Language Signals Preceding Suicide Attempts

Anthony Wood
Qntfy
tony@qntfy.io

Jessica Shiffman
Qntfy
jesse@qntfy.io

Ryan Leary
Qntfy
ryan@qntfy.io

Glen Coppersmith
Qntfy
glen@qntfy.io

ABSTRACT

A terrifying number of people die by suicide each year, and a tragic number of friends and family members suffer with each loss. We derive insights and quantifiable signals from the language of social media users who have previously attempted to take their own lives. We demonstrate a simple machine learning classifier applied to the language of a user's social media posts that could be used as a part of a screening process. Our results suggest that if used in the real world, between one third and one half of the users flagged for further screening would attempt suicide. Remarkable as this technology may be, it is a minuscule portion of a successful and scalable strategy for detection and intervention. The real challenge lies in integrating the technology into the complex human processes that surround effective mental health care and making efficient use of already overburdened clinical resources. Our results suggest that this technology may be useful, but many barriers remain; among them are: significant ethical and legal matters, individuals access to care, and the poor scalability of interventions.

Author Keywords

Mental Health; Suicide; Suicidal Ideation; Twitter; Computational Linguistics; Natural Language Processing

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Suicide is expected to claim the lives of 4.8 million Americans alive today (approximately 1%), and nearly 13 million will attempt suicide [15]. Suicide is a top-10 cause of death in the United States. It is the second leading cause of death for teenagers worldwide, and the leading cause of death for women ages 15-19 worldwide [22].

The design of effective interventions, optimization of scarce clinical resources and effective evaluation of population-level public health initiatives revolves around the ability to detect when someone is in psychological distress. Screening people for suicide risk is a manual process that must be performed by healthcare workers in person. Anecdotally, it is sometimes neglected due to the societal stigma around suicide. Thus, a strategy incorporating a semi-automated and objective screening tool may be warranted. Ultimately, the results here are a small step on the path to a more empowered mental health field. However, only through careful integration with existing human processes and clinical workflows will any benefit come to those suffering or to the clinicians determined to help them. Our main contributions are **[1]** evidence that quantifiable signals relevant to suicidal ideation exists in language usage on social media and **[2]** discussion of the barriers to using such technology in practice.

PRIOR WORK

In recent years, quantified signals relevant to mental health have been found in social media [1, 3, 4, 6, 8, 16, 21]. Most of the work focuses on relatively common conditions like depression, but recently there has been progress looking specifically at suicide [5,7,11,12]. For brevity, we omit detailed discussion of this background work, but refer readers interested in this area (and more generally the intersection of computational linguistics and clinical psychology) to the proceedings of CLPsych workshops [9, 14, 19].

Importantly, language usage provides a rich data source for mental health related signals, and social media provides a rich source of language data and associated metadata.

DATA

For brevity, we describe only the most salient aspects of the data collection and processing. The interested reader will find more details in [2, 5].

Users who Attempted to Take Their Life

We examine users who **publicly** state that they have tried to take their own life, and provide sufficient evidence for a casual reader to identify when they did so. A human annotator read each of these tweets and determined (1) whether the statement appears to be genuine[1] and (2) the date of the attempt. The task was to only label dates for "entirely unambiguous" statements of their last attempt, which yielded perfect agreement on a sample annotated by two of the annotators. This nets 125 users: (1) who have attempted to take their life, (2) disclose the date of their last suicide attempt, and (3) have data available prior to that attempt. For each user who has attempted to take their life, we estimate their age and gender according to their authored content using the lexica made available by the World Well-Being Project [20]. Gender estimates have an accuracy of 91.9% and age estimates correlated with true age at $r = 0.83$. From a large pool

[1]High agreement between annotators has been found for differentiating genuine statements from disingenuous statements on similar tasks involving mental health conditions, $\kappa = 0.77$ [5].

