It's quality and quantity: the effect of the amount of comments on online suicidal posts

Daniel M. Low
Harvard University & MIT
dlow@g.harvard.edu

Kelly L. Zuromski
Harvard University
{kelly_zuromski,dtkessler}@fas.harvard.edu

Daniel Kessler
Harvard University

Satrajit S. Ghosh
MIT & Harvard Medical School
satra@mit.edu

Matthew Nock
Harvard University
nock@wjh.harvard.edu

Walter Dempsey
University of Michigan
wdem@umich.edu

Abstract

Every day, individuals post suicide notes on social media asking for support, resources, and reasons to live. Some posts receive few comments while others receive many. While prior studies have analyzed whether specific responses are more or less helpful, it is not clear if the quantity of comments received is beneficial in reducing symptoms or in keeping the user engaged with the platform and hence with life. In the present study, we create a large dataset of users' first r/SuicideWatch (SW) posts from Reddit (N=21,274), collect the comments as well as the user's subsequent posts (N=1,615,699) to determine whether they post in SW again in the future. We use propensity score stratification, a causal inference method for observational data, and estimate whether the amount of comments —as a measure of social support— increases or decreases the likelihood of posting again on SW. One hypothesis is that receiving more comments may *decrease* the likelihood of the user posting in SW in the future, either by reducing symptoms or because comments from untrained peers may be harmful. On the contrary, we find that receiving more comments *increases* the likelihood a user will post in SW again. We discuss how receiving more comments is helpful, not by permanently relieving symptoms since users make another SW post and their second posts have similar mentions of suicidal ideation, but rather by reinforcing users to seek support and remain engaged with the platform. Furthermore, since receiving only 1 comment —the most common case— decreases the likelihood of posting again by 14% on average depending on the time window, it is important to develop systems that encourage more commenting.

1 Introduction

Suicide is among the leading causes of death in the US and worldwide and the second leading cause of death among youth, 10 to 34 years old (Fortgang and Nock, 2021). Despite a century of scientific research on suicide prevention, suicide rates are similar to what they were 100 years ago because suicide is still very hard to predict and treat (Fortgang and Nock, 2021; Bentley et al., 2020) and because many individuals at risk do not receive optimal treatment due to known perceived barriers such as stigma and cost (Mohr et al., 2010; Andrade et al., 2014). Instead of seeking professional help, many individuals seek support online from peers, which could have negative consequences including learning about suicide methods, receiving ineffective responses, and being subject to suicide contagion (Niederkrotenthaler et al., 2016; Colombo et al., 2016). Therefore, there is an urgent need to model and understand the effect of these responses.

1.1 Prior work

When analyzing the content and quality of peer responses to suicidal posts on social media, prior work has found that responses are generally positive (Naslund et al., 2016; Fu et al., 2013; Li et al., 2015; Jiang et al., 2020; O'dea et al., 2018), suggesting that peer responses may be helpful peer-delivered interventions. One study of the subreddit r/SuicideWatch (SW) found that posters rated the most helpful peer responses as including professional help suggestion (e.g., "Have you tried getting any professional help?"), life meaning (e.g., "Instead of reasons why not to, perhaps think of reasons to live…"), and relationship/loss support (e.g., "I am really sorry for your loss...") (Jiang et al., 2020). Another study found that online users in non-professional suicide forums tend to see reductions in the symptoms of their suicidal ideation when responses to their posts include constructive active listening, collaborative problem-solving, constructive advice, debunking of public suicide myths, alternatives to suicide, stories of lived experience, and recommendations of help services (Niederkrotenthaler et al., 2016). In Jiang et al.

(2020) the most frequent response type was asking questions (e.g., "What kinds of things in the past have you tried that have helped you cope before?"), compared to treatment and medication (e.g., "I recommend medication for anxiety"), which was both the least frequent and lowest-rated response in terms of perceived helpfulness. This finding aligns with studies on intervention messaging engagement (Owens et al., 2011; Whiteside et al., 2014), that rank question-form prompts, or prompts designed to initiate responses (e.g., "Do you want to talk?"), as among the most engaging.

