Exploratory Analysis of Social Media Prior to a Suicide Attempt

Glen Coppersmith Qntfy glen@qntfy.com
Kim Ngo Qntfy, University of Notre Dame kim@qntfy.com
Ryan Leary Qntfy ryan@qntfy.com
Anthony Wood Qntfy tony@qntfy.com

Abstract

Tragically, an estimated 42,000 Americans died by suicide in 2015, each one deeply affecting friends and family. Very little data and information is available about people who attempt to take their life, and thus scientific exploration has been hampered. We examine data from Twitter users who have attempted to take their life and provide an exploratory analysis of patterns in language and emotions around their attempt. We also show differences between those who have attempted to take their life and matched controls. We find quantifiable signals of suicide attempts in the language of social media data and estimate performance of a simple machine learning classifier with these signals as a non-invasive analysis in a screening process.

1 Introduction

Mental health poses a sizable challenge by any metric. An estimated 1 in 4 Americans will contend with a mental health condition in a given year (National Institutes of Health, 2013). Around 1% of people die by suicide, 2.7% attempt suicide, 3.1% make a plan for suicide, and 9.2% are challenged with suicidal ideation (Nock et al., 2008). Tragically, this means roughly 4.8 million Americans alive today will die by suicide, placing suicide among the top ten leading causes of death in the United States (Sullivan et al., 2013). Worldwide, it is the leading cause of death for women age 15-19 and the second leading cause of death for teenagers (World Health Organization and others, 2014). What's worse, the rates of suicide seem to be increasing, up 28% in the civilian population of the United States between 1999 and 2010 (Sullivan et al., 2013).

Despite the magnitude of the challenge posed by suicide, we have a relatively sparse understanding of what precisely gives rise to suicide risk. To prevent suicides, we need a better understanding of the underlying phenomena relating to both the immediate risk of suicide (or *acute suicidal risk*) and the long term risks. For both cases, data is extremely sparse, never in real time, and subject to some bias. Few objective measures exist to measure outcomes, and those that do exist tend to have poor temporal resolution (measured in weeks or months) and are labor intensive. Optimizing intervention efficacy or policy-level strategies is difficult without such data.

Here we explore a novel dataset of social media data from users who have attempted to take their own life. This kind of data has not previously been available in sufficient quantities or at this granularity, so we provide broad intuition and interpretation of trends, rather than testing specific hypotheses. Our primary contributions are: **[1]** We find quantifiable signals of suicide, with sufficient performance and scalability to warrant consideration as part of a screening process. **[2]** We provide intuition about the data via simple visualizations of linguistic content of users prior to a suicide attempt. **[3]** We use automatic emotion classification to uncover interesting patterns in the emotional composition of posts made by users in the time around a suicide attempt. **[4]** Where possible, we tie these phenomena back to existing psychological research. This paper deliberately only scratches the surface of the possible insight encoded in data related to suicide attempts.

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 106–117,
San Diego, California, June 16, 2016.

Quantifying Mental Health: Thanks to the use of vital signs like temperature and blood pressure, cross correlation of various easy-to-observe symptoms and the rapid measurement of blood chemistry, the diagnosis of physical illness has improved radically since 1900. Mental and behavioral healthcare has not benefited in the same way from binary diagnostics. In part, this may be because physical health conditions manifest irrespective of whether the patient is in a diagnostic healthcare setting, while mental health conditions manifest when a person interacts with the rest of their world, making measurement in a laboratory difficult. Social media may seem, at first, to be a strange data source for studying mental health, but there are myriad quantifiable signals within social media that capture how a person interacts with their world. We suggest that data collected in the "white space" between visits with healthcare professionals may be part of a rigorous, scalable, and quantified diagnostic approach to mental and behavioral illness. Language, in particular, has proven to be a potent lens for the analysis of mental health, as evidenced by the wide usage of the Linguistic Inquiry Word Count (Tausczik and Pennebaker, 2010; Pennebaker et al., 2007; Pennebaker et al., 2001) and the depth of publications at the Computational Linguistics and Clinical Psychology workshops (Resnik et al., 2014; Mitchell et al., 2015a; Hollingshead and Ungar, 2016).

Mental Health through Social Media: Social media data is necessarily stored in formats conducive to analysis via computer. This allows for larger sample sizes and higher frequency than anything ever before possible. Collecting the ordinary language of thousands of users over weeks, months or years has become trivial in comparison to the paper based analysis methods of the past.

Work examining mental health conditions that affect a large number of people has proliferated, especially depression (Coppersmith et al., 2015b; 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). Similarly, common psychological phenomena, like personality factors and psychological well-being are now well-studied through emprical analysis of social media data (Schwartz et al., 2013b; Park et al., 2015; Schwartz et al., 2013a). These approaches and survey methods were sufficient to support analysis of relatively common conditions, but are not as effective for rarer ones.

