Improving Mental Health Screening with Predictive and Generative Modeling of Text Messages

ML Tlachac

A Dissertation
Submitted to the Faculty
of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy
in
Data Science

APPROVED:

Elke Rundensteiner, Ph.D., Advisor, Computer Science & Data Science, WPI

Randy Paffenroth, Ph.D., Committee Member, Mathematical Sciences & Data Science, WPI

Dmitry Korkin, Ph.D., Committee Member, Computer Science & Bioinformatics, WPI

Katherine Dixon-Gordon, Ph.D., External Committee Member, Psychology, UMass-Amherst

SUMMARY

Screening for mental illnesses is vital, but traditional screening questionnaires are susceptible to conscious and unconscious bias. In my dissertation, I explore the mental illness screening capabilities of retrospectively harvested text messages. Leveraging lexical category features derived from the text message content of crowd-sourced participants, I trained traditional machine learning models and evaluated their ability to screen for depression and suicidal ideation. For sent texts, I discovered the most recent weeks of texts were more predictive than greater temporal quantities like the last year of texts. I further constructed lexicons with less formal language to improve the depression screening models. For received texts, I identified the 25 percent most prolific contacts as the subset with the messages most predictive of depression. To mitigate privacy concerns, I also explore depression screening potential of text reply latencies and time series of communications. I then collect a new dataset with a larger quantity of call and text logs labeled with depression and anxiety screening scores. Deep learning was more effective at screening for lower score cutoffs while machine learning was more effective at screening for higher score cutoffs. Lastly, I explore the depression screening potential of generated text content. I identify and adopt nine different conditional approaches for sequence generation. I then conduct a comparative evaluation of their ability to generate text messages from depressed and not depressed participants. The transformer-based classifiers proved better able to screen for depression with texts generated by the unconditioned models than the conditioned models, revealing future research opportunities.

PUBLICATIONS

Publications Featured in this Dissertation

1. **ML Tlachac**, Elke Rundensteiner, “Screening for Depression with Retrospectively Harvested Private versus Public Text”, *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, 2020 [1]

2. **ML Tlachac**, Katherine Dixon-Gordon, Elke Rundensteiner, ”Screening for Suicidal Ideation with Longitudinal Text Messages”, *17th IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp 1-4, 2021 [2]

3. **ML Tlachac**, Avantika Shrestha, Mahum Shah, Benjamin Litterer, and Elke Rundensteiner, “Automated Construction of Lexicons to Improve Depression Screening with Text Messages”, in Submission [3]

4. **ML Tlachac**, Ermal Toto, Elke Rundensteiner, “You’re Making Me Depressed: Leveraging Texts from Contact Subsets to Predict Depression”, *16th IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp 1-4, 2019 [4]

5. **ML Tlachac**, Elke Rundensteiner, “Depression Screening from Text Message Reply Latency”, *42nd International Conference of IEEE Engineering in Medicine and Biology Society (EMBC)*, pp 5490-5493, 2020 [5]

6. **ML Tlachac**, Veronica Melican, Miranda Reisch, Elke Rundensteiner, “Mobile Depression Screening with Contact Timeseries”, *17th IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp 1-4, 2021 [6]

7. **ML Tlachac**, Ricardo Flores, Miranda Reisch, Katie Houskeeper, Elke Rundensteiner, “DepreST-CAT: Leveraging Smartphone Call and Text Logs Collected During the COVID-19 Pandemic to Screen for Mental Illnesses”, ACM Proceedings on Interactive, Mobile, Wearable and Ubiquitous Technologies, Accepted [7]

8. **ML Tlachac**, Walter Gerych, Kratika Agrawal, Benjamin Litterer, Nicholas Jurovich, Saitheeraj Thatigotla, Jidapa Thadajarassiri, Elke Rundensteiner, “Text Generation to Aid Depression Detection: A Comparative Study of Conditional Sequence Generative Adversarial Networks”, in Revision [8]

Related Publications Not Featured in this Dissertation

1. **ML Tlachac**, Adam Sargent, Ermal Toto, Randy Paffenroth, Elke Rundensteiner, "Topological Data Analysis to Engineer Features from Audio Signals for Depression Detection", *19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2020 [9]

2. Ermal Toto, **ML Tlachac**, Francis Lee Stevens, Elke Rundensteiner, "Audio-based Depression Screening using Sliding Window Sub-clip Pooling", *19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2020 [10]

3. Ermal Toto, **ML Tlachac**, Elke Rundensteiner, "AudiBERT: A Deep Transfer Learning Multimodal Screening Framework for Depression Classification", 30th ACM International Conference on Information and Knowledge Management (CIKM) Applied Research Track, 2021 (Best Applied Paper) [11]

4. **ML Tlachac**, Ermal Toto, Joshua Lovering, Rimsha Kayastha, Nina Taurich, Elke Rundensteiner, "EMU: Early Mental Health Uncovering Framework and Dataset", 20th IEEE International Conference on Machine Learning and Applications (ICMLA) Special Session Machine Learning in Health, 2021 [12]

5. Ricardo Flores, **ML Tlachac**, Ermal Toto, Elke Rundensteiner, "Depression Screening Using Deep Learning on Follow-up Questions in Clinical Interviews", 20th IEEE International Conference on Machine Learning and Applications (ICMLA), 2021 [13]

6. Saskia Senn, **ML Tlachac**, Ricardo Flores, Elke Rundensteiner, "Ensembles of BERT for Depression Classification", *44nd International Conference of IEEE Engineering in Medicine and Biology Society (EMBC)*, Accepted [14]

7. **ML Tlachac**, Ricardo Flores, Miranda Reisch, Rimsha Kayastha, Nina Taurich, Veronica Melican, Connor Bruneau, Hunter Caouette, Joshua Lovering, Ermal Toto, Elke Rundensteiner, "StudentSADD: Mobile Depression and Suicidal Ideation Screening of College Students during COVID-19", ACM Proceedings on Interactive, Mobile, Wearable and Ubiquitous Technologies, Accepted [15]

8. **ML Tlachac**, Miranda Reisch, Brittany Lewis, Ricardo Flores, Lane Harrison, and Elke Rundensteiner, "Impact of Stereotype Threat on Mobile Depression Screening", in Revision [16]

9. Ricardo Flores, **ML Tlachac**, Ermal Toto, Elke Rundensteiner, "Transfer Learning for Depression Screening from Follow-up Clinical Interview Questions", in Submission to Deep Learning Applications, vol 4, Springer [17]

10. **ML Tlachac**, Ricardo Flores, Ermal Toto, Elke Rundensteiner, “Early Mental Health Uncovering with Short Scripted and Unscripted Voice Recordings”, in Submission to Deep Learning Applications, vol 4, Springer [18]

11. Miranda Reisch, **ML Tlachac**, Ricardo Flores, Ermal Toto, Elke Rundensteiner, “Mental Health Classification Utilizing Multimodal Deep Learning with Mobile Speech Recordings”, In Preparation [19]

Unrelated Publications to this Dissertation

1. David Bevan, Derek Levin, Peter Nugent, Jay Pantone, Lara Pudwell, Manda Riehl, **ML Tlachac**, “Pattern Avoidance in Forests of Binary Shrubs”, *Discrete Mathematics and Theoretical Computer Science*, vol 18(2), pp 1-22, 2016

