Community-level Research on Suicidality Prediction in a Secure Environment: Overview of the CLPsych 2021 Shared Task

Sean MacAvaney
IR Lab, Georgetown University
sean@ir.cs.georgetown.edu
University of Glasgow

Anjali Mittu
University of Maryland
amittu@umd.edu

Glen Coppersmith
Qntfy
glen@qntfy.com

Jeff Leintz
NORC at the University of Chicago
Leintz-Jeff@norc.org

Philip Resnik
University of Maryland
resnik@umd.edu

Abstract

Progress on NLP for mental health — indeed, for healthcare in general — is hampered by obstacles to shared, community-level access to relevant data. We report on what is, to our knowledge, the first attempt to address this problem in mental health by conducting a shared task using sensitive data in a secure data enclave. Participating teams received access to Twitter posts donated for research, including data from users with and without suicide attempts, and did all work with the dataset entirely within a secure computational environment. We discuss the task, team results, and lessons learned to set the stage for future tasks on sensitive or confidential data.

1 Introduction

In natural language processing, and in AI more generally, progress depends on data. The most significant progress on a problem takes place when an entire community is working on the same dataset at the same time; for example, the wide availability of speech recognition today is a result of decades of research using DARPA benchmark datasets and evaluations for speech-related tasks (Juang and Rabiner, 2005).

In healthcare, however, community-level activity is an enormous challenge. Laws and regulations related to data confidentiality create obstacles to access, including significant administrative overhead such as data use agreements and significant technical overhead involving arrangements for secure data distribution, storage, and management (Lane and Schur, 2010). In mental health and particularly crisis detection, missteps like Samaritans Radar raise highly public red flags despite well-intentioned goals (Horvitz and Mulligan, 2015; Resnik et al., 2021). All these legal, regulatory, operational, and public perception risks naturally make potential data providers skittish about data sharing. As a result, important research in healthcare is balkanized, with community efforts scattered among different datasets in *ad hoc* fashion as different teams work with the data they are able to gain access to. Or potentially it doesn't take place at all, as talented researchers go work on other problems where obtaining data is just easier.

Secure data enclaves are one solution to this problem (Lane and Schur, 2010). The key idea in a data enclave is to bring researchers to sensitive data, rather than disseminating data out to researchers. A data enclave provides secure remote access to data using carefully designed statistical, technical, legal and operational controls. Computation on an enclave is done using a copy of the data residing there without full networking access, meaning that nothing can be imported or exported without disclosure review. This does not replace necessary steps like IRB approvals, data use agreements, and record de-identification; for example, data enclave users can still look at private data within the enclave and need to agree not to attempt de-anonymization. However, it drastically simplifies community-level access. A single, comprehensive description of security provisions can be created for data providers and ethical review boards, and data providers need to enter into data use agreements only with the enclave, rather than with individual teams.

To our knowledge, the CLPsych 2021 shared task is a first-of-its-kind endeavor: as far as we know, it is the first time a community-level shared task with sensitive mental health data has been conducted on a data enclave, and more generally shared tasks on sensitive data are rare in the NLP and machine learning communities. In addition, although uses of data enclaves are often centered on the use of analytics tools, in this shared task the environment was designed to support the full arsenal of NLP and machine learning methods. We accomplished this by partnering with NORC at the

Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, pages 70–80
June 11, 2021. ©2021 Association for Computational Linguistics

University of Chicago. Since 2006, the NORC Data Enclave® has served U.S. state and federal agencies, research institutes, foundations, and universities by securely housing and providing remote access to confidential data. In a collaborative project with University of Maryland, NORC has developed the UMD/NORC Mental Health Data Enclave (henceforth the Enclave, for short), a subset of NORC Data Enclave infrastructure designed specifically with the requirements of mental health NLP and machine learning research in mind.

Data for this shared task were provided by Qntfy, which runs OurDataHelps.org, an online platform that permits donations of digital life data (including social media) for the purposes of advancing research in mental health and wellbeing. Individuals come from a range of lived experience with mental health, specifically related to this shared task: individuals who have survived suicide attempts, loved ones of people who have died by suicide, and people who just want to help. For this shared task, Qntfy established a data provider agreement with NORC, and NORC executed data use agreements with the participating teams. The University of Maryland, College Park IRB reviewed and approved a protocol for research with, and sharing of, the OurDataHelps data. The arrangement here therefore exemplifies the advantages of data enclaves discussed above. For the data provider, it was much easier to work out an agreement with just a single entity running an established secure infrastructure, which significantly lowered the bar for sharing data with multiple teams. In addition, NORC's platform and processes for team access, platform security, and import/export review created a far greater level of confidence in privacy controls than sending data out to a large number of far-flung teams with heterogeneous environments. For teams, this provided a rare opportunity to work with sensitive mental health data containing actual outcomes, not proxy data as is more common in social media mental health research and which can be problematic for a variety of reasons (Ernala et al., 2019).