Coppersmith et al. (2014a) introduced methods for examining public data which allowed for more scalable creation of data sets, thus permitting the examination of rarer conditions. Post traumatic stress and schizophrenia are two examples of conditions significantly rarer than depression, whose analysis are possible by these techniques (Coppersmith et al., 2014b; Mitchell et al., 2015b). Suicide and suicidal ideation were more difficult to obtain data for, but some population-level analysis was enabled by anonymous suicide help fora (Kumar et al., 2015; Kiciman et al., 2016). Additionally, Robertson et al. (2012) investigated the role that social media has in suicide clusters (among people in disparate geographies connected online).

At the individual level, techniques similar in nature to Coppersmith et al. (2014a) can provide social media data for users prior to a suicide attempt of sufficient size to allow linguistic analysis (Coppersmith et al., 2015c; Wood et al., 2016). Coppersmith et al. (2015c) was able to automatically separate users who would attempt to end their life from neurotypical controls and further tie signals explicitly back to the psychometrically validated Linguistic Inquiry Word Count categories and existing psychological theories. Furthermore, they found slight but measurable differences between those who would attempt to end their life and those challenged by depression without suicidal ideation. The operative question has been: are there quantifiable markers in an individual's social media content that indicate their current or future risk of acute suicidal crisis?

Biases: The existing methods for assessing the events surrounding suicidal crisis resulting in a suicide attempt are heavily susceptible to recall bias and context bias (Shiffman et al., 2008). People are more likely to remember negatively charged information when they are in a negative mood (Clark and Teasdale, 1982), as when asked to reconstruct information about a suicide attempt. The available information about the events leading up to a suicide attempt are generally based on the self report of peo-

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

Figure 1: Fictitious example tweets of genuine statements of a suicide attempt (top), genuine statements indicating a time (middle) and disingenuous statements (bottom).

ple who survived one or more attempts or the reconstructions of events provided by friends or family members after a traumatic loss. All of these issues pose serious problems for accurate recall, compounded by the effects of biases. Contrastively, social media streams are biased in other ways, often towards self presentation, but recorded in the moment.

Often, treatment progress is assessed using weekly or monthly questionnaires or interviews that require retrospection on the part of the patient. However, retrospective self-report measures are notoriously context dependent and highly influenced by momentary accessible information. Furthermore, the commonly reported tendency toward "backfilling" that often happens when written journals are employed in a therapeutic context is worth noting (Stone et al., 2003). When a patient is asked to keep a paper journal in the space between office visits, they frequently backfill the entries just prior to their appointment from (biased) memory, to please their therapist or appear compliant. Thus, several weeks of mood journaling may be compiled in the waiting area before their visit rather than as they naturally occur. All of these issues pose a problem for reconstructing events surrounding suicidal crisis and make wider generalizations more challenging, bording on speculative. Ideally, analysis of personal social media data in conjunction with more traditional methods may offset the short comings of each method in isolation.

2 Data

We examine data from people who **publicly** state on Twitter that they have tried to take their own life, and provide enough evidence for the casual observer to determine the date of their suicide attempt. Specifically, we have 554 users who stated that they attempted to take their life, 312 of which give an indication of when their latest attempt was. The exact date of their attempt was available for 163 users, and 125 of them had data available prior to the date of their attempt. We do not include any users who have marked their profile as *private*, and for each user we examine only their **public** data, which does **not** include any direct messages or deleted posts.

For each user, a human annotator examined their tweets and verified that **[1]** the user's statement of attempting to take their life appeared genuine[1] **[2]** the user is speaking about their own suicide attempt, and **[3]** that the suicide attempt could be localized in time. See Figure 1 for example tweets.

We estimate the age and gender of each user who attempted to take their life to provide aggregate demographic information from the users in the dataset (see Figure 2) and to allow us to directly control for variability due to age and gender in our analysis. Demographic estimates were derived from the authored content of each user via lexica magnanimously provided by the World Well-Being Project (Sap et al., 2014). Though imperfect (91.9% accuracy for gender, $r = 0.83$ correlation for age), these estimates are informative in aggregate. Notably, there are significantly more women in our data than men, and almost all users are between the age of 15 and 29. This indicates that we do not have a representative sample of the demographics on Twitter, with polling indicating that 37% of adults aged 18 to 29 and 12% of those in middle age are on Twitter (Duggan et al., 2015). Since the older demographic, also at risk for suicide, does not show up in our sample, it suggests that we are primarily capturing the youth at risk for suicide, perhaps because they are more likely to discuss the subject openly.

[1]Previously, annotators have shown high agreement for differentiating between genuine and disingenuous statements involving mental health conditions, $\kappa = 0.77$ (Coppersmith et al., 2015c).

