Left on Read: Reply Latency for Anxiety & Depression Screening

ML Tlachac
Bryant University
Smithfield, RI, USA
mltlachac@bryant.edu

Samuel S. Ogden
California State University Monterey Bay
Seaside, CA, USA
sogden@csumb.edu

ABSTRACT

Mental health is a critical societal issue and early screening is vital to enabling timely treatment. The rise of text-based communications provides new modalities that can be used to passively screen for mental illnesses. In this paper we present an approach to screen for anxiety and depression through reply latency of text messages. We demonstrate that by constructing machine learning models with reply latency features. Our models screen for anxiety with a balanced accuracy of 0.62 and F1 of 0.73, a notable improvement over prior approaches. With the same participants, our models likewise screen for depression with a balanced accuracy of 0.70 and F1 of 0.80. We additionally compare these results to those of models trained on data collected prior to the COVID-19 pandemic. Finally, we demonstrate generalizability for screening by combining datasets which results in comparable accuracy. Latency features could thus be useful in multimodal mobile mental illness screening.

CCS CONCEPTS

• **Applied computing** → *Health informatics*; **Psychology**; • **Information systems** → **Texting**; • **Human-centered computing** → *Mobile phones*; • **Computing methodologies** → *Supervised learning by classification.*

KEYWORDS

mobile health, digital phenotype, metadata, mental health screening

ACM Reference Format:
ML Tlachac and Samuel S. Ogden. 2022. Left on Read: Reply Latency for Anxiety & Depression Screening. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3544793.3563429

1 INTRODUCTION

The prevalence of mental illness is increasing annually with more than a fifth of U.S. adults experiencing mental illness in 2020 [15]. Mental illnesses are debilitating, resulting in an estimated $193.2 billion in lost earnings in the USA [8]. Depression and anxiety are the two most common mental illnesses [7]. As early treatment is important for positive prognosis [6], screening surveys are being deployed more often in clinical care practices [13]. However, this approach can only screen people who seek medical attention. Further, mental health stigma [10] unfortunately means that screening survey scores can be influenced by conscious and unconscious bias.

Consequently, recent research has screened for mental illnesses with a variety of passively harvested modalities, such as smartphone logs [1, 4, 16, 23]. Such logs present a particularly promising solution to increase screening rates given the ubiquity of smartphones. Additionally, SMS text message content has been used to screen for depression [20] and suicidal ideation [17]. Despite encouraging results, message content presents privacy concerns, prompting research to use the metadata without content. In particular, latency of text replies [21] and time series of texts [19] have been used to screen for depression. For a new dataset collected after the start of the COVID-19 pandemic, [18], classification models screened for moderate depression with an F1 of 0.64 and moderate anxiety with an F1 of 0.50 with time series of communication logs.

We hypothesize that text reply latencies in this new dataset may be more useful for anxiety and depression screening than the log time series. As such, we extract reply latency features in order to screen for both depression and anxiety with machine learning models. We further gauge reply latency generalizability for depression screening by combining these new logs with logs collected prior to COVID-19. This research thus provides a more in-depth assessment of the usefulness of text reply latency features to screen for mental illnesses. In this paper, our contributions include:

(1) Assessment of reply latency features for anxiety screening,
(2) Comparison of the anxiety and depression screening ability of text reply latency features for the same set of participants,
(3) Analysis of the impact of COVID-19 on the ability of text reply latency features to screen for depression, and
(4) Assessment of aggregating datasets for depression screening.

2 DATA & MODELING METHODOLOGY

In this research, we leverage datasets [5, 18, 22] containing retrospectively harvested SMS text message logs labeled with mental illness screening scores. In other words, mobile apps administered participants screening surveys while scraping the text logs stored on the phone. In our models, we use logs from the two weeks prior to the participants completing the screening surveys, as that is the time frame captured by the screening surveys [9, 14].

