Helping or Hurting? Predicting Changes in Users' Risk of Self-Harm Through Online Community Interactions

Luca Soldaini*, Timothy Walsh†, Arman Cohan*, Julien Han†, Nazli Goharian*
*Information Retrieval Lab, Georgetown University
†Georgetown University
*{luca, arman, nazli}@ir.cs.georgetown.edu
†{tw614, mh1795}@georgetown.edu

Abstract

In recent years, online communities have formed around suicide and self-harm prevention. While these communities offer support in moment of crisis, they can also normalize harmful behavior, discourage professional treatment, and instigate suicidal ideation. In this work, we focus on how interaction with others in such a community affects the mental state of users who are seeking support. We first build a dataset of conversation threads between users in a distressed state and community members offering support. We then show how to construct a classifier to predict whether distressed users are helped or harmed by the interactions in the thread, and we achieve a macro-F1 score of up to 0.69.

1 Introduction

Suicide is a major challenge for public health. Worldwide, suicide is the 17th leading cause of death, claiming 800,000 lives each year (World Health Organization, 2015). In the United States alone, 43,000 Americans died from suicide in 2016 (American Foundation for Suicide Prevention, 2016), a 30-year high (Tavernise, 2016).

In recent years, online communities have formed around suicide and self-harm prevention. The Internet offers users an opportunity to discuss their mental health and receive support more readily while preserving privacy and anonymity (Robinson et al., 2016). However, researchers have also raised concerns about the effectiveness and safety of online treatments (Robinson et al., 2015). In particular, a major concern is that, through interactions with each other, at-risk users might normalize harmful behavior (Daine et al., 2013). This phenomenon, commonly referred to as the "Werther effect," has been amply studied in psychology literature and observed across various cultures and time periods (Phillips, 1974; Gladwell, 2006; Cheng et al., 2007; Niederkrotenthaler et al., 2009). In the Natural Language Processing (NLP) community, computational methods have been used to study how high profile suicides influence social media users (De Choudhury et al., 2013; Kumar et al., 2015) and study the role of empathy in counseling (Pérez-Rosas et al., 2017) and online health communities (Khanpour et al., 2017; De Choudhury and Kıcıman, 2017). However, most studies about contagious suicidal behavior in online communities are small-scale and qualitative (Haw et al., 2012; Hawton et al., 2014).

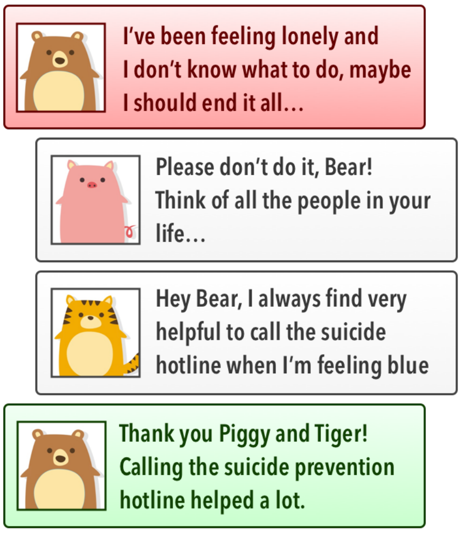
I've been feeling lonely and I don't know what to do, maybe I should end it all...
Please don't do it, Bear! Think of all the people in your life...
Hey Bear, I always find very helpful to call the suicide hotline when I'm feeling blue
Thank you Piggy and Tiger! Calling the suicide prevention hotline helped a lot.

Figure 1: A fictitious example of *flagged* thread with final *green* label (we avoid publishing any text from the dataset in order to preserve users' privacy.)

In this work, we set out to study how users affect each other through interactions in an online mental health support community. In particular, we focus on users who are in a distressed or suicidal state and open a conversation thread to seek support. Our goal is to model how such users respond to interactions with other members in the com-

Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 194–203
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

munity. First, we extract conversation threads by combining the initial post from the distressed user and the set of replies from other users who participate in the discussion. All conversation threads in our dataset are from the 2016 Computational Linguistics and Clinical Psychology (CLPsych) Workshop's ReachOut.com dataset (Milne et al., 2016), a collection of 64,079 posts from a popular support forum with more than 1.8 million annual visits. Then, we build a classifier that, given a thread, predicts whether the user at risk of self-harm or suicide successfully overcomes their state of distress through conversations with other community members. The proposed system achieves up to a 0.69 macro-F1 score. Furthermore, we extract and analyze significant features to explain what language may potentially help users in distress and what topics potentially cause harm. We observe that mentions of family, relationships, religion, and counseling from community members are associated with a reduction of risk, while target users who express distress over family or work are less likely to overcome their state of distress. Forum moderators and clinicians could deploy a system based on this research to highlight users experiencing a crisis, and findings could help train community members to respond more effectively.

