# Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings

**Han-Chin Shing**[1], **Suraj Nair**[1], **Ayah Zirikly**[2,4], **Meir Friedenberg**[3], **Hal Daumé III**[1], and **Philip Resnik**[1]
[1]UMIACS CLIP Laboratory, University of Maryland, College Park, MD
[2]National Institutes of Health, Bethesda, MD
[3]Computer Science Department, Cornell University, Ithaca, NY
[4]Stanford Center for Population Health Sciences, Stanford University, Stanford, CA
{shing,srnair,hal,resnik}@umd.edu
ayah.zirikly@nih.gov, mdf224@cornell.edu

## Abstract

We report on the creation of a dataset for studying assessment of suicide risk via online postings in Reddit. Evaluation of risk-level annotations by experts yields what is, to our knowledge, the first demonstration of reliability in risk assessment by clinicians based on social media postings. We also introduce and demonstrate the value of a new, detailed rubric for assessing suicide risk, compare crowdsourced with expert performance, and present baseline predictive modeling experiments using the new dataset, which will be made available to researchers through the American Association of Suicidology.

## 1 Introduction

The majority of assessment for suicide risk takes place via in-person interactions with clinicians, using ratings scales and structured clinical interviews (Batterham et al., 2015; Joiner et al., 1999, 2005). However, such interactions can take place only after patient-clinician contact has been made, and only when access to a clinician is available. This is no small challenge in many places — in the U.S., for example, nearly 124 million people live in federally designated mental health provider shortage areas, where access to a provider can be difficult even when the person (or someone close to them) knows that clinical help is needed (Bureau of Health Workforce, 2017).

At the same time, people are spending an increasing amount of their time online, and online discussions related to mental health are providing new opportunities for people dealing with mental health issues to find support and a sense of connection; these include Koko, itskoko.com; ReachOut, reachout.com ; 7cups, 7cups.com; Reddit, reddit.com and others. Although many such discussions are peer-to-peer, site moderators often play a crucial role, identifying users who post material indicating imminent risk and the need for intervention.

An emerging subset of the artificial intelligence and language technology communities has been making progress on automated methods that analyze online postings to flag mental health conditions, with the goal of being able to screen or monitor for suicide risk and other conditions (Calvo et al., 2017; Resnik et al., 2014; Milne et al., 2016; Milne, 2017). Some sites have been taking advantage of these methods to add automation to their moderation, in the form of a pipeline from algorithmic risk assessment to human moderator review to preventive action.

With all of these technology-driven developments taking place so quickly, it is easy to forget that *clinician* assessment of suicidality from online writing is a new and largely unstudied problem. To what extent is level of suicide risk discernable from online postings? How are traditional training and experience in assessment brought to bear in the absence of interaction with the person being assessed?

In this paper we investigate risk assessment for online postings using data from Reddit (reddit.com) an online site for anonymous discussion on a wide variety of topics. We focus specifically on users who have posted to a discussion forum called *SuicideWatch*, which, as its name suggests, is dense in postings by people who are considering taking their own lives.[1] We have developed a dataset of users who posted on SuicideWatch, that, by virtue of posting to the forum, were by definition considered potentially at risk. A set of posts was assessed independently by four clinicians who specialize in suicidality assessment. In addition, crowdsource workers assessed a larger set

[1]Titled forums on Reddit are called *subreddits*, but for clarity and generality we sometimes adopt the more common term *discussion forum*.



*Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36
New Orleans, Louisiana, June 5, 2018. 

based on the same detailed instructions. We evaluated levels of inter-rater agreement within and across groups and also looked at differences between groups. In addition, we present initial automatic risk-level classification and screening results for SuicideWatch data using machine learning.

## 2 Dataset

Our approach to data collection is inspired by Coppersmith et al. (2014), who introduced an innovative way to solve for the absence of clinical ground truth when studying mental health in social media. Their approach is to identify users who have produced an overt signal, in social media, indicating they *might* be a positive instance of the relevant condition, and then manually assessing the signal to filter out candidates for which the signal does not appear genuine. They applied this on Twitter by seeking variations of the statement *I have been diagnosed with X*, (where $X$ is *depression*, *PTSD*, or other conditions), and then manually filtering tweets for which the statement was in jest or otherwise not a true indication, e.g. *The Red Sox lost their third game in a row. I've just been diagnosed with depression*. They also collected controls who had not made such statements.

The Coppersmith et al. approach does not yield clinical ground truth, since there is no way to verify an actual diagnosis, nor any way to determine that a control instance might not actually be positive for the condition. However, obtaining clinical data presents extremely challenging procedural burdens, and shared datasets for healthcare are typically orders of magnitude smaller than datasets supporting research in other domains.[2]

We began with a snapshot of every publicly available Reddit posting from January 1, 2008 through August 31, 2015, with partial data from 2006-2007, comprising approximately 42G of compressed data.[3] The "signal" for a user's candidate positive status with respect to suicidality is their having posted in the `/r/SuicideWatch` subreddit, a forum providing "peer support for anyone struggling with suicidal thoughts, or worried about someone who may be at risk".[4] Eliminating users who had fewer than ten total posts across all of Reddit, we had 11,129 users who had posted in SuicideWatch for a total of 1,556,194 posts. Through random sampling we selected 1097 users, of which 934 ultimately were included (see Section 3.2). For these users we extracted not only their SuicideWatch posts, but *all* their Reddit posts available in the snapshot. We also aggregated the data from an equal number of control users who had not posted in any of the mental health subreddits identified by Pavalanathan and De Choudhury (2015), nor in the `/r/schizophrenia` subreddit.[5]

User accounts on Reddit are fundamentally anonymous: when creating a Reddit account, only a user-selected username and password need to be supplied, with e-mail address optional (Reddit, 2018). Since users might have chosen to include potentially identifying information in their usernames, we go a step further and replace usernames with unique numeric identifiers.[6] We discuss privacy and other issues further in Section 6.

