CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts

Ayah Zirikly[1], **Philip Resnik**[2], **Özlem Uzuner**[3,4], and **Kristy Hollingshead**[5]

[1]Rehabilitation Medicine Department, National Institutes of Health, Bethesda, MD, USA
[2]Linguistics and UMIACS CLIP Laboratory, University of Maryland, College Park, MD, USA
[3]George Mason University, Fairfax, VA, USA
[4]Massachusetts Institute of Technology, Cambridge, MA, USA
[5]Florida Institute for Human and Machine Cognition, Pensacola, FL, USA
ayah.zirikly@nih.gov, resnik@umd.edu, ouzuner@gmu.edu, kseitz@ihmc.us

Abstract

The shared task for the 2019 Workshop on Computational Linguistics and Clinical Psychology (CLPsych'19) introduced an assessment of suicide risk based on social media postings, using data from Reddit to identify users at no, low, moderate, or severe risk. Two variations of the task focused on users whose posts to the r/SuicideWatch subreddit indicated they might be at risk; a third task looked at screening users based only on their more everyday (non-SuicideWatch) posts. We received submissions from 15 different teams, and the results provide progress and insight into the value of language signal in helping to predict risk level.

1 Introduction

Predicting risk of suicide is hard. McHugh et al. (2019), reviewing 70 studies, conclude that suicidality cannot be predicted effectively using the standard practice of clinicians asking people in person about suicidal thoughts: 80% of patients who were not already undergoing psychiatric treatment and who died of suicide denied having suicidal thoughts when asked by a general practitioner. They conclude that their study, along with with other recent meta-analyses, "highlight a high degree of uncertainty about the statistical strength of commonly used approaches to suicide risk assessment."

On a similar theme, after carefully reviewing more than three hundred studies, Franklin et al. (2016) conclude that predictive ability for suicidal thoughts and behaviors (STBs) has not improved across 50 years of research. Nock et al. (2019) observe that, in contrast to other fatal problems like flu or tuberculosis, deaths by suicide are as prevalent now as they were a hundred years ago, a lack of progress resulting in large part because "we lack a firm understanding of the fundamental properties of STBs, and when, why, and among whom they unfold" — not least because suicidal thoughts and behaviors rarely occur in a research laboratory.

Coppersmith et al. (2018) offer a powerful example of the information that is available beyond the research laboratory. They observe that for many people the "clinical whitespace" — long intervals between healthcare encounters — is occupied by frequent use of social media, an opportunity for obtaining data "in situ" (Nock et al., 2019), and they demonstrate that this information can be tapped effectively in order to build create automated binary classifiers for screening.

This progress raises two new problems, though. First, when binary screening systems are deployed, the number of people flagged as at risk will far exceed clinical capacity for intervention. So, rather than a binary classification, a finer grained assessment for degree of risk is needed, in order to support decisions about intervention priority. Second, obtaining relevant data for developing, improving, and validating classifiers is extremely difficult. Coppersmith and colleagues, for example, went to considerable effort to obtain donations of private social media data for research on suicide, and these sensitive materials are not easy to share with the broader research community.[1]

With these considerations in mind, we have formulated a new shared task for research community participation, based on a dataset introduced by Shing et al. (2018). In order to address the limits of binary classification, we formulate tasks based on a multi-level assessment of suicide risk

[1] In particular, Coppersmith et al. (2018) have introduced the OurDataHelps.org platform, which permits donors to authorize research access to their data from numerous social media sources, as well as information from wearables and other technologies. The platform has been adapted by their collaborators for research on other mental health topics, as well; for example, UMD.OurDataHelps.org collects data donations for a project focused on schizophrenia.

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 24–33
Minneapolis, Minnesota, June 6, 2019. ©2019 Association for Computational Linguistics

designed for social media, similar in spirit to previous CLPsych shared tasks on four-way assessment of crisis risk in a peer support forum (Milne et al., 2016; Milne, 2017). In order to address ethical access to and sharability of data, we focus on materials collected from Reddit, where posts are public and anonymous, and further de-identified by us; see Section 2. A limitation of the tasks is that we lack information about actual outcomes (suicide attempts or competions); we instead use human annotations of risk level as a starting point. In that regard this year's exercise can be viewed at minimum as establishing face validity for the idea of extracting meaningful signal related to suicidality from Reddit posts, and more optimistically as a step along the path to clinically meaningful predictions.