In summary, our contribution is three-fold:

- We introduce a method for extracting conversation threads initiated by users in psychological distress from the 2016 CLPsych ReachOut.com dataset;

- We construct a classifier to predict whether an at-risk user can successfully overcome their state of distress through conversations with other community members;

- We analyze the most significant features from our model to study the language and topics in conversations with at-risk users.

2 Related Work

2.1 Social media and suicide

There is a close connection between natural language and mental-health, as language is an essential lens through which mental-health conditions can be examined (Coppersmith et al., 2017). At the same time, due to the ubiquity of social media in the recent decades, a huge amount of data has become available for researchers to look at mental health challenges more closely. Suicide and self-harm, which are among the most significant mental health challenges, have been recently studied through analyzing language in social media (Jashinsky et al., 2014; Thompson et al., 2014; Gunn and Lester, 2015; De Choudhury et al., 2016; Coppersmith et al., 2016; Conway and OConnor, 2016) These works exploit various NLP methods to study and quantify the mental-health language in social media. For example, Coppersmith et al. (2015) focused on quantifying suicidal ideation among Twitter users. Through their experiments, they demonstrated that neurotypical and suicidal users can be separated when controlling for age and gender based on the language used on social media. Huang et al. (2015) combined topic modeling, distributional semantics, and specialized lexicon to identify suicidal users on social media. Recently, CLPsych (Hollingshead and Ungar, 2016; Hollingshead et al., 2017) introduced shared tasks to identify the risk of suicide and self-harm in online forums. Through these shared tasks, participants explored various NLP methods for identifying users that are at risk of suicide or self-harm (Milne et al., 2016). Most of the related work in this area uses variations of linear classifiers with features that quantify the language of users in social media. For example, Kim et al. (2016) used a combination of bag-of-words and doc2vec feature representations of the target forum posts. Their system achieved the top score, a macro-average F1 score of 0.42 over four levels of risk. Another successful system utilized a stack of feature-rich random forest and linear support vector machines (Malmasi et al., 2016). Finally, Cohan et al. (2016) used various additional contextual and psycholinguistic features. In a follow-up work, Cohan et al. (2017) further improved the results on this dataset by introducing additional features and an ensemble of classifiers. In addition to these methods, automatic feature extraction methods have also been explored to quantify suicide and self-harm (Yates et al., 2017). In this work, instead of directly assessing a user's risk level based on their own posts, we study how the language of other users affects the level of risk.

2.2 Peer interaction effect on suicide

Beside messages and individuals, researchers have long been interested in how individuals prone to suicidal ideation affect each other. The most prominent examples in this area focused on examining the so-called "Werther effect," i.e. the hypothesis that suicides or attempts that receive press coverage, or are otherwise well-known, cause copycat suicides. Some of the earliest work related to our line of inquiry comes from the sociologist David Phillips who in the 1970s identified an increase in the suicide rate of American and British populations following newspaper publicity of a suicide, and argued that the relationship between publicity and the increase was causal (Phillips, 1974). Other researchers later found similar results in other parts of the world (Gladwell, 2006; Cheng et al., 2007), across various types of media, and involving different types of subjects such as fictional characters and celebrities (Stack, 2003; Niederkrotenthaler et al., 2009).

More recently, researchers have focused on studying instances of the Werther effect in online communities. Kumar et al. (2015) examined posting activity and content after ten high-profile suicides and noted considerable changes indicating increased suicidal ideation. In another work, De Choudhury et al. (2016) performed a study on Reddit users to infer which individuals might undergo transitions from mental health discourse to suicidal ideation. They observe that suicidal ideation is associated with some psychological states such as heightened self-attentional focus and poor linguistic coherence. In a subsequent work, they analyzed the influence of comments in Reddit mental health communities on suicidal ideation to establish a framework for identifying effective social support (De Choudhury and Kıcıman, 2017). In contrast with this work, we focus on studying peer influence on suicidal and self-harm ideation. Online mental health-related support forums, being inherently discussion-centric, are an appropriate platform to investigate peer influence on suicidal ideation.

While most works on the Werther effect focus on passive exposure to print, broadcast, or online media across large populations, other research has studied the contagion of suicide within smaller social clusters (Hawton et al., 2014; Haw et al., 2012). Recently, researchers have observed similar behavior in online communities (Daine et al., 2013). However, research in this area tends to be qualitative rather than quantitative, thus ignoring the possibility of leveraging linguistic signals to prevent copycat suicides.