## 3 Annotation

For purposes of annotation, we began with the temporally ordered sequences of posts on SuicideWatch for each of the 934 users. In order to facilitate crowdsourced as well as expert annotation, we divided sequences of more than five SuicideWatch posts for a single user into multiple annotation units containing up to five posts each, yielding a total of 982 annotation units. (For example, a user with 12 posts would yield three annotation units of their first 5 posts, next 5 posts, final 2 posts.)

[2] Access to healthcare data in the U.S. is governed by the Healthcare Insurance Portability and Accountability Act, or HIPAA. Resnik (2017) has argued that, owing to the fact that the law was written without anticipating the importance of large scale, community-wide research datasets, the state of the art in clinical natural language processing is significantly behind the state of the art in other domains. For example, the widely used Enron email corpus contains 1.2 million emails (Klimt and Yang, 2004); in contrast, the SemEval-2017 Clinical TempEval shared task used 400 manually de-identified clinical notes and pathology reports from cancer patients at the Mayo Clinic (Bethard et al., 2017).

[3] `https://www.reddit.com/r/datasets/comments/3mg812/full_reddit_submission_corpus_now_available_2006/`

[4] `https://www.reddit.com/r/SuicideWatch/`, which henceforth we refer to simply as SuicideWatch

[5] Our full set: addiction, alcoholism, Anger, bipolarreddit, BPD (Bederline Personality Disorder), depression, DPDR (depersonalization, derealization), EatingDisorders, feelgood, getting_over_it , hardshipmates, mentalhealth, MMFB (MakeMeFeelBetter), panicparty, psychoticreddit, ptsd, rapecounseling, schizophrenia, socialanxiety, StopSelfHarm, SuicideWatch, survivorsofabuse, traumatoolbox.

[6] For example, a hypothetical user could choose the username `maryjanesmith1973.collegepark`, identifying name, birth year, and location.

In order to determine user-level risk, we consider a user to have the highest risk associated with any of their annotation units.

We defined a four-way categorization of risk adapting Corbitt-Hall et al. (2016) (who provided lay definitions based on risk categories in Joiner et al. (1999)): **(a) No Risk (or "None")**: I don't see evidence that this person is at risk for suicide; **(b) Low Risk**: There may be some factors here that could suggest risk, but I don't really think this person is at much of a risk of suicide; **(c) Moderate Risk**: I see indications that there could be a genuine risk of this person making a suicide attempt; **(d) Severe Risk**: I believe this person is at high risk of attempting suicide in the near future.[7]

We then defined two sets of annotator instructions. The *short* instructions, intended only for experts, simply presented the above categorization and asked them to follow their training in assessing patients with suicide risk. A *long* set of instructions was similar in intent to Corbitt-Hall et al. (2016), but whereas their instructions focused on three risk factors (*thoughts of suicide*, *planning*, and *preparation*), we identified four families of risk factors: *thoughts* includes not only explicit ideation but also, e.g., feeling they are a burden to others or having a "fuck it" (screw it, game over, farewell) thought pattern; *feelings* includes, e.g., a lack of hope for things to get better, or a sense of agitation or impulsivity (mixed depressive state, Popovic et al. (2015)); *logistics* includes, e.g., talking about methods of attempting suicide (even if not planning), or having access to lethal means like firearms; and *context* includes, e.g. previous attempts, a significant life change, or isolation from friends and family.[8]

In both sets of instructions, annotators were also asked to label the post (if there are more than one) that most strongly supports the judgment, and they were told that choices should never be downgraded: if an earlier post suggests a person is at severe risk ("I'm going to kill myself"), and a later post suggests the risk has decreased ("I've decided not to kill myself"), the higher risk should be chosen along with the severe-risk post as the basis for the judgment.

[7]These correspond roughly to the *green*, *amber*, *red*, and *crisis* categories defined by Milne et al. in CLPsych ReachOut shared tasks (Milne et al., 2016; Milne, 2017).

[8]We will of course be happy to share our instructions with other researchers.

### 3.1 Expert Annotation

We selected 245 users at random to create a set of 250 annotation units that were labeled independently by four volunteer experts in assessment of suicide risk.[9] These included a suicide prevention coordinator for the Veteran's Administration; a co-chair of the National Suicide Prevention Lifelines Standards, Training and Practices Sub-Committee; a doctoral student with expert training in suicide assessment and treatment whose research is focused on suicidality among minority youth; and a clinician in the Department of Emergency Psychiatry at Boston Childrens Hospital. Two of these experts received the detailed long instructions, and the other two were given the short instructions.