2. **ML Tlachac**, Elke Rundensteiner, Kerri Barton, Scott Troppy, Kerri Beaulac, Shira Doron, “Predicting Future Antibiotic Susceptibility using Regression-based Methods on Longitudinal Massachusetts Antibiogram Data”, *11th International Conference on Health Informatics (HealthInf)*, pp 103-114, 2018

3. **ML Tlachac**, Elke Rundensteiner, Kerri Barton, Scott Troppy, Kerri Beaulac, Shira Doron, Jian Zou, “CASSIA: An assistant for identifying clinically and statistically significant decreases in antimicrobial susceptibility,” *15th IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 389-392, 2018

4. **ML Tlachac**, E Rundensteiner, TS Troppy, K Beaulac, S Doron, K Barton, “Predictive Modeling of Emerging Antibiotic Resistance Trends”, *Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science, Springer*, vol 1024, pp 348-366, 2019

5. Susmitha Wunnava, Xian Qin, Tabassum Kakar, **ML Tlachac**, Xiangnan Kong, Elke Rundensteiner, S Sahoo, S De, “Multi-layered Learning for Information Extraction from Adverse Drug Event Narratives”, *Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science, Springer*, vol 1024, pp 421-446, 2019

6. **ML Tlachac**, Elke Rundensteiner, Kerri Barton, T Scott Troppy, Kerri Beaulac, Shira Doron, “Anomalous Antimicrobial Susceptibility Trend Identification”, *42nd International Conference of IEEE Engineering in Medicine and Biology Society (EMBC)*, pp 5880-583, 2020

7. Mallak Alkhathlan, **ML Tlachac**, Lane Harrison, Elke Rundensteiner, “Honestly I Never Really Thought About Adding a Description: Why Highly Engaged Tweets are Inaccessible”, IFIP Conference on Human-Computer Interaction (INTERACT), Springer, 2021

8. Mallak Alkhathlan, **ML Tlachac**, Lane Harrison, Elke Rundensteiner, “Improving Image Accessibility by Combining Haptic and Auditory Feedback”, In Submission

ACKNOWLEDGMENTS

Professor Elke Rundensteiner deserves special acknowledgement for serving as my graduate advisor, funding me as a research assistant, and providing me with new opportunities. Professors Randy Paffenroth and Katherine Dixon-Gordon have been valuable mentors and collaborators. I also thank Professor Dmitry Korkin for offering guidance as a member of my PhD thesis committee. Additionally, I must acknowledge my undergraduate research advisors, Professors Manda Riehl and Peter Hart-Brinson, for supporting me and inspiring me to pursue research.

I thank my PhD student peers at WPI. Ermal Toto was a patient mentor and an invaluable collaborator on the Emutivo project. I am grateful to Sam Ogden and Brittany Lewis for their emotional and technical support. Francisco Castro is an inspiration who challenges me to improve by example. I am glad to have had the opportunity to collaborate with Ricardo Flores, Avantika Shrestha, Walter Gerych, Jidapa Thadajarassiri, Susmitha Wunnava, Tabassum Kakar, and Mallak Alkhathlan. I thank Tom Hartvigsen, Caitlin Kuhlman, Luke Buquicchio, Peter VanNostrand, Eric Vertina, Tonwei Ren, Shiya Cao, Geri Dimas, and Mia Barger for also supporting me.

I appreciate Professor Emmanual Agu, Ada Dogrucu, Alex Perucic, Anabella Isaro, and Damon Ball laying the foundation for the Emutivo project. I had the pleasure of mentoring Adonay Resom, Adonay Resom, Maurice Flannery, Yufei Gao, Yuxin Wu, Adam Sargent, Joseph Caltabiano, Myo Thant, Nicolas Pingal, Yosias Seifu, Yared Taye, Joshua Lovering, Nina Taurich, Rimsha Kayastha, Veronica Melican, Connor Bruneau, Hunter Caouette, Karthika Suresh, Kratika Agrawal, Junying Li, Matthew Dzwil, Katie Houskeeper, Benjamin Litterer, Saitheeraj Thatigotla, Mahum Shah, Nicolas Jurovich, Mairead O'Neill, Jyalu Wu, Madeline Halley, Lillian Garfinkel, Brian Phillips, Dante Amicarella, and Saskia Senn on the Emutivo project. In particular, I acknowledge Miranda Reisch who has shared my enthusiasm for this research during our two year collaboration.

I am forever grateful for the unfaltering support of my parents Jacki and Steve for supporting me during my academic journey. I thank my brother Matthew for challenging me to improve. Lastly, I acknowledge my dog Bumper for always finding ways to make me smile.

FUNDING

This research was supported by the following grants:

- US Department of Education P200A150306: GAANN Fellowships in Computer Science to Support Data-Driven Computing Research
- US Department of Education P200A180088: GAANN Fellowships in Computer Science to Support Data-Driven Computing Research
- NSF 1910880 III Small: Outlier Discovery Paradigm
- AFRI Grant 1023720

My conference registration and travel were supported by the following grants:

- WPI Graduate Student Travel Fund, 2019
- NSF travel grant for IEEE BHI, 2019

Results were obtained in part using a high-performance computing system acquired through NSF MRI grant DMS-1337943 to WPI.