2.1 Reply Latency Feature Engineering

We define text message reply latency as the number of seconds between when the participant received a text from a particular contact and when the participant sent a text to that particular contact. For each participant and contact combination, we consider only pairs of received and sent texts that occurred consecutively. We extract reply latencies from all such consecutive pairs in the two weeks preceding the completion of the screening surveys,

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09…$15.00
https://doi.org/10.1145/3544793.3563429

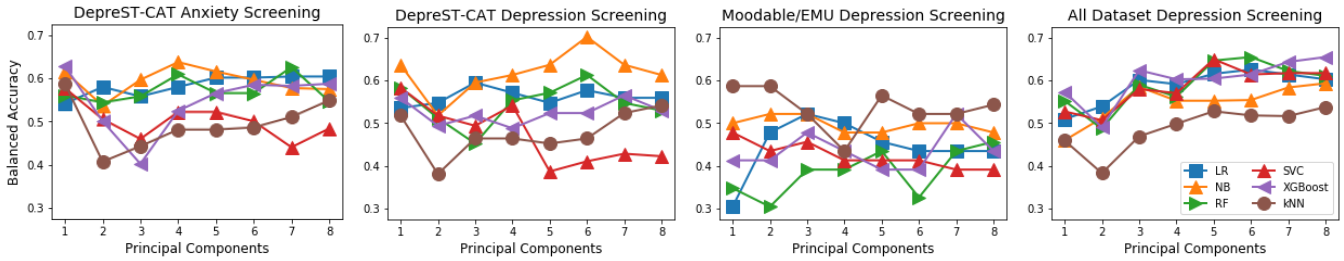

Figure 1: Results of screening for moderate anxiety and moderate depression with text message reply latency features.

regardless of the response time. In practice, text logs would only be used to screen people who text relatively often. As such, we consider participants who have at least 7 replies in the last 14 days.

The reply latencies from each participant form a distribution. We extract the minimum, 10% quantile, 25% quantile, 50% quantile, 75% quantile, 90% quantile, and maximum seconds from the reply latency distributions. To this set of features, we also include the number of contacts and the number of replies. As there is high correlation among the features [21], we perform principal component analysis (PCA) [12] to extract up to eight principal components. The PCA transformation is learned on the training data and applied to the test data. We also upsample the training data to balance classes.

2.2 Datasets Containing Labeled Text Logs

In this research, we use two datasets. Both leverage the Patient Health Questionnaire-9 (PHQ-9) [9] for depression screening labels. This popular nine question survey [13] asks participants to rate the frequency of depression symptoms over the past two weeks on four point Likert scales. One of the datasets also uses the General Anxiety Disorder-7 (GAD-7) [14] which asks seven questions gauging the frequency of anxiety symptoms. The PHQ-9 scores range from 0 to 27 while the GAD-7 scores range from 0 to 21, with the cutoff for moderate depression and anxiety being 10 [9, 14].

The first dataset was collected approximately a year after the start of the COVID-19 pandemic from Prolific [11] crowd-sourced workers, the Call and Text log (CAT) subset of the DepreST dataset [18] contains logs labeled with both screening scores. Of the 49 participants with at least 7 replies, 28 (57.1%) screened positive for moderate depression and 26 (53.1%) screened positive for moderate anxiety. The 49 DepreST-CAT participants collectively sent 2396 replies in the two weeks prior to completing the screening surveys.

The second dataset was collected in 2017-2019 prior to the COVID-19 pandemic from Mechanical Turk [2], the Moodable and EMU datasets [5, 22] contain logs labeled with PHQ-9 scores. We treat them as a single dataset, as is common in prior work [17, 19–21]. Of the 46 participants with at least 7 replies, 23 (50.0%) screened positive for moderate depression. The 46 participants collectively sent 3780 replies in the two weeks prior to completing the PHQ-9.

2.3 Classification Methodology & Evaluation

The main goal of the machine learning models is to screen DepreST-CAT participants for moderate anxiety (GAD-7≥10) and moderate depression (PHQ-9≥10). Additionally, we screen for moderate depression in participants from both individual and aggregated datasets. To do so, we use a representative sample of classification methods [12]: Gaussian Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), Support Vector Classifier (SVC), k-Nearest Neighbor (kNN), and eXtreme Gradient Boosting (XGBoost) [3]. All models were trained using the default parameters [3, 12].

To maximize the amount of training data, we employ a leave-one-out cross-validation evaluation strategy. In this approach, data from all but one participant is used to train the model. The trained model then makes a prediction for the participant in the test set. This process is repeated until there is a prediction for each participant. The number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) predictions are tallied for each model configuration. We consider the best models[1] to be those that maximize balanced accuracy, which is the average of sensitivity and specificity. We also report on the sensitivity, specificity, and F1.