Table 1 shows Krippendorff's $\alpha$ pairwise among the experts, indicating the set of instructions they used as (S)hort or (L)ong. The average of 0.812 satisfies the conventional reliability cutoff for chance-corrected agreement ($> 0.8$, Krippendorff (2004)), which is to our knowledge the first result demonstrating inter-rater reliability by clinical experts for suicide risk based on social media postings. Inter-rater reliability for the pair receiving short instructions was substantially lower (0.768), demonstrating the value of our detailed rubric based on explicitly identified risk factors.

We generated consensus user-level labels based on the expert annotations using a well known model for inferring true labels from multiple noisy annotations (Dawid and Skene, 1979; Passonneau and Carpenter, 2014), including consensus for the pairs receiving long instructions (*Long Experts*), short instructions (*Short Experts*), and consensus among all four experts. Table 2 summarizes the data, partitioning categories according to the all-experts consensus.

| Krippendorff $\alpha$ | exp_L1 | exp_L2 | exp_S1 | exp_S2 |
|---|---|---|---|---|
| exp_L1 | 1 | 0.837 | 0.804 | 0.823 |
| exp_L2 | - | 1 | 0.808 | 0.831 |
| exp_S1 | - | - | 1 | 0.768 |
| exp_S2 | - | - | - | 1 |

Table 1: Krippendorff's $\alpha$ pairwise among experts

[9]Random selection was from the set of crowdsource-annotated users obtained in Section 3.2, ensuring that all expert annotations would be accompanied by crowdsourced annotations. Recall that a user's label is the highest-risk label assigned for any of that user's annotation units, if there are more than one.

| | # users | avg # words | avg # posts |
|---|---|---|---|
| None | 36 | 175 | 1.08 |
| Low | 50 | 247 | 1.46 |
| Moderate | 115 | 281 | 1.37 |
| Severe | 44 | 259 | 2.05 |

Table 2: Expert annotation dataset statistics.

### 3.2 Crowdsourced Annotation

We created a task on CrowdFlower (crowdflower.com) using the long instructions. We restricted participation to high performance annotators (as determined by the CrowdFlower platform) and who also agreed with our annotations on seven clear test examples. Although we began with 1,097 users to annotate, crowdsourcer participation tailed off at 934.[10] After discarding any annotation unit labeled by fewer than three annotators, our data comprises 865 users and 905 annotation units. We used CrowdFlower's built-in consensus label as the crowdsourced label for each unit.[11] Krippendorff's $\alpha$ for inter-annotator agreement of the crowdsourcers for user labels is $0.554$.

### 3.3 Annotation Disagreements

To investigate the quality of annotation across and within groups of crowdsourcers and experts, we begin by treating it as a human prediction task. Table 3 shows the macro F1 score using all-experts consensus labels as ground truth, with different human consensus values as the prediction. These pattern as one would expect, decreasing from experts with long instructions, to experts with short instructions relying on (varied) training, and we hypothesize that the much lower performance of crowdsourcers arises both because they have less training than experts, and because they are less mission-driven in their motivations and therefore are likely to feel a lower committment to the task.

Nonetheless, it is worth noting that there is clear value in the crowdsourced annotations. Table 4 shows a confusion matrix measuring crowdsourcers' consensus against the all-experts consensus, and it appears that most of the errors involve erring on the side of caution, misclassifying more than half of the low-risk users as having higher risk, and misclassifying a large number of moderate risk users (no imminent threat of a suicide attempt) as having severe (imminent) risk. In settings where the goal is to flag users for more careful review and possible intervention, false positives seem likely to be the preferred kind of error.[12]

Table 5 shows the confusion matrix for experts receiving short versus long instructions, which may be illuminating for scenarios in which trained clinicians perform assessment using social media posts but do not take the time to apply the long-instructions rubric or do not do so consistently. We observe the same trend toward erring in the direction of false positives, and it is notable that *no* severe-risk users (based on the long-instruction consensus) are assigned to no risk or even low risk by the short-instructions consensus.

| | Long Experts | Short Experts | CrowdFlower |
|---|---|---|---|
| All Experts | 0.8367 | 0.7173 | 0.5047 |

Table 3: Macro F1 scores for consensus human predictions on the 245 users labeled by both experts and crowdsourcers, using all-experts consensus as ground truth

| | | Crowdflower | | | |
|---|---|---|---|---|---|
| | | None | Low | Moderate | Severe |
| All Experts | None | 29 | 1 | 1 | 5 |
| | Low | 11 | 13 | 20 | 6 |
| | Moderate | 6 | 11 | 47 | 51 |
| | Severe | 1 | 1 | 8 | 34 |

Table 4: All Experts vs. Crowdsourcers

| | | Short Experts | | | |
|---|---|---|---|---|---|
| | | None | Low | Moderate | Severe |
| Long Experts | None | 36 | 1 | 1 | 0 |
| | Low | 5 | 16 | 34 | 3 |
| | Moderate | 1 | 0 | 56 | 14 |
| | Severe | 0 | 0 | 17 | 61 |

Table 5: Long Experts vs. Short Experts

[10] We conjecture that, with fewer jobs left available, annotators were less inclined to go through the detailed instructions and test because there was less for them to get paid for.