TABLE OF CONTENTS

Publications iv

Acknowledgments viii

Funding ix

List of Tables xvi

List of Figures xviii

List of Acronyms xx

Chapter 1: Introduction and Motivation 1

1.1 Motivating Mental Illness Screening 1

1.1.1 Mental Illness Prevalence 2

1.1.2 Mental Illness Cost 2

1.1.3 Barriers to Diagnosis 3

1.1.4 Impact of Diagnostic Technology 3

1.2 State-of-the-art in Mental Illness Screening Research 4

1.2.1 Screening Modalities 4

1.2.2 Screening Methods 4

1.3 Challenges of Screening with Text Messages 5

1.4 Dissertation Research 6

1.4.1 Mental Illnesses Screening with Text Message Content 7

1.4.2 Mental Illnesses Screening with Text Message Logs 7

1.4.3 Generating Data for Mental Illness Screening 8

1.5 Dissertation Overview 8

Chapter 2: Background & Related Work 10

2.1 Predictive Modeling 10

2.1.1 Feature Selection/Reduction Techniques 10

2.1.2 Evaluation Strategies 11

2.1.3 Data Balancing 11

2.1.4 Traditional Machine Learning Methods 12

2.1.5 Ensemble Machine Learning Methods 14

2.1.6 Transfer Learning Methods 15

2.1.7 Evaluation Metrics for Predictive Models 16

2.1.8 Statistical Analyses. 17

2.2 Generative Modeling 18

2.2.1 Generative Adversarial Newtork (GAN) 18

2.2.2 Evaluation Metrics for Generative Models 19

2.3 Mental Health Screening Surveys 21

2.3.1 Patient Health Questionnaire-9 (PHQ-9) 21

2.3.2 General Anxiety Disorder-7 (GAD-7) 22

2.3.3 Additional Screening Surveys 22

2.4 Related Mental Illness Screening 23

2.4.1 Participant Recruitment 23

2.4.2 Mental Illness labeling Strategies 24

2.4.3 Screening with Text Data 26

2.4.4 Screening with Audio Data 27

2.4.5 Screening with Prospective Smartphone Sensor Data 30

2.5 Datasets Containing Retrospectively Harvested Smartphone Data 34

2.5.1 Mood Assessment Capable (Moodable) Dataset 34

2.5.2 Early Mental Health Uncovering (EMU) Dataset 36

2.5.3 Student Suicidal and Depression Detection (StudentSADD) Dataset 36

2.5.4 Screening Results Without Text Messages 36

2.5.5 My Research in Context 37

Chapter 3: Mental Illness Screening with Text Message Content 38

3.1 Participant Depression Screening with Private Versus Public Text 39

3.1.1 Dataset 40

3.1.2 Methodology 40

3.1.3 Depression Screening Results 43

3.1.4 Feature Importance Results 45

3.1.5 Discussion 47

3.2 Suicidal Ideation with Longitudinal Text Messages 49

3.2.1 Text Message Data 49

3.2.2 Machine Learning Methodology 50

3.2.3 Suicidal Ideation Screening Results 51

3.2.4 Discussion 53

3.3 Constructing Lexicons for Depression Screening 55

3.3.1 Lexicon Construction Methodology 55

3.3.2 Feature Extraction Methodology 58

3.3.3 Feature Selection and Machine Learning Methodology 59

3.3.4 Important Features for Depression Screening 64

3.3.5 Discussion 64

3.4 Depression Screening with Texts from Contact Subsets 65

3.4.1 Data 65

3.4.2 Contact Subset Creation 66

3.4.3 Text & Emotion Feature Engineering 67

3.4.4 Machine Learning Methodology 67

3.4.5 Results 68

3.4.6 Discussion of Implications & Considerations 69

3.5 Outlook 70

Chapter 4: Depression Screening from Text Message Logs 72

4.1 Introduction 73

4.2 Reply Latencies for Depression Screening 73

4.2.1 Data 73

4.2.2 Principal Component Analysis for Feature Reduction 75

4.2.3 Machine Learning Methodology for Depression Classification 75

4.2.4 Screening Results 76

4.2.5 Discussion 78

4.3 Communication Time Series for Depression Screening 79

4.3.1 Text and Call Log Data 79

4.3.2 Time Series Construction 80

4.3.3 Machine Learning Methodology 80

4.3.4 Results 81

4.3.5 Discussion 83

4.4 Collecting Text and Call Logs for Mental Illness Screening 84

4.4.1 DepreST-CAT Collection Methodology 85

4.4.2 The Data Collection Application 85

4.4.3 Participants Recruitment, Eligibility, and Compensation 86

4.4.4 Call and Text Log Screening Methodology 87

4.4.5 Collection Results: DepreST-CAT Participants and Smartphone Logs . . . 90

4.4.6 Machine Learning and Deep Learning Screening Results 92

4.4.7 Discussion 99

4.5 Outlook 101

Chapter 5: Depression Screening with Generative Modeling 104

5.1 Conditional Generative Adversarial Networks 105

5.1.1 Existing Conditional Generative Models 106

5.1.2 Methodological Overview 107

5.1.3 Sequence GANs 107

5.1.4 Design Dimension: Weighting Strategies 108

5.1.5 Design Dimension: Feedback Mechanisms 110

5.1.6 Datasets 113

5.1.7 Experimental Setup 114

5.1.8 Machine Experimental Evaluation 114

5.1.9 Human Experimental Evaluation 115

5.1.10 Experimental Results 116

5.1.11 Discussion 119

5.2 Outlook 120

Chapter 6: Concluding Thoughts 123

6.1 Contributions & Impact 124

6.1.1 Small Datasets 124

6.1.2 Text Message Privacy 125

6.1.3 Predictive Data Identification 126

6.1.4 Text Message Linguistics 126

6.2 Implementation of Mobile Screening 126

6.3 Current and Future Research Directions 127

6.3.1 Current Screening with Text Content Research 127

6.3.2 Current Screening with Text Logs Research 128

6.3.3 Current Generated Text Research 128

6.3.4 Available Future Research Directions 128

References 130

LIST OF TABLES

2.1 PHQ-9 score interpretation [92, 94]. 22

3.1 Number of participants and average messages M per participant ± standard deviation (std) for all datasets. 41

3.2 Models with the highest average F1 scores for each tweet dataset leveraged chi-squared selected features. 44

3.3 Models with the highest average F1 scores for each text dataset leveraged chi-squared selected features. 44

3.4 The number of text messages sent by participants in the interval and cumulative weeks prior to reporting suicidal ideation with the ninth item of the PHQ-9. 50

3.5 The suicidal ideation screening ability of the model configurations with highest average metrics for each subset of texts. For each metric, the average and standard deviation of the 100 scores for the model configuration are displayed. 52

3.6 The 98 words in the Empath's pre-existing *vacation* category [129]. 56

3.7 The average ± standard deviation of the metrics for the models configurations with the highest average F1 scores for each lexicon. The p-values (P) are derived from the t-tests that compare the F1 scores from each lexicon with the F1 scores from the bag-of-words (B) and default Empath lexicon (D). The models that are statistically significantly better than the baselines based on these t-tests are marked with asterisks. The lexical categories were generated with seed words that were the 1 to 5 closest (c), first (f), and random (r) words to the names of the categories. The six best models are bolded. 60

3.8 Comparison of the average F1 scores for each method trained on features from the most successful lexicons in Table Table 3.7. The p-values (P) are derived from the t-tests that compare the F1 scores from each of the best lexicons with the F1 scores from the bag-of-words (B) and default Empath lexicon (D). 63

4.1 Best model configuration for each method with PCA. 76

4.2 Number of participants with each type of data. 80

4.3 Comparison of model configurations with the highest average $F1$ scores for text logs. 83

4.4 Comparison of model configurations with the highest average $F1$ scores for call logs. 83

4.5 The number of DepreST-CAT participants (N) who screen positive for depression and anxiety at PHQ-9 and GAD-7 cutoffs in the mild (5) to severe (15) score range. 92

4.6 Comparison of the average ± standard deviation of the F1 scores of the **traditional machine learning** models with different screening cutoffs. These models used the first principal component constructed from the features of two weeks of data. . . . 96

4.7 Comparison of the average ± standard deviation of the F1 scores of the **ensemble machine learning** models with different screening cutoffs. These models used the first principal component constructed from the features of two weeks of data. . . . 97

4.8 Comparison of the average ± standard deviation of the F1 scores of the **deep learning** models with different screening cutoffs and two weeks of data. 98

5.1 **Clinical Interviews:** Average ± standard deviation of evaluation metrics for the five conditional models with the same weighting strategy and feedback mechanism. For the unconditioned models, these values were calculated for five models trained on depressed sentences and five models trained on not depressed sentences. The machine evaluation accuracy was calculated with BERT classifiers. The accuracy of both human evaluation tasks are also displayed. 117

5.2 **Text Messages:** Average ± standard deviation of evaluation metrics for the five conditional models with the same weighting strategy and feedback mechanism. For the unconditioned models, these values were calculated for five models trained on depressed texts and five models trained on not depressed texts. The machine evaluation accuracy was calculated with BERT classifiers. The accuracy of both human evaluation tasks are also displayed. 117