Although there are relatively few studies examining the impacts of specific peer responses on suicidal posts, numerous other studies have highlighted messaging content features that lead to engagement among participants in online interventions. For example, prior work has found that people tend to be attracted to, and engage with, intervention-related messages that suggest they are cared about, show concern or caring for the individual, are personalized to them (even if personalization is automated), provide information on resources, and that others experiencing similar symptoms found helpful (Aguilera and Berridge, 2014; Whiteside et al., 2014). Furthermore, participants of such message-based interventions tend to engage with messaging that encourages help-seeking (e.g., "Sometimes family can be supportive during tough times") (Pisani et al., 2018), provides psychoeducational information (Jaroszewski et al., 2019; Murray et al., 2015) or that provides validation and normative feedback (e.g., "It's okay to feel angry") (Owens et al., 2011).

De Choudhury and Kiciman (2017) used stratified propensity score matching to estimate which ngrams make it more likely that users posting on 14 mental health subreddits (e.g., r/Depression, r/ptsd, r/mentalhealth) would later post on SW for the first time in comparison to those who never end up making the transition to this overall higher-severity subreddit. They modeled different ngrams in the comments as treatments (i.e., interventions on a given outcome), ngrams appearing prior were covariates, and posting on SW was the binary outcome.

While these studies have identified the distribution, quality, and potential effects of specific types of comments, it is not clear whether receiving more or less comments overall is helpful in keeping the user engaged.

1.2 Current study

SW posts receive 0 to many peer responses in the form of comments. Here we study the effect of receiving few or many replies to suicidal posts. We estimate the likelihood of posting again on SW as function of how many comments a post received. This could be thought of as estimating the treatment effect given a particular dosage in a randomized trial. There are multiple possible hypotheses:

- H0: The amount of comments is not associated with whether users will post on SW again.

- H1: Receiving more comments *decreases* the likelihood of posting again on SW. Plausible reason: receiving more comments is more helpful, that is, it reduces symptoms which would result in a decreased need to post in SW in the future. It could also indicate something very different; namely that receiving more non-professional responses is ineffective or harmful, and therefore users tend to not to use the forum in the future. If receiving more comments decreases the likelihood of posting again in SW (H1), it seems unlikely that this would be due to the effectiveness of peer responses from mostly untrained individuals. Helping reduce an individual's suicidal thoughts and behaviors (STBs) is very challenging and seems to require more substantial treatments such as cognitive behavioral therapy, mindful emotion awareness, generating cognitive flexibility, and emotion exposure (Bentley et al., 2020). Even when undergoing formal treatment, about half of individuals with depression relapse within 2 years of treatment (Steinert et al., 2014). Therefore, viewing responses as ineffective might be the most likely explanation to a decrease in returning to SW as users receive more responses.

- H2: Receiving more comments *increases* the likelihood of posting again in SW. Plausible reason: it seems comments would be reinforcing support-seeking behavior and not reducing symptoms, because if they were reducing STBs, then users would be less likely to post again. However, it is possible that users that post in SW again are returning to report signs of improvement, which we will will check by comparing the amount of mentions of STBs

	25%	50%	75%	max
Prior posts	0	3	15	4436
Subs. posts	1	9	43	9566
Subs. SW posts	0	0	1	235
Comm. per post	1	2	5	261

Table 1: Descriptive statistics per user. Minimum = 0 for all rows. SW: r/SuicideWatch; Subs.: subsequent; comm.: comments.

between first and second posts using custom lexicons.

2 Methods

2.1 Dataset

We obtained 24,401 SW posts from 2018 and 2019 from the Reddit Mental Health Dataset (Low et al., 2020). We then used the Reddit pushshift API [1] to download all posts by those users until 2021-07-20 (N=1,615,699). We removed users that had deleted their accounts, posts that were deleted by their users as well as posts removed by moderators. For every user, we located their first SW post (N=21,274 unique posts and users) and downloaded the corresponding comments (N=102,394) (see Table 1). First SW posts range from 2010-09-14 to 2019-04-21 and we know whether users posted again in SW or not until 2021-07-20, a minimum of 27 months, which we consider a large enough time window to estimate comments' effects on initial SW posts.

