Quantifying Suicidal Ideation via Language Usage on Social Media

Glen Coppersmith
glen@qntfy.io

Ryan Leary
ryan@qntfy.io

Eric Whyne
eric@qntfy.io

Tony Wood
tony@qntfy.io

Abstract

Suicide is a large and growing problem, yet relevant data to draw informed decisions and assess intervention strategies is sorely lacking, and often at least two years out of date. We analyze publicly available data to assess the viability of using it to provide more timely information. We examine quantifiable signals related to suicide attempts and suicidal ideation in the language of social media data. Our data consists of Twitter users who have attempted suicide and age- and gender-matched neurotypical controls and similarly matched clinically depressed users. We apply simple language modeling techniques to separate those users automatically, and examine what quantifiable signals allow them to function, tying them back to psychometrically validated concepts related to suicide. We then use these scalable classifiers with public social media data and open government data to suggest some direction for future epidemiological research. All this research is done with public data, though we take great care to protect the privacy of the users.

Key Words: Mental Health, Social Media, Suicidal Ideation, Suicide

1. Introduction

Suicide is a large and growing problem worldwide, but data to inform decisions is decided lacking. When data is available, it is often significantly delayed. Suicide is the second leading cause of death for teenagers and the leading cause of death for women ages 15-19 worldwide [World Health Organization et al., 2014], and among the top ten leading causes of death in the United States [Sullivan et al., 2013]. Certain groups are particularly at risk, with veterans having an elevated risk compared to the civilian population [Blow et al., 2012]. Moreover, the rates of suicides seem to be growing, with an increase of 28% in the civilian population of the United States between 1999 and 2010, with some groups notably higher than that (e.g., whites increased 40% and Native Americans 65%) [Sullivan et al., 2013].

While approximately 1% of people die by suicide, an estimated 2.7% attempt suicide, 3.1% make plans for suicide and 9.2% think about ending their own life, termed *suicidal ideation* [Nock et al., 2008]. Ultimately we seek ways to prevent suicides, but a better understanding of the underlying phenomenon is required to facilitate the design of and evaluation of effective interventions. This data is difficult to obtain, and when available it is typically delayed by at least two years – for example consider the BRFSS [Centers for Disease Control and Prevention (CDC), 2010]. Public and health policy decisions need higher quality population-level data in a more reasonble timeframe to be able to adequately plan and adjust responses. At the individual patient level, mental health professionals seeking to identify those at risk of suicide could benefit from a better understanding of suicidal ideation and intervention strategies augmented by automatic means to identify it. Intervention immediately prior to a suicide attempt (e.g., means restriction) is time consuming and costly. Earlier intervention aimed at the underlying causes of suicide could significantly scale the ability of mental healthcare professionals to help those at risk of suicide.

Broadly, we find and examine quantifiable signals related to suicide attempts and suicidal ideation in the language of social media data. Introspection on those quantifiable signals provides some insight as to how they work and how they might connect to known psychological phenomena related to suicidal ideation. We then go on to demonstrate what is made possible by scalable quantifiable measures of mental health, such as these classifiers, and insight they can provide us at both and individual and population level.

Why Social Media? Social media may *prima facie* seem like a strange choice to seek mental health signals, but upon closer inspection, it seems to be the digital version of more venerable, well-established methods. Having patients or their relatives keep journals of mood, thoughts, feelings, and occurrences has been a frequently-used technique in psychology research. These journals were often manually transcribed and codified by researchers or clinicians to get quantifiable information regarding the hypotheses studied – a time-consuming and expensive type of study to undertake, often yielding only tens of subjects. Social media, however, is providing the raw materials for the largest journaling study to date, with tens of millions of users with publicly available data.

Social media may be more useful for the study of mental health than journaling-type studies of yester-year. Physical health ailments can often be easily measured in a doctor's office or emergency room, but the equivalent mental health ailments are more difficult to assess, because many of the effective causes and symptoms relate to the patient's interaction with the rest of the world – almost by definition everything that happens *outside* of interaction with traditional healthcare professionals. Furthermore, data gathered via social media is already in digitized form, making it conducive to automated analysis. Looking forward, it also provides the technical means to interact with users. This opens up interesting avenues for scalable interventions (e.g., connecting a suicidal person to resources or peer support), but striking the balance between privacy and intervention will not be easy. While technology can support such interventions, it appears that the general population is not amenable to such interventions at the expense of privacy. Despite the potential for lives saved, the recent events surrounding the Samaritan's Radar App[1] are a cautionary tale.

Mental Health and Social Media There has been an explosion of recent work examining mental health signals through social media. Most of the work has focused on pervasive mental health concerns and psychological states, for example detecting depression [Schwartz et al., 2014, Resnik et al., 2013, De Choudhury et al., 2013a, De Choudhury et al., 2013b, Rosenquist et al., 2010, Ramirez-Esparza et al., 2008, Chung and Pennebaker, 2007], examining personality factors [Schwartz et al., 2013b, Park et al., 2015], or assessing psychological well-being [Schwartz et al., 2013a].

Alternate methods for obtaining data related to mental health conditions were introduced by Coppersmith et al., which widened the aperture of possible conditions to investigate [Coppersmith et al., 2015a, Coppersmith et al., 2014a, Coppersmith et al., 2014b]. These techniques enabled analysis of rarer conditions like schizophrenia [Mitchell et al., 2015], which affects an estimated 1% of the population of the United States [The National Institute of Mental Health, 2015]. Those rates are roughly equivalent to the suicide rate, and many times smaller than the estimated rate of suicidal ideation [Nock et al., 2008].