The shared task itself involved assessment of suicide risk via prediction of suicide attempts, based on the natural language of users on Twitter. There were two subtasks: Subtask 1 involved assessing suicide risk given 30 days of tweets prior to the date of an attempt (or a corresponding date when no attempt was made), and Subtask 2 involved assessing suicide risk given the prior six months of tweets.

A set of 21 teams signed up and were onboarded on the Enclave. A total of five teams ultimately submitted systems by the deadline. All teams have been given several months of additional access and support on the Enclave, in order to permit continued experimentation. We are hopeful that results obtained during this extended time period will lead to publications beyond CLPsych.

In this overview paper, we provide not only a summary overview the shared task itself, in terms of the research problem and participating teams' findings about predicting suicide risk from Twitter data, but also a retrospective analysis of conducting a shared task in a secure enclave, including lessons learned and recommendations for future tasks of this kind.[1]

2 Background and Related Work

A number of recent articles discuss the use of NLP, machine learning, and social media in service of mental health. As important motivating background, a meta-analysis by Franklin et al. (2017) concludes that prediction of suicidal thoughts and behaviors has not improved in fifty years, encouraging a shift to algorithmic and machine learning approaches. Schafer et al. (2021) provide significant empirical support for this view via another meta-analysis looking specifically at traditional theory-driven versus machine learning approaches to prediction of suicide risk, demonstrating that the latter are significantly more effective at prediction.[2] Naslund et al. (2020) and Lee et al. (2021) provide overviews that include thoughtful, big-picture commentary on research and clinical applications for mental health taking advantage of NLP, machine learning, and social media. Resnik et al. (2021) offer an overview of issues more specifically focused on using naturally occurring language as a source of evidence in suicide prediction.

One running theme throughout discussions of

[1] We would be happy to discuss logistical issues, and share details and specific language from our IRB protocol, data provider, and data use agreements, in order to facilitate others who would like to organize shared tasks similar to this one. Interested readers should contact clpsych-2021-shared-task-organizers@googlegroups.com.

[2] In regard to the goals of prediction versus scientific explanation and understanding, it is worth noting the argument by Yarkoni and Westfall (2017) that psychology research as a whole, including research with explanatory goals, would benefit by taking a predictive approach.

this kind involves the availability of data to work with, and the interplay, or even tension, between the need for research and the need to respect privacy and other ethical considerations. Horvitz and Mulligan (2015) provide one short, useful discussion specifically focused on data and privacy, and Benton et al. (2017) and Chancellor et al. (2019) discuss ethical issues specifically with regard to social media and work on mental health. Lane and Schur (2010) provide a valuable entry point to the concept of data enclaves as a way to balance the need for data access in order to make progress in healthcare with respect for patient privacy — this concept ties in directly with the call by Schafer et al. (2021) for community-level mental health datasets to be easily available for research so that the predictive ability of models can be compared and research can be replicated. Those kinds of comparisons and replications are instrumental in modern data-driven research because without them it is impossible to gain insight into which approaches are most promising or to rule out the possibility that apparent differences are related to idiosyncratic differences in data.

Related, the most current paradigms in NLP and machine learning involve both general-purpose pre-training and task-specific fine-tuning. To some extent, pre-training data may capture generalizations about language that transfer well to problems in the mental health space. However, many off-the-shelf language resources that are commonly used, such as BERT (Devlin et al., 2019), are built from sources such as books and Wikipedia entries. These may translate poorly to systems dependent on social media posts from Twitter, Facebook, or an online discussion forum. It is well known that systems perform better when they are trained on materials similar to the materials the system will run on (Alsentzer et al., 2019; Beltagy et al., 2019). Therefore using task-specific data from immediately relevant sources as training data for social media based mental health tasks is a high priority that requires attention.

Another theme found in related literature involves the nature and quality of the variables being predicted. The sensitivity of mental health data has led to a proliferation of proxy variables taken from publicly available data rather than ground-truth clinical variables or real-world outcomes (e.g. De Choudhury and De, 2014; Coppersmith et al., 2014; Yates et al., 2017; Shing et al., 2018; Cohan et al., 2018; Thorstad and Wolff, 2019). As two particularly well known and influential examples, Coppersmith et al. (2014) infer mental health diagnoses of Twitter users by looking for publicly self-reported diagnoses, and De Choudhury et al. (2016) infer mental health progressions to suicidal ideation by examining when Reddit users shift from mental health subreddits to the SuicideWatch subreddit. Such data tend to have the advantages of being readily accessible and large in size. However, Ernala et al. (2019) note a variety of problems and limitations in using proxies rather than clinically grounded variables. Coppersmith et al. (2018) offer a rare exception in this kind of work, using an ethical process of data donation to obtain social media data *with outcomes* for research on prediction of suicide attempts; our shared task is based on a subset of their data.

3 Data

We briefly describe our data sources, and how we constructed the shared task datasets for binary classification tasks.