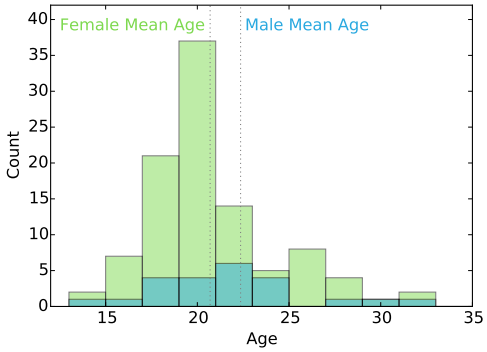

Figure 1. Histogram of the ages of users who have previously attempted to take their life. Females in green, and males in blue. The mean age of each gender is denoted by vertical lines.

of random English users, we select a matched control user of the same estimated gender and the closest estimated age for each user who has attempted to take their own life. We proceed as if these control users are neurotypical (i.e., have not attempted to take their own life), though this assumption is likey untrue. We expect that our neurotypical users are contaminated with 4-8% of users have attempted to take their life (5-10 users), commensurate with the rates in the general population [15]. All data was accessed via the public Twitter API, this **excludes** private (direct) messages, users who elected to be *private*, or posts the user deleted.

Figure 1 indicates that the users examined in this study are primarily women between the ages of 15 and 29, estimated as above – a group at high risk for attempting to take their own life [13, 15]. Thus, our findings, may not generalize to other at-risk groups (e.g., middle-aged white males). Interestingly, the group examined here are digital natives, which should inform intervention strategies, especially since much of the activities of this group are digitally mediated.

SUICIDAL IDEATION SCORES FROM LANGUAGE

We create simple classifiers from the language generated by the users who have attempted to take their lives and their age- and gender-matched controls. We deliberately use simple classifiers conducive to introspection and analysis (character n-gram language models), and only minimally optimize parameters, since the primary objective here is illustrative application. More complex models and incorporation of additional non-linguistic features will likely yield higher performance numbers, as demonstrated elsewhere [2].

Briefly, we train a pair of models, one from the language of users who have attempted to take their own life, and one from a their age- and gender-matched controls. These models estimate the probability that a given segment of language was generated by each group of users. When presented with novel text (i.e., not text used to train either model), the model estimates the probability that the text was generated by someone who has attempted to take their life, and the probability it was generated by a control user. Whichever probability is higher is considered the decision for the model on this given segment of text (and has an associated score). In practice, each Tweet

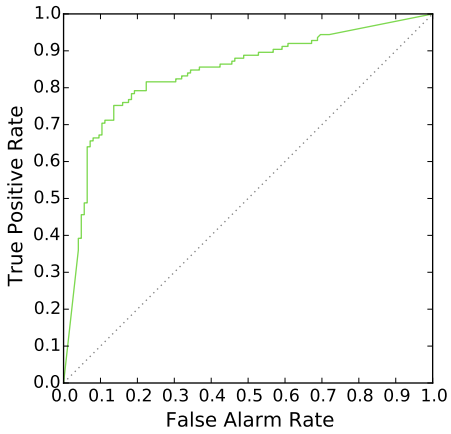

Figure 2. ROC curve for separating those who attempted to take their life from their matched controls. False alarms are on the x-axis and true positives on the y-axis, as we vary the sensitivity of the model.

gets a score, and there are many schemes for how to aggregate those Tweet-level scores to a single number to calculate raw performance statistics, omitted for brevity[2].

EXPERIMENTS

We demonstrate that there are sufficient quantifiable signals to create models capable of separating users who have attempted to take their life from their age- and gender-matched controls. Then, to provide qualitative intuition and grounding of the models in existing psycholinguistic theory, we then examine the language that these models depend on to function.

Quantified Signals

To demonstrate that quantifiable signals relevant to suicidal ideation exist in social media, we build simple language-based classifiers to separate users who have attempted to take their life from their age- and gender-matched controls. We examine only data **prior** to a user's suicide attempt, in line with the envisioned application of this technology.

For a single performance number, at 10% false alarms (ostensibly neurotypical users identified as at risk to take their life), our models correctly identify 70% of the users who do attempt to take their life. For a more detailed view of the true positive/false alarm tradeoff, a ROC curve for this simple linguistic model, see Figure 2. For brevity, many parameter settings are roughly equivalent and yield performance distinctly better than chance (dotted gray line) but far from perfect (which would go from lower left, to upper left, to upper right). This is taken as evidence that the models find some quantifiable signals relevant to suicidal ideation automatically from language usage on social media.