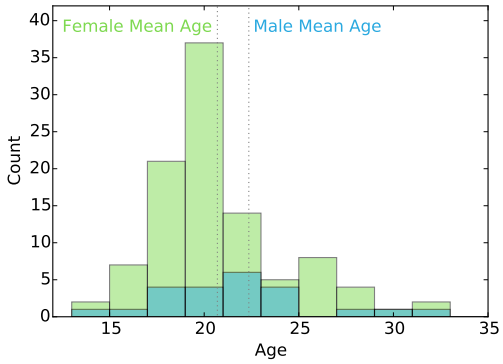

Figure 2: Histogram of the ages of users who attempted to take their life. Females are in green, and males in blue. The mean age of each gender is denoted by vertical lines.

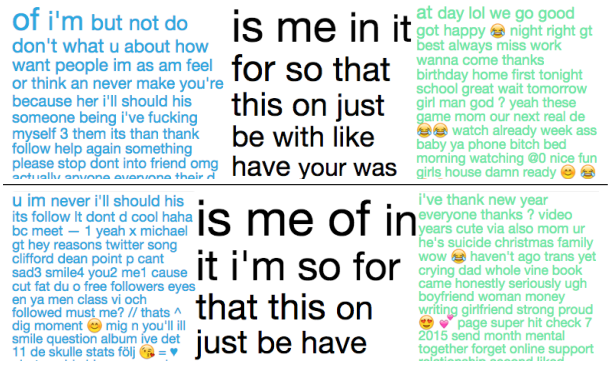

Figure 3: Venncloud visualizing the differences in language usage between the groups examined here. The top cloud compares users who attempt to take their life (left) with neurotypicals (right). The bottom compares users who attempt to take their life prior to (left) and after (right) their attempt. Larger words occur more frequently in the corpus.

For each user who has attempted to take their life, we draw an age- and gender-matched control from a large pool of random English users. We find a user of the same estimated gender and the smallest difference in esitmated age. It is likely that 4-8% of these (assumed) neurotypical control users have or will try to take their life, given the rates of suicide attempts in the population (Nock et al., 2008). This contamination will only serve to weaken our models and obscure trends. We make no attempt to remedy this and the results should be treated as lower bounds.

3 Exploration of Language Data

First, we visualize the linguistic differences in our populations via simple and straightforward methods

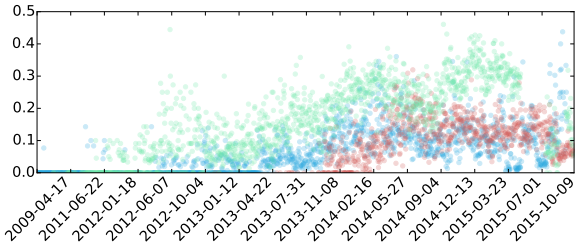

Figure 4: Proportion of tweets containing an emoji (y-axis), by date (x-axis). Neurotypicals in green, users prior to their suicide attempt in blue and after their attempt in red.

to provide intuition about the sort of information available and glean insight as to how this data might relate to existing psychological theory. In all cases, we compare (1) users who have tried to take their life to their matched neurotypical controls and (2) users prior to and after they attempt to take their life.

Venncloud: Figure 3 show Venncloud comparing word usage in our populations. As explanation, consider the top Venncloud which compares users prior to their attempt to take their life (left) with neurotypicals (right). This examines language at the level of *tokens*, which here is either a single word, emoticon, or symbol. Each token can only show up once in the visualization, so if the token is used with higher probability by neurotypical users, it is displayed on the right. If it is used with higher probability by users who tried to take their life (only examining data prior to that attempt), it is displayed on the left. Tokens that occur with approximately the same probability are displayed in the middle. For a more detailed description, see Coppersmith and Kelly (2014). A few interesting phenomena emerge from this simple analysis: **[1]** neurotypicals use emoticons and emoji with much higher probability than a user prior to a suicide attempt (also see Figure 4), **[2]** users are more likely to talk about suicide *after* an attempt than before it, **[3]** users prior to a suicide attempt use more self-focused language, replicating similar findings in those challenged with depression (Chung and Pennebaker, 2007; Coppersmith et al., 2014a; Coppersmith et al., 2015a), **[4]** users prior to a suicide attempt are more likely to employ automatic means of tracking their followers (as most uses of the token "followers" are from the automatic output of these applications).

Figure 4 indicates that neurotypicals (green) use

emoticons and emoji with a higher frequency than those who attempt suicide, before (blue) or after (red) that attempt. For each day where we have at least 10 tweets, we calculate the proportion of tweets for each group that contains an emoticon or an emoji. Interestingly, neurotypicals and people who attempt suicide seem to adopt emoji around the same time, starting in 2012, but neurotypicals use them more.

4 Methods

We are primarily concerned with drawing two comparisons here. First, what observable differences exist between users who attempt to take their life and the rest of the (neurotypical) population? Second, what observable differences exist between users prior to and after a suicide attempt?