Finally, computational linguists have also investigated the use of empathy in counseling (Pérez-Rosas et al., 2017) and online health communities (Khanpour et al., 2017). Both works focused on classification of empathetic language. In the first, linguistic and acoustic features are used to identify high and low empathy in conversations between therapists and clients. The second leverages two neural models (one convolutional, the other recurrent) to identify empathetic messages in online health communities; the two models are combined to achieve a 0.78 F1 score in detecting empathetic messages. Unlike this work, their model focuses on predicting how empathy affects the next message from an at-risk user rather than modeling the entire conversation.

3 Identifying Flagged Threads

3.1 Methodology

To study the effect of peer interaction on mental health and suicidal ideation, we leverage conversation threads from the 2016 CLPsych ReachOut.com dataset (Milne et al., 2016). ReachOut.com is a popular and large mental health support forum with more than 1.8 million annual visits (Millen, 2015) where users engage in discussions and share their thoughts and experiences. In online forums, users typically start a conversation by creating a discussion thread. Other community members, including moderators, may then reply to the initial post and other replies in the thread. The post contents can be categorized into two severity categories, *flagged* and *green* (Milne et al., 2016). The *flagged* category means that the user might be at risk of self-harm or suicide and needs attention, whereas the *green* category specifies content that does not need attention. Our goal is to investigate how the content of users changes over the course of discussion threads they initiate.

In particular, we focus on threads in which the first post was marked as *flagged*, and use subsequent replies to predict the change in status for the user who initiated the thread. More formally, let t_i be a thread of posts $\{p_{i_0}, \ldots, p_{i_m}\}$, u_T the user who initiated the thread t_i (which we will refer to as "target user" for the reminder of this paper), and $\{u_{P_1}, \ldots, u_{P_o}\}$ users who reply in thread t_i (we will refer to these users as "participating users"). Let l_{i_j} be the label for post p_{i_j}, where l_{i_j} = *flagged* if the author of the post is in a distressed mental state and requires attention, and

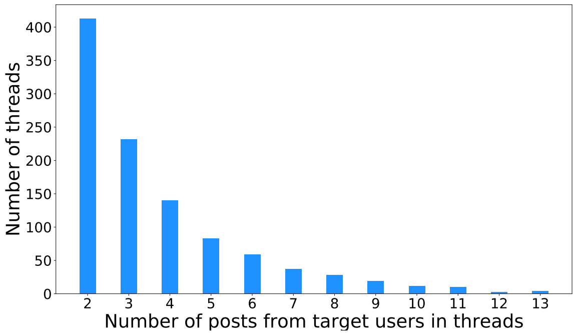

Figure 2: Distribution of threads in the dataset with respect to the number of times the target user posted in their own thread. The majority of threads contain two or three posts from the target user (the initial flagged posts and up to two replies.)

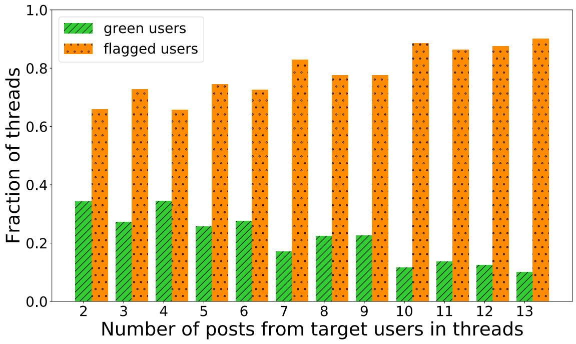

Figure 3: Distribution of labels from the final status of target users with respect to the number of times the target user posted in the thread. We observe a strong negative and statistically significant correlation between the number of target user posts and the likelihood of a final status of *green* (Pearson's correlation, $\rho = -0.91$, $p < 0.0001$.)

l_{i_j} = *green* if the author is not in a state of distress. Given a post p_{i_j} in t_i authored by target user u_T who is initially in a distressed state (indicated by $l_{i_0} =$ *flagged*), our goal is to predict the state of u_T at post p_{i_j} (i.e., we want to predict l_{i_j}).