3 RESULTS OF SCREENING MODELS

3.1 Anxiety versus Depression Screening

Using DepreST-CAT logs, we are able to assess the ability of reply latency features to screen for anxiety for the first time. Gaussian Naive Bayes achieved the highest balanced accuracy with four principal components. With TP=23, FP=14, FN=3, and TN=9, this model has a balanced accuracy of 0.64, sensitivity of 0.88, specificity of 0.39, and F1 of 0.73. XGBoost with only the first principal component is comparable with a balanced accuracy of 0.62.

DepreST-CAT also has depression screening labels for the same participants, allowing us to compare depression and anxiety screening capabilities of reply latency. Like for anxiety, Gaussian Naive Bayes achieved the highest balanced accuracy when screening for depression; it required six principal components. With TP=26, FP=11, FN=2, and TN = 10, the best model has a balanced accuracy of 0.70, sensitivity of 0.93, specificity of 0.48, and F1 of 0.80.

Reply latencies proved to be slightly more useful for depression screening than anxiety screening. When screening for moderate anxiety, the F1 of 0.73 with reply latencies is notably superior to the previous F1 of 0.50 with log time series [18]. Likewise, when screening for moderate depression, the F1 of 0.80 with reply latencies is much better than the F1 of 0.64 achieved with log time series [18]. Notably, our analysis only used the subset of participants with text replies. As both the anxiety and depression screening models had high sensitivity and low specificity, latency features could be useful to determine who should receive further mental illness screening.

[1]Code and features will be available at https://github.com/mltlachac/UbiComp2022.

3.2 Before COVID-19 versus After COVID-19

Collected prior to the COVID-19 pandemic, Moodable and EMU average 82.2 replies per participant. In contrast, the recently collected DepreST-CAT averages only 48.9 replies per participant. This may be indicative of changing communication trends due to the COVID-19 pandemic and increase in alternative messaging platforms [24].

Despite the lower average number of messages per participant, the most successful depression screening model with the DepreST-CAT data achieved a balanced accuracy that was 0.21 higher than the most successful model with the Moodable and EMU data. For the before COVID-19 datasets, a kNN model achieved the highest balanced accuracy using the first principal component. With TP=13, FP=9, FN=10, and TN=14, the model has a balanced accuracy of 0.59, sensitivity of 0.57, specificity of 0.61, and F1 of 0.58.

We observed that the data collected after the start of the COVID-19 pandemic was more successful at depression screening than data collected prior. Even with just one principal component, Gaussian Naive Bayes with DepreST-CAT achieves a balanced accuracy of 0.64, sensitivity of 0.89, specificity of 0.38, and F1 of 0.76 with TP=25, FP=13, FN=3, and TN=8. While the DepreST-CAT model has higher sensitivity, the Moodable/EMU model has higher specificity.

3.3 Individual versus Aggregated Datasets

To further demonstrate potential for generalization, we compare the depression screening performance of the aforementioned models with that of models trained with all logs. Using all datasets, an RF model with 6 principal components achieved the highest balanced accuracy of 0.66. However, both RF and SVC models with 5 principal components are almost as successful with a balanced accuracy of 0.65. The RF model has a sensitivity of 0.73 and a specificity of 0.57 whereas the SVC model has a sensitivity of 0.69 and a specificity of 0.62. While the best balanced accuracy of 0.70 was achieved with the DepreST-CAT dataset, the second best balanced accuracy of 0.66 was achieved with all datasets. Thus, despite the differences in the datasets, using an aggregated dataset still yields comparatively impressive depression screening results.

4 DISCUSSION & FUTURE OPPORTUNITIES

In this paper, we explored the feasibility of using reply latency of direct text-based messages to screen for anxiety and depression. While the DepreST-CAT dataset created new modeling opportunities and effectively doubled the number of participants with reply latencies, the lack of participants remains the main limitation for research in this domain. We adopted a leave-one-out cross-validation strategy to combat this limitation but that does not remove the need for data from more participants to train deployable models.

The retrospective nature of the data collection means the data can not be biased by study awareness, but participants regrettably could have deleted texts prior to submitting data. We further assumed that participants submitted data from their personal smartphones. We demonstrated that using aggregated datasets is a valid approach to developing more generalized models for screening for mental illness. Future work could thus compare the mental illness screening ability of reply latencies extracted from retrospective and prospective logs.