Preprocessing: The processing of unedited language data prior to the application of any machine learning or visualization techniques often have significant effects on the outcome. Here, for each tweet we replace all usernames in the text with the single token "@", and replace all URLs with the single token "*". For example "Check out `https://OurDataHelps.org` powered by @Qntfy ! :)" would be "Check out * powered by @ ! :)" after preprocessing. All emoticons and emoji remain intact and are treated as single characters. While many types of linguistic analysis examine the content and topics of documents, we are equally interested in content and context. Here, we diverge from most natural language processing, which often dismiss many very frequently used words as uninteresting, and remove them from analysis (sometimes referred to as "filler" or "stop" words). Previous work has demonstrated (and frequently replicated) that some of these words (e.g., first person and third person pronouns) hold psychological meaning, and thus should be included in analysis (Pennebaker, 2011; Chung and Pennebaker, 2007). Likewise, lemmatizing or stemming words may also remove information about how the author experiences the world, such as whether their language is future- or past-focused.

Character Language Models: For classification, we prefer simple, straightforward methods that provide scores at a per-tweet level. Here, we use character n-gram language models followed by logistic regression via scikit-learn (Pedregosa et al., 2011). These models are particularly appropriate for social media given their robustness to creative spelling, missing spaces, and other eccentricities that result from short, unedited text (Coppersmith et al., 2014b; McNamee and Mayfield, 2004). We use character n-grams up to length 5 (so tokens might include "suici" and "uicid"). Spaces, punctuation, emoticons, emoji, and generic username and url tokens ("@" and "*" respectively) are included as characters. Logistic regression allows us to learn how strongly each of these character n-gram tokens are associated with the populations examined. We use this method to: **[1]** compare those who attempted to take their life against neurotypicals, **[2]** compare language before and after a suicide attempt, and **[3]** n-way classification of emotions. All performance measures are calculated via 10-fold cross validation.

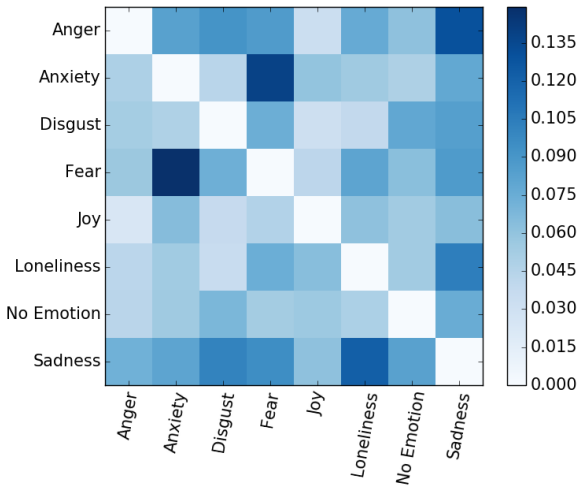

Figure 5: Confusion matrix for emotion classifier, denoting the proportion of tweets from the emotion on the row that are misclassified as the emotion on the column. Diagonals (representing correct classifications) have been removed to better illustrate the misclassifications. Thus, *sadness* is most frequently misclassified as *loneliness* while *fear* and *anxiety* are most confusable.

Emotional States: To estimate emotional states from social media posts, we collected a novel corpus with automatically induced emotion labels, as inspired by Mohammad (2012). These methods might be used to detect emotional states that indi-

cate high risk for suicidal crisis. Detection of hypomanic states (associated with multiple attempts) (Bryan et al., 2008) and elevated levels of guilt or shame have been found among some populations at risk for suicide (Ray-Sannerud et al., 2012). Hashtags provide implicit labels of emotion (excluding any tweet that also has #SARCASM or #JK) – a tweet that contains #ANGER is labeled *anger*, but not one that contains #ANGER #SARCASM. We diverged from past work and focused on emotions more directly related to suicide and psychological phenomena, as well as an automatically-induced *no emotion* category. We used up to 40,000 tweets from each label, selected from a random Twitter sample collected during 2015. For each tweet, we removed the hashtag label from the text, and trained a character n-gram language model.

Inclusion of a *no emotion* label calls for a slightly more complicated training procedure, as these training tweets were selected simply because they lacked an explicit emotional hashtag. Many of the tweets in this category do express an emotion. Creating *no emotion* training data using tweets that lack an explicit emotion hashtag results in the *no emotion* label being particularly contaminated by tweets expressing emotions. This, in turn leads the classifier to frequently misclassify emotional tweets as having *no emotion*. This would skew the performance of the classifier when used beyond training and skew the estimates of accuracy of the classifier (since many tweets labeled and evaluated as *no emotion* actually have emotional content). Thus, we employ semi-supervised learning to decrease the effect of this contamination: We train the model once with 40k random tweets we label as *no emotion*, then use this initial model to score each of a second set of *no emotion* tweets. Any tweet in this second set of ostensibly *no emotion* tweets that is classified by the inital model as having any emotion is removed, since it is likely to be a contaminating emotion-bearing tweet. A random (second) subset of 40k tweets are then selected from those that remain. The model we use for analysis is trained with this cleaner (second) set of 40k *no emotion* tweets.

