A Appendix

A.1 StudentSADD data description

The Student Suicidal Ideation and Depression Detection (StudentSADD) dataset includes 345 sessions of data from 302 unique student participants with modalities including participant demographics, PHQ-9 scores, responses to a text prompt, unscripted audio recordings, scripted audio recordings, Twitter data (whether or not the participant has Twitter and their Twitter username), and retrospective phone data (calendar, call logs, text logs, contacts and GPS). Participants were given access to an Android app and Website Survey. They could choose which version of the survey took. All questions were the same, however, phone data was only collected from students who used the Android app. All publicly available data is stored in CSV files.

Responses to the text prompt, demographic questions, phq-9 questions, willingness to share, and timestamps of submission by participants, can be found in summaryDataStudentSADD.csv. Participants were asked to share their favorite place (text prompt); demographic info, specifically, age, gender, student status, race/ethnicity, whether or not they have had prior depression treatment, whether they were remote, and whether or not they had covid; responses to 9 depression screening questions with four point Likert scales; GPS data (if shared); and Twitter information, such as whether or not they have Twitter and their username. Participant willingness to share audio and phone data was also included in this file. The sessions from the same student are also marked as copies in this CSV file.

The file phoneDataStudentSADD.csv includes data collected from those participants who used the Android app and elected to share phone data. Each phone modality could be shared or denied. Phone data modalities that could be elected to be shared were calendar entries, call logs, text logs, and contact entries. All names and phone numbers in these data modalities were one-way hashed. The file includes calendar and contacts entries, for which the count can be extracted. The call and text logs include a one-way hashed address (phone number) of the sender/receiver, as well as the timestamp and size/length of each correspondence.

The file scriptedTranscript.csv contains scripted audio transcripts and the file unscriptedTranscript.csv contains unscripted audio transcripts. These CSV files also have the PHQ-9 and item-9 scores of the corresponding participants. Feature and embedding representations of the audio data will be released instead of the raw audio data to protect student privacy. The files scriptedSMILE.csv and unscriptedSMILE.csv contain examples of extracted audio features.

Further, detailed results for machine learning models that use the StudentSADD data will be released. The files baselineScripedML.csv, baselineUnscriptedML.csv, baselineTranscriptML.csv, and baselineTextML.csv contain examples of detailed machine learning model results.

84 IDs were designated as a test set to evaluate the machine learning models trained on this data. The Test and Train IDs can be found below. In addition to the training IDs from unique students, we have also included the IDs from duplicated entries. Note, not all IDs shared every modality.

Test IDs: [1607712777, 292, 2613, 1610640355, 1607494599, 1607040811, 1608492986, 6390, 396, 1607734901, 1607350992, 1608992344, 1609903202, 74, 7159, 4698, 7547, 4441, 1607097951, 8479, 8170, 4707, 7516, 1609174124, 1608853150, 8516, 1611424664, 2843, 1607040596, 1953, 1607772081, 1608564004, 2627, 1607217921, 1607118643, 1607314413, 1609887404, 1608335387, 4098, 1607046006, 1608242917, 8918, 1607131299, 9754, 1607262842, 1607273026, 2478, 1607536408, 1607291545, 1608707232, 1609941585, 1608200497, 1610630377, 7711, 1607810287, 9934, 1608850448, 4041, 1609166629, 1608168856, 1607572897, 6831, 1608586814, 1608588581, 2837, 8180, 1608631410, 1607051003, 3830, 4879, 1608920128, 1607019351, 8181, 3473, 1608335906, 1607738757, 1608770486, 7564, 1607495239, 1609983150, 1607397061, 1607696074, 103, 2222]

Unique Train IDs: [4769, 1607928177, 1607269923, 7755, 4598, 1607807806, 1608741452, 3323, 1610110670, 1607133044, 9745, 1607291670, 5245, 4442, 319, 1607133218, 1607010270, 1608587203, 1609256130, 1608582258, 5028, 1609771771, 5229, 3517, 1608595561, 1608048050, 1607410780, 528, 1607134906, 3102, 1607555727, 1609887167, 3985, 7256, 3523, 1607289708, 1609890222, 850, 1608917024, 5047, 1608061691, 4782, 1608062276, 1056, 1611517276, 1607636681, 1607891972, 5571, 1609052616, 1607927243, 2525, 4353, 1610818662, 8640, 1607559849, 6706, 1608624428, 1607968838, 1608672132, 552, 1608537399, 1610381937, 1608607986, 381, 1608589576, 3920, 1608059746, 1609027319, 1607357022, 1607691623,

1609899907, 1608470962, 8791, 1610380419, 3064, 1609473849, 1607712704, 1609887249, 1609888813, 1608588103, 1244, 7279, 1607339125, 1607712682, 8472, 1269, 1607045076, 1607365865, 1846, 191, 1811, 1608702785, 1609049435, 5330, 1607257348, 1609890530, 3278, 1608586899, 1607939718, 2430, 1609893292, 60, 1607270186, 6336, 8650, 1608495626, 1608586953, 2121, 1607295286, 896, 1609889389, 1607560754, 6548, 6580, 1607440988, 1609111416, 1607807159, 8663, 1607129044, 6658, 1607799213, 3933, 1608596696, 1608663032, 1610791060, 1607135820, 1607413039, 1607659758, 1608487726, 4859, 1609142183, 1607276888, 7452, 1607368510, 1607266081, 2623, 1608416516, 2128, 3227, 5881, 6510, 1609166843, 7569, 1607712793, 1608850996, 3273, 1607939838, 9986, 3302, 1607206195, 1609082904, 1607510222, 7612, 1607022963, 1607051040, 1607719324, 1608849324, 1607642639, 1607104225, 705, 1608506424, 1608188073, 8018, 8085, 4755, 1611704179, 1607193886, 7007, 3041, 4001, 1552, 1716, 1608053349, 1608572299, 1608051417, 1607712784, 836, 1607929944, 1607795480, 1608200317, 415, 3662, 1610109929, 2496, 8550, 6868, 1608587385, 1608591490, 7370, 4549, 7505, 1879, 1876, 1608003341]