2 Data

2.1 Source dataset

We derived our shared task data from the dataset introduced by Shing et al. (2018). Shing et al. began with a collection intended to contain essentially every publicly available Reddit posting from its beginning in 2005 into summer 2015, and identified a subset of users potentially at risk by extracting all users who had posted to the *r/SuicideWatch* subreddit.[2] The process was analogous to the data collection method pioneered by Coppersmith et al. (2014) for a variety of mental health conditions, where an explicit signal for candidate (potentially relevant) Twitter users was defined by specifying a self-report pattern, e.g. *I have been diagnosed with [condition]*, and then matching posts were reviewed manually to identify candidates where the signal does not appear genuine, such as sarcastic or joking references. For the suicidality dataset, posting on SuicideWatch constituted the signal, and Shing et al. (2018) collected 11,129 candidate users on SuicideWatch, accounting for a total of 1,556,194 posts across Reddit, along with a comparable number of control users who did not post on SuicideWatch.[3]

2.2 User-level annotation

As discussed in more detail by Shing et al. (2018), annotation involved the assessment of risk for a randomly selected subset of 621 users on a four-level scale, based on their SuicideWatch posts. A detailed set of annotation instructions drawing on prior literature (Joiner et al., 1999; Corbitt-Hall et al., 2016), created in consultation with suicide prevention experts, identified four families of risk factors, described as follows:

- *Thoughts* includes not only explicit ideation but also, e.g., feeling they are a burden to others or having a "f*** 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;
- *Context* includes, e.g. previous attempts, a significant life change, or isolation from friends and family.

Using this assessment scheme, Shing et al. obtained annotations both from experts and from crowdsource workers for a randomly selected subset of users based on their SuicideWatch postings, assigning one of the following risk levels (a to d):

(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;

[2] The r/SuicideWatch subreddit, `https://www.reddit.com/r/SuicideWatch/`, is a forum providing "peer support for anyone struggling with suicidal thoughts, or worried about someone who may be at risk". Henceforth we refer to it simply as SuicideWatch.

[3] It is worth noting that, subsequent to Shing et al.'s collection and annotation, Gaffney and Matias (2018) reported on an analysis showing that the widely used Baumgartner Reddit collection, which Shing et al. had used as their starting point, has a number of gaps and limitations. However, Gaffney and Matias identify the greater risks as pertaining to user history analyses, network analysis, or comparison of participation across communities. They posit lower risk from coverage gaps for machine learning work on predictive modeling, commenting, "since the purpose of this kind of machine learning research is to make inferences about out-of-sample observations rather than to test hypotheses about a population, such research may be less sensitive to variation due to missing data."

(d) Severe Risk: I believe this person is at high risk of attempting suicide in the near future.

It is important to note that this process produced risk assessments at the level of individual users, not of individual posts. Inter-rater reliability was achieved for experts (Krippendorff's $\alpha = 0.81$) (to our knowledge the first published demonstration of reliability for clinical assessment of suicidality based on social media), along with fair agreement among crowdsourcers (Krippendorff's $\alpha = 0.55$). Analysis of the results also showed that when crowdsource workers make mistakes relative to experts' judgments, they tend to err on the side of caution — a good thing in a setting where false positives are a better kind of error than false negatives.

In the absence of data about outcomes (see discussion in Section 6), we expect the expert annotations to represent "truth" more accurately than crowdsourced judgments. However, for the shared task we elected to create both training and test data using the crowdsourced annotations, rather than using expert judgments as test data. We made this choice for two reasons. First, at least this first time creating a shared task on Reddit suicidality assessment, we wished to avoid the extra difficulties encountered in machine learning when there are mismatches between the training set and the test set. Second, we anticipate the possibility of repeating this shared task, and would like to lay the groundwork for a head-to-head comparison of results; obtaining crowdsourced judgments to create fresh test data will be considerably more practical than obtaining more expert judgments.