Emotion classification from statements in isolation is a very difficult task, even for humans, as evidenced by low inter-annotator agreement (e.g., 47% agreement between three annotators in Purver and Battersby (2012)). Additionally, the emotions that are conveyed are also often mixed, making a single label insufficiently descriptive. For further analysis of performance and errors of the emotion classifier, see the Technical Appendix.

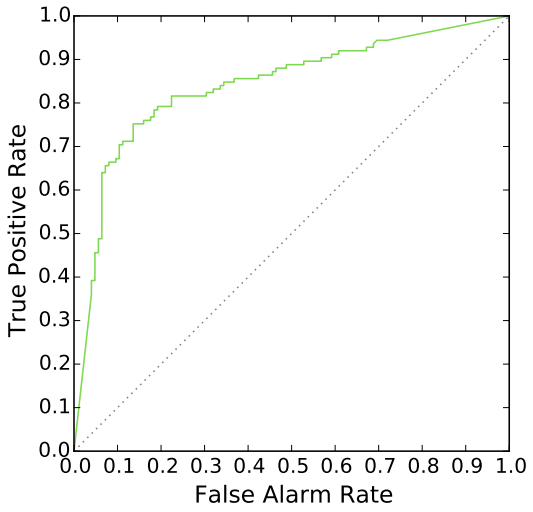

Figure 6: ROC curve for separating users who attempted to take their life from matched neurotypicals.

Briefly, we assessed classification accuracy of this 8-way classifier with 10-fold cross validation, with a resulting F1 of 53. While not directly comparable, reported state of the art results for 2- and 6-way classification range between 45 and 65 (though some treat the task as a multi-level classification problem, *emotion-detection* followed by *emotion-discrimination*, reporting F1 separately and further complicating comparisons) (Mohammad and Kiritchenko, 2015; Purver and Battersby, 2012). The confusion matrix for all the emotions examined can be found in Figure 5.

5 Results

We demonstrate that quantifiable signals relevant to suicide can be found in social media data with simple analysis, then put them in the context of performance in a realistic setting. We use techniques conducive to introspection to facilitate comparison with existing psychological literature.

Quantifiable Signals: To find quantifiable signals of suicide risk, we build character n-gram language models to separate users who have attempted to take

their life from their matched neurotypicals. Specifically, we examine only data prior to each user's suicide attempt. A ROC curve denoting the tradeoff between false alarms (neurotypical users misidentified as at risk to take their life) and true hits (users who will attempt to take their life, identified as such) can be seen in Figure 6.

For a single point of performance for comparison, note that at roughly 10% false alarms, we correctly identfy about 70% of those who will try to take their life. Extrapolating from this likely performance in the real world is not entirely straightforward, but a worthy exercise. We can assume that in our neurotypical population of 15-29 year olds, 4-8% of users will (or have) tried to take their life (Nock et al., 2008; Kann et al., 2014). Thus, the size of the neurotypical population is likely to be more than ten times the size of the at-risk population.

If we were to use this simple method to screen 1000 people aged 15-29, we would expect 40-80 of them (4-8%) to attempt to take their life at some point in time. For simplicity, we will use 6% or 60 users. If we were to randomly select users for additional screening, we would expect that 6% of them will go on to try to take their life – a hit rate of 6%. Straightforward application of the example operating point to 1000 person population would be expected to yield 42 (70% of 60) at risk individuals and 94 (10% of 940) neurotypical flagged for additional screening – a hit rate of 30%.

Our sample of neurotypicals are likely contaminated by users who have or will attempt to take their life, so our estimates of false-alarms may be inflated due to this contamination. In the best-case scenario, these at-risk neurotypical users were flagged correctly, so we reduce our false alarm estimates accordingly. Thus an upper-bound on our performance would be if we consider that 6% of the neurotypical population is currently classified as false alarms, but are actually true hits. Factoring them out would yield a false alarm rate of just 4%, so this optimistic operating point would identify the same 42 at-risk people as above, and 38 (4% of 940) neurotypical users for additional screening – a hit rate of 58%.

In sum, a screening tool for people aged 15-29 based on these simple methods could identify a group for additional screening for which between 30 and 60% would be at risk for a suicide attempt. While more optimization remains to be done, this strongly suggests that technology-assisted screening is within the realm of the possible.