Suicide and suicidal ideation has been less well studied via social media. There has been some analysis of a suicide support forum and how the Werther or Papageno effects might be quantified using this data [Kumar et al., 2015]. The operative question is whether

[1] `http://www.samaritans.org/how-we-can-help-you/supporting-someone-online/samaritans-radar`

there are significant markers that can be inferred from a user's social media stream that predict, with high specificity, an impending suicide attempt. There has also been some work investigating the role that social media has in suicide clusters (among people in disparate geographies connected online) [Robertson et al., 2012].

While insightful at the level of individuals, some of the most powerful and interesting use cases of this sort of research inform population-level or epidemiological questions in a scalable manner. For example, previous work has demonstrated how web search queries can measure population level mental health trends [Yang et al., 2010, Ayers et al., 2013, Althouse et al., 2014]. Similar approaches using geolocated social media data have been able to predict heart disease mortality from language usage [Eichstaedt et al., 2015]. The power of these approaches are complementary to traditional survey methods, particularly in regards to better geographic and temporal granularity [Centers for Disease Control and Prevention (CDC), 2010].

2. Data

Finding data on individuals challenged with mental illness has traditionally been difficult, compounded by the deeply personal nature of one's mental health and the stigma placed upon mental health by society. Increasingly, however, people are turning to the anonymity and pseudo-anonymity of the Internet to discuss mental health issues, seeking help, seeking information, or seeking to help others similarly suffering. We previously used self-stated diagnoses to explore mental health conditions [Coppersmith et al., 2015a, Coppersmith et al., 2014a, Coppersmith et al., 2014b], and we apply similar methodology here to find users who discuss past suicide attempts on Twitter. To maintain the privacy of the individuals in the dataset, we do not present direct quotes from any data, nor any identifying information. All human annotators are full or part time employees of Qntfy, no data was ever presented to users outside of Qntfy, or via any crowdsourcing platforms.

For illustrative examples of tweets discussing past suicide attempts, see Table 1. Annotators read each of these tweets to determine if (1) the user is discussing their own suicide attempt and (2) if the exact date of the attempt can be determined. Provided that (1) and (2) are true, we use Twitter's API to obtain up to the last 3200 public tweets from this user and assert that they meet the following additional criteria before being included in this study: they have at least 100 tweets prior to the date of their suicide attempt, they tweet primarily in English, and have their privacy settings such that their tweets are public. We use data from 2014 to August 2015.

Control users were sampled from Twitter's 1% "spritzer" stream. Specifically, we took all users who tweeted publicly at least once during a two week period in November 2014, and gave them each equal probability of inclusion. We excluded any users who did not primarily tweet in English, who had less than 100 tweets (largely new or inactive accounts), more than 20k tweets (often automated or pseudo-automated accounts), or whose tweets were private or protected. Our depression users were drawn from the population made available for the CLPsych 2015 Shared Task [Coppersmith et al., 2015b]. Briefly, the selection criteria was similar to that used for people who attempt suicide. These were users who stated that they had been diagnosed at some point in their life with depression, validated by a human annotator. Importantly, the annotator is validating that *statement* that the user was diagnosed with depression appears genuine, rather than the *diagnosis* being genuine – no clinical evaluation took place to verify that these people *are* clinically depressed. We expect some of of this population of depressed users have or will attempt suicide. Thus, this population of users is contaminated with people who attempt suicide, so all comparisons made between depressed users and and people who attempt suicide will

Included: Author's suicide attempt descriptions with discernable date in the past
It's been 3 months since my suicide attempt, and I couldn't be happier!
November 16, 2012 was my last suicide attempt, if you must know.
I wish my suicide attempt on new years succeeded. blah.
Excluded: suicide attempt descriptions of others or with no discernable past date
It's hard to believe it's been almost a year since my last suicide attempt.
My brother's suicide attempt on halloween last year was just too much for me.
It's not worth it, @USER, I'm going to kill myself tomorrow.

Table 1: Paraphrased illustrative examples of suicide attempts included (top) and excluded (bottom) from our analysis.

be an underestimate of the true capability here.

Age and Gender: Estimation and Matching To isolate the signals most relevant to the suicide and suicidal ideation, we make every effort to eliminate other strong confounding signals. Psychology studies frequently match the age and gender of their subjects to remove confounds induced by those demographics. Though we do not have access to the true age and gender of these public Twitter users, we can estimate them using techniques and lexica graciously made available by the World Well-Being Project [Sap et al., 2014]. Specifically, we concatenate all of the tweets from a given user and obtain score for gender ($\hat{g}(u) \in [0, 1]$). All users above 0.5 are labeled Female, while those below 0.5 are labeled Male. Users where $\hat{g}(u) = 0.5$ would be randomly assigned a gender label, though this did not occur in practice. Similarly, we obtained an estimate for age ($\hat{a}(u)$).

See Figure 1 for a histogram of the estimated ages of our users who attempt suicide, separated by estimated gender. Notably, these users appear to be primarily young and female. This is not representative of all suicides seen throughout the country, where the most at-risk are middle-aged males. We discuss the specific implications of this in depth later.