Because of this experimental setting and the limited manual annotation of the CLPsych 2016 dataset, we cannot exclusively leverage the annotations provided with the dataset. In fact, only 42 conversation threads with an initial post marked as *flagged* can be extracted using the 1,227 manually-annotated posts. Therefore, we supplement the manual annotations with automatic labels extracted using the high performance system described by Cohan et al. (2017)[1]. This system achieves an F1 score of 0.93 identifying *flagged* vs *green* posts on the CLPsych 2016 test set. Such high performance makes this system appropriate for annotating all posts as either *flagged* or *green* without the need for additional manual annotation. To do so, we first obtain the probability of *green* being assigned by the system to each of the 64,079 posts in the dataset, and then label as *green* all posts whose probability is greater than threshold $\tau = 0.7751$. We choose this value of τ because it achieves 100% recall on *flagged* posts in the 2016 CLPsych test set (therefore minimizing the number of users in emotional distress who are classified as *green*), while still achieving high precision (0.91, less than 1% worse than the result reported by Cohan et al. (2017).)

Using the heuristic described above with the automatic post labels, we obtain 1,040 threads, each containing between 2 and 13 posts from the target user, including the initial post (mean=3.67, median=3, mode=2, stdev=2.15). The distribution of threads with respect to the number of posts by the target user is shown in fig. 2. We exclude threads containing less than two posts from target users, as we cannot assess the impact of interaction between target and participating users. Similarly, we exclude threads containing no posts from participating users. On average, each thread contains 6.62 replies (median=6, mode=5, stdev=5.67) from 4.76 participating users (median=4, mode=4, stdev=2.45).

In fig. 3, we report the distribution of labels for the final posts of target users in relation to the number of times target users post in a thread. Interestingly, we observe that the more a target user engages with participating users, the less likely their final post is labeled *green* (Pearson's correlation, $\rho = -0.91$, $p < 0.0001$.) After manually reviewing fifty of the longest threads in the dataset, we hypothesize that, in these cases, participating users struggle to connect with target users for a meaningful two-way conversation, thus failing to ameliorate any distress. This suggests that in order to effectively classify the final status of target users, language and topics from target and participating users should be modeled separately, as the mental state of a target user is not only influenced by the replies they receive, but it is also expressed through the the target user's intermediate posts.

3.2 Ethical Considerations

Ethical considerations for our effort closely mirror

[1] The annotations for all posts in ReachOut.com were graciously provided by the authors of this work.

those described by Milne et al. (2016) in constructing and annotating the original dataset. Particular care should be placed on minimizing any harm that could arise from a system deployed to notify clinicians of crises. When an individual is identified as having a moment of crisis, direct contact might aggravate their mental state. False alarms are also usually undesirable and can be distressful especially for people with a history of mental health struggles. To minimize risk, additional precautions should be taken. Examples of such measures include notifying the forum users of the automated system and explaining its purpose and functionality, and asking the users (and their clinicians) for permission to make contact during a crisis.

4 Classifying Flagged Threads

In this section, we describe the system and features we propose for classifying the final status of initially *flagged* threads. Based on the analysis of conversation dynamics in threads discussed in the previous section, we model the target and participating users in a thread separately. As such, for every thread t_i from target user u_T, we first partition t_i in subsets *Posts*(t_i, u_T) of posts authored by u_T and *Posts*$(t_i, \overline{u_T})$ of all posts in t_i authored by participating users. Then, we extract the following identical groups of features from subsets *Posts*(t_i, u_T) and *Posts*$(t_i, \overline{u_T})$:

- **LIWC**: We consider 93 indicators in the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2015). Previous research found these features to be effective in capturing language patterns for distressed mental state (Kumar et al., 2015; Milne et al., 2016; Cohan et al., 2017; Yates et al., 2017). In contrast with other efforts, we consider LIWC features for participating users.

- **Sentiment and subjectivity**: We consider sentiment and subjectivity of posts from target and participating users. We extract these features using the TextBlob tool[2].

- **Topic modeling**: To investigate the conversation topics, we perform LDA topic modeling (Blei et al., 2003) on a subset of the CLPsych ReachOut.com dataset. LDA analysis has been successfully used to study how users talk about their mental health conditions online (Resnik et al., 2015). In particular, we use posts from the "Wellbeing" and "Tough Times" sub-forums to build a model that captures mental health topics. We exclude the "Hang out" and "Introduction" subforums to prevent topic drift. We used LDAvis (Sievert and Shirley, 2014) to inspect and tune the number of topics. We tested models with 5 to 50 topics, and ultimately settled on 10 through empirical evaluation. We use the LDA implementation of Gensim[3] to compute the model.

- **Textual features**: To more precisely capture the content of posts, we consider single-word tokens extracted from posts as features. We remove stopwords, numbers, and terms appearing in less than 5 posts. When representing posts, we weight terms using *tf-idf*.

Besides the sets of features shared between target and participating users, we also consider the following signals as features: likelihood of the first reply in the thread being *green*, as assigned by the classifier by Cohan et al. (2017); number of posts in the thread from the target user; number of posts in the thread from participating users; number of posts in the thread from a moderator.