5.3 **Movie Reviews:** Average ± standard deviation of evaluation metrics for the five conditional models with the same weighting strategy and feedback mechanisms. For the unconditioned models, these values were calculated for five models trained on positive reviews and five models trained on negative reviews. The machine evaluation accuracy was calculated with BERT classifiers. The accuracy of both human evaluation tasks are also displayed. 118

LIST OF FIGURES

2.1 Machine learning results with features from Moodable text messages [41]. 35

3.1 Comparison of texts and tweets for Moodable participant 7237 with PHQ-9 = 7. . . 39

3.2 PHQ-9 scores of participants with messages within last 364 days. 40

3.3 Comparing highest average F1 scores for best methods on text datasets. 45

3.4 Boxplots of 100 LR models on 14 days of text messages. 46

3.5 Use frequency of top 10 chi-squared features from 14 days of texts. 47

3.6 Comparison of the suicidal ideation screening ability for the model configurations with the highest average AUC, F1, and accuracy. 51

3.7 Suicidal ideation screening ability for the SVC models using the one prior week of texts with different quantities of chi-squared selected features. 52

3.8 The 17 top chi-squared selected features for the prior week of texts; 15 are word category frequencies. For each feature, the normalized values for all 57 participants are displayed. 54

3.9 The number of words in the categories for each lexicon. 57

3.10 The results of the model configurations with the the highest average F1 scores for each lexicon. Bag-of-words (BOW) and Reddit 5c used 3 chi-squared selected features while the other lexicons used only 1 chi-squared selected feature. The lexical categories were generated with seed words that were the 1 to 5 closest (c), first (f), and random (r) words to the category names. 61

3.11 F1 scores of 100 Gaussian Naive Bayes models for every contact subset. 69

4.1 Text response latency within the last 14 days for each participant. 74

4.2 Pearson correlation between features. 75

4.3 100 trials of XGBoost models with depth 2. 77

4.4 Coefficients for each principal component (PC). 77

4.5 Top two principal components. 78

4.6 PHQ-9 distributions of participants who submitted text and call logs. 79

4.7 Time series representing the count of texts for an example participant with PHQ-9= 7. The different plots are for different aggregation intervals. For the 4 hour aggregation interval, there are 6 samples per day whereas for the 24 hour aggregation interval, there is only 1 sample per day. 81

4.8 Comparison of model configurations with the highest average $F1$ scores. 82

4.9 An example of the feature engineering process for two weeks of call and text logs for participant Q6N7PT68V with a PHQ-9 score of 12 and a GAD-7 score of 8. We repeat this process for the logs from 4, 8, and 16 weeks prior to data submission. These features are then used in the traditional machine learning models. Not all extracted features are relevant for log time series, such as zero crossing rate (ZCR), so they do not contribute to the principal components. 88

4.10 The PHQ-9 and GAD-7 scores of the 369 DepreST-CAT participants. 91

4.11 The distribution of depression (PHQ-9) and anxiety (GAD-7) screening scores for the 369 DepreST-CAT participants. 92

4.12 The average F1 of the traditional machine learning models screening for depression and anxiety at different screening score cutoffs. The models displayed use two weeks of data and the first principal component. 93

4.13 The average F1 of the deep learning models screening for depression and anxiety at different screening score cutoffs. The models displayed use two weeks of data. . 94

5.1 The feedback mechanisms include: (a) a discriminator that performs a single task, (b) a discriminator that performs two tasks, and (c) a discriminator that performs a single task and a classifier that performs a single task. The realism prediction in (b) and (c) only consider the text while the consistency prediction in (a) considers the realness of the concatenated text and label. 111

LIST OF ACRONYMS

AdaBoost Adaptive Boosting

BERT Bidirectional Encoder Representations from Transformers

BLEU bilingual evaluation understudy

CES-D Center for Epidemiological Studies Depression Scale

CNN Convolutional neural network

cSeqGAN Conditional Sequence Generative Adversarial Network

DAIC-WOZ Distress Analysis Interview Corpus Wizard-of-Oz

DepreST-CAT Depression Stereotype Threat Call and Text logs

DTW Dynamic Time Warping

EMA ecological momentary assessment

EMU Early Mental Health Uncovering

EVC Ensemble Voting Classifier

FN false negative

FP false positive

GAD-7 General Anxiety Disorder-7

GAN Generative Adversarial Network

GRU Gated Recurrent Unit

kNN k-Nearest Neighbors

LIWC Linguistic Inquiry and Word Count

LR Logistic Regression

LSTM Long Short Term Memory

MODMA Multi-modal Open Dataset for Mental-disorder Analysis

Moodable Mood Assessment Capable

MTurk Amazon Mechanical Turk

NB Gaussian Naive Bayes

NER named entity recognition

NLL Negative log likelihood

PHQ-9 Patient Health Questionnaire-9

POS part of speech

RF Random Forest

RNN Recurrent neural network

SeqGAN Sequence Generative Adversarial Network

StudentSADD Student Suicidal Ideation and Depression Detection

SVC Support Vector Classifier

SVM Support Vector Machine

TN true negative

TP true positive

TSFEL Time Series Feature Extraction Library

VAE Variational AutoEncoder

XGBoost Extreme Gradient Boosting

CHAPTER 1
INTRODUCTION AND MOTIVATION

Advances in technology are revolutionizing the approach to healthcare problems. Machine learning algorithms are being used with increasing frequency for image classification tasks within healthcare given their ability to perform as well or better than clinicians [20]. Within the field of psychiatry, machine learning algorithms can provide recommendations that guide patient diagnosis, prognosis, and treatment. As such, incorporating machine learning into psychiatric decision support systems can be very beneficial for patients and healthcare providers.

Accurate prognosis and effective treatment require correct diagnoses. While research is being conducted in all three of these crucial areas, applying machine learning models to aid psychiatric diagnostics will likely also benefit prognosis and treatment research. Machine learning has been applied to neuroimaging data to make diagnostic determinations for many mental illnesses [20]. Such research is notably useful for differential diagnoses; pattern detection on magnetic resonance imaging reduced bipolar depression being misdiagnosed as unipolar depression from the standard 75% to only 31% [21, 20]. While neuroimaging may be useful for diagnostic purposes, requiring it would make psychiatric diagnoses inaccessible to many people. As such, researchers in the last decade have begun to explore the diagnostic ability of less burdensome modalities such as videos [22], voice recordings [23], social media posts [24, 25], and smartphone sensor data [26].

1.1 Motivating Mental Illness Screening

Machine learning can be leveraged to diagnose many medical conditions, as evidenced by the research conducted on the Mimic-III dataset [27] which consists of many medical tests labeled with multiple diagnoses. However, psychiatry is an especially important medical field for machine learning aided diagnoses due to the prevalence, cost, and diagnostic barriers of mental illness.

1.1.1 Mental Illness Prevalence

A national survey of 67,625 individuals from 2019 shows these illnesses are experienced by over 20 percent of U.S. adults annually [28]. Specifically, the annual prevalence is 19.1% for anxiety, 7.8% for depression, 3.6% for post-traumatic stress disorder, 2.8% for bipolar disorder, 1.4% for borderline personality disorder, and 1.2% for obsessive compulsive disorder. While this is the most comprehensive mental illness survey, the results should be considered lower thresholds for annual mental illness prevalence in the U.S. given the data collection process.

