[131] Danuta Wasserman, Christina W Hoven, Camilla Wasserman, Melanie Wall, Ruth Eisenberg, Gerg Hadlaczky, Ian Kelleher, Marco Sarchiapone, Alan Apter, Judit Balazs, Julio Bobes, Romuald Brunner, Paul Corcoran, Doina Cosman, Francis Guillemin, Christian Haring, Miriam Iosue, Michael Kaess, Jean-Pierre Kahn, Helen Keeley, George J Musa, Bogdan Nemes, Vita Postuvan, Pilar Saiz, Stella Reiter-Theil, Airi Varnik, Peeter Varnik, and Vladimir Carli. School-based suicide prevention programmes: the {SEYLE} cluster-randomised, controlled trial. *The Lancet*, 385(9977):1536 – 1544, 2015.