Emotional Posts: We scored each tweet with an emotion classifier, and examined the relative composition of each user's tweets by week, for three months on either side of a user's suicide attempt. Figure 7 shows the percentage of each user's tweets each week that contained a given emotion. Time (by week) on the x-axis and percentage of tweets with that emotion on the y-axis. The day of the suicide attempt and the week following it are included at $x = 0$, indicated by the dotted vertical line. The colored dot indicates the median percentage across all users who attempted to take their life, and the error bars indicate one standard deviation above and below the median. The equivalent median from the neurotypical population is included as a solid horizontal line, with one and two standard errors above and below indicated by dashed and dotted horizontal lines respectively. The median emotional percentage of the users who attempted to take their life for the three months prior to a suicide attempt is indicated by a colored horizontal line left of 0. Similarly, for the three months after the attempt.

Since our analysis is largely exploratory, and not hypothesis-driven, it behooves us to take results that might otherwise be considered statistically significant with a higher degree of skepticism. A reasonable guideline for interpreting these plots to account for the many comparisons made is to consider differences where the error bars are entirely non-overlapping. While other more subtle differences may exist, they should be the subject of more principled and hypothesis-driven experiments. With that lens, some stark differences remain.

Interestingly, while users appear to have a markedly higher incidence of tweets tagged with *anger* and *sadness* prior to the attempt, they fall to levels more in line with neurotypicals after an attempt. A few weeks prior to the suicide attempt there is a marked increase in the percentage of *sadness* tweets and then a noticeable increase in *anger* and *sadness* the week following the suicide attempt (to include the day of the attempt). Some examples of tweets from the day of the suicide attempt and tagged as *anger* or *sadness* are shown in Figure 8,

My parents admitted they ignore my mental health, I am so pissed off now. I'm only good for being a verbal punching bag. Why can't I find my damn pills so I can just fucking overdose?
I listed reasons I should die and reasons I should not die. I had no reasons not to die. I found 7 reasons to die. two people next to each other in the same room can be in totally separate places, one of the saddest truths I'm totally pathetic even the scars from my attempts are pathetic

Figure 8: Example tweets labeled with *anger* (top) and *sadness* (bottom) from the day of a suicide attempt.

as an illustration of what signals may be driving that change. In some of these tweets, the depth of emotion is more complex than is captured by these simplistic labels – some indicate that the author is angry at themselves and the situation they find themselves in, perhaps in line with the *guilt* and *shame* found by Bryan et al. (2013).

Contrasting *anger* and *sadness*, the percentage of *fear* and *disgust* tweets appear in line with neurotypicals prior to their attempt, yet they decrease to levels below neurotypicals after the attempt. They also appear to have a consistently lower amount of tweets that convey *loneliness*, which decreases further after their attempt. There are a number of apparent single-week shifts away from neurotypical or away from the users who have attempted to take their life, though drawing conclusions on any of them would be prematurely speculative. These should serve as grist for more directed studies in the future. No interesting trends were observed for *anxiety* so it was omitted for brevity.

People who attempt to take their life tend to have a higher overall proportion of tweets estimated to be emotional, and that proportion tends to *increase* after their attempt. Intriguingly, this finding seems (at first blush) at odds with the results from the Vennclouds and Figure 4, where users who attempted suicide used emoticons and emoji less frequently than neurotypicals. Taken together, these might indicate that though users who attempt suicide express more emotion, they do so with words rather than emoticons or emoji – perhaps suggesting a depth of emotion that are not adequately served by the vast array of emoji.

Volume: Finally, some interesting changes in the overall volume of activity are illustrated in Figure 9. Users who attempt to take their life generate tweets at a level higher than neurotypicals prior to their attempt, but after their attempt appear to return to levels commensurate with neurotypicals. One possible explanation for this might be an implicit or explicit call for help, though deeper analysis is certainly required.

6 Caveats and Limitations

When drawing conclusions from this work, there are some caveats and limitations to keep in mind, any of which may affect the generalizability of the findings – all suggesting future, more controlled studies. All the people investigated here survived their suicide attempt, so there may be systematic differences between those in our dataset and those who die by suicide. Similarly, we have no verification of the attempts of these users, though the data has face validity with existing research on suicide. The data explored here is primarily from women aged 15-29. While this is a group at elevated risk for suicide, their behavior, motivations, and stressors are likely significantly different from other at-risk groups (e.g., transgendered individuals or middle-aged males). Furthermore, these users self identify and present themselves as challenged with a highly stigmatized issue in a very public manner. It is clear this is a subpopulation separate from neurotypical controls. We cannot be sure, however, exactly how different this population might be from the larger cohort who has attempted to take their life.