These performance numbers are based on artificially balanced data sets, as appropriate for finding quantifiable signals, but not for the clinical setting. In this population, 4-8% of the users are expected to attempt to take their life, and 17% are challenged with suicidal ideation [10, 15]. Thus, the negative class (i.e., neurotypical controls) are likely to be a bit

[2]See [5] for more details.

more than ten times that of the positive class (i.e., those who attempt to take their life). In practice, given the contaminated training sample it is impossible to have an accurate estimate of performance in the clinical setting, but we can suggest bounds on the performance of this simple classifier. Let us assume we have a population of 1000 people aged 15-29, where we expect 17%, or 170 people to be challenged by suicidal ideation and 6%, or 60 people to attempt to take their life, leaving 940 who do not. Random selection for additional screening would result in a hit rate commensurate with the rate in the population: 6%. The worst-case scenario is that there is no contamination in the neurotypical population, so all users there are true negatives (as we assume in the ROC). We would then expect 42 (70% of 60) at risk individuals and 94 (10% of 940) neurotypical would be selected for additional screening – a hit rate of 30%. The best-case scenario is that the system is correctly identifying the contaminating users in the neurotypical population, so 6% of our false alarms are actually at-risk rather than neurotypical. This optimistic operating point is instead 4% false alarms (10% - 6%) at 70% precision – identifying 58 at risk individuals and 56 (6% of 940) not-at-risk for additional screening – a hit rate of 51%. We can roughly conclude that between one third and one half of those flagged for additional screening by these simple measures would be actually at risk to attempt suicide.

There is much more exploration and tuning needed before use in a clinical setting, but this does provide impetus for discussions of how such technology would be most useful and what the implications of its use are.

What Language Drives the Model?

To provide some grounding for how these models relate to existing psychological research, we use the models to score the words that comprise the categories in the Linguistic Inquiry Word Count (LIWC), a psychometrically validated lexicon [18]. We score each category with the models and examine those with the highest differences between those who have attempted to take their own life and their matched controls. Some example categories more likely to be generated by users who have attempted to take their life are DEATH, HEALTH, SAD, THEY, I, SEXUAL, FILLER, SWEAR, ANGER, and NEGATIVE EMOTIONS. Some of these categories have an obvious connections to suicide and suicidal ideation. Intriguingly, some have been examined for their role in other psychological phenomenon [17]. For example, THEY is used when talking about a group that the author is not part of, and I is used by the author in self reference. Exactly what these changes mean will require deeper analysis and more controlled studies, but one potential explanation that fits with existing suicidality research is that these changes might indicate increased social isolation in those that attempted to take their life. Furthermore, some of these same changes in language have been found for other mental health conditions often concomitant with suicidal ideation and suicide attempts [3].

BALANCING PRIVACY WITH INTERVENTION

Assuming that technology will be able to identify some people at risk for taking their own life (imperfectly), what happens next? How should the person's privacy be balanced against preventing them from taking their own life? This technology would scale arbitrarily as an imperfect screening tool, but that would put additional burden on already-overburdened clinical resources. Cautionary tales are also to be found there as some users will feel that their privacy has been invaded, as with Samaritans Radar[3]. Since advertisers continually perform the same calculations to serve ads and receive little complaint from the public, the difference seems to be in providing this information to another arbitrary human, without the consent of the user being examined.

For users willing to opt-in to such analysis, concerns of privacy and informed consent can be easily managed. This suggests effective screening is possible for these users, but that strategy will fail to reach a significant portion of those most at risk. To reach population segments who will not opt-in will likely require an opt-out procedure. This is an extremely delicate path which requires significantly more consideration and nuance than possible here. Some consideration of the opt-out path include **[1]** public discussion and consensus around balancing privacy with saving lives, **[2]** assuring that information would only reach trained and vetted individuals (e.g., clinicians or peer-support personnel), **[3]** anonymous communication channels (e.g., `7cups.com`) to ease the burden of care and the barriers erected by stigma, prejudice, and discrimination, and **[4]** adjustment of the legal requirements for clinicians to take action when made aware that someone is at risk – such requirements with their duty of care would immediately outstrip their ability to provide such care. Ultimately, though, to save the most lives requires an opt-out procedure, which is at odds with the general public's view on privacy.