2.3 Reddit posts and metadata

For our tasks, the evidence we have about users' mental state comes from their Reddit posts. Information provided to participant teams included post_id (a unique identifier for the post), user_id (a unique numeric identifier for the user who authored the post), timestamp (time the post was created, encoded as a Unix epoch), subreddit (the name of the subreddit where the post appeared), post_title (title of the post) and post_body (text contents of the post).[4]

As discussed further in Section 7, although Reddit data are publicly available and the site was created specifically for anonymous posting, discussions on the platform nonetheless need to be viewed as sensitive and subject to careful ethical consideration (Benton et al., 2017; Chancellor et al., 2019). For that reason, a number of steps were taken to further remove identifying information from the dataset for the shared task.

First, although Reddit is a site for anonymous discussion, it is possible for users to put identifying information in their self-selected user names; although most select names like *awesomeprogrammer*, in principle nothing on the site would prevent someone from naming herself *mary-smith-UMDsophomore-born7July2002*. Therefore the dataset replaces the self-selected user names with arbitrary numeric identifiers for the user_id.

Second, automatic processing was performed on post titles and bodies, to replace IP addresses, email addresses, URLs, and person entities with special tokens.[5] For example, a processed post body might resemble this made-up example: *Taking a great class from _PERSON_ _PERSON_. If you want to learn more about it drop me a line at _EMAIL_ or check it out at _URL_.*

In addition, we filtered out all posts containing Arabic using the langdetect library.[6] We also performed data-cleaning steps to remove encoding issues or special string sequences that tokenizers such as spaCy's would fail to handle.

3 Tasks

Teams participated in one or more of the following three tasks.

- Task A is about risk assessment: the task simulates a scenario in which there is already online evidence that a person might be in need of help (e.g., because they have posted to a relevant online forum or discussion, in this case r/SuicideWatch), and the goal is to assess their level of risk from what they posted. This task uses the smallest amount of data, with each user typically having no more than a few SuicideWatch posts.

- Task B is the same risk assessment problem as task A, but in addition to the SuicideWatch posts (which identify that they may need help), teams can also use the users posts

[4] Unix epochs are a widely used standard for encoding time. Any timestamp is represented as the number of seconds that have passed since 00:00:00 Thursday, 1 January 1970, Coordinated Universal Time (UTC), minus leap seconds.

[5] We used spaCy for named entity recognition.

[6] https://pypi.org/project/langdetect/

elsewhere on Reddit (which might tell you more about them or their mental state). On average each user we collected data for has more than 130 posts on Reddit, and the subreddit categories are wildly diverse, from *Accounting* to *mylittlepony* to *SkincareAddiction* to *zombies*.

- Task C is about screening. This task simulates a scenario in which someone has opted in to having their social media monitored (e.g., a new mother at risk for postpartum depression, a veteran returning from a deployment, a patient whose therapist has suggested it) and the goal is to identify whether they are at risk even if they have not explicitly presented with a problem. Here predictions are made only from users posts that are not on SuicideWatch.

For all tasks, we provided participating teams with training and test data using an 80-20 split. In order to keep the original labels' distribution in the split, we applied the proportional training/test split separately for each label. The statistics of the data are shown in Tables 1 and 2. Note the large number of posts in tasks B and C, which makes these two tasks more challenging given the extra information and noise the participants have about each user.

	train	test	total
a	127	32	159
b	50	13	63
c	113	28	141
d	206	52	258
control	497	124	621
total	993	249	1242

Table 1: Number of users in training and test data

	Task A	Task B	Task C
train	919	57015	56096
test	186	9610	14231

Table 2: Number of posts for each task per split

4 Shared task submissions

Fifteen teams participated in at least one task, with 12 participating in task A, 11 in task B, and 8 in task C. Each team was permitted to submit up to 3 runs per task, and each identified a primary system that would be used in the official results and rankings. The full number of submissions we received for tasks A, B, and C were 33, 28, and 22, respectively. Teams were given in total (training and testing) about four weeks to develop their systems, generating predictions on test data during a roughly week-long interval at the end. Table 3 shows the participating teams and the tasks they submitted to, with per-task rankings (see Section 5).