The national survey found minority groups are disproportionately impacted by mental illnesses [28], yet individuals in these groups are least likely to trust researchers [29]. Further, participation barriers in mental illness studies include concerns about confidentiality and stigma [29], which may have contributed to the overall response rate of 45.8% [28]. In addition to preventing participation, these barriers may also have led participants to downplay symptom severity. The national survey [28] also excluded groups of individuals likely to have high rates of mental illness: homeless, military, and institutionalized individuals.

1.1.2 Mental Illness Cost

Mental illnesses are very costly medically, socially, and economically. In 2010, mental illnesses globally accounted for 823 billion US dollars in direct costs and 1671 billion US dollars in indirect costs [30]. Estimates indicate the cost of mental illnesses on the global economy will exceed 6 trillion US dollars in 2030. Depression, low back pain, and headache disorders were the three leading causes of disability in 2017 [31].

Mental illnesses increase risk of developing physical diseases [32]. For example, depression increases the risk of developing cardiovascular and metabolic diseases by 40% [32]. Additionally, psychological autopsies reveal 90% of people who died by suicide experienced mental illness symptoms [33]. Suicide was the tenth leading cause of death in the US in 2018 [34]. For the population in the age range of 10 to 34, suicide was the second leading cause of death [34].

Mental illnesses are a barrier to social equality. As mentioned, minority groups are disproportionately impacted by mental illnesses [28]. Further, people with mental illnesses tend to struggle financially as mental illness symptoms can interfere with schooling and employment. As a result, they are least likely to be able to afford the direct and indirect costs associated with obtaining mental illness diagnoses and treatment.

1.1.3 Barriers to Diagnosis

Mental illnesses are currently diagnosed through a series of interviews with a trained mental health professional. Currently, screening surveys are used to determine if an individual needs to be evaluated by a trained mental health professional. Unfortunately, these screening surveys are often viewed as cumbersome and intrusive [35]. Additionally, they require honest self-reflection and are therefore susceptible to both conscious and unconscious bias.

On average, it takes 11 years from symptom onset to receive treatment [36]. Further, many people with mental illnesses remain undiagnosed. It is estimated that 25 percent of individuals with depression never receive a formal diagnosis [37]. In addition to financial barriers, diagnostic barriers include fear of stigma, lack of symptom recognition, and inability to access medical resources [37]. Mental illness symptoms may also be barriers to obtaining a diagnosis. For example, people with depression are less likely to seek help and more likely to delay seeking help [38].

1.1.4 Impact of Diagnostic Technology

Widely deployed passive mental health screening technologies could have a huge impact on the healthcare system. If such technology is adopted, it would reduce the time between symptom onset and treatment. Early diagnosis is important to achieve positive health outcomes [39] which is why universal screening is recommended for adults by the US Preventative Services Task Force [40]. Early diagnostics are required for timely treatment, which would reduce the likelihood of developing physical diseases and experiencing financial troubles related to mental illnesses. Mobile screening technologies can overcome the aforementioned barriers posed by traditional screening surveys to make screening more universally accessible.

1.2 State-of-the-art in Mental Illness Screening Research

1.2.1 Screening Modalities

There are a number of datasets collected and released to the research community promote the development of mental illness screening technologies. These datasets include video recordings [22], voice recordings [41, 42], and smartphone sensor data [43, 26, 44]. Transcripts can be derived from the video and voice recordings with varied verbal content. Further, there are also private datasets which cover many different modalities including voice [23], transcribed interviews [45, 46], essays [47], social media posts [24, 25], smartphone sensor data [26, 48], wearable sensors [49], environmental sensors [50], and environmental audio [51]. The mental illness labels for these datasets originate from clinical diagnosis, screening surveys, and self-declaration.

Notably missing in these analysis is extracting diverse feature from text logs [26, 41, 44, 52] and text message content [41] to screen for mental illnesses. While analysis on public social media posts is common [24, 25] to screen for depression, the private nature of text messages should make texts more predictive of mental illnesses. This is especially true as people often cultivate an internet persona and thus social media posts may only be reflective of that persona. Further, texting popularity [53] makes text logs and content promising modalities for universal passive screening.

1.2.2 Screening Methods

Correlation is a common analysis performed on these datasets to determine the relationship between features and mental illness labels [46, 51, 26, 44]. Traditional machine learning models that predict the mental illness labels are also common. These models include support vector classifiers [54, 55, 56, 57, 49, 48, 41], regressions [58, 55, 49], random forests [59, 57, 49, 41], Naive Bayes classifiers [55], and k-nearest neighbors [41]. Recently, deep learning techniques have started to be applied to screen for mental illnesses [60, 61, 62]. More novel methods have been applied to the related tasks of affective computing and sentiment analysis [63].

Traditionally, the field of psychopathology preferred traditional machine learning over deep learning for three reasons [20]: the applicability to smaller datasets, the inherent privacy resulting from storing the data as features, and model explainability. However, recent advances in deep learning have resulted in models that can handle smaller datasets and offer explainable results. For instance, transfer learning [64] leverages large quantities of unlabeled data so only a small quantity of labeled data is required to train an effective predictive model. Research has also been conducted comparing what humans and neural networks focus on when classifying text [65]. As such, it is worth exploring the applicability of deep learning models to the task of mental illness screening.

1.3 Challenges of Screening with Text Messages

There are a number of challenges in developing universal passive screening technologies. In order to build a model successful at predicting mental illnesses, it is important to overcome the issues arising from small datasets in this domain and identify predictive data. Additionally, any modality selected to use for passive screening will come with unique challenges. Text messages specific challenges involve privacy concerns and linguistic qualities.

Small dataset challenges. The small quantities of available data with mental health labels makes it challenging to construct reliable models that can achieve high screening success [42, 26, 66, 48, 67]. The small quantities of available data originate from two sources. First, it is challenging to recruit participants for a longitudinal study, especially ones that requires sharing private data. Thus the existing datasets are small in size [66, 48, 67] and may not be representative. Many participants also leave such studies before completion [51]. Second, the available datasets are not compatible with each other given the differences in collected modalities and mental illness labels. Further, sharing data across research teams to increase data quantity is hampered by privacy concerns.

Text message privacy challenges. Text messages in particular face unique privacy challenges. In a willingness to share survey completed by 202 crowd-sourced participants [41], 41% indicated willingness to share text messages. Of the 11 proposed modalities, the only modality fewer participants indicated willingness to share was browser history. For comparison purposes, 65% of participants were willing to share voice samples, making it the modality participants were most willing to share. We suspect the percents would vary based on the population surveyed, especially among minority groups [29] and students [48]. Crowd-sourced participants may also delete sensitive messages from their phones before sharing data [41], reducing the predictive value.

Predictive data identification challenges. A challenge in building effective screening models is identifying modalities that carry strong mental illness signals. Further, it is important to identify what subset of the data is most useful for screening purposes. This is particularly challenging for messaging data which has temporal, directional, and conversational aspects. Additionally, as the text messages are not in response to a particular prompts, the number of messages to include in models must be considered. Further, the feature engineering strategies must produce features that capture the mental illness signal from the texts.