For 33% of users (N=6,954), posting on SW was their first time posting on Reddit, and 10% (N=2,176) posted on Reddit only once before. After their first SW post, 12% of users (N=2,482) posted in SW again within a week, and 36% of users (N=7,749) posted in SW again within a year (see Fig. 1). For our analysis, we chose the following time windows in days: 7, 14, 21, 30, 60, 90, 180, 365, because we expect effects from the amount of comments received to be most present sooner (e.g., 1 or 2 weeks) and then fade with larger time windows (e.g., 6 months, 1 year), when other factors may play a larger role regarding whether users post again.

20% of users (N=4,207) never posted on Reddit again. For some, death is a possible outcome given posts often describe imminent suicidal plans (e.g., "Hello, I keep it short because i will kill myself today. (...)"), but not posting again can also be related to not finding Reddit helpful or symptoms improving and not needing any further help. Therefore, not posting on Reddit again should be considered an ambiguous outcome in that it can be positive or negative. We removed post authors' replies from the count of comments received. Half the posts in SW receive up to 2 comments (see Table 2). Each comment is counted, even if they are from the same commenter.

[1] https://github.com/pushshift/api

N	Posts receiving N comments	Proportion
0	1777	0.08
1	5193	0.24
2	3689	0.17
3	2421	0.11
4	1725	0.08
5	1313	0.06
6	939	0.04
7	721	0.03
8	542	0.03
9	445	0.02
10+	2509	0.12

Table 2: Amount of posts with N amount of comments.

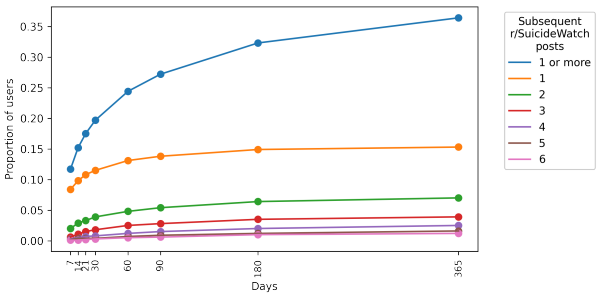

Figure 1: Proportion of users that post again in r/SuicideWatch.

2.2 Estimating the effect of comments through propensity score stratification

Given a causal model such as the one depicted in Fig. 2, we can infer the causal effect of treatment. Whether a user will re-post on SW again depends not only on the amount of comments they received (the treatment) but also on the severity and content of their post (e.g., more severe posts may capture more suicidality which would make it more likely to post again on SW), which in turn may also influence how many comments they will receive. In an ideal setting, treatment is randomized so that the treatment assignment mechanism is ignorable

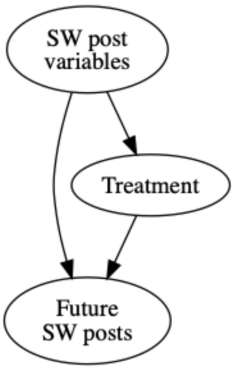

Figure 2: A directed acyclic graph of the model with Treatment (T; indicator variable of N comments), Outcome (Y; discrete variable of how many times the user posted in SW in the future) and Confounders (C; 48 linguistic variables extracted from the SW post). We expect the content of the post to influence the amount of comments received (T) and the likelihood to post in SW again (Y). SW: r/SuicideWatch.

and independent of counterfactual outcomes. In observational settings, however, treatment is not randomly assigned. Under the causal model in Fig. 2, treatment assignment is ignorable given the extracted linguistic variables because exposure to treatment is no longer confounded by post characteristics. In this case, the propensity score, the probability of treatment exposure given the observed covariates, is the basis for adjusting for confounding (Lunceford and Davidian, 2004):

$$\hat{e}(C) = P(T = 1|C)$$

where T is treatment (here, for simplicity, it is a binary indicator) and C is the vector of confounders. Under the causal model, $T \perp\!\!\!\perp C|\hat{e}(C)$, so that individuals from either treatment group with the same propensity score are "balanced". The treatment indicates receiving a certain amount of comments; therefore, we built one model per treatment from 1 to 10+ comments, and each treatment is compared to receiving 0 comments. We next discuss how we exploit the causal model to estimate the treatment effect using propensity score stratification.