3.1 Data sources

We began with data donated to the OurDataHelps.org platform, discussed in greater detail by Coppersmith et al. (2018). Donations to the platform include data from people who have survived a suicide attempt, data from people who died by suicide that has been donated by loved ones, and data donated by people who have not attempted suicide but want to help. When donations take place, a questionnaire is filled out that collects basic demographic data and mental health history. This includes the number of past suicide attempts and dates associated with them, although dates are not provided in all cases.

Although the platform permits collection of a wide range of data, including, for example, social media, fitness, and wearable data, in this shared task we restricted our attention to Twitter data and a subset of basic information from the questionnaire. Only publicly available tweets are used, typically visible to friends and family, and these were de-identified before being provided to the Enclave.

On the Enclave, participants also had access to a copy of the UMD Reddit Suicidality Dataset (Shing et al., 2018; Zirikly et al., 2019). This dataset was used by one of the teams (NUSIDS) in their submission.

In addition, a non-sensitive practice dataset using the shared task data format was provided to participants so they could work on developing and debugging their systems outside of the Enclave. It was based on a modified version of the `depression-detection` dataset (Wang et al., 2019).[3]

3.2 Users with Suicide Attempts

In the version of the data we began with, there are 3,631 users, 1,613 of whom attempted (and possibly died by) suicide. From this version, we imposed several filters. We only considered users who had donated Twitter data and who had reported their gender and date of birth in the questionnaire, in order to match users with a suicide attempt to a control user. If a user had attempted suicide, we only included them if they had a date associated with the attempt, a necessary restriction in order to examine tweets in the time period leading up to the attempt. For users with multiple attempts, we only considered the most recent attempt having a date. Filtering in this way left 250 users with suicide attempts, associated dates, and data prior to the attempt. For Subtask 1, we restricted the set to users who had made posts in the 30 days prior to their suicide attempt, a total of 68. For Subtask 2, we restricted the set to users who had made Twitter posts during the six months prior to the attempt, which included a total of 97 users. Teams were provided with anonymized user IDs, the date of the most recent suicide attempt (if applicable), and a list of the user's de-identified tweets from the applicable time span.

3.3 Control Users

Similar to Coppersmith et al. (2018), we included a set of control users matched one-to-one with users who had attempted suicide, based on having the same gender, similar age (within 5 years), and similar number of tweets. These criteria resemble previous matching in the 2015 CLPsych shared task (Coppersmith et al., 2015) and in Coppersmith et al. (2018). Age and gender are common controls in the mental health space, and we chose to match using a similar number of tweets so that corresponding users in the dataset would be represented by similar quantities of social media evidence. For each user with a suicide attempt, we found a match by first finding all users matching age and gender, then selecting the user with the closest number of tweets. Tweets taken from the control user were from the same time frame as their match who had an attempt in order to minimize differences in context, such as tweets about world events.

	Subtask 1	Subtask 2
Total # of Users	114 / 22	164 / 30
Users Under 30	104 / 15	138 / 23

Table 1: The total number of users in each subtask and the number of users under the age of 30. The numbers in the table are given as (training set) / (test set)

	Subtask 1	Subtask 2
Female	118	168
Male	12	20
Non-Binary	4	4
Other	2	2

Table 2: The distribution of gender across all users.

Table 1 shows the final number of users in each subtask and Table 2 shows the age distribution of users. In the shared task, we saved 15% of the users for the test set; these numbers are shown in the table. For both subtasks, most of the users were female between the ages of 18 to 24 and most of the users were under the age of 30. Within the time period, for Subtask 1, users had an average of 24 tweets per person and in Subtask 2, there were an average of 102 tweets per person.

4 Baseline

A baseline system was provided to shared task participants to use or build upon.[4] Baseline preprocessing includes several standard steps. First, we removed all URLs, user mentions, and emojis from the tweets. Whenever a user's tweet includes an image, GIF, or link, the links are removed. We tokenized the tweets using the Twitter-specific Twikenizer and removed stopwords from the tweets' text using the default SpaCy (Honnibal et al., 2020) stopword list.[5] Last, we split hashtags into the words they are made up of: first, we try to split by camel-case or by underscores; if that fails, we use a method from HashTagSplitter, attempting to split into the smallest subset of real words.[6]

[3] https://github.com/seanmacavaney/clpsych2021-shared-task/tree/main/practice-dataset

[4] https://github.com/anjmittu/clpsych2021-shared-task-baseline

[5] https://github.com/Guilherme-Routar/Twikenizer

[6] https://github.com/matchado/HashTagSplitter

The baseline classification model used logistic regression with the default parameters from SciKit Learn (Pedregosa et al., 2011), employing unigram and bigram count vectors.

5 The Enclave

As discussed in the introduction, data-driven research in mental health, and healthcare more generally, faces significant obstacles owing to important concerns about privacy and data confidentiality. Data enclaves offer a potential solution (Lane and Schur, 2010).

NORC at the University of Chicago, an independent, non-profit research institution, took on the operational aspects of running this shared task on their data enclave. Significant time was spent working with Qntfy, who were responsible for providing the OurDataHelps data, and the shared task organizers, to develop the data provider agreement, data use agreements, operational policies, supporting infrastructure, and technical and operational support for the organizers and shared task teams.