[11] See *Confidence Score* https://success.crowdflower.com/hc/en-us/articles/202703305-Getting-Started-Glossary-of-Terms

[12] Performance differences between experts and non-experts require more study. For example, Homan et al. (2014) found that two novice annotators were *more* likely to assign their expert's "low distress" tweets to the "no distress" category. Conversely, on a related but coarser-grained categorization task, Liu et al. (2017) find "some evidence that multiple crowdsourcing workers, when they reach high inter-annotator agreement, can provide reliable quality of annotations".

## 4 Baseline Experimentation

In addition to making progress on human assessment of suicide risk in social media, our goal in this work is also to create new resources for automated methods. Since this is a new dataset, we provide some initial predictive performance figures using machine learning methods, with the intent that these will be improved upon by the community once we make the dataset available.

We distinguish the tasks of *risk assessment* and *screening*. Risk assessment is the assignment of a risk category for someone for whom risk is already believed to exist (e.g. a patient with signs of depression at intake, a suicidal patient being monitored, an individual posting to SuicideWatch), i.e. the machine equivalent of the human assessments in Section 3. For risk assessment, the data to be categorized comprises all of a user's postings on SuicideWatch, just as in the human assessment.

Screening is the identification of potential risk in individuals for whom no potential risk had yet been established (e.g. a new mother, a patient visiting their primary care physician, a person posting in everyday social media forums not related to mental health). We treat screening as a binary classification task, distinguishing positive (at-risk) versus control as in Coppersmith et al. (2014) and others, and this therefore requires data from control users. We define our potential population of positive users as the 865 for whom we obtained crowdsourced ratings. Since this is a screening task, the data to be classified is their postings on *other* Reddit forums (also excluding mental health forums) prior to that first SuicideWatch posting.[13] Control users are selected at random excluding users who posted on SuicideWatch or any other mental health forum.

To explore the extent to which evidence of suicidality may be attenuated at greater temporal distance from the first SuicideWatch posting, we evaluate sets of posts starting 7 days, 5 days, 2 days, and 1 day before that posting. The equivalent time periods are defined for control users by randomly choosing a post as the endpoint and selecting sets of posts starting 7, 5, 2, and 1 day before that one.

### 4.1 Preprocessing

We replace every instance of a URL with the token *url*, and we normalize numbers by substituting with @, preserving the shape of number. (E.g. 123 → @, whereas 12.3 → @.@.) We also convert emojis and emoticons to their corresponding text. Posts are then tokenized and lemmatized using Spacy.[14]

### 4.2 Feature Engineering

We employ the following features.

**Bag of words**. We represent the post title as a bag of words vector, including unigrams and bigrams from the title with tf-idf weighting.[15]

**Empath** (Fast et al., 2016). We use the normalized frequency of Empath lexical categories, exploring both the use of all 200 Empath-generated lexical categories and depression-based lexical categories (e.g. love, sympathy, irritability, nervousness, etc.).

**Readability.** We included Automated Readability Index (ARI) (Senter and Smith, 1967), Gunning fog index (Gunning, 1952), SMOG index (Mc Laughlin, 1969), Coleman-Liau index (Coleman and Liau, 1975), Flesch Reading Ease (Farr et al., 1951), Flesch-Kincaid Grade Level (Kincaid et al., 1975), LIX and RIX (Anderson, 1983).

**Syntactic features.** We include the proportion of transitive verbs (out of all verbs), the proportion of active verbs, proportion of passive verbs, proportion of active verbs with "I" as subject, proportion of passive verbs with "I" as subject, and proportion of transitive verbs with "me" or "myself" as object.

**Topic model posteriors.** We used Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to infer a 20-topic model on the training set using each post body as a document, in order to use the set of topic posteriors as features, which has proven useful in previous work (Resnik et al., 2015).[16]

**Word embeddings.** We compute 300-dimensional embeddings for the entire Reddit corpus using a SkipGram model with negative sampling of size 15, sampling rate 1e-5, window

[13] This definition of a positive user for screening is of course noisy; effectively in this first pass we are adopting Coppersmith et al.'s strategy but using the signal evidence without further filtering. We plan to use the risk labels for filtering in future work, e.g. defining a positive instance only as someone whose risk level is moderate or severe, which is why we limit our universe population here to those for whom we have risk ratings. See also Liu et al. (2017) on aggregation of annotator labels for supervised learning in this domain.

[14] https://spacy.io/

[15] All other features are extracted from the body of the post.

[16] We used Gensim, https://radimrehurek.com/gensim/models/ldamulticore.html.

size 5, and discarding any words that occur fewer than 5 times. We calculate the embedding of a post body by averaging the embeddings of all its words.

**Linguistic Inquiry and Word Count (*LIWC*).** The category frequency for each LIWC category (Tausczik and Pennebaker, 2010) using the post body's lemmas.

**Emotion features (*NRC*).** The count of *emotion* tokenized lemmas occurring in the post body based on the NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013). The emotions included are anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.

**Mental disease lexicon (*mentalDisLex*).** The maximum count of the post body's tokens or lemmas that match entries in the mental disease lexicon introduced by Zirikly et al. (2016).

The feature vector for each user is the average of the feature vectors from the relevant set of a user's posts, which differs depending on the task.