In this section, we list the common preprocessing steps that the teams used prior to training and testing. Additionally, we descibe the approaches followed (machine learning models and features if applicable) in Sections 4.2 and 4.3. In section 6, we provide more details about the top systems per task.

Team	A	B	C
Affective_Computing	7	7	
ASU (Ambalavanan et al., 2019)	2		5
CAMH †	5	2	2
Chen et al. (2019)		4	
CLaC (Mohammadi et al., 2019)	1	5	1
CMU (Allen et al., 2019)	8		
IBM data science (Morales et al., 2019)	12	10	4
IDLab (Bitew et al., 2019)	4		
JXUFE †	9	8	
SBU-HLAB (Matero et al., 2019)	3	1	3
TsuiLab (Ruiz et al., 2019)		3	
TTU (Iserman et al., 2019)	6		8
UniOvi-WESO (Hevia et al., 2019)	10	11	
uOttawa †		9	7
USI-UPF (Ríssola et al., 2019)	11	6	6

Table 3: CLPsych 2019 participating teams and rankings (no paper is available for teams indicated with †)

4.1 Data preprocessing

The most common preprocessing steps that teams followed was removing stop words and punctuation, in addition to lowercasing. Some teams opted to remove the special deidentification tokens (e.g., *_PERSON_*, *_URL_*), and to apply number normalization or removal. Some filtered out posts that contain more than thirty _PERSON_ special tokens. An interesting preprocessing step suggested ordering the posts by timestamp, following the intuition that recent posts have more impact on the risk assessment. Additionally, some teams aggregated the name of the subreddit to the post for task B and C (Ambalavanan et al., 2019). Most teams employed commonly used tokenizers such as spaCy (Honnibal and Montani, 2017); an exception is Ríssola et al. (2019), who used Ekphrasis (Baziotis et al., 2017), a tool set that is tailored for social media data. Iserman et al. (2019) applied two-stage error spelling correction using edit distance from augmented dictionary entries.

4.2 Approaches

4.3 Model inputs

The submitted systems used a wide range of input representations on the post and user level. We can distinguish several main categories:

- Embeddings on the word, sentence or document (post/posts) level. In addition to GloVe and word2vec, we mostly see the more recently introduced contextualized embedding techniques such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018).
- Lexicon-based features. Teams used dictionaries mainly to capture emotions represented in the user's posts. Examples of dictionaries used are NRC (Mohammad, 2017) and LIWC (Tausczik and Pennebaker, 2010). These features were generally represented as the normalized count of post per category. Other lexicons were employed to capture user-level features including age and gender (Sap et al., 2014), and assessment of the Big-5 personality traits (Schwartz et al., 2013).
- N-gram features, mainly in the form of unigrams with TF-IDF weighting.
- Meta-features such as the time when the post was made available (i.e. timestamp) or the post's subreddit (Tasks B and C).
- Topic models such as LDA (Blei et al., 2003) and Empath (Fast et al., 2016).

We also see keywords to identify certain behaviors such as motivations linked to suicidality using a set of keywords; clinical findings in terms of UMLS (Bodenreider, 2004) keywords in the posts, flagging the suicide-related unique identifiers (CUIs); and language style similarity between posts in the same subreddit.

4.4 Models

The submissions for the shared task range from conventional machine learning approaches to deep neural network models. Support vector machines (SVM) and logistic regression are frequently used, in addition to the occasional decision tree and random forests approach. These approaches often involve feature engineering, where we see a wide variety and extensive combinations of the features mentioned above (Section 4.3).

The neural network models, on the other hand, depend mainly on embeddings, though teams opted to use the embeddings output from the language models in different ways. Many teams finetune the embeddings on either the full training data, the SuicideWatch subset, or on each of the title and body of the posts to create separate language models. Some teams used models that were pre-trained on Wikipedia and some other large corpora as-is in their system.