CAVEATS

While the rates from people aged 15-29 are high, they are relatively stable over the last few years. This means that this data may not be helpful in addressing the climbing suicide rate (largely driven by middle-aged males). Our analysis is based on users who publicly discuss a sensitive subject, which some consider taboo. These users may differ from the general population of those challenged with suicidal ideation or who might take their own lives in systematic ways. Furthermore, while our sample may be large for some suicide-related studies, it is at best medium-sized by most statistical, natural language processing, and Internet research standards.

CONCLUSION

These caveats notwithstanding, our results indicate interesting possibilities for the role that technology can play in generating additional insight into suicide and suicidal ideation. They provide further evidence of the power of quantifiable signals present in language usage. Interestingly, digitally mediated interventions facilitated by this technology might be particularly effective for this digitally-native group. Most importantly, they bring into stark relief the critically important concerns around privacy and the ethics of intervention. These questions will not be easily or quickly resolved, and likely should be a matter of public discussion and debate. Ultimately, we must find a reasonable tradeoff between the need

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

to preserve privacy and the need to intervene with those at risk of taking their own life.

ACKNOWLEDGMENTS

The authors would like to thank April Foreman, Bart Andrews, William Schmitz, and the #SPSM group for their insight and open-mindedness in addressing suicide prevention. We would also like to thank the anonymous reviewers whose comments greatly strengthened this work.

REFERENCES

1. Chung, C., and Pennebaker, J. The psychological functions of function words. *Social Communication* (2007).
2. Coppersmith, G., Dredze, M., and Harman, C. Quantifying mental health signals in Twitter. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology* (2014).
3. Coppersmith, G., Dredze, M., Harman, C., and Hollingshead, K. 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*, North American Chapter of the Association for Computational Linguistics (Denver, Colorado, USA, June 2015).
4. Coppersmith, G., Harman, C., and Dredze, M. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)* (2014).
5. Coppersmith, G., Leary, R., Whyne, E., and Wood, T. Quantifying suicidal ideation via language usage on social media. In *Joint Statistics Meetings Proceedings, Statistical Computing Section*, JSM (2015).
6. De Choudhury, M. Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd International Workshop on Socially-Aware Multimedia* (2013).
7. Dinakar, K., Jones, B., Lieberman, H., Picard, R., Rose, C., Thoman, M., and Reichart, R. You too?! mixed-initiative LDA story matching to help teens in distress.
8. 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. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science 26*, 2 (2015), 159–169.
9. Hollingshead, K., and Ungar, L., Eds. *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 2016.
10. Kann, L., Kinchen, S., et al. Youth risk behavior surveillance – united states, 2013.
11. Kiciman, E., Kumar, M., Coppersmith, G., Dredze, M., and De Choudhury, M. 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 (2016).
12. Kumar, M., Dredze, M., Coppersmith, G., and De Choudhury, M. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM conference on Hypertext and hypermedia*, ACM (2015).
13. Mendez-Bustos, P., Lopez-Castroman, J., Baca-García, E., and Ceverino, A. Life cycle and suicidal behavior among women. *The Scientific World Journal 2013* (2013).
14. Mitchell, M., Coppersmith, G., and Hollingshead, K., Eds. *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 2015.
15. 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. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry 192*, 2 (2008), 98–105.
16. Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., and Seligman, M. E. Automatic personality assessment through social media language. *Journal of personality and social psychology 108*, 6 (2015), 934.
17. Pennebaker, J. W. The secret life of pronouns. *New Scientist 211*, 2828 (2011), 42–45.
18. Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. *The development and psychometric properties of LIWC2007*. LIWC.net, Austin, TX, 2007.
19. Resnik, P., Resnik, R., and Mitchell, M., Eds. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Baltimore, Maryland, USA, June 2014.
20. Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., Ungar, L. H., and Schwartz, H. A. Developing age and gender predictive lexica over social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014), 1146–1151.
21. 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. Characterizing geographic variation in well-being using tweets. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)* (2013).
22. World Health Organization, et al. *Preventing suicide: A global imperative*. World Health Organization, 2014.