### 4.3 Risk Assessment

A user's relevant posts for risk assessment, from which the user-level feature vector is constructed, are the set of all of their posts on SuicideWatch. Using the CrowdFlower consensus as labels for the training set (620 users) and the all-experts consensus label as ground truth in the test set (245 users), we explored the use of supervised multiclass classification to detect the risk level of a user using support vector machines (SVM) in scikit-learn (Pedregosa et al., 2011). For standardizing data, we use max absolute scaling to scale every feature to lie in $[-1, 1]$. We used 5-fold cross validation on training data in order to explore both RBF and linear kernels, as well as to optimize the SVM's C parameter. We obtained a macro-averaged F1 score on test data of 0.46 with macro-averaged precision and recall scores being 0.48 and 0.53 respectively.[17]

### 4.4 Screening

We conduct screening experiments looking at evidence within $t$ days before the "signal" (i.e. the first SuicideWatch post), where $t$ could be 1, 2, 5, or 7 days. For control users, a random post is chosen as the point from which $t$ is determined. A user's relevant posts for screening, from which the user-level feature vector is constructed, include all of their posts during the relevant time interval on all Reddit forums *excluding* SuicideWatch or mental health forums. A user is excluded if they have no posts during the relevant interval.

Using these criteria, Table 6 shows the training and test set sizes, including number of positive and negative instances. Dataset size increases with the width of the time interval since, for example, there are more people who post within two days before the signal as compared to within just one day of the signal.

| $t$ | Train | | Test | |
|---|---|---|---|---|
| | positive | negative | positive | negative |
| 1 | 2024 | 1951 | 229 | 208 |
| 2 | 2806 | 2597 | 304 | 293 |
| 5 | 4184 | 3688 | 458 | 398 |
| 7 | 4763 | 4112 | 524 | 457 |

Table 6: Screening datasets

We explore the same set of classifiers as we did for the risk assessment part above. Again, we use F1 score on the test set as an evaluation metric. We also report macro averaged precision and recall scores. Binary classification is performed with results shown in Table 7.[18]

| | **Time Period (t)** | | | |
|---|---|---|---|---|
| | 1 | 2 | 5 | 7 |
| F1 | 0.66 | 0.65 | 0.65 | 0.66 |
| Precision | 0.70 | 0.67 | 0.67 | 0.67 |
| Recall | 0.68 | 0.66 | 0.66 | 0.66 |

Table 7: Screening results

#### 4.4.1 User-level Convolutional Neural Networks Assessment Classifier

In additional to our baseline classifier for the assessment task, we explored using a convolutional neural network (CNN), since CNNs are effective in many NLP tasks, especially text classification problems like sentence-level sentiment analysis (Kim, 2014; Flekova and Gurevych, 2016). We adopt a similar CNN architecture to the one introduced in Kim (2014) due to its popularity and ease of scalability to multiple tasks and strong results on many datasets. Figure 1 depicts the structure of our CNN architecture, where the input of the network is the concatenation of all user's posts and

[17] We also experimented with logistic regression and XGBoost, with substantially inferior results.

[18] For these experiments logistic regression and XGBoost had performance very similar to SVM.

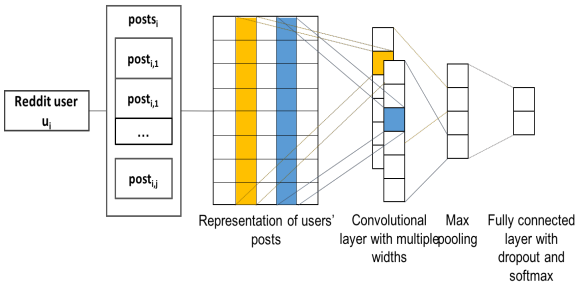


Figure 1: User-level CNN architecture

can be descibed as:

$$posts_{i,1:k} = post_{i,1} \oplus post_{i,2} \ldots \oplus post_{i,k} \quad (1)$$

Here $\oplus$ is the concatenation operator, $i$ represents $user_i$ and $k$ is the number of posts by $user_i$. Whereas a single post is the concatenation of the pre-trained word vectors (as introduced in 4.2), and can be defined as:

$$post_{i,j} = vec_{i,j,1} \oplus vec_{i,j,2} \ldots \oplus vec_{i,j,|W|_j} \quad (2)$$

Where $post_{i,j}$ represents the post $j$ of $user_i$, $vec_{i,j,\ell}$ is the embedding representation of $word_\ell$ in $post_j$ and $|W|_j$ is the number of words in $post_j$. We apply a filter window= $\{3, 4, 5\}$ words, where employing this filter to all the possible windows would represent a feature map $c$. On the resulting $c$, we apply max pooling (Collobert et al., 2011) and take the maximum feature as the representative one. Finally, we pass the output to a softmax layer to generate the label probability distribution. The neural model's performance yields a macro F1-score of 0.42 on the test data. Although the performance of SVM surpasses the CNN model, we opt to report CNN results as a deep learning baseline for this dataset, a reference for further research in this direction.