In order to define an age- and gender-matched set of controls for use in the subsequent experiments, for each person who attemped suicide we find the user in the control group who has the same gender label and the closest age estimate. This selection was done without replacement. We repeated this procedure to find age- and gender-matched users from the depression group. The control and depression groups had significantly more users than the people who attempt suicide group, so a very close match can be found in each, with differences smaller than the reported resolution of the estimators ($\pm$ 2 days for control users, $\pm$ 3 months for depression users).

Public Geolocated Tweets To examine geographic trends related to suicide and suicidal ideation, we use public geolocated Twitter data. Sample users who publicly tweet and tag their posts with their current location. We aim to examine suicide and suicidal ideation at the state level, but users tend to move between states with some frequency. We simplify the user's movement into an estimate of which state they spend the most time in, termed their *Home Range* [Worton, 1989]. Specifically, we include users in this study who (1) primarily tweets in English, (2) have at least 30 geolocated tweets in the United States (3) have at least 100 tweets overall, and (4) have at least 30% of their geolocated tweets in a single

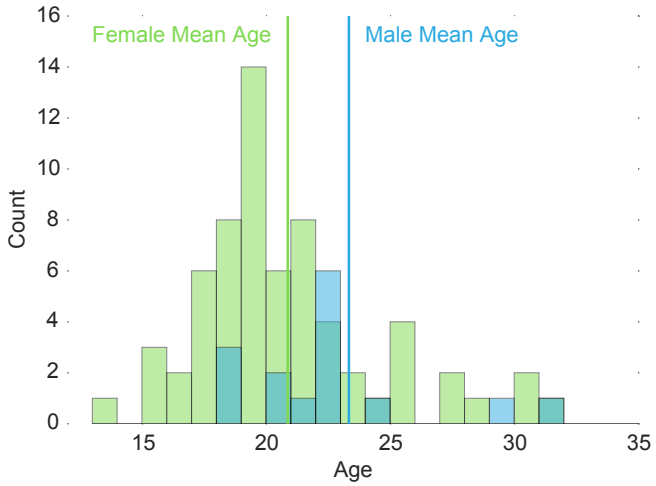

Figure 1: Histogram of the age of users with past suicide attempts, separated by gender. Females are in green, males are in blue. The mean age of each gender is denoted by the vertical lines.

state.

CDC Statistics for Deaths by Suicide We obtain data for suicide decedents by state and age from the Center for Disease Control[2] for 2013. Combined with demographic data from the 2010 Census we estimate the rate of suicides by age and gender for each state. Notably, these dates do not directly align with the Twitter data we use, but they are the most up-to-date national data available.

3. Methods

We deliberately use relatively simple and intuitive classifiers that are conducive to introspection and analysis. More powerful and complicated machine learning methods are available, which would yield higher performance numbers with the same data (generally at the cost of interpretatability). Likewise, there are a number of calibration and parameter tuning steps that could be applied to improve performance. Furthermore, there are many more relevant signals in this data than the language around suicide and suicidal ideation that we examine here. These experiments and results are meant to be illustrative of what is possible, along with accompanying insight as to what the relevant aspects of language are (*how* it is possible) rather than a reporting on the precision of maximally tuned machine learning approaches.

Character n-gram Language Models Language models are frequently employed to estimate the likelihood of a given sequence of words. This is done often by examining a moving window of n words (n-gram). Traditionally each word is treated as a token, but previous work indicates that treating each character as a token creates classifiers that capture some of the creative language use and emoticons frequently found on social media and are afforded a modicum of robustness to misspellings [McNamee and Mayfield, 2004, Coppersmith et al., 2015a]. Our character n-gram language models use sequences of up to 5

[2] `http://wonder.cdc.gov/ucd-icd10.html`

characters as features, which has been shown elsewhere to capture some signals relevant to mental health [Coppersmith et al., 2015b, Mitchell et al., 2015]. For illustration, in this sentence we would observe the sequences: "for i", "or il", "r ill", " illu", "illus" and so on. Briefly[3], we normalize the text by removing all retweets and links containing a URL (as these are often not authored by or about the user), lowercasing all characters, replacing all usernames with a single "@" token and replacing all URLs with a single "*" token. To train the model, we use data from a training corpus of users. For each user in the corpus (e.g., each person who attempt suicide), for every tweet from that user, the model tabulates the number of each of these sequences observed. After training, the model can produce a likelihood score that an arbitrary text was generated by the model estimated (denoted here as $CLM(t)$).

Specifically, we create one language model from the text of the people who attempted suicide, prior to the date of their suicide attempt (CLM_s), and we create a second language model with the contrasting class (e.g., matched neurotypical controls; CLM_c). To obtain a score for an arbitrary text t we measure the probability of each sequences of characters in t up to length n under CLM_s and CLM_c and compare their relative probabilities:

$$c(t) = \frac{log(CLM_s(t)) - log(CLM_c(t))}{|t|}$$

If it is more probable that the text was generated by class s (person who attempted suicide), $c(t)$ would be positive. Likewise for class t and negative scores.

Aggregators The CLMs produce scores for each tweet, which provide for a temporally-local, but noisy, measure of how much this language looks like that of someone prior to a suicide attempt. Since some of the symptoms change over time, examining a moving window of tweets, aggregated in some way, may provide a less-noisy estimate of symptoms at that point in time, a la [Resnik et al., 2015]. We treat each user's tweets as an ordered list (T) and score each tweet ($t_i \in T$) with the CLMs ($c(\cdot)$).