We used the *dowhy v0.6* package implementation of propensity score stratification (Sharma et al., 2019). The analysis used the following steps: (1) train a logistic regression classifier with L2 penalty to predict T from C. For every post, compute the propensity score based on the estimated logistic regression model; (2) based on the score, stratify (i.e., rank scores so that high propensity scores of treated and untreated are in the same stratum or subgroup of similar samples, and split evenly with at least 11 samples per strata); (3) compute the per stratum average treatment effect (ATE) as the weighted average of outcome differences per strata; (4) estimate the overall ATE by a weighted sum of the differences of sample means across strata where the weight is proportional to the number of observations falling in each stratum. See Lunceford and Davidian (2004) for technical details.

To validate estimates, we include two methods from the *dowhy* package: (1) data subset refuter: we make sure estimates do not change in comparison to the mean estimate run on multiple subsets of the data (similar to cross-validation); (2) placebo treatment refuter: randomly assign a covariate as a treatment and re-run the analysis, the estimate of which should be close to 0 with a high p-value (e.g., above 0.05).

2.3 Feature extraction

To capture the content of the post, we extracted 39 Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015) variables including anxiety, anger, sadness, personal pronouns, and swearing. These were chosen because they reflect emotional and psychological processes. Nine additional lexicons were created to capture different subtypes of suicidal thoughts and behaviors (active, passive, self-harm) as well as stressors (domestic stress and violence, substance use, financial stress, and mental health; see Appendix A.1 for details).

3 Results

3.1 The effect of amount of comments on future SW posts

Receiving 1 comment *decreased* the likelihood of posting on SW again by 14% on average across the time windows. As comments increase, so does the likelihood of posting on SW (increasing the likelihood up to 16% on average), providing evidence for hypothesis H2, especially up until 5 comments and again with 9 or more comments (see Fig. 3). The mean changes in likelihood over time windows (and range) for the different amount of comments are 1: -14% [-23% – -8%]; 2: -13% [-21% – -8%]; 3: -6% [-9% – -4%]; 4: 0% [-3% – 1%]; 5: 8% [1% – 13%]; 6: 7% [-2% – 28%]; 7: -4% [-9% – -1%]; 8: -1% [-12% – 8%]; 9: 6% [-5% – 11%]; and 10+: 16% [0% – 27%].

To validate these estimates, using validation

method 1 (data subset refuter), we found the mean error between (a) the original estimate and (b) the mean estimate (and standard deviation) made after multiple re-runs using subsets of the data increases with larger time windows: 7 days: 0.003 (0.016); 14 days: 0.001 (0.019); 21 days: 0.002 (0.023); 30 days: 0.002 (0.022); 60 days: 0.02 (0.059); 90 days: 0.021 (0.062); 180 days: 0.044 (0.124); 365 days: 0.054 (0.154). Therefore, it seems the effect of comment dosages is not captured in such large time windows (i.e., whether a user posts in SW within 180 or 365 days is affected less by past comment experience). Furthermore, all estimates were validated by method 2 (placebo treatment refuter) as randomly assigning a confounder as a treatment resulted in effects near 0 with p-values >= 0.37. Overall, it is important to note that *smaller* time windows may have less precision given there have been less cases of one of the outcomes (reposting in SW) since it is more likely to post in SW as more time passes (see Fig. 1), while *larger* time windows may have less precision since the effect of the decision to post in SW again may be influenced less by the amount of comments received a long time ago.

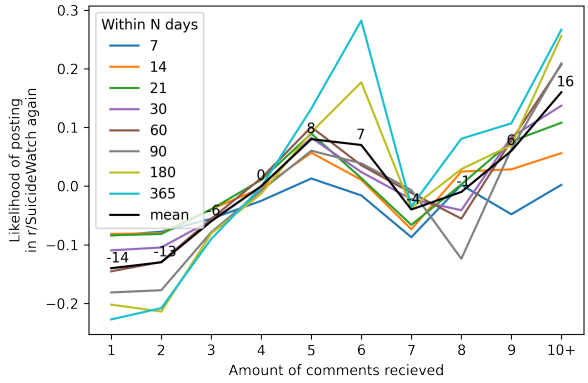

Figure 3: Likelihood of posting in r/SuicideWatch as function of the amount of comments (dosage) received within different time windows. Mean percentage values are displayed.