Text message linguistic challenges. While we hypothesize that text messages will be a very valuable screening modality, they do have some unique linguistic challenges. The informal language often used in text messages and the short nature of text messages poses issues for existing text-based systems which expect longer passages of formal text [68]. Further, many existing natural language processing tools are trained on third person narratives [69, 70] instead of first person narratives, which further negatively impacts their performance text messages.

1.4 Dissertation Research

In my research, I explore the mental illness screening ability of text messages. This is an extremely underutilized modality in the domain of passive mental illness screening. Prior to my research, only one dataset existed that contained text message content with a mental health label [41]. While the initial depression screening results were not promising [41], my research has since proved that the text messages in this dataset carry strong mental illness signals [1]. My dissertation research involves screening for depression and suicidal ideation with text message content, screening for depression and anxiety with text logs, and generating text to screen for depression.

1.4.1 Mental Illnesses Screening with Text Message Content

While it is common to screen for depression with tweet content [25], it is rare to screen for depression with text message content [41]. Thus, we compare the depression screening capabilities of machine learning models that use features derived from tweets and text messages [1]. This involved comprehensive feature engineering as well as experimentation to determine the best longitudinal quantity of messages. As such, this paper tackles the predictive data identification challenge and text message linguistic challenges. Overall, this research motivates the use of text messages as a mental illness screening modality.

We replicated the feature engineering process on text messages sent within different time periods prior to reporting suicidal ideation [2]. We then compare the suicidal ideation screening ability of these different sets of texts, addressing the predictive data identification challenge. We further tackle the text message linguistic challenges by constructing alternative lexicons comprised of less formal terms [3]. We then compare the depression screening capabilities of features from different lexicons. The alternative lexicons can be used by other research with informal text data. Lastly, we extract features from the messages sent by different subsets of contacts for depression screening [4], which also addresses the predictive data identification challenge. This research is novel as it suggests received text content as well as sent text content can be useful in screening models.

1.4.2 Mental Illnesses Screening with Text Message Logs

Given the text message privacy challenges associated with text content, I further explore the predictive ability of text message logs without content. We extract four novel sets of features from reply latencies [5], average number of texts/calls, number of unique contacts, and average length of texts/calls [6]. We construct time series by calculating the latter three sets over different time aggregation intervals for incoming, outgoing, and all messages. We then use these feature sets in machine learning models to screen for depression. Thus, this research addresses the predictive data identification challenges as well as the privacy challenges.

To additionally address the small dataset challenge, we design and deploy a mobile collection app to amass a larger set of call and text logs with depression and anxiety screening scores [7]. We then construct machine learning and deep learning models with different length time series constructed from these logs to screen for depression and anxiety at different score cutoffs. Thus, we identify which logs are most useful for screening which further tackles the predictive data identification and privacy challenges. This dataset is a valuable resource to the research community.

1.4.3 Generating Data for Mental Illness Screening

We also addressed the small dataset challenge by generating data to increase data quantity. Due to the size of available datasets, we generate text messages and transcripts from depression and not depressed participants. We use an adversarial approach to generation which requires a generator a discriminator. We construct a family of nine conditional sequence generation methods by combining different generators and discriminators. I use transformer-based classifiers to compare the screening ability of the text from the unconditioned and conditioned generative models.

1.5 Dissertation Overview

Chapter 2: Mental Illness Screening with Text Message Content

1. Predictive modeling including feature selection, data balancing, methods, evaluation metrics.

2. Generative modeling including methods and evaluation.

3. Mental health screening questionnaires for depression, suicidal ideation, and anxiety.

4. Related mental illness screening research including participant recruitment strategies, data labeling strategies, and screening with a variety of data modalities.

Chapter 3: Mental Illness Screening with Text Message Content

1. I conduct comprehensive feature engineering to compare the depression screening abilities of tweets and text messages [1]. I also identify the temporal quantity most useful for screening.

2. I assess the suicidal ideation screening capabilities of text messages sent within different time periods prior to reporting suicidal ideation [2].

3. I construct alternative lexicons containing less formal language [3]. These lexicons are used to engineer text features which are used in depression screening models.

4. I use messages sent by different subsets of contacts to screen for depression [4].

Chapter 4: Mental Illness Screening with Text Message Logs

1. I extract text message reply latencies to screen for depression [5].

2. I construct time series from text and call logs [6]. I then use features extracted from these logs in machine learning models to screen for depression.

3. I collect a larger dataset of smartphone logs [7]. In addition to machine learning models, I use deep learning models with the log time series to screen for depression and anxiety.

Chapter 5: Generating Text for Mental Illness Screening

1. I identify and adopt nine conditional adversarial models to generate text [8]. I then use transformer-based classifiers to compare their ability to generate text with depression labels.

Chapter 6: Concluding Thoughts

1. Research overview for text content, text logs, and text generation.

2. Contributions to the small dataset, privacy, data identification, and linguistic challenges.

3. Implementation within and outside of clinical settings.

4. Future research with text content, text logs, and text generation.

CHAPTER 2

BACKGROUND & RELATED WORK

2.1 Predictive Modeling

2.1.1 Feature Selection/Reduction Techniques

There are many feature selection and feature reduction techniques. These techniques are important to reduce complexity and overfitting which can increase the testing performance of the models. While some machine learning methods incorporate feature selection, we perform feature selection before utilizing the machine learning methods so each method receives the same input features. In addition to the machine learning benefits, the selected features can also offer important domain insight. The features are normalized prior to feature selection/reduction. We use two feature selection/reduction techniques, both of which are implemented with Scikit-learn [71].

Principal Component Analysis (PCA). A popular feature reduction technique, PCA calculates successive combinations, known as principal components, from the original features such that each principal component explains the maximum amount of variance not already explained [72]. Thus, each subsequent principal component explains less variance than the prior principal component. PCA is notably not influenced by the target variable. While PCA typically calculates linear combinations of the original features, PCA can also be used for non-linear dimensional reduction [73]. We use kernel PCA with a Gaussian kernel, which we denote as kPCA, in some experiments. Note, the abbreviation PCA shall forthwith be used to denote PCA with linear combinations.

Chi-Squared Feature Selection. The chi-squared statistic measures the dependence between a feature and the target variable [74]. Thus, the features selected by the chi-squared feature selection technique are those with the highest chi-squared statistics.

2.1.2 Evaluation Strategies

Cross-Validation. For cross-validation, the data is split into a user-specified number of folds. One fold is designated as the testing data and the rest of the data is used to train the model. This is repeated until each fold has been used as the testing data. Thus, if there are f folds, f models are trained. The metrics for all f models are averaged to evaluate the performance of that method.

Leave-one-out. The leave-one-out evaluation strategy is a form of cross-validation where the training data consists of all but one data instance. Thus, if there are n data instances, the training set consists of $n - 1$ data instances. Each of the n trained models make a single prediction for a total of n predictions. This is strategy is very computationally expensive on larger datasets, but is preferred for smaller datasets as the size of the training set is maximized.