Duplicated Train IDs: [1607315588, 1607087749, 518, 6280, 1607315467, 1607497867, 1608359052, 8468, 1607124379, 6941, 1607510942, 1607921053, 1611575587, 1608201126, 1607639591, 4521, 1607315498, 1607639340, 1607785646, 1609353141, 1607453496, 1607293881, 1607674683, 5948, 1608683584, 1608087234, 3267, 1607802179, 7109, 838, 9034, 1868, 1608683738, 8284, 95, 1609154655, 1607159523, 1607124838, 1607088999, 7912, 1607802347, 1607764720, 1607785968]

A.2 Data and code access for reviewers

To access the StudentSADD data, reviewers can use the login information (provided in the submission system) at our project website: `https://emutivo.wpi.edu/data/`

Upon publication, we will grant similar access to other researchers who request our data and agree to the terms of the data licence in Section A.3. This approach is standard for datasets in this domain [32; 34]. This is not a static dataset, as we will continue to add more data features and embedding representations. We have inquired about obtaining a DOI for StudentSADD through our institutional library.

To access the code used to generate the results in this paper, reviewers can navigate to our public github repository: `https://github.com/mltlachac/StudentSADD`

A.3 StudentSADD Dataset - End User License Agreement

(`https://emutivo.wpi.edu`)

The person in request may download and use this database only after signing and returning this agreement form. By signing this document, the user agrees to the following terms:

Commercial and academic use

The database is made available for research purposes only. Any commercial use of this data is forbidden.

Redistribution

The user may not distribute the database or parts of it to any third party.

Publications

The use of data for illustrative purposes in publications is allowed. Publications include both scientific papers and presentations for scientific/educational purposes. In this case, the identity of the subjects should be protected (no release of identifiable information for subjects).

Citation

All publications reporting on research using this database have to acknowledge this by citing the following article:

ML Tlachac, Ricardo Flores, Miranda Reisch, Rimsha Kayastha, Nina Taurich, Veronica Melican, Connor Bruneau, Hunter Caouette, Ermal Toto, Elke Rundensteiner, "StudentSADD: Mobile De-

pression and Suicidal Ideation Screening of College Students during the Coronavirus Pandemic",
in Submission at Neural Information Processing System (NeurIPS) 2021 Datasets and Benchmarks
Track

For specific software output that is shared as part of this data, the user agrees to respect the individual software licenses and use the appropriate citations as mentioned in the documentation of the data.

EULA changes

Worcester Polytechnic Institute is allowed to change these terms of use at any time. In this case, users will be informed of the changes and will have to sign a new agreement form to keep using the database.

Warranty

The database comes without any warranty. In no event shall the provider be held responsible for any loss or damage caused by the use of this data.

Name   Date   Signature

Organization   Address

A.4   Data management plan

Data description and formats

The Student Suicidal Ideation and Depression Detection (StudentSADD) dataset contains 245 sessions submitted through a mobile app by 302 students over between August 2020 and January 2021. The data includes PHQ-9, demographics, retrospective phone data, text prompt, audio recordings, location history, and tweets. The audio is stored as WAV files. The remaining data is stored as text in CSV files.

Data archiving, access and sharing, and data preservation

Data will be stored in the file systems of Worcester Polytechnic Institute (WPI). The stored data will be protected with disk mirroring, daily backups and other means. Full-time system administrators will monitor the security and availability of these systems. Appropriate access control and other security policies and mechanisms will be put in place to protect the integrity, security, privacy, confidentiality and other rights or requirements. For access by and sharing with the greater research community and general public, research group websites will be used, and personally identifiable information (PII) will be appropriately removed or anonymized from the essential data used for our research. To protect privacy, confidentiality or other rights/requirements, while at the same time ensuring scientific reproducibility and verifiability, the data will be summarized in the forms of intermediate statistics, and made publicly available, or when requested from other researchers.

Data privacy

Data will be shared only under rules specified by our IRBs, and only when properly anonymized. The systems on which the data will be stored will have access restricted to the project members via standard filesystem permissions management. The server on which the data is stored is within the University's restricted access datacenter. Files that will be made publicly available will not contain any identifiable location information.

Policies and provisions for re-use, re-distribution

Any data collected will be in compliance with the IRB protocols of Worcester Polytechnic Institute. Data gathered for this project may be reused in other, related, research projects conducted by the PIs or graduate students. It is expected that other researchers in machine learning in mental health would be interested in our dataset. Requests for the data would be handled as per Section III. These datasets cannot be used for commercial applications or purposes, or changed and resubmitted without the PIs' permission and are subject to the intellectual property policies of WPI.

Rights and Obligations

The Principal Investigators (PIs) will be responsible for the implementation of the Data Management Plan. WPI owns the technology developed at each individual university by each university's employees.