We checked whether the second time a user posts in SW, variables capturing suicidality reduced as a function of the amount of comments received. In Fig. 4 we show the difference in suicidality general lexicon values between each user's second and first post. The whiskers are set at the 5 and 95 percentiles; therefore 90% of the changes are near 0. This indicates that language about suicidality, as approximated by lexicons, remain relatively the same the second time they post for 90% of users. See Appendix A.2 for similar results using other variables.

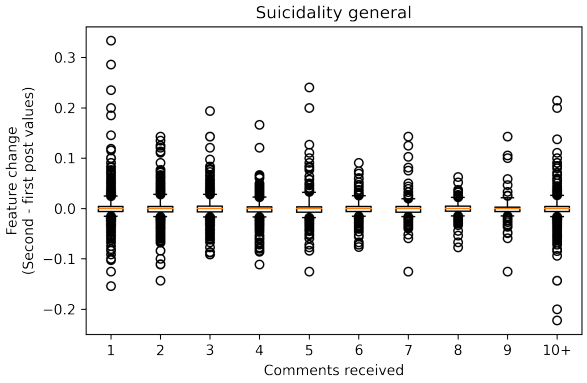

Figure 4: Boxplots showing differences in lexicon values between each user's second and first post. Variable values do not change considerably as 90% within-participant differences remain near 0.

4 Discussion

Non-professional online support groups for suicide are often considered dangerous in that untrained peers may invalidate emotions or inspire more suicidal thoughts and behaviors (Niederkrotenthaler et al., 2016). However, prior research on Reddit has shown responses are generally positive (Jiang et al., 2020). This may be due to strict moderation that removes responses containing guilt-tripping (e.g., one SW guideline asks to report "abuse or 'tough love' including any guilt-tripping like 'suicide is selfish' or 'think of your loved ones'"), pro-suicide comments, or descriptions of suicide methods. The question remains whether a few comments is enough to keep the user engaged with peers and whether responses actually reduce symptoms indefinitely.

A positive outcome after a suicidal post would be that either the comments decrease the poster's suicidal ideation or that the user engages by responding to comments and returning when symptoms persist. We estimated the effect of dosing comments on the likelihood of posting in SW again after having posted for the first time. In the case of receiving a single comment, first-time SW users are 14% less likely on average to post again on SW (range: 23% – 8%). The plausible explanation is that the user did not receive enough support and was not reinforced to use the platform again. As users received more comments they are more likely to post in SW again,

supporting hypothesis 2 (H2) at least up until 5 comments and again for 9 or more comments. Receiving 4-8 comments had mean likelihoods closer to 0, which would mean first-time SW posters are equally likely to return.

Given how difficult treating suicidal thoughts and behaviors has shown to be (Bentley et al., 2020), it is likely that positive peer responses may only provide in-the-moment reduction given effective treatment for STBs include cognitive behavioral therapy, mindful emotion awareness, generating cognitive flexibility, emotion exposure and learning skills that prevent crises for when therapy is discontinued (Bentley et al., 2020). Therefore, positive comments on Reddit that include validating emotions, providing resources towards more professional help such as crisis hotlines, and keeping the user socially supported (Jiang et al., 2020) should result not in less likelihood of returning but rather in re-engaging with the platform when or if symptoms return. Furthermore, out of the users posting on SW a second time, it does not seem to be the case that users are generally returning to express signs of improvement; we found measures of suicidality concerns (as approximated through custom lexicons) remain similar between first and second posts (see Fig. 4 and Appendix section A.2).