## 5 Related Work

There is an extensive clinical literature on suicidality assessment (e.g. Batterham et al. (2015); Joiner et al. (1999, 2005)), but very little specifically looking at assessment of suicidality based on social media content. This is a new topic that has received very little study to date in the clinical literature, with prior work focusing on non-clinican rather than clinician judgments (Egan et al., 2013; Corbitt-Hall et al., 2016). Griffiths et al. (2010) present a review of randomized controlled trials involving internet interventions for depression and anxiety disorders. Lind et al. (2017) offer a comprehensive discussion of crowdsourcing, using CrowdFlower, as a means for obtaining coding of latent constructs in comparison with content analysis.

Calvo et al. (2017) and Guntuku et al. (2017) present reviews of NLP research in which social media are used to identify people with psychological issues who may require intervention, and Conway and O'Connor (2016) provide a shorter survey focused on public health monitoring and ethical issues, highlighting the annual Workshop on Computational Linguistics and Clinical Psychology (CLPsych), initiated in 2014, as a forum for bridging the gap between computer science researchers and mental health clinicians (Resnik et al., 2014). Recent CLPsych shared tasks using data from the ReachOut peer support forums have provided opportunities for exploration of technological approaches to risk assessment and crisis detection (Milne et al., 2016; Milne, 2017); see also Yates et al. (2017).

Although predictive modeling for risk assessment is a burgeoning area, a key challenge for work on mental health in social media is connecting the clinical side with available social media datasets. Combining ground truth health record data with social media data is rare, with Padrez et al. (2015) representing a promising exception; they found that nearly 40% of 5,256 Facebook and/or Twitter users who were approached in a hospital emergency room consented to share both their health record and social media data for research.[19] Approximations of clinical truth are more common, e.g. self-report of diagnoses in social media (Coppersmith et al., 2014), or observed user behaviors such as posting on SuicideWatch (De Choudhury et al., 2016). Coppersmith et al. (2015, 2016) employed the Twitter data collection method of Coppersmith et al. (2014) to discover Twitter users with self-stated reports of a previous suicide attempt in order to identify valuable signal and support automated classification.

In work similar to the work we report here, Vioulès et al. (2018) applied a similar data collection approach to Coppersmith et al., searching Twitter for tweets containing key phrases based on risk factors and warning signs identified by the American Psychiatric Association and the American Association of Suicidology. They defined a four-

[19] Interestingly, participants agreeing to social media access were only slightly younger on average than those who declined (29.1 $\pm$ 9.8 versus 31.9 $\pm$ 10.4 years old.

category scale for distress and 500 tweets were annotated by researchers, with a subset of 55 validated by a psychologist. They achieved 69.1% and 71.5% chance-corrected agreement using Cohen's kappa and weighted kappa, respectively, with Fleiss kappa of 78.3% for the 55 tweets with three annotators; for automated classification they explored eight text classifiers and a variety of features, with their best performing combination for four-way classification achieving an F-measure of 0.518.

## 6 Dataset Availability and Ethical Considerations

The research we report was approved by the University of Maryland's Institutional Review Board (IRB). As Benton et al. (2017) discuss, human subjects research using previously existing data falls into a category exempted from the requirement of full IRB review as long as the data are either from publicly available sources or they do not provide a way to recover the identity of the subjects. In our case, the data are publicly available *and* from a site where users are anonymous. As an extra precaution we replace Reddit usernames with numeric identifiers.

Benton et al. (2017) point out that even exempt research needs to be reviewed by an IRB to make an exemption determination. In addition, they discuss the importance of taking particular care with sensitive data. In order to ensure appropriate standards are met, we will be making our dataset available to other researchers through the American Association of Suicidology (AAS), on organization whose mission is to promote the understanding and prevention of suicide and support those who have been affected by it.[20] AAS will provide governance in which researchers submit requests for access, with panel review ensuring, for example, that proper IRB procedures have been followed, that the researchers will provide appropriate protections for sensitive data, and that there will be no linkage of the dataset to other sites that could jeopardize user anonymity.

## 7 Conclusion

Assessing someone's suicide risk via social media has potential for enormous impact. In the U.S. alone, 124 million people live in areas where a mental health provider shortage is officially recognized (Bureau of Health Workforce, 2017). At the same time, online interaction is increasingly the norm; as of 2016, 68% of all U.S. adults were Facebook users (with high participation across all categories of age, education, income, and geography) with more than half of all U.S. adults actually visiting the site at least once per day.[21]

The context for this work is one in which the reliability of clinical assessment for suicidality is a real problem even when direct contact with the patient is available: clinicians are often using some kind of structured interview but also going on instinct, with attendant risks of bias, and most clinicians have not had specialized training for dealing with high risk populations, many of whom are underserved and with special characteristics such as veterans or substance abusers (R. Resnik, 2016). Reliably coded datasets are important for development and testing of machine learning methods, and such datasets also have the potential to help improve training methods for people engaged in suicide prevention (Tony Wood, Chair of the Board of Directors of the American Association of Suicidology, personal communication).

Against that backdrop, we have created a new dataset for research on risk assessment for suicidality based on social media, which includes expert ratings for 245 users and crowdsourced ratings for a superset of 865 users. We found that inter-rater agreement among experts is very good, with consistency particularly encouraged using detailed instructions specifying classification criteria. We also looked at differences in consistency when ratings are provided by experts using their own experience and judgment rather than following detailed instructions, and non-expert crowdsourcers.