We experiment with several classification algorithms, which we compare in section 5.1, while we present the outcome of our feature analysis in section 5.2.

5 Results

5.1 Classification

We report results of the final state classification in table 1. We train all methods shown here on all features described in section 4 and we test with five-fold cross validation. We observe that all methods perform better than a majority classifier baseline. In particular, XGBoost (Chen and Guestrin, 2016) achieves the best precision, but a simple logistic regression model with ridge penalty achieves better F1 score. This suggests that the former might suffer from overfitting due to the limited size of the training data. The classification outcome of the logistic regression model is statistically different than all other models (Student t-test, $p < 0.001$, Bonferroni-adjusted.)

In fig. 4, we report results of a variant of receiving operating characteristic (ROC) analysis designed to handle probabilistic labels. Recall that

[2] https://textblob.readthedocs.io/

[3] https://radimrehurek.com/gensim/

Method	Pr	Re	F1
Majority classifier	37.69	50.00	42.98
Logistic regression	71.64	**68.16**	**69.10**
Linear SVM	70.40	61.41	62.83
XGBoost	**75.72**	63.95	66.06

Table 1: Performance of several classification methods of determining the final state of a target user in a *flagged* thread. The logistic regression is statistically different from all other models (Student t-test, $p < 0.001$, Bonferroni-adjusted.)

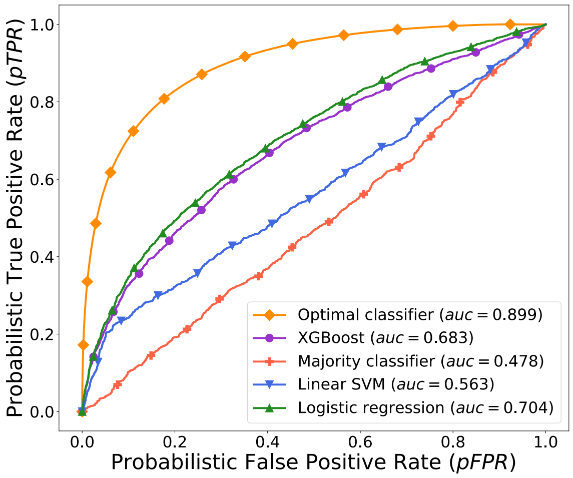

Figure 4: Probabilistic receiving operating characteristic (Burl et al., 1994) for the classification methods. Because the labels obtained from (Cohan et al., 2017) are real number in the range $[0, 1]$, results evaluated on them represent lower bounds on performance of classifiers. The optimal ROC achievable by any classifier is shown in orange (optimal $AUC = 0.899$.)

the labels obtained from Cohan et al. (2017) are real numbers in the range $[0, 1]$ representing the likelihood of each post of being *green*. While the labels can be turned into binary labels (as done to compute precision, recall, and F1 score reported in table 1), doing so ignores the uncertainty associated with probabilistic labels. Techniques to modify ROC analysis to consider probabilistic labels have been proposed in the literature. We consider the variant introduced by Burl et al. (1994) which was recently used to evaluate cohort identification in the medical domain from web search logs (Soldaini and Yom-Tov, 2017). Results in fig. 4 largely mimic performance shown in table 1, with the logistic regression model outperforming all other classifiers and achieving 78.3% of the area under the curve (AUC) of the optimal classifier (i.e., a classifier that always predicts the exact value of probabilistic labels.) While the ROC curve for the XGBoost is comparable to the logistic regression classifier, we observe that the SVM model achieves similar performance to the two only for high confidence samples (bottom left corner), and its performance declines sharply when more positive samples are inferred.

Feature set	Pr	Re	F1
Only features from target users	69.72	62.40	63.93
Averaged features from target and participating users	60.04	58.90	59.31
Separate symmetric features for target and participating users	**71.64**	**68.16**	**69.10**

Table 2: Comparison of different strategies for extracting features from target and participating users. The best feature set (separate symmetric features) is significantly better than the other two (Student t-test, $p < 0.001$, Bonferroni-adjusted.)

Finally, results shown in table 2 motivate our approach for separately modeling features for target and participating users. We observe that using separate symmetric features for the two groups of users not only improves upon using features from target user posts alone, it also outperforms averaging features extracted from the two groups together (Student t-test, $p < 0.001$, Bonferroni-adjusted.) This empirically confirms our hypothesis that language and topics from target and participating users should be modeled separately.