Train/Test. For this evaluation strategy, the data is randomly split into one training and one testing set. The testing set is typically about a fourth to a third of the overall dataset. In larger datasets, a single train/test split is typically sufficient. For smaller datasets, often multiple random train/test splits are used to ensure model robustness, which is considered as leave-group-out cross-validation with replacement. This differs from traditional cross-validation since the assignment of a data instance in a prior split has no influence on the assignment of the data instance in a future split. Sometimes a single train/test split may be used for smaller datasets when leveraging computationally expensive methods or for comparison purposes on benchmark datasets [75].

2.1.3 Data Balancing

For classification tasks, it is important to train models on balanced data to avoid biasing models towards predicting the majority class. We thus upsample or downsample the training set.

Downsampling. In the training data, random instances of the majority class(es) are sampled until all classes are the same size, which is the size of the smallest class. This sampling method results in a smaller dataset, which can become problematic when there are minimal instances of the minority class. Additionally, as the majority class(es) are sampled, not all instances are represented.

Upsampling. In the training data, additional instances of the minority class(es) are added until all classes are the same size. This is most commonly achieved by duplicating instances of the minority class(es), which is our approach. Alternatively, data generation techniques can be applied to create fake instances of the minority class(es). Upsampling is advantageous for smaller datasets as the data size is increased. However, the duplication or generation of minority class instances can result in overfitting when training machine learning models.

2.1.4 Traditional Machine Learning Methods

Traditional machine learning methods tend to be preferred in healthcare [76] due to their interpretability, anonymization inherent from storing the input as features, and applicability to datasets with minimal participants. This representative selection of supervised machine learning models includes parametric methods that assume the data follows a specific distribution and non-parametric methods that do not make any assumptions regarding the distribution of the data. The models are all implemented with Scikit-learn [71].

k-Nearest Neighbors (kNN). This method identifies the k nearest neighbors in the feature space and classifies the new data instance as the most represented class among the k nearest neighbors. The principal is that similar data should be of a similar class. For binary classification, k is typically an odd number to avoid ties. A benefit of kNN is that it is non-parametric. The main downside of kNN is that it is sensitive to outliers and mislabeled data. Additionally, the kNN algorithm does not scale well to larger datasets as the distance must be calculated between the new data instance and all other points to identify the k nearest neighbors.

Gaussian Naive Bayes (NB). Gaussian Naive Bayes belongs to a family of Naive Bayes probabilistic classifiers. The principal of Naive Bayes classifiers is that the probability of a class c given feature vector X can be calculated with Bayes' Theorem in Equation 2.1. We include Naive Bayes as it is known to be good at text classification. We select the simplest Naive Bayes algorithm which assumes the data follows a Gaussian distribution.

$$P(c|X) = \frac{P(X|c)P(c)}{P(X)} \tag{2.1}$$

Logistic Regression (LR). This method builds a model by analyzing the statistical impact of each feature on each class. A new data instance is assigned a class based on which class to which it has the highest probability of belonging. LR assumes the data is linearly divisible and struggles to capture more complex relationships.

Support Vector Classifier (SVC). The support vector machines (SVM) algorithm identifies hyperplanes that best divides the data into classes. Before calculating the hyperplanes, the data can be mapped into different kernels which allows this method to make different assumptions regarding the distribution of the data. Gaussian and linear kernels are most popular, but Scikit-learn [71] also supports Sigmoid and polynomial kernels. A new data instance is classified by the SVC based on where it falls in relation to the decision boundaries. This makes SVC robust to outliers and over-fitting. However, similar to kNN, this method does not scale well to larger datasets.

Decision Trees. We do not use decision trees in our experiments as they are relatively inaccurate and prone to over-fitting. They are also very sensitive to outliers and mislabled data, especially for smaller datasets. Rather, we use ensemble methods that leverage decision trees but are more robust. As decision trees are non-parametric, the ensemble methods leveraging them are also non-parametric. The decision tree algorithm selects the most important feature based on an impurity metric and then creates a rule to divide the data based on this feature. This is iteratively repeated for the divided groups of data reach a certain purity.

2.1.5 Ensemble Machine Learning Methods

These ensemble methods have the tendency to outperform individual traditional machine learning models. However, this can be task dependent. As these ensemble methods are often computationally expensive to train, we tend to not include these ensemble methods for final experiments if they do not perform well in initial exploration. The exception to this is random forests, as we include this robust method in replacement of decision trees to represent tree-based algorithms. Except for Extreme Gradient Boosting (XGBoost) [77], the models are implemented with Scikit-learn [71].

Random Forest (RF). This ensemble method trains many decision trees independently on subsets of the training data created by randomly sampling the training data with replacement. Each decision tree then output a classification for a new data instance. The classification decision of the random forest is the class that was predicted by the majority of the individual decision trees. This makes random forests very robust against over-fitting.

Ensemble Voting Classifier (EVC). This classifier intakes user specified machine learning methods. A model is trained independently for each of these methods. Each of these models predicts the class of a new data instance. The classification decision of the voting classifier is the class that was predicted by the majority of the models.

Adaptive Boosting (AdaBoost). This method is compatible with many classifiers. It starts by training a user specified classifier with the original training data. The instances in the training data that were incorrectly classified by the trained classifier are identified. This information helps inform the training of the next classifier. By sequentially training classifiers, it is able to adjust for prior mistakes. The vote of each classifier is weighted by the accuracy of that classifier on the training data. This sequential training makes AdaBoost robust to outliers, though the process tends to result in overfitting. The default classifiers are very shallow decision trees and thus we use AdaBoost with these decision trees.

Extreme Gradient Boosting (XGBoost). This newer scalable tree boosting system is very popular [77]. To avoid overfitting like other gradient boosting systems, XGBoost applies regularization which penalizes more complex models. There are also other beneficial modifications to the XGBoost algorithm [78], such as optimization that allows it to train more quickly than other gradient boosting systems. As such, it often trains quicker and yields better results than the other tree-based ensemble methods we include in our experiments.

2.1.6 Transfer Learning Methods

While there are many deep learning methods, most require large quantities of training data. Transfer learning models are able to overcome this obstacle for classification tasks by training on unlabeled data, thus requiring only a few training epochs with labeled data to produce a tailored classification model. While traditional machine learning models have been so far favored in the field of psychopathology [76], transfer learning promises to achieve unprecedented results on diagnostic tasks and therefore worth exploring.

Bidirectional Encoder Representations from Transformers (BERT). BERT is a deep transfer learning model designed to generate text embeddings [70]. Specifically, it pretrains representations from unlabeled text data. Then the embedding model can be fine-tuned on labeled data with just a couple epochs to generate embeddings that can be used in prediction models. The original BERT model [70] was trained on two corpuses of documents: BooksCorpus which contains 800 million words and English Wikipedia which contains 2.5 billion words. Pretraining on unlabeled text corpuses allow for even smaller datasets to excel at natural language processing tasks. There are now variants of BERT that employ different pretraining strategies or datasets. Most relevant are RoBERTa (Robustly Optimized BERT) [79] which has demonstrated success at mental illness classification [80] and BERTweet [81] which was pretrained on a large corpus of English tweets. BERT models are implemented with Hugging Face [82].

2.1.7 Evaluation Metrics for Predictive Models

We evaluate the machine learning models by considering a subset of $F1$, precision, recall, specificity, area under the ROC Curve (AUC), and accuracy. These evaluation metrics consider the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).