Some limitations of the work thus far are worth noting. One is that we have so far limited ourselves to Reddit, which may have particular characteristics that fail to generalize; in particular, evidence suggests that users show different behavior when posting anonymously, with both positive and negative implications (Christopherson, 2007; De Choudhury and De, 2014).

A second limitation is that, without health records, outcomes, or even self-report questionnaires from the users whose postings were as-

[20] http://www.suicidology.org/about-aas/mission

[21] www.pewinternet.org/2016/11/11/social-media-update-2016/

sessed, we cannot validate clinician assessments; nor are we able to provide clinical evidence for improved validity using the detailed assessment instructions. Outcomes data would clearly be preferable if it were available; for example, Pokorny (1983) and Goldstein et al. (1991) attempt prediction of suicide using a wide range of variables and clinical measures for thousands of psychiatric inpatients. However, outcomes data are very difficult to obtain at scale; both of those studies failed at individual-level prediction, and Pokorny (1983) attributes that result in part to the low base rate of the positive instances. At the same time, it is worth noting that with some exceptions, e.g. phyiological evidence like tumors or seizures, psychiatric diagnosis is largely a pattern recognition task performed by clinicians. For example, dyslexia and schizophrenia are diagnosed via clinician assessment, and Alzheimer's disease cannot be definitively determined until post-mortem examination of the brain. We would therefore argue that, within the domain of mental health, good modeling of clinician risk assessment has the potential for high impact even without prediction of outcomes.

What this study provides is evidence that reliable clinician risk-assessment ratings for social media users are achievable, along with initial evidence that the detailed instructions can improve consistency — presumably helping to compensate for variation in training and experience — when human experts are assessing a person's risk level on the basis of their posting to a suicidality support forum. In addition, the results support cautious optimism regarding the ability of non-experts to make (or at least contribute to) risk assessment judgments; cf. pioneering work by Snow et al. (2008) showing that many natural language annotation tasks can achieve expert-level performance by combining multiple crowdsourced judgments.

A third limitation is that we have so far focused primarily on assessment when there is already reason to believe someone may be at risk, as signalled by their posting to the SuicideWatch forum. This *risk assessment* task, analogous to other tasks like CLPSych's ReachOut shared tasks (Milne et al., 2016; Milne, 2017), is different from the task of *screening*, where a wider net is cast in order to identify people who might not even know they have a problem. The two tasks are likely to differ in important ways. Fortunately, the data we have collected includes posts from SuicideWatch and control users in forums completely unrelated to mental health and therefore is amenable to research on screening, as well. This is one of the avenues we are currently pursuing, beginning with the very preliminary exploration presented in Section 4.4. In addition to the risk assessment dataset, we plan to also take similar steps to make the broader screening dataset available to other researchers in order to foster more rapid progress.

Finally, from a technical perspective, we have only just begun to tap the potential of the dataset. For example, metadata associated with posts includes potentially valuable temporal information (Coppersmith et al., 2015), and we also have not yet explored the value of the annotators' selecting the post that most strongly supports their judgment. In addition, the classification results here are just an initial exploration of the problem; for example, we plan to follow Vioulès et al. (2018) in exploring hierarchical rather than four-way classification, which yielded substantial improvements, and we are exploring the role of hierarchical attention networks (Yang et al., 2016) as a way to cut through noise to identify the most relevant signals. We look forward to other researchers joining us in order to foster more rapid progress.

## Acknowledgments

This research was supported in part by a University of Maryland MPower Seed Grant and by the National Institutes of Health. The authors wish to thank the anonymous reviewers for their thoughtful guidance, as well as thanking Bart Andrews, Jennifer Battle, Julie Bindeman, Craig Bryan, Glen Coppersmith, Darcy Corbitt-Hall, April Foreman, Kimberly O'Brien , Rebecca Resnik, William ("Bill") Schmitz Jr., Hannah Szlyk, and Tony Wood, for their incredibly helpful discussions and, in many cases, contributions of time and attention above and beyond the call of duty. Any errors are, of course, our own.

## References

Jonathan Anderson. 1983. Lix and rix: Variations on a little-known readability index. *Journal of Reading* 26(6):490–496.

Philip J Batterham, Maria Ftanou, Jane Pirkis, Jacqueline L Brewer, Andrew J Mackinnon, Annette Beautrais, A Kate Fairweather-Schmidt, and Helen Christensen. 2015. A systematic review and evaluation of measures for suicidal ideation and behaviors in

population-based research. *Psychological assessment* 27(2):501.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. pages 94–102.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. Semeval-2017 task 12: Clinical tempeval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. pages 565–572.

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

Bureau of Health Workforce. 2017. Designated health professional shortage areas: Statistics, first quarter of fiscal year 2018, designated HPSA quarterly summary. Health Resources and Services Administration (HRSA) U.S. Department of Health & Human Services, https://ersrs.hrsa.gov/ReportServer?/HGDW_Reports/BCD_HPSA/BCD_HPSA_SCR50_Qtr_Smry&rs:Format=PDF.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering* 23(5):649–685.

Kimberly M Christopherson. 2007. The positive and negative implications of anonymity in internet social interactions:on the internet, nobody knows youre a dog. *Computers in Human Behavior* 23(6):3038–3056.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60(2):283.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

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

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pages 51–60.

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. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. pages 106–117.