All aspects of the shared task on the Enclave were run using exactly the same procedures as for NORC's traditional Data Enclave clients, such as government agencies working with confidential databases. Teams that worked on the shared task executed a data use agreement with NORC and then were "onboarded" to the Enclave, being provided with account logins, passwords, documentation, procedures for uploading and export (both requiring human review of the material entering or leaving the Enclave), and contacts and procedures for technical support.

The Enclave environment includes two main parts. The first part is a secure virtual desktop (using Citrix), accessed via the Data Enclave login page through an internet browser. The second part of the Enclave is NORC's Mental Health Data Enclave (MHDE) Cluster on Amazon Web Services (AWS). From within the secure Citrix desktop, participants use PuTTY ssh to reach a gateway machine on this cluster. They can run code there or submit batch jobs using the Slurm cluster management and job scheduling system.[7] The AWS environment is configured to spin up a new instance for the duration of the job and then spin it down when completed, conserving compute resources to save cost.

Crucially, the Enclave is a closed environment. Neither the secure desktop nor the AWS cluster permit access to the Internet. It is not possible to `scp` or `sftp` data. It is not possible to open a socket in a program that connects externally. It is not possible to print, print screen, or even to copy/paste to or from the external environment.

The NORC Data Enclave's data security model integrates a portfolio approach with the Five Safes framework (Ritchie, 2017) to harden the security posture. This means that bringing materials in, such as code, data, or other resources, requires an import request process. Each request triggers a robust review process to provide safe passage of confidential micro-data and ensure imported material does not contain any virus or code aimed at disabling the capabilities or facilitating unauthorized access. In order to set up the Enclave environment and hopefully speed up this process for shared task participants, it was pre-loaded with major Python packages and tools (more than 4000 of them), the shared task baseline code, and shared task data; see further discussion in Section 8.

Similarly, as a data custodian for restricted data (e.g. confidential micro-data for federal, state and commercial clients), NORC must ensure that any data leaving the NORC Data Enclave is safe and free of inappropriate disclosures. This means that there is a request-based procedure for exporting any material from the Enclave, with formal review criteria that include both dataset-specific criteria and general guidelines applied globally across all requests.

6 Submissions

Each team was permitted up to three submissions for each subtask (30 days and 6 months). In each subtask, the numbered submissions for each team distinguish the "primary" submission (numbered 1) from additional contrastive runs (numbered 2 and 3). In total, we received 30 submissions, with five teams providing three runs each for both subtasks.

NUSIDS (Zagatti et al., 2021). For the shared task, NUSIDS designed SHTM, a Self-Harm Topic Model, which combines standard Latent Dirichlet Allocation (LDA) with a self-harm dictionary. This was tested using a combination of the shared task data, along with the practice dataset and the UMD Reddit Suicidality Dataset. In their submission to the task, the team used a combination of an LSTM and term feature vectors with SHTM-based fea-

[7] https://slurm.schedmd.com/

Team (Sub.)	F_1	F_2	TPR	FAR	AUC
NUSIDS (1)	0.583	0.648	0.700	0.636	0.645
NUSIDS (2)	0.615	0.714	0.800	0.727	0.664
NUSIDS (3)	0.300	0.300	0.300	0.636	0.373
ScyLab (1)	0.526	0.481	0.455	**0.273**	0.678
ScyLab (2)	0.526	0.481	0.455	**0.273**	0.678
ScyLab (3)	0.421	0.385	0.364	0.364	0.636
sentimenT5 (1)	0.455	0.455	0.455	0.545	0.438
sentimenT5 (2)	0.500	0.472	0.455	0.364	0.616
sentimenT5 (3)	0.571	0.656	0.727	0.818	0.413
SoS (1)	0.286	0.278	0.273	0.636	0.264
SoS (2)	0.400	0.377	0.364	0.455	0.529
SoS (3)	0.364	0.364	0.364	0.636	0.397
UlyaLamia (1)	**0.692**	**0.763**	**0.818**	0.545	0.702
UlyaLamia (2)	0.522	0.536	0.545	0.545	0.409
UlyaLamia (3)	0.636	0.636	0.636	0.364	**0.740**
Our baseline	0.636	0.636	0.636	0.364	0.661

Table 3: Results of participating systems and our baseline for Subtask 1 (30 days). The best result for each metric is listed in bold.