The most commonly used neural architecture is convolutional neural networks (CNN) on the user or post level, where in the latter case an aggregation step is needed to produce the final outcome. Other frequently employed architectures were long-short term memory (LSTM) networks or recurrent neural networks (RNNs) and LSTMs with an attention mechanism. Some teams experimented with multichannel neural networks in a multi-task learning setting.

5 Results

5.1 Metrics

The official metric used in this shared task is the macro-averaged F1 score. This metric was also used in previous CLPsych shared tasks that classified online posts into one of four labels (Milne et al., 2016; Milne, 2017); as a way of defining a single figure of merit, macro-averaging treats each class as contributing equally to performance, which helps avoid performance on a single class dominating the result when there is class imbalance (cf. Table 1).

In addition, we adopt two metrics introduced in those previous shared tasks, derived from systems' four-way classifications: *urgent* is the accuracy in making the binary distinction between a, b vs. c, d, and *flagged* is the accuracy in distinguishing b, c, d from a. The reasoning behind these metrics lies in real-world use cases one might encounter. A system that is good at identifying *urgent* posts can be viewed as a first step in potentially time-sensitive triage (erring on the side of caution by including moderate as well as severe risk), while a system that is good at *flagged* distinctions helps avoid wasting valuable human effort on no-risk cases.[7]

For each of the three tasks we report official rankings based on the primary system identified by the team. Additionally, in Section 6 we report on the best run in terms of macro-F1 score, whether primary or not.

Tables 4, 5, and 6 provide the results of the primary runs of the participating teams for each of the three tasks, ranked by highest macro-F1 score. For tasks A and C, the CLaC team (Mohammadi et al., 2019) ranks first with a combination of conventional and neural models: an SVM is employed at the end of the pipeline, where it acts as a meta-classifier on top of a set of CNN, Bi-LSTM, Bi-RNN and Bi-GRU neural networks. However, for both of those tasks, the primary runs do not generate the best *unofficial* macro-F1 score on the test set: a different variation on the CLaC approach, in which SVM uses as input both the neural features and the predicted class probabilities from the SVM, yields the best macro-F1 score, 0.533 for task A as compared with 0.481 for the primary system. On the other hand, the CAMH system, which uses a stacked parallel CNN with LIWC and a universal sentence encoder (Cer et al., 2018), produced the best unofficial F1 score for task C: 0.278 as compared to 0.268 for the CLaC primary system.

For task B the best official score is 0.457, obtained by the HLAB team, where the system used logistic regression with features from SuicideWatch and non-SuicideWatch language that were processed separately. The best unofficial F1 score (0.504) is also obtained by HLAB system, using BERT features generated separately from SuicideWatch and non-SuicideWatch posts.

6 Discussion

In comparing the results of tasks A and B, we note that systems, especially the top systems, perform comparably in terms of predicting the severe risk label (*d*). This suggests that, in general, information about all the other Reddit posts by a user does not necessarily add noise that hurts the performance, but rather, in some instances, it might have positive impact. Surprisingly identifying severe risk posts in task C yields good results given that the set of available posts excludes SuicideWatch and other mental health subreddits. However, the overall F1 score is low, which suggests that future work should focus on correctly classifying the non-severe risk labels (*c* and *b*). Across tasks, classifying label *b* has a low performance, which is mainly due to its smaller training size in comparison to the other labels. Additionally, and as expected, all systems are better at predicting the two extreme labels (*d* and *a*) as opposed to the medium-risk labels(*c* and *b*).

As a way to augment the training data and to benefit from other available datasets, Hevia et al. (2019) experimented with including ReachOut data from the CLPsych 2016 and 2017 shared tasks (Milne et al., 2016; Milne, 2017). Unfortunately, adding this dataset resulted in slightly worse performance. Although both datasets adopt a four-way scale, the annotation guidelines are different and there is no guaranteed one-to-one mapping between the two.

One of the interesting findings from the different systems is that severe-risk users appear to use a distinct vocabulary in comparison to the rest of the labels. This would support the intuition of building separate language models for SuicideWatch and non-SuicideWatch, or special features that can benefit from emotion and mental-health related lexicons.