Let the ith tweet in the time-ordered list T be t_i. We examine b sequential tweets in a window (i.e., the bandwidth), letting the window be $W_i = c(t_i), c(t_{i+1}), ..., c(t_{i+b})$. For each W_i we aggregate the scores $f(W)$, which yields $|T|-b$ aggregated scores for each list T. Some simple choices for $f(\cdot)$ are: (1) mean, (2) median, and (3) the proportion of tweets classified as more likely to be generated by the suicide attempter model than the control model. This produces a range of aggregated scores for how "suicide-attempter-like" a user's tweets look at all points in time. For experiments where we need a single score for each user, we can further aggregate these local scores for some notion of how bad the worst *local* time period was. Specifically, we use *mean*, *median*, or *max* (denoted $s(\cdot)$), all perform roughly equivalently for the data considered here. We abbreviate this double aggregation as: $\psi(T) = s(f(W_i \forall i \in [0, |T| - b]))$.

Figure 2 illustrates the scores for attempters and controls over time. Note that the lines for the attempters (blue) are often above zero, while those of the controls (green) are often below zero, as one would expect if the CLM was finding quantifiable signal useful for making this distinction. Also note that there is a significant amount of variation present, and that many of the users do cross the 0 line (switching between "more-neurotypical-like" and "more-suicide-attempter-like").

Detecting Marked Increases Time has a significant role in understanding mental health, and temporally localized patterns may be more powerful than analysis that spans longer

[3]For a more detailed description, see [Coppersmith et al., 2014a].

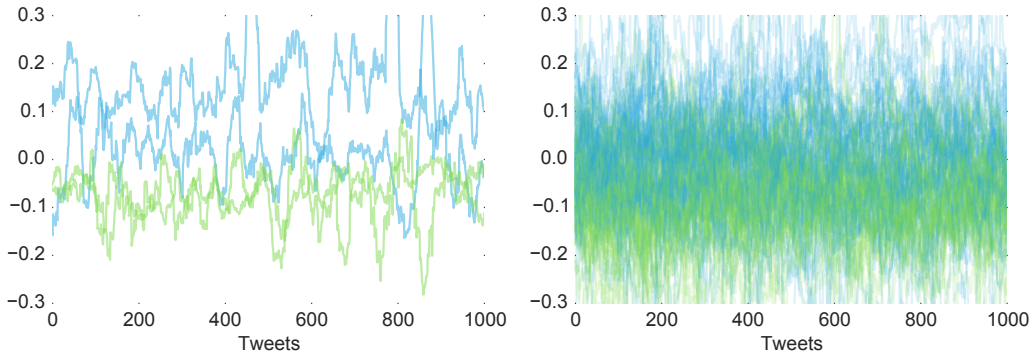

Figure 2: Timeseries CLM scores for suicide attempters (in blue) and neurotypicals (in green), $b = 30$ and $s = mean$. Aggregated CLM score ψ on the y-axis, temporally ordered tweets on the x-axis. The left plot is an example of two attempters with their matched controls, the right plot shows all attempters and matched controls.

periods of time [Resnik et al., 2015]. Many conditions ebb and flow in severity, the importance of which is highlighted a common suicide prevention scenario: the decision to enact an intervention is based on not only a person's current severity, but that current temporally-local severity compared to that person's historic severity. Thus, we will examine the time immediately preceding a suicide attempt specifically and separately, to glean information about what quantifiable signals exist. In particular, we examine k tweets immediately prior to a person's suicide attempt, and compare them to that person's previous tweets. We compare the highest score obtained by the aggregated methods above ($\psi(\cdot)$) with the highest point observed any time previously: $\Psi(T, k) = \max(\psi(T_i, i \in [k, |T|])) - \max(\psi(T_i, i \in [0, k]))$ as a way to estimate the amount to which their symptoms worsened over this time period.

4. Experiments

All experiments, unless otherwise noted, are the result of ten-fold cross-validation. Performance measures are calculated across all folds, so each user is used exactly once to evaluate performance, and exactly nine times as training data in the creation of the classifiers.

Can we separate people who attempt suicide from neurotypical controls? To find quantifiable signals of suicidal users, we build machine learning classifiers to separate them via their language usage from age- and gender-matched controls. Performance for this comparison is shown as a ROC curve (false alarms on the x-axis and the precision on the y-axis) in Figure 3. Many of the parameter selections (e.g., selections for f, s, and b) yield roughly equivalent rates of precision and false alarms, all of which are distinctly better than chance prediction. This should be interpreted as evidence that we are finding quantifiable signals relevant to separating people who attempt suicide from neurotypical controls, based on their language alone.