Keeping users engaged on a platform is extremely important. STBs can be chronic and have very different longitudinal trajectories, with one study finding the median onset for suicidal ideation occurs 1 to 5 years prior to attempting (Millner et al., 2017). Furthermore, taking into account the type of ideation as we did in this work (active vs. passive vs. self-harm) seems to be key for more accurate estimation given more active STBs where individuals describe planning future attempts (Millner et al., 2017). Furthermore, it is also worth considering how many samples are available for estimation. Compared with trying to estimate the effects of specific ngrams (i.e., the content of comments), estimating effects of the amount of comments as was done here will tend to have many more samples for a more accurate estimation (e.g., the amount of times "it gets better" is responded is likely lower than the amount of posts receiving 1 to 10+ comments).

4.1 Limitations and future work

There are several limitations resulting from the model's assumptions. For instance, there are multiple unobserved confounders that prevent us from claiming the comments causally affect subsequent posting behavior (e.g., life crises or stressors, psychiatric or psychological treatments); however, the fact that individuals return to SW would likely indicate that they found it useful somewhat independent of unobserved confounders. Therefore, it is important to try to understand variables that make users post again even if we cannot know why they may have worsened or improved. Furthermore, the stable unit treatment value (SUTVA, i.e., that user's outcomes are affected only by the treatment they receive and not by treatments other users receive) may be violated given possible network effects between posters receiving comments (e.g., reading other posts and comments may affect posting or re-posting behavior). Our current implementation does not provide confidence intervals for estimates nor the distribution for propensity scores, which will be important future work. Additional covariates could include the user's post frequency, amount of posts before first post to SW, and age of the account. Using nonproprietary features instead of LIWC could help reproducibility. Additional future work can include looking into which confounder variables seem to be most important in driving the effect. Ultimately, surveying SW users could help better determine why users decide to return or disengage with the platform.

4.2 Conclusion

We provide empirical evidence that receiving few comments decreases the likelihood of posting on SW again by 14% on average when receiving 1 comment, which is concerning given it is the most common case (24% of SW posts). Receiving more comments increases the likelihood of using SW again by 16% on average when receiving 10+ comments. Receiving many comments can be considered a positive treatment by keeping users engaged with peer-support and hence with life. It is therefore critical to better promote peer responses and develop systems to reinforce SW posters to return and seek resources.

Acknowledgements

DML was supported by a RallyPoint Fellowship and National Institute on Deafness and

Other Communication Disorders T32 training grant [5T32DC000038-28]. KLZ was supported by National Institute of Mental Health grant K23MH120439. The work was supported by a gift to the McGovern Institute for Brain Research at MIT. MN receives publication royalties from Macmillan, Pearson, and UpToDate. He has been a paid consultant in the past year for Microsoft Corporation, the Veterans Health Administration, Cerebral, and for a legal case regarding a death by suicide. He is an unpaid scientific advisor for Empatica, Koko, and TalkLife.

References

Adrian Aguilera and Clara Berridge. 2014. Qualitative feedback from a text messaging intervention for depression: benefits, drawbacks, and cultural differences. *JMIR mHealth and uHealth*, 2(4):e3660.

Laura Helena Andrade, J Alonso, Z Mneimneh, JE Wells, A Al-Hamzawi, G Borges, E Bromet, Ronny Bruffaerts, G De Girolamo, R De Graaf, et al. 2014. Barriers to mental health treatment: results from the who world mental health surveys. *Psychological medicine*, 44(6):1303–1317.

Kate H Bentley, Shannon Sauer-Zavala, Kimberly T Stevens, and Jason J Washburn. 2020. Implementing an evidence-based psychological intervention for suicidal thoughts and behaviors on an inpatient unit: Process, challenges, and initial findings. *General hospital psychiatry*, 63:76–82.

Gualtiero B Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. 2016. Analysing the connectivity and communication of suicidal users on twitter. *Computer communications*, 73:291–300.

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Rebecca G Fortgang and Matthew K Nock. 2021. Ringing the alarm on suicide prevention: a call to action. *Psychiatry*, pages 1–4.

King-wa Fu, Qijin Cheng, Paul WC Wong, and Paul SF Yip. 2013. Responses to a self-presented suicide attempt in social media. *Crisis*.

Adam C Jaroszewski, Robert R Morris, and Matthew K Nock. 2019. Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. *Journal of consulting and clinical psychology*, 87(4):370.