Team (Sub.)	F_1	F_2	TPR	FAR	AUC
NUSIDS (1)	0.684	0.812	**0.929**	0.786	0.663
NUSIDS (2)	0.703	**0.823**	**0.929**	0.714	0.648
NUSIDS (3)	0.649	0.759	0.857	0.786	0.480
ScyLab (1)	0.769	0.704	0.667	**0.067**	0.809
ScyLab (2)	0.769	0.704	0.667	**0.067**	0.791
ScyLab (3)	**0.815**	0.764	0.733	**0.067**	**0.844**
sentimenT5 (1)	0.467	0.467	0.467	0.533	0.618
sentimenT5 (2)	0.516	0.526	0.533	0.533	0.591
sentimenT5 (3)	0.727	0.769	0.800	0.400	0.720
SoS (1)	0.429	0.411	0.400	0.467	0.444
SoS (2)	0.533	0.533	0.533	0.467	0.640
SoS (3)	0.400	0.400	0.400	0.600	0.502
UlyaLamia (1)	0.595	0.671	0.733	0.733	0.582
UlyaLamia (2)	0.581	0.592	0.600	0.467	0.564
UlyaLamia (3)	0.645	0.658	0.667	0.400	0.569
Our baseline	0.710	0.724	0.733	0.333	0.764

Table 4: Results of participating systems and our baseline for Subtask 2 (6 months). The best result for each metric is listed in bold.

tures. Submissions varied in the hyper-parameters of the model (e.g., window size and number of topics), as well as the training data.

ScyLab (Gamoran et al., 2021). The ScyLab submission used Bayesian modeling over features grounded in domain knowledge. These features included behavioral information learned by Twitter activity, Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) based features using priors from Eichstaedt et al. (2018) and other dictionary-based approaches. The submissions varied the distributions for the priors and hyper-parameters (type of regression) for the logistic-regression model.

sentimenT5 (Morales et al., 2021). SentimenT5 took different approaches in their submissions to explore the performance of simple traditional models versus fine-tuned deep learning models. In both Subtasks 1 and 2, they submitted results from gradient-boosted classifiers. One used syntax features and the other character TF-IDF features. For Subtask 1, they also submitted results from a contextualized language model classifier, and, for Subtask 2, a voting ensemble method.

SoS (Wang et al., 2021). Team SoS introduced the C-Attention Network, which uses latent feature information implicitly in the embeddings. This was compared with submissions using KNN and SVM classifiers. Latent features included using Doc2vec embeddings (Lau and Baldwin, 2016). Hand-crafted features included emotion lexicons, part-of-speech tags, and a custom dictionary that models various stages of suicidal behavior.

UlyaLamia (Bayram and Benhiba, 2021). In the UlyaLamia submissions, the authors were motivated by real-life applicability of their model to use tweet-level classification. The team's submissions used a majority voting approach over individual tweets. In order to pick which machine learning method to use, the team experimented with multiple methods tuned on the training data using a leave-one-out strategy. Their final submissions were the top methods from the leave-one-out results.

7 Results

We evaluated each system in terms of F_1, F_2 (favoring recall), True Positive Rate (TPR), False Alarm (Positive) Rate (FAR), and Area Under the ROC Curve (AUC). We use F_1 score as the primary evaluation metric, though it is valuable to consider all metrics for a complete view of the system performance.

We present the results of the submissions in Tables 3 and 4. In Subtask 1, Team UlyaLamia ranked highest in F_1, F_2 and TPR; however, their FAR was higher than the baseline and in the middle of the other team's submissions. Team UlyaLamia was also the only team to exceed the baseline F_1 score, with NUSIDS being the next closest team. In Subtask 2, Team ScyLab ranked highest in F_1, FAR, and AUC. Their strongest submission beat or met

	UlyaLamia			ScyLab			sentim.			SoS			NUSIDS		
label	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	8	10	7	6	6	4	1	3	4	8	4	20	1	1	5
1	11	18	12	5	5	7	22	2	2	20	21	10	2	3	3
1	10	16	9	16	16	11	19	4	6	16	5	18	7	6	14
1	1	20	1	10	10	9	11	19	20	13	8	2	22	22	22
1	4	9	4	15	15	17	3	16	18	7	18	14	3	4	12
1	15	12	11	1	1	1	17	11	13	2	9	15	18	18	18
1	19	15	15	12	12	15	18	12	14	9	3	8	5	5	9
1	2	21	2	3	3	10	13	13	15	17	1	6	19	19	19
1	17	8	20	13	13	13	16	7	9	5	22	19	14	13	16
1	3	3	3	2	2	5	14	14	16	18	10	16	17	15	17
1	12	5	14	22	22	18	21	22	22	19	11	9	8	8	11
0	5	6	5	4	4	2	6	8	10	4	7	22	6	16	1
0	9	11	8	8	8	6	7	6	7	15	19	13	12	12	7
0	7	2	10	7	7	3	4	9	11	21	16	5	15	14	10
0	16	17	17	11	11	8	2	20	3	11	6	11	10	9	6
0	20	7	18	19	19	21	12	10	12	6	13	7	9	7	13
0	14	13	16	17	17	12	8	21	1	10	2	17	13	11	8
0	13	19	13	20	20	19	20	1	21	22	12	12	11	10	2
0	21	1	21	14	14	14	15	15	17	12	15	1	16	17	15
0	6	14	6	18	18	16	9	18	5	3	14	21	20	20	20
0	18	22	19	21	21	20	10	5	19	1	17	4	4	2	4
0	22	4	22	9	9	22	5	17	8	14	20	3	21	21	21

Figure 1: Rank comparison of the submissions for Subtask 1. A label of 1 indicates users with suicide attempts. Ranks closer to 1 indicate a higher score (more likely to have made a suicide attempt) given to the user. Rows are sorted by label, then median rank.