Interestingly, we note that most submitted systems over-predict label *d* when the correct label is *c*. This confirms the value of reporting the *urgent* F1 score, noting that, in some instances, the distinction between the two labels is hard even for the crowdsourcers (Shing et al., 2018). Additionally, a number of the false positives observed concern users seeking advice for a relative or a friend as opposed to themselves. This suggests that building models specifically to separate such cases would be of value.

7 Ethical considerations

Mental health is a sensitive subject area, and work on technology for mental health using social media has broad implications. Benton et al. (2017) and Chancellor et al. (2019) provide thoughtful and comprehensive consideration of ethical issues. Informed by their discussions we focus here on several key ethical considerations for this shared task and how we handled them.

[7] Similarly to the previous shared tasks with four-way labeling, we exclude the no-risk label *a* in evaluation for screening task C. However, macro-F1 score is calculated over all four labels for tasks A and B.

team	accuracy	macro-f1	(flagged) f1	(urgent) f1	(d) f1	(c) f1	(b) f1	(a) f1
CLaC	0.504	**0.481**	**0.922**	0.776	0.543	0.4	0.244	0.737
ASU	0.544	0.477	0.882	0.826	0.655	0.281	**0.316**	0.656
SBU-HLAB	0.56	0.459	0.842	0.839	**0.692**	0.235	0.25	0.658
IDLab	0.544	0.445	0.852	0.789	0.673	0.292	0.167	0.649
CAMH	0.528	0.435	0.897	0.783	0.623	0.327	0.083	0.708
TTU	0.504	0.402	0.902	0.844	0.6	0.14	0.2	0.667
Affective Computing	**0.592**	0.378	0.92	**0.862**	0.685	0.065	0	**0.762**
CMU	0.472	0.373	0.876	0.773	0.545	0.302	0	0.646
JXUFE	0.464	0.364	0.882	0.779	0.571	0.217	0.087	0.582
UniOvi-WESO	0.512	0.312	0.897	0.821	0.633	0.062	0	0.553
USI-UPF	0.376	0.291	0.753	0.707	0.475	**0.408**	0	0.281
IBM data science	0.432	0.178	0.861	0.788	0.594	0	0	0.118

Table 4: Official results of task A primary systems ordered by macro-F1 score

team	accuracy	macro-f1	(flagged) f1	(urgent) f1	(d) f1	(c) f1	(b) f1	(a) f1
SBU-HLAB	**0.56**	**0.457**	0.821	**0.816**	**0.699**	0.245	0.25	0.634
CAMH	0.512	0.413	**0.91**	0.812	0.598	0.226	0.105	**0.721**
TsuiLab	0.408	0.37	0.789	0.603	0.506	0.264	0.205	0.507
Chen et al.	0.424	0.358	0.83	0.738	0.478	0.14	0.182	0.633
CLaC	0.416	0.339	0.843	0.718	0.549	0.185	0.069	0.554
USI-UPF	0.336	0.311	0.743	0.667	0.439	0.089	**0.417**	0.299
ASU	0.368	0.261	0.765	0.691	0.536	0.151	0	0.358
JXUFE	0.36	0.259	0.798	0.694	0.508	**0.298**	0	0.231
uOttawa	0.448	0.253	0.787	0.71	0.596	0	0	0.418
IBM data science	0.416	0.212	0.82	0.738	0.566	0	0	0.28
TTU	0.416	0.148	0.848	0.775	0.591	0	0	0

Table 5: Official results of task B primary systems ordered by macro-F1 score

7.1 Participants and research oversight

Social media posts are a window into people's thoughts and often into details of their lives. This has enormous value in understanding and predicting mental health, but it stands in tension with concerns about privacy, and formalized ethical standards only address these issues to a limited extent. The dataset used in this shared task was derived from previously existing, publicly available material on Reddit, and we obtained an Institutional Review Board (IRB) determination that work using the material constitutes "secondary research for which consent is not required", including the ability to share the dataset with other researchers, under the U.S. Federal Policy for the Protection of Human Subjects.[8] However, we also took several additional steps regarding participant protection and research oversight.