Meng Jiang, Brooke A Ammerman, Qingkai Zeng, Ross Jacobucci, and Alex Brodersen. 2020. Phrase-level pairwise topic modeling to uncover helpful peer responses to online suicidal crises. *Humanities and Social Sciences Communications*, 7(1):1–13.

Ang Li, Xiaoxiao Huang, Bibo Hao, Bridianne O'Dea, Helen Christensen, and Tingshao Zhu. 2015. Attitudes towards suicide attempts broadcast on social media: an exploratory study of chinese microblogs. *PeerJ*, 3:e1209.

Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, and Satrajit S Ghosh. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.

Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960.

Alexander J Millner, Michael D Lee, and Matthew K Nock. 2017. Describing and measuring the pathway to suicide attempts: A preliminary study. *Suicide and Life-Threatening Behavior*, 47(3):353–369.

David C Mohr, Joyce Ho, Jenna Duffecy, Kelly G Baron, Kenneth A Lehman, Ling Jin, and Douglas Reifler. 2010. Perceived barriers to psychological treatments and their relationship to depression. *Journal of clinical psychology*, 66(4):394–409.

Melanie CM Murray, Sara O'Shaughnessy, Kirsten Smillie, Natasha Van Borek, Rebecca Graham, Evelyn J Maan, Mia L van der Kop, Karen Friesen, Arianne Albert, Sarah Levine, et al. 2015. Health care providers' perspectives on a weekly text-messaging intervention to engage hiv-positive persons in care (weltel bc1). *AIDS and Behavior*, 19(10):1875–1887.

John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and SJ Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2):113–122.

T Niederkrotenthaler, M Gould, G Sonneck, S Stack, and B Till. 2016. Predictors of psychological improvement on non-professional suicide message boards: content analysis. *Psychological medicine*, 46(16):3429–3442.

Bridianne O'dea, Melinda R Achilles, Mark E Larsen, Philip J Batterham, Alison L Calear, and Helen Christensen. 2018. The rate of reply and nature of responses to suicide-related posts on twitter. *Internet interventions*, 13:105–107.

Christabel Owens, Paul Farrand, Ruth Darvill, Tobit Emmens, Elaine Hewis, and Peter Aitken. 2011. Involving service users in intervention design: a participatory approach to developing a text-messaging in-

tervention to reduce repetition of self-harm. *Health Expectations*, 14(3):285–295.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Anthony R Pisani, Peter A Wyman, Kunali Gurditta, Karen Schmeelk-Cone, Carolyn L Anderson, and Emily Judd. 2018. Mobile phone intervention to reduce youth suicide in rural communities: field test. *JMIR mental health*, 5(2):e10425.

Amit Sharma, Emre Kiciman, et al. 2019. DoWhy: A Python package for causal inference. https://github.com/microsoft/dowhy.

Christiane Steinert, Mareike Hofmann, Johannes Kruse, and Falk Leichsenring. 2014. Relapse rates after psychotherapy for depression–stable long-term effects? a meta-analysis. *Journal of Affective Disorders*, 168:107–118.

Ursula Whiteside, Anita Lungu, Julie Richards, Gregory E Simon, Sarah Clingan, Jaeden Siler, Lorilei Snyder, and Evette Ludman. 2014. Designing messaging to engage patients in an online suicide prevention intervention: survey results from patients with current suicidal ideation. *Journal of medical Internet research*, 16(2):e42.

A Appendix

A.1 Feature extraction

We used the following LIWC features: word count, personal pronouns, 1st person singular pronouns, negations, comparisons, interrogatives, affective processes, anxiety, anger, sadness, family, friends, insight, causation, discrepancy, tentative, certainty, perceptual processes, body, health, sexual, ingestion, achievement, power, reward, risk, past focused, present focused, future focused, work, leisure, home, money, religion, death, informal, swear words, question markers, exclamation marks. We also built custom lexicons to count tokens in posts and normalized the counts by the posts's word count (see Table 1 for examples). Lexicons were created based on most frequent ngrams in r/SuicideWatch, r/Lonely, and r/PersonalFinance, and a glossary from the US National Institute on Drug Abuse [2]. STB general included tokens that could not be strictly classified into the other STB lexicons. Mental health included words related to common non-STB disorders, symptoms, and complaints. Lexicons are freely available within the code [3].