Darcy J Corbitt-Hall, Jami M Gauthier, Margaret T Davis, and Tracy K Witte. 2016. College students' responses to suicidal content on social networking sites: an examination using a simulated Facebook newsfeed. *Suicide and life-threatening behavior* 46(5):609–624.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics* pages 20–28.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

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 2016 CHI conference on human factors in computing systems*. ACM, pages 2098–2110.

Katie G Egan, Rosalind N Koff, and Megan A Moreno. 2013. College students responses to mental health status updates on Facebook. *Issues in mental health nursing* 34(1):46–51.

James N Farr, James J Jenkins, and Donald G Paterson. 1951. Simplification of Flesch reading ease formula. *Journal of applied psychology* 35(5):333.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, pages 4647–4657.

Lucie Flekova and Iryna Gurevych. 2016. Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 2029–2041.

Rise B Goldstein, Donald W Black, Amelia Nasrallah, and George Winokur. 1991. The prediction of suicide: Sensitivity, specificity, and predictive value of a multivariate model applied to suicide among 1906 patients with affective disorders. *Archives of general psychiatry* 48(5):418–422.

Kathleen M Griffiths, Louise Farrer, and Helen Christensen. 2010. The efficacy of internet interventions for depression and anxiety disorders: a review of randomised controlled trials. *Medical Journal of Australia* 192(11):S4.

Robert Gunning. 1952. The technique of clear writing .

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18:43–49.

Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. 2014. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pages 107–117.

Jr Thomas E Joiner, Rheeda L Walker, Jeremy W Pettit, Marisol Perez, and Kelly C Cukrowicz. 2005. Evidence-based assessment of depression in adults. *Psychological Assessment* 17(3):267.

Jr Thomas E Joiner, Rheeda L Walker, M David Rudd, and David A Jobes. 1999. Scientizing and routinizing the assessment of suicidality in outpatient practice. *Professional psychology: Research and practice* 30(5):447.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

J Peter Kincaid, Jr Robert P Fishburne, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*. Springer, pages 217–226.

Klaus Krippendorff. 2004. Reliability in content analysis. *Human communication research* 30(3):411–433.

Fabienne Lind, Maria Gruber, and Hajo G Boomgaarden. 2017. Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication methods and measures* 11(3):191–209.

Tong Liu, Qijin Cheng, Christopher M Homan, and Vincent Silenzio. 2017. Learning from various labeling strategies for suicide-related messages on social media: An experimental study. *ACM International Conference on Web Search and Data Mining Workshop on Mining Online Health Reports* .

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading* 12(8):639–646.

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 Linguistics and Clinical Psychology*. Association for Computational Linguistics, San Diego, CA, USA, pages 118–127. http://www.aclweb.org/anthology/W16-0312.

D.N. Milne. 2017. Triaging content in online peer-support: an overview of the 2017 CLPsych shared task. Available online at http://clpsych.org/shared-task-2017.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29(3):436–465.

Kevin A Padrez, Lyle Ungar, Hansen Andrew Schwartz, Robert J Smith, Shawndra Hill, Tadas Antanavicius, Dana M Brown, Patrick Crutchley, David A Asch, and Raina M Merchant. 2015. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. *BMJ Qual Saf* pages bmjqs–2015.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics* 2:311–326.

Umashanthi Pavalanathan and Munmun De Choudhury. 2015. Identity management and mental health discourse in social media. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, pages 315–321.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Alex D Pokorny. 1983. Prediction of suicide in psychiatric patients: report of a prospective study. *Archives of general psychiatry* 40(3):249–257.

Dina Popovic, Eduard Vieta, Jean-Michel Azorin, Jules Angst, Charles L Bowden, Sergey Mosolov, Allan H Young, and Giulio Perugi. 2015. Suicide attempts in major depressive episode: evidence from the bridge-ii-mix study. *Bipolar disorders* 17(7):795–803.

Reddit. 2018. Reddit privacy policy. Downloaded March 22, 2018, https://www.reddit.com/help/privacypolicy/.

Philip Resnik. 2017. The (in)ability to triangulate in data driven healthcare research. Presentation, SBS Decadal Survey - Workshop on Culture, Language, and Behavior, National Academies of Sciences, Engineering, and Medicine.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015. The university of maryland clpsych 2015 shared task system. In *CLPsych@ HLT-NAACL*. pages 54–60.

Philip Resnik, Rebecca Resnik, and Margaret Mitchell, editors. 2014. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Baltimore, Maryland, USA. http://www.aclweb.org/anthology/W/W14/W14-32.

Rebecca Resnik. 2016. Psychological assessment: The not good enough state of the art. Presentation, Veterans Affairs Suicide Prevention Innovations Conference (VASPI).

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pages 254–263.

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

M Johnson Vioulès, Bilel Moulahi, Jérôme Azé, and Sandra Bringay. 2018. Detection of suicide-related posts in twitter data streams. *IBM Journal of Research and Development* 62(1):7–1.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1480–1489.

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*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2968–2978. https://www.aclweb.org/anthology/D17-1322.

Ayah Zirikly, Varun Kumar, and Philip Resnik. 2016. The GW/UMD CLPsych 2016 shared task system. In *CLPsych@ HLT-NAACL*. pages 166–170.