Which quantifiable signals are the model using? To provide intuition for what quantifiable signals the classifiers are finding, we examine how they score words from the linguistic inquiry word count (LIWC), a psychometrically validated lexicon [Pennebaker et al., 2007]. LIWC provides a mapping from English words to a psychological category related to the word. For example, words like *alive*, *dead*, *kill*, *mourn* and *suicide* are members of

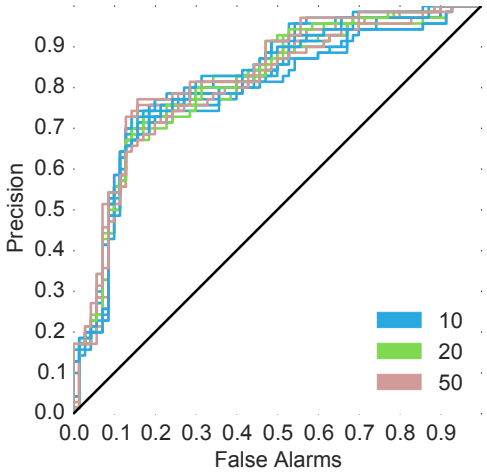

Figure 3: ROC curve of the performance for separating people who attempt suicide from their age- and gender-matched controls. Each curve denotes a particular combination of $c \in$ [mean, median], $s \in$ [mean, median], and $b \in$ [10, 20, 50]. Color denotes b, while differences between c and s were minimal and are excluded for clarity.

the DEATH category, all words related to death and dying. For each LIWC category L, containing words $l_0, ..., l_n$, we compute the median of the scores: $c(l_i)$ for $i \in [0, n]$, and show the ordered differences in the top of Figure 4. This can be interpreted as how language from a given psychological category will be scored by our classifiers – a large positive score indicates the category is used more frequently by people who attempt suicide than neurotypical controls, a large negative score indicates the category is used less frequently by the people who attempt suicide, and a value near zero indicates minimal differences between people who attempt suicide and neurotypical controls in their use of this category. The LIWC categories for which we observed large median positive differences (i.e., used more often by people who attempt suicide) are DEATH, HEALTH, SAD, THEY, I, SEXUAL, FILLER, SWEAR, ANGER, and NEGATIVE EMOTIONS. Some of these categories seem obvious to connect with suicide and suicidal ideation (e.g., DEATH and NEGATIVE EMOTIONS), and some are less obvious, but have been linked to a number of psychologically interesting phenomena (e.g., THEY, I, and FILLER) [Pennebaker, 2011]. The general trend that many of the LIWC categories is above zero might indicate that the people who attempt suicide group is more likely to talk about psychologically-interesting phenomenon which is not entirely surprising, though more in-depth analysis is warranted before any conclusions are drawn. Interestingly, ASSENT has a marked decrease, which includes words typically used in conversation in response to a person like *agree* and *yes*, which might indicate that people who attempt suicide are engaged in fewer conversations. Post-hoc analysis of the data supports this idea, indicating that people who attempt suicide have a smaller proportion of their tweets directed to another user (as is common in conversation on Twitter) than the depression or the control populations. We cannot immediately measure the level of response or reciprocity in conversation that their tweets receive from others, so deeper questions of conversational dynamics within this population will be relegated to future work.

Can we separate people who attempt suicide from depressed users? Since suicide and depression often co-occur in users, we apply a similar approach to determine whether the signals we find are specific to suicidality or merely capturing signals relevant to depression.

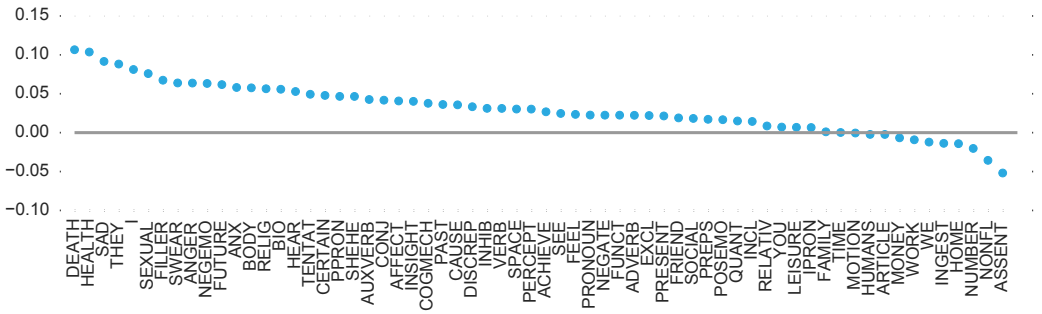

Figure 4: Median scores ($c(\cdot)$) for all words in each LIWC category. Higher scores indicate category is more strongly associated with people who attempt suicide than neurotypical controls (DEATH being the highest).

We train classifiers to distinguish between suicidal users and age- and gender-matched users who have previously stated a diagnosis of depression, drawn from the population of [Coppersmith et al., 2015b]. Recall that since we have no way to assert that the depression users have not attempted suicide this experiment is likely significantly contaminated in this manner. When coupled with the relatively small samples sizes we investigate, such contamination likely significantly degrades performance. Given that the contamination would work against the classifiers presented here, this should be considered a lower bound on performance.

In Figure 5, the left ROC curve denotes the classifier's performance at separating these two groups, given all the data prior to a user's suicide attempt and the data from the age- and gender-matched control. When we examine the relative change in classifier scores just before a suicide attempt (specifically if they markedly increase), rather than the user's whole history, we see an increase in performance (the right ROC curve in Figure 5). Specifically, for $k \in [10, 20, 30]$, we scored each user according to $\Psi(T, k)$. Interestingly, the comparison between performance for $\psi(T)$ (left) and $\psi(T, k)$ (right), indicates that there is more quantifiable information present in comparing those last k tweets (a limited temporal window) with the rest of the user's history.