First, although a key characteristic of Reddit is its focus on anonymity (Gutman, 2018), users retain the ability to volunteer identifying information. As discussed in Section 2.3, therefore, we implemented additional, conservative measures for automatic de-identification to reduce the possibility of including any identifying information in either metadata or text data. In informal review of two sets of 100 randomly sampled postings from our training data, after de-identification — one from all postings and the other just from SuicideWatch — we found zero instances of personally identifying information in either text or metadata.

In addition, in order for teams to participate in the shared task, we required them (a) to provide evidence that their *own* organization's IRB (or equivalent ethical review board) had reviewed and approved their research activity using the dataset, (b) to provide a data management plan including provisions for appropriate protection of the data, and (c) to affirm that all team members had read Benton et al. (2017) and were committed to its broad ethical principles.[9] Mindful of Chancellor et al.'s call to include key stakeholders in the research process, the design of participant applications and their reviewing took place in consultation with clinical and domain experts at the American Association of Suicidology.

[8] https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html

[9] Teams' papers in this proceedings may or may not explicitly have mentioned IRB or ethical review, but it can be presumed in all cases to have been done.

team	accuracy	macro-f1	(flagged) f1	(urgent) f1	(d) f1	(c) f1	(b) f1
CLaC	0.673	**0.268**	0.671	**0.625**	**0.527**	**0.189**	0.087
CAMH	0.613	0.226	**0.673**	0.599	0.497	0.048	**0.133**
SBU-HLAB	**0.69**	0.176	0.587	0.554	0.465	0.065	0
IBM data science	0.435	0.165	0.554	0.455	0.329	0.097	0.069
ASU	0.597	0.159	0.63	0.575	0.396	0.082	0
USI-UPF	0.5	0.136	0.377	0.297	0.291	0.115	0
uOttawa	0.52	0.129	0.541	0.485	0.386	0	0
TTU	0.222	0.118	0.542	0.489	0.353	0	0

Table 6: Official results of task C primary systems ordered by macro-F1 score

7.2 The role of predictive models

Social media's window into the "clinical whitespace" (Coppersmith et al., 2018) offers the potential to identify and intervene with people who do not or cannot receive attention through conventional healthcare interactions. At the same time, algorithmic prediction of suicidality creates new challenges, such as creating potentially stigmatizing labels for false or even true positives, or generating an overwhelming number of new cases requiring intervention.

We cannot hope to address these issues in a single shared task, but we did have them in mind when designing it. Our view, informed by research in other domains, is that the most substantial, rapid progress on a problem takes place when a community is constructed around a common task with common data, even when the task and data are not perfect. (As is the case here, for example, in starting with crowdsourced judgments; see Section 2.2.) The way to understand tradeoffs and consequences involving false negatives and false positives is to build systems that make predictions, and then to involve clinicians and other practitioners in discussion of what those systems do, and how this relates to the real-world need — which makes CLPsych, as the venue for this shared task, just as important as the shared task itself.

8 Conclusions

The CLPsych 2019 shared task succeeded in its primary aims, which were to elicit community interest and effort in the problem of suicidality assessment using social media, and to lay solid foundations for work on this problem that will ultimately lead to deployable technology. The best results here show strong performance in culling out, among users who have posted to Reddit's SuicideWatch forum, those who are in urgent need of attention, and, conversely, in distinguishing people who might need attention from those who are at no risk. We also see a solid start on the even more challenging problem of identifying users in need of attention from more ordinary posts that do *not* come from SuicideWatch. On evalution of finer grained, four-way classification we find that the medium risk categories (low and moderate, as opposed to no risk or severe risk) are more challenging for systems, just as they are more difficult for human judges (Shing et al., 2018).

We aim to address some of the limitations of the present shared task in the near future. Although crowdsourced judgments permit easily repeatable evaluations, we also hope to facilitate community-level evaluation against expert judgments. We are also working on the creation of secure community infrastructure for research on sensitive mental health data, in order to reduce practical obstacles and reduce data privacy concerns by bringing researchers to the data, rather than disseminating data out to researchers. Our ultimate goal is to create an environment where rapid progress can be achieved by combining the benefits of large scale, publicly available, annotated data, as explored here, with private social media and associated outcomes obtained using fully consented, donated data (e.g. via OurDataHelps.org, Coppersmith et al. (2018)).