7 Conclusion

The caveats above notwithstanding, we have provided an empirical analysis of the language usage of people prior to a suicide attempt, to be used as grist for further exploration and research. Ideally, even these simple analyses can provide a foundation for non-invasive screening and interventions to prevent suicides. However, significant challenges exist in applying this technology broadly in ways that preserve privacy and maintain a high standard of care

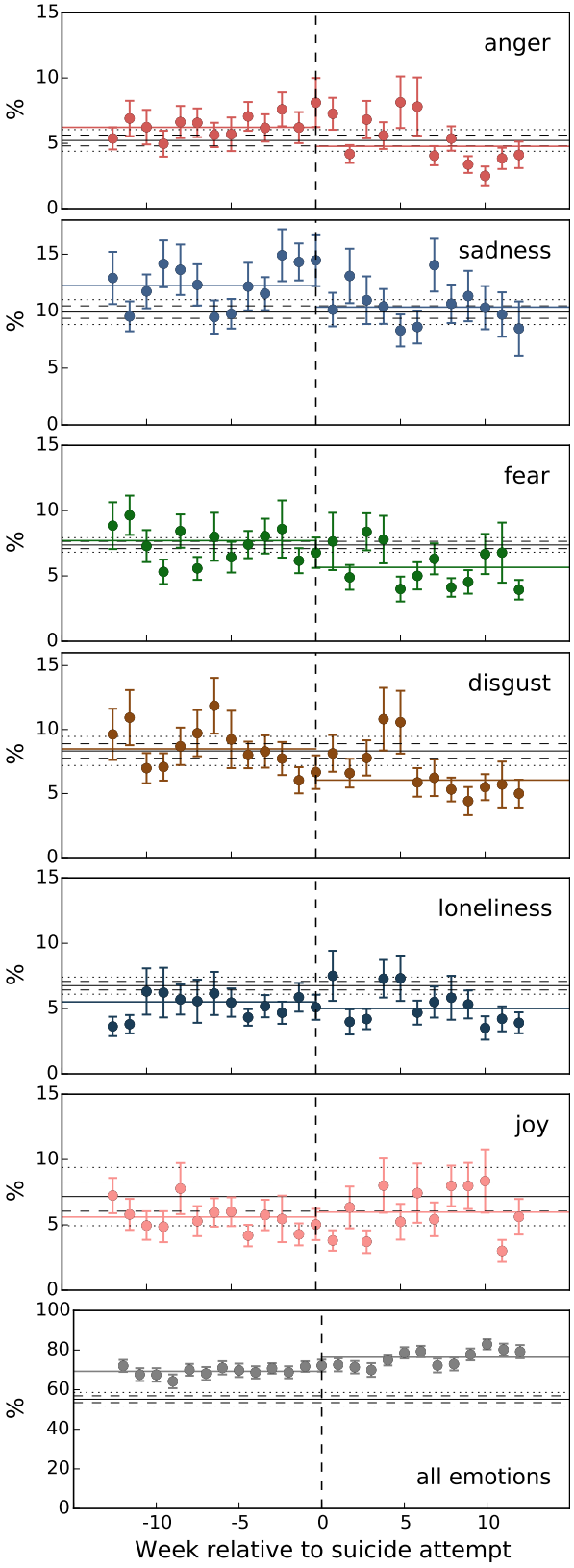

Figure 7: Emotion-labeled tweets from users who tried to take their life.

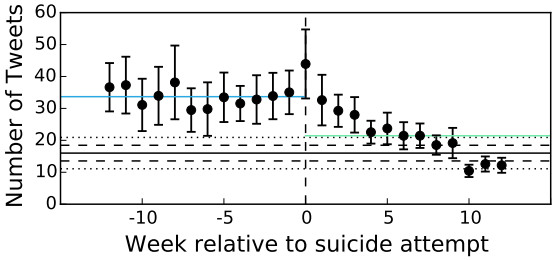

Figure 9: Volume of tweets from users who tried to take their life (dots), surrounding their suicide attempt. All features of the plot are equivalent to those in Figure 7.

using limited clinical resources. Despite the potential lives saved, the general population may not be amenable to its use given the perceived cost to privacy, as reaction to the Samaritan's Radar[2], suggests. However, **opt-in** continual analysis of social media data may be a reasonable method for ecological momentary assessment and for monitoring psychological and behavioral state over time. For further discussion of the ethics, privacy, and practical considerations around interventions using this kind of technology, see Wood et al. (2016).

Suicide is a large and looming challenge, claiming a tragic number of lives each year. Given the societal stigma, discrimination, and prejudice associated with it, finding data to better understand the risk of suicide has been a consistent challenge. Our analysis here suggests some future directions for exploration, along with providing some quantified insight into the phenomena of acute suicidal risk. It is a small but important step towards improved outcomes and lives saved.

Acknowledgments

The authors would like to thank April Foreman, Bart Andrews, Bill Schmitz and the #SPSM community for their continued intellectual and emotional contributions to the improvement and empowerment of suicide prevention. We would also like to thank the anonymous reviewers for improving the quality of this manuscript, and the organizers of CLPsych for their contributions to this research community.