This should not be interpreted as performance at *predicting* a suicide attempt, however, since the post-hoc and conditional nature of this analysis (starting from the known suicide attempt and working backwards) does not even try to assess how common such increases might be in times that do not include a suicide attempt. The important point to take from this experiment is that our classifiers are able to differentiate people who attempt suicide from depressed individuals better than random, despite the contaminated and small data.

Can we suggest areas for future epidemiological research? We use our classifiers and public data to estimate the rates of suicidal ideation for regions in the United States. The experiments above provide some evidence that our techniques for estimating suicidal ideation from language are viable, though more validation is warranted. For the nonce, let us assume that these are reasonable: at best, this is actually informative to suicide prevention, at worst it is illustrative of what is possible with a more precise models. We build a single classifier for separating people who attempt suicide from neurotypical controls using all people who attempt suicide and their age- and gender-matched controls (i.e., not ten-fold cross validation).

Figure 6 shows a choropleth of the United States, each state colored by estimates of

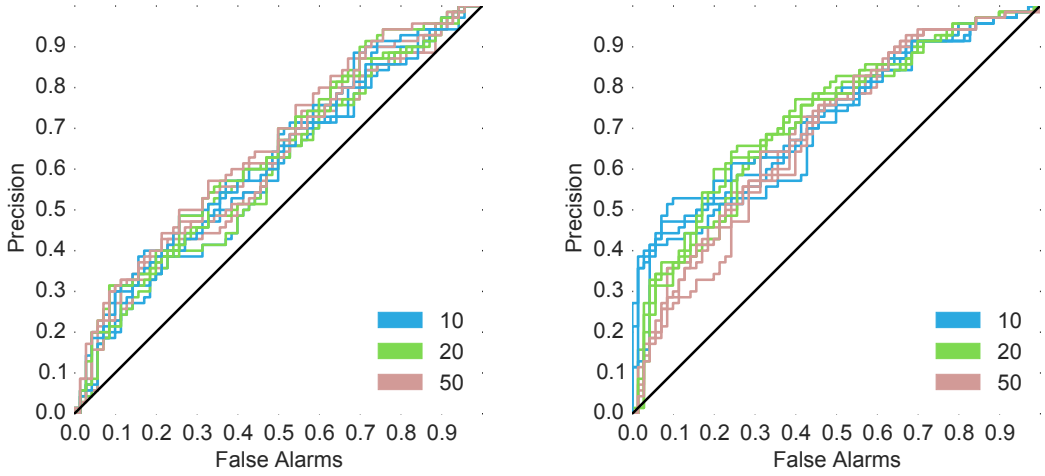

Figure 5: ROC curve of the performance for separating people who attempt suicide from their age- and gender-matched depression users. The plot on the left uses all data prior to a user's suicide attempt to make the determination, $\psi(T)$, while on the right we examine the difference between the k last tweets and all prior tweets, $\Psi(T, k)$. Color denotes b, which has a much larger and systematic effect on performance than the selection of f, s (left), and k (right). For simplicity, the range of values for f, s, and k is not differentiated visually.

death by suicide (top), estimated suicidal ideation (middle), and the ratio of suicide decedents to ideators (bottom). All data is for women 15-24 (which is the majority population in our dataset). For each state, we have at least 100 Twitter users (estimated to be women aged 15-24), with a mean of 147. The suicide deaths are taken from the CDC and normalized by the census population of each state. The estimates of suicidal ideation are the proportion of Twitter users geolocated to each state that our classifiers estimate are suicidal ideators, divided by the number of Twitter users estimated to be neurotypical controls. The ratio of suicide decedents to ideators is effectively the top plot (those lost to suicide), divided by the middle plot (those who ideate about suicide – the at risk population). Thus the states with high ratios in the bottom plot are places where the population has a higher propensity to progress from ideation to die by suicide – perhaps where risk factors are present. Conversely, the states with low ratios are places where the population has a lower propensity of the same – perhaps where protective factors are present. Analysis of this sort, even with noisy estimates, can suggest areas for future epidemiological research, or perhaps to estimate the efficacy of interventions and mental health systems in near real time.

5. Discussion

Here we analyzed the language of Twitter users prior to a suicide attempt, as compared to age- and gender-matched control users. We compared people who attempt suicide to both neurotypical controls and users who have previously stated that they were diagnosed with depression. In both cases, quantifiable linguistic differences were found.

Primarily the population examined here was estimated to be females between the ages of 15 and 24. Thus, we demonstrated how these classifiers, combined with public social media data and open government data can be fruitfully combined to indicate where fertile ground for future epidemiological research might be, pertaining to this population.