Acknowledgments

The second author was supported in part by an Amazon Machine Learning Research Award and a UMB-UMCP Research and Innovation Seed Grant. The authors gratefully acknowledge important discussions with Kate Loveys, Kate Niederhoffer, Rebecca Resnik, April Foreman, Jonathan Singer, Beau Pinkham, and Tony Wood, as well as Amazon and the University of Maryland Center for Health-related Informatics and Bioimaging (CHIB) for their generous sponsorship of CLPsych this year.

References

Kristen Allen, Shrey Bagroy, Alex Davis, and Tamar Krishnamurti. 2019. ConvSent at CLPsych 2019 task a: Using post-level sentiment features for suicide risk prediction on Reddit. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Ashwin Karthik Ambalavanan, Pranjali Dileep Jagtap, Soumya Adhya, and Murthy Devarakonda. 2019. Using contextual representations for suicide risk assessment from internet forums. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

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.

Semere Kiros Bitew, Giannis Bekoulis, Johannes Deleu, Lucas Sterckx, Klim Zaporojets, Thomas Demeester, and Chris Develder. 2019. Predicting suicide risk from online postings in Reddit the UGent-IDLab submission to the CLPysch 2019 shared task A. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

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

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (FAT* '19)*.

Lushi Chen, Abeer Aldayel, Nikolay Bogoychev, and Tao Gong. 2019. Similar minds post alike: Assessment of suicide risk by hybrid language and behavioral model. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

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, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10:1178222618792860.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*, pages 4647–4657. ACM.

Joseph C Franklin, Jessica Ribeiro, Kathryn Fox, Kate Bentley, Evan Kleiman, Xieyining Huang, Katherine Musacchio, Adam Jaroszewski, Bernard Chang, and Matthew Nock. 2016. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological bulletin*, 143. DOI 10.1037/bul0000084.

Devin Gaffney and J Nathan Matias. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PloS one*, 13(7):e0200162.

Rachel Gutman. 2018. Reddit's case for anonymity on the Internet. *The Atlantic*.

Alejandro González Hevia, Rebeca Cerezo Menéndez, and Daniel Gayo-Avello. 2019. Analyzing the use of existing systems for the CLPsych 2019 shared task. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Micah Iserman, Taleen Nalabandian, and Molly E. Ireland. 2019. Dictionaries and decision trees for the 2019 CLPsych shared task. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Thomas E Joiner, Jr, 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.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, and Mohammadzaman Zamani. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Catherine M McHugh, Amy Corderoy, Christopher James Ryan, Ian B Hickie, and Matthew Michael Large. 2019. Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. *BJPsych open*, 5(2).

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*, pages 118–127, San Diego, CA, USA. Association for Computational Linguistics.

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. 2017. Word affect intensities. *arXiv preprint arXiv:1704.08798*.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. CLaC at clpsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Michelle Morales, Danny Belitz, Natalia Chernova, Prajjalita Dey, and Thomas Theisen. 2019. An investigation of deep learning systems for suicide risk assessment. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Matthew K. Nock, Francesca Ramirez, and Osiris Rankin. 2019. Advancing Our Understanding of the Who, When, and Why of Suicide Risk. *JAMA Psychiatry*, 76(1):11–12.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

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.

Esteban A. Ríssola, Diana Ramírez-Cifuentes, Ana Freire, and Fabio Crestani. 2019. Suicide risk assessment on social media: USI-UPF at the CLPsych 2019 shared task. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Victor Ruiz, Lingyun Shi, Jorge Guerra, Wei Quan, Neal Ryan, Candice Biernesser, David Brent, and Fuchiang Tsui. 2019. CLPsych2019 shared task: Predicting users suicide risk levels from their Reddit posts on multiple forums. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36. Association for Computational Linguistics.

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.