5.2 Features analysis

We report the result of an ablation study on feature groups in table 3. For all runs, we always include the group of “shared” features detailed in section 4. Overall, we observe that the method of including all feature sets outperforms all other runs. We note, however, that the addition of most feature sets only sums to a modest improvement (up to 7.6% in F1 score) over LIWC features alone, which confirms previous observations on the effectiveness of LIWC in modeling mental health language (Kumar et al., 2015; Milne et al., 2016; Cohan et al., 2017; Yates et al., 2017).

Beside LIWC, we found LDA and sentiment features to be moderately effective for user mental state classification. On the other hand, we found

Feature group	**Pr**	**Re**	**F1**
LIWC	67.85*	63.12*	64.26*
LIWC + Sentiment	68.34*	63.80*	64.93*
LIWC + Sentiment + LDA topics	70.95	67.73	68.83
LIWC + Sentiment + LDA topics + Tokens	**71.64**	**68.16**	**69.10**

Table 3: Ablation study of feature groups. Results marked by * indicate runs that are significantly different from the best method (Student *t*-test, $p < 0.001$, Bonferroni-adjusted.)

token features to have a limited impact on the performance of the system, improving the overall F1 score by just 0.35%. Their contribution was also found to be insignificant (Student *t*-test, $p = 0.19$.) We attribute this result to the fact that, compared to other features, non-sentiment terms used by target or participating users represent a much weaker signal for modeling the change in self-harm risk that interests us.

We report the most significant features for each feature group in table 4. For each group of features, we report the top three positive and negative features for target and participating users. In order to improve readability, feature weights are ℓ_2-normalized with respect to their group (token features are one to two orders of magnitude smaller than the other groups.) For LDA features, we report a list of significant terms for topics, as well as possible interpretations of them, in table 5.

When analyzing LIWC features for the target users, we note that mention of support communities (e.g., religion), internet slang (netspeak), and talk about leisure activities correlate with a decrease in risk by the end of a thread. Conversely, filler language (which can sometimes indicate emotional distress (Stasak et al., 2016)), mention of family, and swearing are all associated with target users remaining in a *flagged* status at the end of a thread. While participants share some positive LIWC features with target users (e.g., religion), we notice that mention of home and family are, in this case, associated with a positive outcome. To explain this difference, we sampled 20 threads in which target or participating users mentioned “family” or “home.” We empirically observe that, in a majority of cases, when target users mention family it is because they have trouble communicating with or relating to them. On the other hand, when participating users mention family and home it is usually to remind target users of their relationships with loved ones. While not conclusive, this observation suggests a possible explanation of the difference.

Compared to LIWC categories, we found the scores assigned to LDA topics more difficult to explain. As shown in table 5, while some topics have clearly defined subjects, others are harder to interpret. However, we note that most topics reported in table 4 as having a high positive or negative weight for our best classifier have a clear interpretation. For example, topics #8, #7, and #3 are about school and counseling, thanking the ReachOut.com community, and family. Among negative topics, discussion of weight loss (topic #10), or work and university (topic #9) are associated with *flagged* states. Interestingly, topic #7 has a positive weight for target users (this is expected: users who thank other users for their help are more likely to have transitioned to *green* by the end of a thread), but it has a negative weight for participants. We could not find a plausible intuition for the latter observation. Similar to LIWC features, we found that family is correlated with a decrease in risk by the end of a thread when mentioned by participating users (topic # 3), while the opposite is true for target users.

We analyze the importance values associated with tokens. We note that observations of these features are less likely to be conclusive, given their limited impact on classification performance (table 3.) Nevertheless, we observe that features in this category either represent emotional states that are also captured by sentiment features (e.g., *hope, proud, scared*) or relate to topics discussed in the previous section (e.g., *school*, *thanks*.)

Finally, while not reported in table 4, we also study the importance of sentiment and subjectivity features. We observe that sentiment positively correlates with both target and participating users ($+0.902$ and $+0.629$, respectively). High levels of objectivity by target users correlate with a decrease in risk of harm by the end of a thread ($+0.510$), while the model finds that objectivity by participating users is not predictive of a *green* final state (assigned weight: $+0.033$.)