While text message latency is only a pertinent screening modality for people who text relatively frequently, latency features could be extracted from any type of direct message. Future research could thus compare the mental illness screening ability of reply latencies from different types of messages, thus further exploring the generalizability of reply latency. Additionally, future research could combine reply latency features with features engineered from other mobile sensors to improve mental illness screening.

ACKNOWLEDGMENTS

We thank E Rundensteiner, M Reisch, K Houskeeper, E Toto, and other Emutivo researchers at WPI for data collection contributions.

REFERENCES

[1] Mehdi Boukhechba, Alexander Daros, Karl Fua et al. 2018. DemonicSalmon: Monitoring mental health and social interactions of college students using smartphones. *Smart Health* 9 (2018), 192–203.
[2] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5.
[3] Tianqi Chen, and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *22nd ACM SIGKDD*. 785–794.
[4] Prerna Chikersal, Afsaneh Doryab et al. 2021. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. *ACM TOCHI* 28, 1 (2021), 1–41.
[5] Ada Dogrucu, Alex Perucic et al. 2020. Moodable: On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data. *Smart Health* 17 (2020), 100–18.
[6] Aron Halfin. 2007. Depression: The Benefits of Early and Appropriate Treatment. *American Journal of Managed Care* 13, 4 (2007).
[7] Ronald Kessler, Wai Chiu, Olga Demler et al. 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry* 62, 6 (2005), 617–27.
[8] Ronald Kessler, Steven Heeringa et al. 2008. Individual and societal effects of mental disorders on earnings in the United States: results from the national comorbidity survey replication. *American J of Psychiatry* 165, 6 (2008), 703–11.
[9] Kurt Kroenke, Robert Spitzer, and Janet Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 16, 9 (2001), 606–13.
[10] Lisa A Martin, Harold W Neighbors, and Derek M Griffith. 2013. The experience of symptoms of depression in men vs women: analysis of the National Comorbidity Survey Replication. *JAMA psychiatry* 70, 10 (2013), 1100–1106.
[11] Stefan Palan, and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.
[12] Fabian Pedregosa, Gaël Varoquaux et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine Learning research* 12 (2011), 2825–30.
[13] Margot Savoy, and David O'Gurek. 2016. Screening your adult patients for depression. *Family practice management* 23, 2 (2016), 16–20.
[14] Robert L Spitzer, Kurt Kroenke et al. 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Int Med* 166, 10 (2006), 1092–97.
[15] Substance Abuse and Mental Health Services Administration. 2021. Key substance use and mental health indicators in the United States: Results from the 2020 National Survey on Drug Use and Health. (2021).
[16] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare et al. 2017. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing* 11, 2 (2017), 200–13.
[17] ML Tlachac, Katherine Dixon-Gordon, and Elke Rundensteiner. 2021. Screening for Suicidal Ideation with Text Messages. In *IEEE EMBS BHI*. 1–4.
[18] ML Tlachac, Ricardo Flores, Miranda Reisch et al. 2022. DepreST-CAT: Retrospective Smartphone Call and Text Logs Collected during the COVID-19 Pandemic to Screen for Mental Illnesses. *ACM IMWUT* 6, 2 (2022), 1–32.
[19] ML Tlachac, Veronica Melican, Miranda Reisch et al. 2021. Mobile depression screening with time series of text logs and call logs. In *IEEE EMBS BHI*. 1–4.
[20] ML Tlachac, and Elke Rundensteiner. 2020. Screening for depression with retrospectively harvested private versus public text. *IEEE J-BHI* 24, 11 (2020), 3326–32.
[21] ML Tlachac, and Elke A Rundensteiner. 2020. Depression screening from text message reply latency. In *42nd IEEE EMBC*. 5490–5493.
[22] ML Tlachac, Ermal Toto, Joshua Lovering et al. 2021. Emu: Early mental health uncovering framework and dataset. In *20th IEEE ICMLA*. 1311–18.
[23] Rui Wang, Fanglin Chen, Zhenyu Chen et al. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *ACM UbiComp*. 3–14.
[24] Britta Wetzel, Rüdiger Pryss, Harald Baumeister et al. 2021. "How Come You Don't Call Me?" Smartphone Communication App Usage as an Indicator of Loneliness and Social Well-Being across the Adult Lifespan during the COVID-19 Pandemic. *Int J of Environmental Research and Public Health* 18, 12 (2021).