Some important caveats and limitations on this work: First, our analysis was aimed

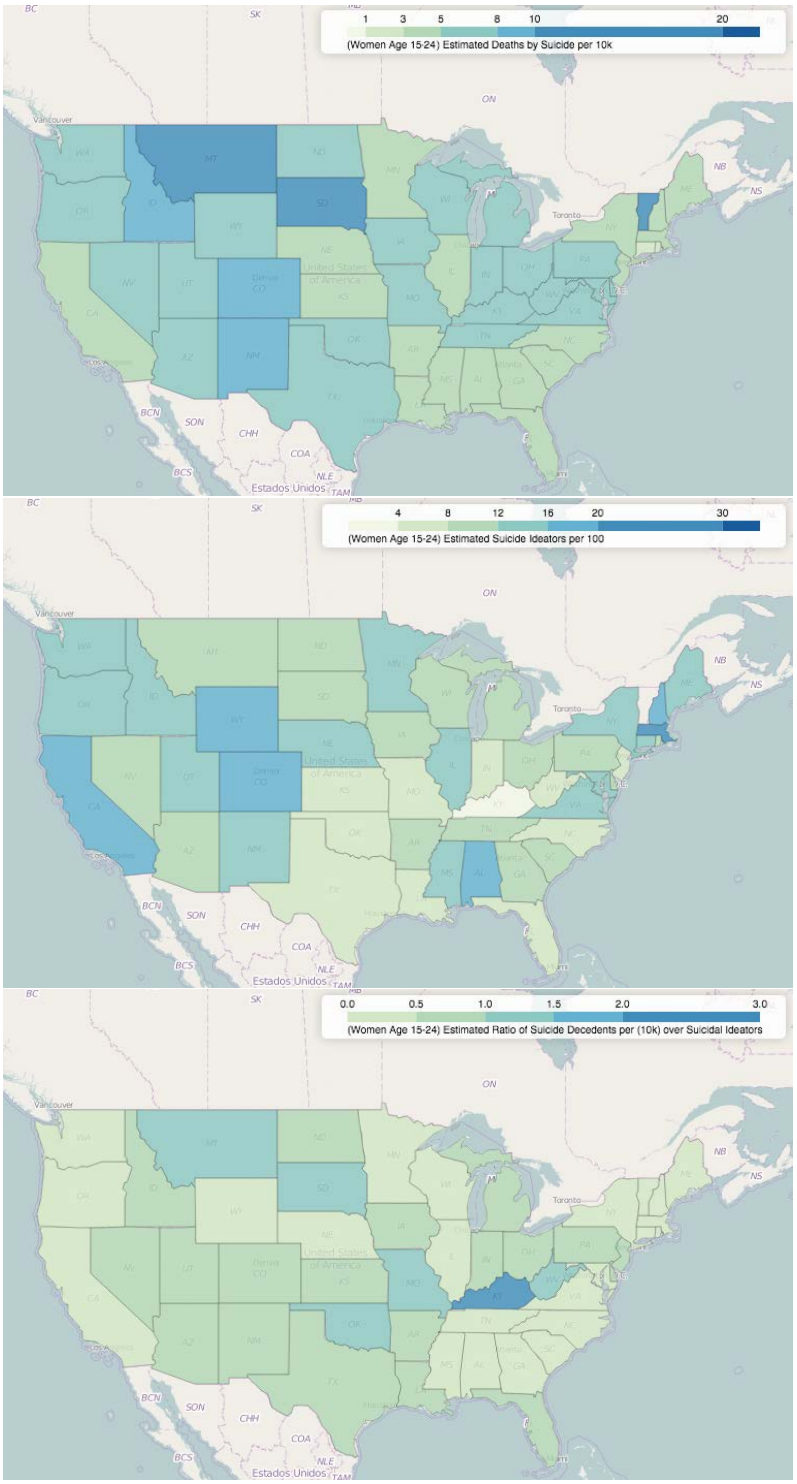

Figure 6: Choropleths, all showing data for women age 15-24. Top: estimate of completed suicides per 10000 people. Middle: proportion of Twitter users that are suicidal ideators (estimated via our classifer) Bottom: Ratio of completed suicides over estimated ideators.

at insight, introspection, and illustrative usage rather than maximally performant machine learning. We expect applications of more complicated and tuned techniques would yield higher performance with this same data. Second, these techniques seem to capture a different demographic than is most at risk for suicide in the United States (middle aged, white, males). However, the demographics we primarily examine here are also at high risk. Third, there is some inherent selection bias in that all our population has elected to engage on Twitter, which not is a representative sample of the country as a whole. More poignantly, our users have elected to discuss their mental health (an often stigmatized subject) publicly on Twitter. All the analysis we discussed was on Twitter data, but there is evidence that these methods are applicable accross social media platforms, e.g., [Preotiuc-Pietro et al., 2015].

Despite these caveats, we see this as an important step towards better understanding how suicidal ideation is manifest in the language use on social media. Our investigation indicates a few promising next steps and future directions for both individual and epidemiological research, hopefully resulting in better, earlier interventions and fewer lives lost to suicide. The inherent power of quantified and scalable measures available at finer geographic and temporal resolution holds great promise for filling in the vast gaps in the data available to drive informed decisions of public and health policy. Most importantly, though, this indicates potential for getting estimates of suicide-related data in real time. In many cases, errorful estimates of the current truth is more useful than an accurate estimate of what the truth was two years ago.

Acknowledgements The authors would like to thank April Foreman, Bart Andrews, Bill Schmitz, Craig Harman and the dedicated group at #SPSM (`http://spsmchat.com`) for their insight throughout this research.

References

[Althouse et al., 2014] Althouse, B. M., Allem, J.-P., Childers, M. A., Dredze, M., and Ayers, J. W. (2014). Population health concerns during the United States' great recession. *American Journal of Preventive Medicine*, 46(2):166–170.

[Ayers et al., 2013] Ayers, J. W., Althouse, B. M., Allem, J.-P., Rosenquist, J. N., and Ford, D. E. (2013). Seasonality in seeking mental health information on Google. *American Journal of Preventive Medicine*, 44(5):520–525.