LIWC				LDA				Tokens			
target users		participants		target user		participants		target users		participants	
+0.423	religion	+0.528	home	+0.471	topic #4	+0.397	topic #8	+0.847	*thanks*	+0.541	*proud*
+0.339	netspeak	+0.478	family	+0.245	topic #8	+0.303	topic #6	+0.614	*hope*	+0.521	*value*
+0.295	leisure	+0.436	religion	+0.210	topic #7	+0.272	topic #3	+0.570	*didn't*	+0.509	*dreams*
...		...		...		...		...		...	
-0.126	filler	-0.447	swear	-0.624	topic #9	-0.756	topic #10	-0.994	*anymore*	-0.582	*ready*
-0.392	family	-0.333	sexual	-0.217	topic #3	-0.274	topic #7	-0.651	*scared*	-0.572	*trying*
-0.354	swear	-0.248	money	-0.092	topic #1	-0.088	topic #1	-0.598	*I'm*	-0.477	*school*

Table 4: Top positive and negative features for each feature group. Scores are ℓ_2-normalized with respect to their group. For LDA, we report a list of significant terms for each topic, as well as possible interpretations, in table 5.

Topic	Significant terms by relevance (Sievert and Shirley, 2014), $\lambda = 0.6$	Potential interpretation
1	*like, think, life, need, know, will, i've, feeling*	?
2	*help, people, reachout, talk, need, support*	discussions about ReachOut
3	*important, home, people, family, need, back*	family
4	*think, time, want, now, today, people, help*	?
5	*know, feel, great, negative, day, work, person*	?
6	*didn't, who, like, know, feeling, will, life*	?
7	*thanks, want, life, feel, hey, great, guys*	thanking other ReachOut users for their support
8	*talk, school, time, self, counsellor, seeing*	therapy, school
9	*work, time, paycheck, study, uni, goal*	work, study
10	*see, body, hard, care, health, weight*	weight loss

Table 5: Significant terms for the LDA topics computed over the ReachOut forum. Significance is determined using *relevance* (Sievert and Shirley, 2014).

5.3 Conversation Threads Analysis

Beside analyzing individual features, we also present an overview of how aspects of threads correlate with performance of the classifier. We observe that there exists a positive correlation between the length of a thread and the performance of the classifier. Such correlation is significant for both *green* ($\rho = 0.33$, $p < 0.05$) and *flagged* ($\rho = 0.37$, $p < 0.05$) conversation threads, and likely explained by observing that longer threads may contain more information about the mental state of target users. We note that the average standard deviation of target user state within a thread is 0.32 (median=0.19), which suggests that some target users oscillate between *green* and *flagged* states in a conversation. However, we observe no correlation between variance and model performance (p-value= 0.44.)

While encouraging, we recognize the limitations of our current approach. Online data is a great resource to study the language of mental health, but it often lacks granularity. This is not an issue for long-trend studies, but it poses issues when trying to model language around suicidal ideation, given the short duration of suicidal crises (Hawton, 2007). While analyzing conversations that were incorrectly classified by our system, we also noted that several target users transitioned between states without any meaningful interaction at all with participating users. Our intuition is that the mental state of a target user is significantly influenced by passively reading other threads, interacting over a secondary channel like private messages, and experiencing offline events, none of which are available as inputs to our system, thereby limiting its accuracy.

6 Conclusions

In recent years, the number of online communities designed to offer support for mental health crises has grown significantly. While users generally find these communities helpful, researchers have shown that, in some cases, they could also normalize harmful behavior, discourage professional treatment, and instigate suicidal ideation. We study the problem of assessing the impact of interaction with a support community on users who are suicidal or at risk of self-harm.

First, using the 2016 CLPsych ReachOut.com corpus, we build a dataset of conversation threads between users in distress and other members of the community. Then, we construct a classifier to predict whether an at-risk user can successfully overcome their state of distress through conversations with other community members. The classifier leverages LIWC, sentiment, topic, and textual features. On the dataset introduced in this paper, it achieves a 0.69 macro-F1 score. Furthermore, we analyze the effectiveness of features from our model to gain insights from the language used and the topics appearing in conversations with at-risk users.

References

American Foundation for Suicide Prevention. 2016. Suicide statistics 2016. *AFSP; New York, NY.*

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Michael C Burl, Usama M Fayyad, Pietro Perona, and Padhraic Smyth. 1994. Automated analysis of radar imagery of venus: handling lack of ground truth. In *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, volume 3, pages 236–240. IEEE.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Andrew TA Cheng, Keith Hawton, Tony HH Chen, Amy MF Yen, Jung-Chen Chang, Mian-Yoon Chong, Chia-Yih Liu, Yu Lee, Po-Ren Teng, and Lin-Chen Chen. 2007. The influence of media reporting of a celebrity suicide on suicidal behavior in patients with a history of depressive disorder. *Journal of affective disorders*, 103(1):69–75.

Arman Cohan, Sydney Young, and Nazli Goharian. 2016. Triaging mental health forum posts. *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 143–147.

Arman Cohan, Sydney Young, Andrew Yates, and Nazli Goharian. 2017. Triaging content severity in online mental health forums. *Journal of the Association for Information Science and Technology*.