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

References

Craig J Bryan, Leigh G Johnson, M David Rudd, and Thomas E Joiner. 2008. Hypomanic symptoms among first-time suicide attempters predict future multiple attempt status. *Journal of clinical psychology*, 64(4):519–530.

Craig J Bryan, Chad E Morrow, Neysa Etienne, and Bobbie Ray-Sannerud. 2013. Guilt, shame, and suicidal ideation in a military outpatient clinical sample. *Depression and anxiety*, 30(1):55–60.

Cindy Chung and James Pennebaker. 2007. The psychological functions of function words. *Social Communication*.

David M Clark and John D Teasdale. 1982. Diurnal variation in clinical depression and accessibility of memories of positive and negative experiences. *Journal of abnormal psychology*, 91(2):87.

Glen Coppersmith and Erin Kelly. 2014. Dynamic wordclouds and Vennclouds for exploratory data analysis. In *Association for Computational Linguistics Workshop on Interactive Language Learning and Visualization*, June.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014a. Quantifying mental health signals in Twitter. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014b. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 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, June. North American Chapter of the Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 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*.

Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. 2015c. Quantifying suicidal ideation via language usage on social media. In *Joint Statistics Meetings Proceedings, Statistical Computing Section*. JSM.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th ACM International Conference on Web Science*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Maeve Duggan, Nicole B Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. 2015. Social media update 2014. *Pew Research Center*, 19.

Kristy Hollingshead and Lyle Ungar, editors. 2016. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Association for Computational Linguistics, San Diego, California, USA, June.

Laura Kann, Steve Kinchen, et al. 2014. Youth risk behavior surveillance – united states, 2013.

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

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM conference on Hypertext and hypermedia*. ACM.

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

Margaret Mitchell, Glen Coppersmith, and Kristy Hollingshead, editors. 2015a. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Association for Computational Linguistics, Denver, Colorado, USA, June.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015b. 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, June. North American Chapter of the Association for Computational Linguistics.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255,

Montréal, Canada, 7-8 June. Association for Computational Linguistics.

National Institute of Mental Health National Institutes of Health. 2013. Statistics: Any disorder among adults. http://www.nimh.nih.gov/statistics/1ANYDIS_ADULT.shtml. [Online; accessed 2013-03-05].

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. 2008. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry*, 192(2):98–105.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, and Matthieu Perrot Édouard Duchesnay. 2011. scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic inquiry and word count: LIWC2001*. Erlbaum Publishers, Mahwah, NJ.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. 2007. *The development and psychometric properties of LIWC2007*. LIWC.net, Austin, TX.

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

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 482–491, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nairan Ramirez-Esparza, Cindy K. Chung, Ewa Kacewicz, and James W. Pennebaker. 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)*.

Bobbie N Ray-Sannerud, Diana C Dolan, Chad E Morrow, Kent A Corso, Kathryn E Kanzler, Meghan L Corso, and Craig J Bryan. 2012. Longitudinal outcomes after brief behavioral health intervention in an integrated primary care clinic. *Families, Systems, & Health*, 30(1):60.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 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.

Philip Resnik, Rebecca Resnik, and Margaret Mitchell, editors. 2014. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Baltimore, Maryland, USA, June.

Lindsay Robertson, Keren Skegg, Marion Poore, Sheila Williams, and Barry Taylor. 2012. An adolescent suicide cluster and the possible role of electronic communication technology. *Crisis*.

J. Niels Rosenquist, James H. Fowler, and Nicholas A. Christakis. 2010. Social network determinants of depression. *Molecular psychiatry*, 16(3):273–281.

Maarten Sap, Greg Park, Johannes C. Eichstaedt, Margaret L. Kern, David J. Stillwell, Michal Kosinski, Lyle H. Ungar, and H. Andrew Schwartz. 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.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Richard E. Lucas, Megha Agrawal, Gregory J. Park, Shrinidhi K. Lakshmikanth, Sneha Jha, Martin E. P. Seligman, and Lyle H. Ungar. 2013a. Characterizing geographic variation in well-being using tweets. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013b. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8(9).

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through Facebook. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32.

Arthur A Stone, Saul Shiffman, Joseph E Schwartz, Joan E Broderick, and Michael R Hufford. 2003.

Patient compliance with paper and electronic diaries. *Controlled clinical trials*, 24(2):182–199.

E Sullivan, Joseph L Annest, F Luo, TR Simon, and LL Dahlberg. 2013. Suicide among adults aged 35–64 years, united states, 1999–2010. *Center for Disease Control and Prevention, Morbidity and Mortality Weekly Report*.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Anthony Wood, Jessica Shiffman, Ryan Leary, and Glen Coppersmith. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

World Health Organization et al. 2014. *Preventing suicide: A global imperative*. World Health Organization.