	UlyaLamia			ScyLab			sentim.			SoS			NUSIDS		
label	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	5	2	4	3	3	2	3	3	2	9	5	1	11	22	9
1	26	5	25	7	5	8	9	11	11	23	28	2	7	7	2
1	8	4	8	9	11	1	28	4	3	7	20	15	14	11	12
1	12	21	11	4	4	5	16	19	18	3	2	12	2	2	18
1	25	28	28	2	2	3	20	20	19	5	27	11	3	3	5
1	13	26	12	6	8	4	11	15	15	30	19	3	4	4	7
1	6	12	9	22	22	28	27	2	22	1	10	6	20	17	10
1	14	14	13	10	9	22	5	6	5	21	8	8	13	13	16
1	10	11	6	30	30	11	25	7	27	25	17	14	15	16	14
1	16	15	16	8	7	7	14	16	16	11	14	16	16	15	19
1	29	1	29	13	13	15	8	17	17	6	24	29	28	28	22
1	2	9	2	17	17	29	29	28	28	26	4	17	12	12	23
1	19	23	27	1	1	10	22	22	20	15	1	18	5	5	24
1	28	22	26	20	25	12	2	5	4	22	21	7	21	19	25
1	1	24	1	11	10	6	26	26	25	2	12	24	29	29	29
0	11	6	7	24	23	19	10	1	7	12	11	22	8	10	8
0	18	10	17	23	21	16	4	8	8	8	30	9	10	9	4
0	4	3	5	28	29	13	12	13	13	10	22	21	9	8	21
0	9	20	10	5	6	9	24	9	6	18	18	13	22	23	15
0	27	13	22	15	15	14	30	30	30	20	6	4	1	1	28
0	15	17	18	18	19	24	6	10	10	13	9	25	17	14	13
0	21	7	23	16	16	25	17	21	29	16	16	10	6	6	3
0	17	29	14	25	24	17	1	12	12	17	7	30	19	21	11
0	24	18	24	19	18	21	21	25	24	19	29	19	18	18	6
0	3	16	3	21	20	26	18	23	21	14	26	26	26	25	20
0	22	19	21	26	26	20	7	29	1	24	25	5	23	26	17
0	20	8	19	29	28	30	19	24	23	29	13	23	24	20	1
0	7	25	15	12	12	23	23	27	26	28	23	20	27	27	27
0	23	27	20	27	27	27	15	18	9	4	15	27	25	24	26
0	30	30	30	14	14	18	13	14	14	27	3	28	30	30	30

Figure 2: Rank comparison of the submissions for Subtask 2. A label of 1 indicates users with suicide attempts. Ranks closer to 1 indicate a higher score given to the user. Rows are sorted by label, then median rank.

the baseline in every metric and was notably low in their FAR. Five submissions came close or beat the baseline in F_1 score in Subtask 2.

The methods used by teams in the shared task had difficulties performing well in both subtasks. Given shorter-term information starting 30 days prior to an attempt, tweet-specific language (UlyaLamia) performed beste, but dictionary-based methods (e.g., ScyLab) worked best with the longer-term evidence (6 months prior to an attempt).

To gain a better understanding of the differences between the submissions, we plot the ranks of each test user for both subtasks in Figures 1 and 2. From these figures, we can see that some users easily classified by most systems, while others were notably difficult. For instance, in the last positive (label=1) row in Figure 2 (Subtask 2), the majority of systems were (incorrectly) very confident that the user did not make a suicide attempt. Nevertheless, three submissions gave this user the highest or second-highest likelihood. These results suggest that an ensemble method may be beneficial for this task.

This task is notably similar to Coppersmith et al. (2018), who performed experimentation including OurDataHelps.org data with similar restrictions, matching criteria, and the same binary outcomes. They found that a longer history of tweets led to slightly better predictions, but, unlike our shared task, they did not find a significant increase in performance between using tweets 90 to 0 days prior to an attempt and using tweets 180 to 90 days prior. In Coppersmith et al. (2018), the AUC score using tweets 30 days prior to an attempt is .89 and the AUC score using tweets six months prior to an attempt is .93.

At the same time, it is important to note that those results are not directly comparable to the present task, given differences in dataset size and composition. Coppersmith et al. (2018) used more OurDataHelps data, and this was augmented with a dataset of users who had made publicly self-stated suicide attempts, building on work in Coppersmith et al. (2016). In total, Coppersmith et al. (2018) performed their experimentation using a dataset containing 418 users with suicide attempts, compared to this task's 97 users.