[Blow et al., 2012] Blow, F. C., Bohnert, A. S., Ilgen, M. A., Ignacio, R., McCarthy, J. F., Valenstein, M. M., and Knox, K. L. (2012). Suicide mortality among patients treated by the veterans health administration from 2000 to 2007. *American journal of public health*, 102(S1):S98–S104.

[Centers for Disease Control and Prevention (CDC), 2010] Centers for Disease Control and Prevention (CDC) (2010). Behavioral risk factor surveillance system survey data.

[Chung and Pennebaker, 2007] Chung, C. and Pennebaker, J. (2007). The psychological functions of function words. *Social Communication*.

[Coppersmith et al., 2014a] Coppersmith, G., Dredze, M., and Harman, C. (2014a). Quantifying mental health signals in Twitter. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

[Coppersmith et al., 2015a] Coppersmith, G., Dredze, M., Harman, C., and Hollingshead, K. (2015a). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, Colorado, USA. North American Chapter of the Association for Computational Linguistics.

[Coppersmith et al., 2015b] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (2015b). CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the Shared Task for the NAACL Workshop on Computational Linguistics and Clinical Psychology*.

[Coppersmith et al., 2014b] Coppersmith, G., Harman, C., and Dredze, M. (2014b). Measuring post traumatic stress disorder in Twitter. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[De Choudhury et al., 2013a] De Choudhury, M., Counts, S., and Horvitz, E. (2013a). Social media as a measurement tool of depression in populations. In *Proceedings of the 5th ACM International Conference on Web Science*.

[De Choudhury et al., 2013b] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013b). Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[Eichstaedt et al., 2015] Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., Weeg, C., Larson, E. E., Ungar, L. H., and Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2):159–169.

[Kumar et al., 2015] Kumar, M., Dredze, M., Coppersmith, G., and De Choudhury, M. (2015). Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM conference on Hypertext and hypermedia*. ACM.

[McNamee and Mayfield, 2004] McNamee, P. and Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1-2):73–97.

[Mitchell et al., 2015] Mitchell, M., Hollingshead, K., and Coppersmith, G. (2015). Quantifying the language of schizophrenia in social media. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, Colorado, USA. North American Chapter of the Association for Computational Linguistics.

[Nock et al., 2008] Nock, M. K., Borges, G., Bromet, E. J., Alonso, J., Angermeyer, M., Beautrais, A., Bruffaerts, R., Chiu, W. T., De Girolamo, G., Gluzman, S., et al. (2008). Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry*, 192(2):98–105.

[Park et al., 2015] Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., and Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

[Pennebaker, 2011] Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, 211(2828):42–45.

[Pennebaker et al., 2007] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. LIWC.net, Austin, TX.

[Preotiuc-Pietro et al., 2015] Preotiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H. A., and Ungar, L. (2015). The role of personality, age and gender in tweeting about mental illnesses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, NAACL*.

[Ramirez-Esparza et al., 2008] Ramirez-Esparza, N., Chung, C. K., Kacewicz, E., and Pennebaker, J. W. (2008). The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. In *Proceedings of the 2nd International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[Resnik et al., 2015] Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A., and Boyd-Graber, J. (2015). The University of Maryland CLPsych 2015 shared task system. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, Colorado, USA. North American Chapter of the Association for Computational Linguistics.

[Resnik et al., 2013] Resnik, P., Garron, A., and Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1348–1353.

[Robertson et al., 2012] Robertson, L., Skegg, K., Poore, M., Williams, S., and Taylor, B. (2012). An adolescent suicide cluster and the possible role of electronic communication technology. *Crisis*.

[Rosenquist et al., 2010] Rosenquist, J. N., Fowler, J. H., and Christakis, N. A. (2010). Social network determinants of depression. *Molecular psychiatry*, 16(3):273–281.

[Sap et al., 2014] Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., Ungar, L. H., and Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.

[Schwartz et al., 2014] Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., Kosinski, M., and Ungar, L. (2014). Towards assessing changes in degree of depression through Facebook. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

[Schwartz et al., 2013a] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E. P., and Ungar, L. H. (2013a). Characterizing geographic variation in well-being using tweets. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[Schwartz et al., 2013b] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., and Ungar, L. H. (2013b). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8(9).

[Sullivan et al., 2013] Sullivan, E., Annest, J. L., Luo, F., Simon, T., and Dahlberg, L. (2013). Suicide among adults aged 35–64 years, united states, 1999–2010. *Center for Disease Control and Prevention, Morbidity and Mortality Weekly Report*.

[The National Institute of Mental Health, 2015] The National Institute of Mental Health (2015). Schizophrenia. http://www.nimh.nih.gov/health/topics/schizophrenia. [Online; accessed 2015-03-04].

[World Health Organization et al., 2014] World Health Organization et al. (2014). *Preventing suicide: A global imperative*. World Health Organization.

[Worton, 1989] Worton, B. J. (1989). Kernel methods for estimating the utilization distribution in home-range studies. *Ecology*, 70(1):164–168.

[Yang et al., 2010] Yang, A. C., Huang, N. E., Peng, C.-K., and Tsai, S.-J. (2010). Do seasons have an influence on the incidence of depression? The use of an internet search engine query data as a proxy of human affect. *PLOS ONE*, 5(10):e13728.