[2] https://www.drugabuse.gov/drug-topics/commonly-used-drugs-charts

[3] Code and lexicons: https://github.com/danielmlow/playbook_comments

Lexicon	Tokens
STB general	suicid, shitty life, i deserve to die, crisis hotline, wasting space
STB active	jump into traffic, kill myself, commit suicide, hang myself
STB passive	don't want to wake up, wish i was never born, thinking about death, i'm ready to go
STB self-harm	cut myself, slit my wrists, burn my, self harm
Mental health	hospitalized, psychiatr, stress, insomnia, worried, xanax
Loneliness and isolation	lonely, no one cares, i miss my, any friends, quarantine, am single
Domestic stress and violence	divorc, violence, husband, single parent, push, scream, fight
Financial stress	my credit, loan, rent, mortgage, bills, salary, job, poverty, evict
Substance use	clean, rehab, sober, relaps, withdraw, 12 step, adderal, alcohol, hangover, cocaine

Table 1: Examples of custom lexicons. STB: suicidal thoughts and behaviors.

A.2 Differences between first and second post

See Figures A.1, A.2, and A.3 for within-user differences for different suicidality variables.

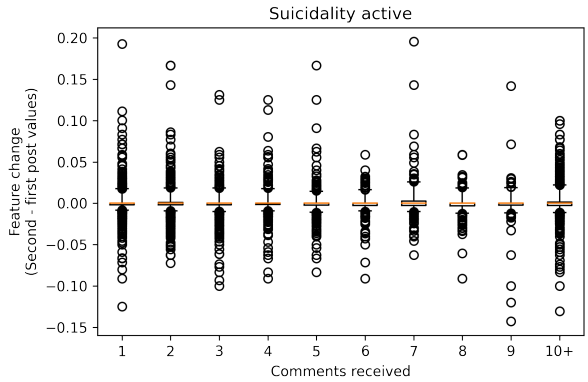

Figure A.1: Boxplots showing differences in lexicon values between each user's second and first post. Variable values do not change considerably as 90% within-participant differences remain near 0.

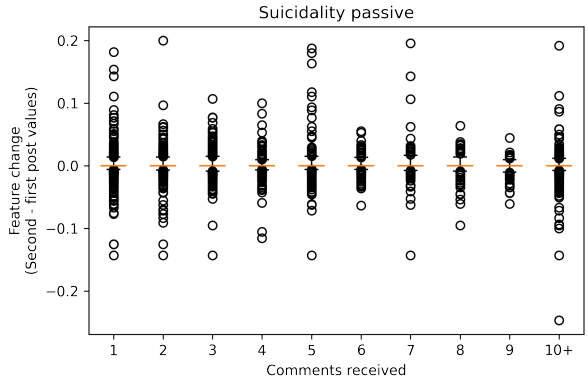

Figure A.2: Boxplots showing differences in lexicon values between each user's second and first post. Variable values do not change considerably as 90% within-participant differences remain near 0.

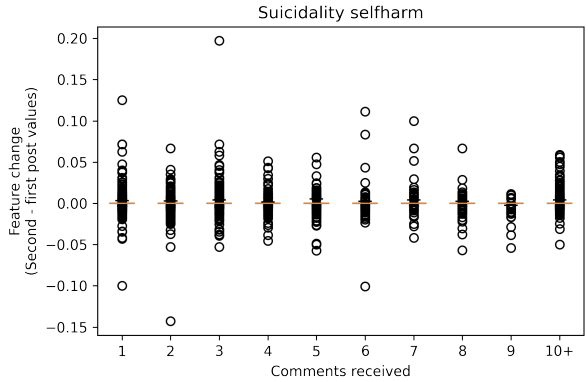

Figure A.3: Boxplots showing differences in lexicon values between each user's second and first post. Variable values do not change considerably as 90% within-participant differences remain near 0.