8 Enclave Lessons Learned

We solicited feedback from all registered teams (both those who submitted results and those who did not) regarding the shared task experience. This discussion and our lessons learned for the future are informed by their comments.

Onboarding. Shared tasks are bursty by nature,

the first burst involving participants getting started. In contrast, the ongoing operations of a data enclave involve a more continuous scheduling process for new user account requests. This led to challenges in the onboarding process. As noted in Section 5, procedures for this shared task were identical to the procedures used when serving organizations like government agencies, with not one fewer *i* dotted, not one fewer *t* crossed. This meant that teams experienced longer than expected delays between completing their paperwork and actually being able to begin work on the Enclave. We would recommend more lead time in the future, leaving significant time for account requests and also having teams prioritize which members need access first.

Importing code and dependencies. Similarly, data enclaves require strict import policies and procedures; every import request is treated as though it could contain highly confidential data, a virus, or disabling code. Again, the bursty nature of shared task activity created challenges. Despite our attempts to anticipate and pre-load software and data resources that were likely to be needed (informed by an earlier survey of people engaged in CLPsych-related work), the burst of requests as teams got started created long delays as teams waited for their code and software dependencies to come online. Workarounds, such as recreating code manually, were complicated by the inability to copy/paste inside the environment.

Time zones. The CLPsych 2021 Shared Task received global interest, with teams participating on several continents. However, data enclaves rarely provide 24/7 support. While having a diverse set of teams work on the task is indispensable, having support concentrated in a single U.S. time zone disproportionately affected those working outside the U.S. We anticipate that these issues could be mitigated in part by greater lead time (again), and also by streamlining processes to require fewer round trips of communication.

Slurm and Notebooks. These days, many prefer to conduct NLP research in an interactive setting using Jupyter Notebooks. While these were supported on the head node of the cluster, they were not available when running jobs on compute nodes, including those with GPU resources. This is worth considering. While such an arrangement would run through one's compute budget faster (as compute nodes would remain running), the interactive benefits may be a tradeoff that teams are willing to make, and this would also avoid batch-job overhead for those who do not require the capabilities offered by a scheduler like Slurm.

Connectivity and Enclave Maintenance. Like any well supported infrastructure, the Enclave requires regular maintenance and has occasional downtime. Scheduled maintenance was easy to plan for, but unplanned downtime can be a real challenge in deadline-driven activities like a shared task.

Despite these challenges, which certainly gave rise to some frustration, a number of teams expressed gratitude for being able to work on data that would otherwise be unavailable, and others expressed that they were pleased with the overall responsiveness and speed of the Enclave. Some also expressed appreciation for having had ample of compute credits for conducting their experiments.[8]

If there is a unifying theme in our lessons learned, it is that the challenges we encountered are connected almost entirely with the gap between the typical flexibility of experimental computational work in NLP, particularly in the compressed time frame of a shared task, versus the more extended, carefully centralized, step-by-step, controlled processes that take place on a data enclave. But of course that's the whole point: those same careful, centralized processes are the things that guard against inappropriate use and disclosure of sensitive data.

As a particular note for the future, more advance planning and communication with participants would alleviate several of these challenges, especially onboarding and importing code and dependencies. For this shared task, we chose to prioritize allowing participants to start working on the task sooner, rather than requiring teams to commit long before they would begin work and start going through a more structured and scheduled process to prepare the Enclave with their specific team-level requests. We attempted to preload needed libraries and tools onto the Enclave even before teams began to register — but we could not predict all of the tools and resources participants would want, so even with our efforts there was still a gap. And although we tested the onboarding process and coding experience, any new, diverse group of people is going to discover unanticipated issues when using

[8] AWS credits supporting this activity were provided by Amazon.

a large production environment for a new purpose.

That said, it is worth noting that a time-bounded shared task is just one model for this type of collaborative work. In other domains, it is not uncommon for community shared activity to take place over the longer term, e.g use of the MIMIC dataset (Johnson et al., 2016) in research on electronic health records. A shorter-term, bursty event like a shared task may be the wrong model when navigating between the requirements of flexible research and the requirements of data privacy — many challenges would be mitigated if participants were not all attempting to meet the same deadline. Therefore, an alternative paradigm to consider would involve a more gradual intake of participants, reducing the backlogs and avoiding bottlenecks in account creation and handling of initial import requests. This would would also allow participants to more freely work in their own time zone, and factor in downtimes in their schedule.

9 Conclusion

In this effort, we introduced a mental health shared task using sensitive language data in a secure data enclave that offered broad NLP and machine learning capabilities. Participants conducted studies on the prediction of suicide risk based on tweets, using donated data containing actual outcomes rather than proxy data and matching individuals who attempted suicide with control users. Participants built systems that were able to achieve high predictive power (up to 0.823 F_1 score), while carefully balancing true positives and false alarms. Through the shared task, we learned more about the challenges of conducting such a task in an enclave environment, leading to observations that will help set the stage for future efforts of this kind.