Mike Conway and Daniel OConnor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology*, 9:77–82.

Glen Coppersmith, Casey Hilland, Ophir Frieder, and Ryan Leary. 2017. Scalable mental health analysis in the clinical whitespace via natural language processing. In *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on*, pages 393–396. IEEE.

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

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt.

Kate Daine, Keith Hawton, Vinod Singaravelu, Anne Stewart, Sue Simkin, and Paul Montgomery. 2013. The power of the web: a systematic review of studies of the influence of the internet on self-harm and suicide in young people. *PloS one*, 8(10):e77555.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. AAAI.

Munmun De Choudhury and Emre Kıcıman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Proceedings of the AAAI International Conference on Web and Social Media (ICWSM)*, volume 2017, page 32.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 34rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '16. ACM.

Malcolm Gladwell. 2006. *The tipping point: How little things can make a big difference*. Little, Brown.

John F Gunn and David Lester. 2015. Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death. *Suicidologi*, 17(3).

Camilla Haw, Keith Hawton, Claire Niedzwiedz, and Steve Platt. 2012. Suicide clusters: a review of risk factors and mechanisms. *Suicide and life-threatening behavior*.

Keith Hawton. 2007. Restricting access to methods of suicide: Rationale and evaluation of this approach to suicide prevention. *Crisis*, 28(S1):4–9.

Keith Hawton, Louise Linsell, Tunde Adeniji, Amir Sariaslan, and Seena Fazel. 2014. Self-harm in prisons in england and wales: an epidemiological study of prevalence, risk factors, clustering, and subsequent suicide. *The Lancet*, 383(9923):1147–1154.

Kristy Hollingshead, Molly E. Ireland, and Kate Loveys, editors. 2017. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Vancouver, BC.

Kristy Hollingshead and Lyle Ungar, editors. 2016. *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, San Diego, CA, USA.

Xiaolei Huang, Xin Li, Tianli Liu, David Chiu, Tingshao Zhu, and Lei Zhang. 2015. Topic model for identifying suicidal ideation in chinese microblog. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 553–562.

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

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 246–251.

Sunghwan Mac Kim, Yufei Wang, Stephen Wan, and Cecile Paris. 2016. Data61-csiro systems at the clpsych 2016 shared task. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 128–132, San Diego, CA, USA. Association for Computational Linguistics.

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 & Social Media*, pages 85–94. ACM.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting post severity in mental health forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 133–137, San Diego, CA, USA. Association for Computational Linguistics.

Doug Millen. 2015. Reachout annual report 2013/2014.

David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 118–127.

Thomas Niederkrotenthaler, Benedikt Till, Nestor D Kapusta, Martin Voracek, Kanita Dervic, and Gernot Sonneck. 2009. Copycat effects after media reports on suicide: A population-based ecologic study. *Social science & medicine*, 69(7):1085–1090.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1426–1435.

David P Phillips. 1974. The influence of suggestion on suicide: Substantive and theoretical implications of the werther effect. *American Sociological Review*, pages 340–354.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107.

Jo Robinson, Georgina Cox, Eleanor Bailey, Sarah Hetrick, Maria Rodrigues, Steve Fisher, and Helen Herman. 2016. Social media and suicide prevention: a systematic review. *Early intervention in psychiatry*, 10(2):103–121.

Jo Robinson, Sarah Hetrick, Georgina Cox, Sarah Bendall, Alison Yung, and Jane Pirkis. 2015. The safety and acceptability of delivering an online intervention to secondary students at risk of suicide: findings from a pilot study. *Early intervention in psychiatry*, 9(6):498–506.

Carson Sievert and Kenneth Shirley. 2014. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.

Luca Soldaini and Elad Yom-Tov. 2017. Inferring individual attributes from search engine queries and auxiliary information. In *Proceedings of the 26th International Conference on World Wide Web*, pages 293–301. International World Wide Web Conferences Steering Committee.

Steven Stack. 2003. Media coverage as a risk factor in suicide. *Journal of Epidemiology & Community Health*, 57(4):238–240.

Brian Stasak, Julien Epps, and Nicholas Cummins. 2016. Depression prediction via acoustic analysis of formulaic word fillers. *Polar*, 77(74):230.

Sabrina Tavernise. 2016. Us suicide rate surges to a 30-year high.

Paul Thompson, Craig Bryan, and Chris Poulin. 2014. Predicting military and veteran suicide risk: Cultural aspects. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–6, Baltimore, Maryland, USA. Association for Computational Linguistics.

World Health Organization. 2015. *World health statistics 2015*. World Health Organization.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.