Acknowledgments

The shared task organizers would like to express deep gratitude to the individuals who donated data to OurDataHelps, without whom this research would not be possible. The organizers are also immensely grateful to all the participants for their efforts and patience; to the NORC partners and personnel (particularly co-author Jeff Leintz, Ron Jurek, Kyle Stufflebaum, Ramon Castillo, Rachel Miller, Sundeep Bhatia, Wesley Hale, Kim Le, John Nieszel, Jason Keller, and the Data Enclave Manager team) for their tremendous contributions and their willingness to step out onto the bleeding edge in making the Enclave and this shared task happen; to Tim Mulcahy, Scot Ausborn, and Christian Ilie for foundational discussions and effort getting the UMD/NORC Enclave collaboration off the ground; to co-author Glen Coppersmith, Tony Wood, Alex Yelskiy, and the rest of the Qntfy team for their leadership in suicide-related research and collecting and sharing OurDataHelps donated data; to Alexander Hoyle for technical assistance with AWS configuration; to Julia Lane for useful background on data enclaves; to NAACL for its support of CLPsych; and to the creators of the `depression-detection` github repository. This shared task received internal financial support at NORC and was also supported in part by Amazon through an AWS Machine Learning Research Award and by a University of Maryland AI + Medicine for High Impact (AIM-HI) Challenge Award.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP 2019)*.

Ulya Bayram and Lamia Benhiba. 2021. Determining a person's suicide risk by voting the short-term history of tweets for CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2021)*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*.

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.

Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018.

SMHD: A large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*.

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 (CLPsych 2014)*, pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (CLPsych 2015)*, pages 31–39.

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.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Tony Wood. 2016. Exploratory data analysis of social media prior to a suicide attempt. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (CLPsych 2016)*.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

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*, pages 2098–2110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16.

Joseph C. Franklin, Jessica D. Ribeiro, Kathryn R. Fox, Kate H. Bentley, Evan M. Kleiman, Xieyining Huang, Katherine M. Musacchio, Adam C. Jaroszewski, Bernard P. Chang, and Matthew K. Nock. 2017. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*.

Avi Gamoran, Yonatan Kaplan, Almog Simchon, and Michael Gilead. 2021. Using psychologically-informed priors for suicide prediction in the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2021)*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Eric Horvitz and Deirdre Mulligan. 2015. Data, privacy, and the greater good. *Science*, 349(6245):253–255.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Biing-Hwang Juang and Lawrence R Rabiner. 2005. Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, 1:67.

Julia Lane and Claudia Schur. 2010. Balancing access to health data and privacy: a review of the issues and approaches for the future. *Health services research*, 45(5p2):1456–1467.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*.

Ellen E Lee, John Torous, Munmun De Choudhury, Colin A Depp, Sarah A Graham, Ho-Cheol Kim, Martin P Paulus, John H Krystal, and Dilip V Jeste. 2021. Artificial intelligence for mental healthcare: Clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

Michelle Morales, Prajjalita Dey, and Kriti Kohli. 2021. Team 9: A comparison of simple vs. complex models for suicide risk assessment. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2021)*.

John A Naslund, Ameya Bondre, John Torous, and Kelly A Aschbrenner. 2020. Social media and mental health: Benefits, risks, and opportunities for research and practice. *Journal of technology in behavioral science*, 5(3):245–257.

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.

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

Philip Resnik, April Foreman, Michelle Kuchuk, Katherine Musacchio Schafer, and Beau Pinkham. 2021. Naturally occurring language as a source of evidence in suicide prevention. *Suicide and Life-Threatening Behavior*, 51(1):88–96.

Felix Ritchie. 2017. The 'five safes': a framework for planning, designing and evaluating data access solutions. In *Data for Policy 2017: Government by Algorithm? (Data for Policy)*. Zenodo.

Katherine M Schafer, Grace Kennedy, Austin Gallyer, and Philip Resnik. 2021. A direct comparison of theory-driven and machine learning prediction of suicide: A meta-analysis. *PLoS one*, 16(4):e0249833.

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 (CLPsych 2018)*, pages 25–36.

Robert Thorstad and Phillip Wolff. 2019. Predicting future mental illness from social media: A big-data approach. *Behavior research methods*, 51(4):1586–1600.

Ning Wang, Fan Luo, Yuvraj Shivtare, Varsha Badal, K.P. Subbalakshmi, R. Chandramouli, and Ellen Lee. 2021. Learning models for suicide prediction from social media posts. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2021)*.

Susan Wang, Labiba Kanij Rupty, Mahfuza Hu mayra Mohona, Aarthi Alagammai, Munira Omar, and Marwa Qabee. 2019. Depression detection using Twitter data - group project for udacity private and secure ai project showcase. https://github.com/swcwang/depression-detection.

Tal Yarkoni and Jacob Westfall. 2017. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*.

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 (EMNLP 2017)*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Guilherme Augusto Zagatti, Sujatha Das Gollapalli, and See-Kiong Ng. 2021. Suicide risk prediction by tracking self-harm aspects in tweets. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2021)*.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2019)*.