The $F1$ score, Equation 2.2, is the balance between precision and recall. Precision, Equation 2.3, is the portion of positive predictions that are positive instances. Recall, Equation 2.4, is the portion of positive instances that are predicted as positive. Given the focus on tp, $F1$ is very useful for classification within the medical domains. For most of the papers that comprise this dissertation, I consider the best method configuration to be the one that maximizes the $F1$ score.

$$F1 = \frac{2(precision)(recall)}{precision + recall} \tag{2.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

Recall (also known as sensitivity) and specificity are also often used in medical domains. Specificity, Equation 2.5, is the portion of negative predictions that are negative instances. Recall, the true positive rate, needs to be maximized to ensure all neurodivergent individuals are identified as neurodivergent. Specificity, also known as the true negative rate, needs to be maximized to ensure neurotypical individuals are not identified as neurodivergent. Thus, a useful diagnostic model has to maximize both of these metrics.

$$Specificity = \frac{TN}{TN + FP} \tag{2.5}$$

AUC determines the classification ability of the model at different thresholds by calculating the area under the ROC curve which is formed by plotting recall in Equation 2.4 and false positive rate (FPR) in Equation 2.6. For two classes, it is the average of sensitivity and specificity, which is also referred to as balanced accuracy. We report on this metric for comparison purposes as it is commonly used to evaluate machine learning models for message classification [24].

$$FPR = \frac{FP}{FP + TN} \tag{2.6}$$

Accuracy, Equation 2.7, is the portion of correctly classified instances. As tp and tn are given the same weight, accuracy is an inappropriate metric to assess an unbalanced dataset and therefore often less useful in medical domains. Like AUC, this metric appears in related literature [24].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.7}$$

When leveraging different train/test splits of the data, we often repeat the experimental procedure to ensure robust results. We evaluate the machine learning models by considering the average evaluation metrics of the experiments with the same experimental parameters. Regardless of the evaluation strategy, we often consider the best model to be the one that maximizes the (average) $F1$ score given this metric focuses on tp. However, we still report on other metrics for these best models to understand their diagnostic utility and for comparison purposes.

2.1.8 Statistical Analyses.

There are two statistical analyses that are useful for mental health detection research and are featured within the related literature. Below, I give a brief explanation of these two analyses and mention some potential uses within the domain.

Correlation. A correlation is a statistical relationship between two variables. Specifically, performing a Pearson correlation yields a coefficient that measures the strength of linear relationship between two variables [71]. The values range from -1 to 1 with a 0 indicating no correlation. A p-value indicates the statistical significance of the calculated correlation. Correlation analyses are often used to find linear relationships in the data. Within the domain of mental health detection, correlation analyses may be performed instead of training machine learning models. This type of test can also be used to determine the amount of co-linearity between two variables in a dataset.

Statistical t-tests This two-sided statistical test assumes that two independent groups of values have the same average [71]. In addition to the calculated t-statistic, the t-test returns a p-value which measures the statistical significance of the results. If the p-value is less than 0.05, the two independent groups of values are considered to have statistically significantly different means with a confidence of 95%. A smaller p-value is considered more statistically significant. This t-test allows us to compare values between partitions of the data. For example, it allows us to determine if mental illness scores are statistically significantly different between different demographic groups in a dataset. Further, by repeating experiments multiple times, we can utilize a t-test to determine if the metrics for two different experimental procedures are statistically significantly different.

2.2 Generative Modeling

2.2.1 Generative Adversarial Newtork (GAN)

GANs are very popular models to generate data. Introduced in 2014, the original GAN paper [83] has already been cited over 40 thousand times. GANs are composed of a generator and disciminator engaged in a minimax game. The generator attempts to fool the discriminator by generating new data instances. The discriminator attempts to differentiate between the real and generated data instances. As such, the discriminator trains the generator to produce more realistic generated data. GANs [83] was originally demonstrated on images. In addition to research on improving GANs for image generation [84], GANs has been expanded to generate audio [85] and text [86].

To make GANs [83] applicable to text, Sequence Generative Adversarial Network (SeqGAN) [86] had to make some modifications. For instance, the limited dictionary space when using discrete tokens makes it difficult for the original GAN model to update the generator and a sequence can only be evaluated once it is completely generated by the original GAN discriminator. Thus, SeqGAN [86], considers the generation process as a sequential decision-making process and employs a Monte Carlo search to approximate intermediate rewards as the sequence is being generated [86]. There have a number of SeqGAN variants attempting to improve text generation [87], especially for longer texts. The two most promising are LeakGAN [88] in which a hierarchical generator receives high-level feature representations from the discriminator and RelGAN [89] which, among other additions, contains a relational memory based generator. On the COCO image caption dataset, the BLUE-2 score (defined in subsection 2.2.2) is 0.745, 0.746, and 0.849 for SeqGAN, LeakGAN, and RelGAN, respectively [89].

The main competition for GANs are Variational AutoEncoder (VAE). These models learn the distribution of the data and then sample from that distribution to generate new data instances [90]. There seems to be a general consensus that GANs produce more realistic output than VAEs Due to the training process, VAE models should generate very uniform instances whereas the instances generated by GANs can be more diverse. For the purposes of generating realistically diverse text messages, this makes GANs is a more attractive option. Generated texts could be used to both augment and replace real text messages.

2.2.2 Evaluation Metrics for Generative Models

There are three evaluation strategies commonly used by text generation models, all of which were leveraged to evaluate RelGAN [89]. Negative log likelihood NLL loss is an output of GANs models. Specifically, NLL_{gen}, defined in Equation 2.8, is used to assess the diversity of the generated samples with a score closer to 0 being desirable. This metric is calculated with the generated sentence distribution P_θ and the real sentence distribution P_r.

$$NLL_{gen} = NLL_{test} = -E_{r_{1:T}\ P_r} logP_\theta(r_1, \ldots, r_T) \tag{2.8}$$

To evaluate the linguistic quality of the generated texts and similarity to the real texts, seqGAN variants have adopted bilingual evaluation understudy (BLEU). This metric was designed to evaluate machine text translation of sentences [91], though it can also compare the similarities between corpuses of texts. The BLEU score measures the similarity between texts with values ranging between 0 and 1. A value of 1 is achieved by comparing identical texts.

The modified precision score P_n in Equation 2.9 and brevity penalty BP in Equation 2.10 are needed to calculate the BLUE score in Equation 2.11 [91]. In these questions, c is the generated translation length, r is the real text corpus length, and $n \in 1, \ldots, N$ is the number of sequential words referred to in this domain as $n - grams$. Note, in Equation 2.9, C and $n - gram$ are from the generated texts while C' and $n - gram'$ are from the real texts. The $MaxRefCount$ is the maximum number of times a word in the generated text occurs in any real text. The $Count_{clip}$ is the smaller number between the $Count$ and $MaxRefCount$.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n - gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count(n - gram')} \tag{2.9}$$

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{1-\frac{r}{c}}, & \text{if } c \leq r \end{cases} \tag{2.10}$$

$$BLEU = BP\dot{e}xp(\sum_{n=1}^{N} w_n log(p_n) \tag{2.11}$$

In the SeqGAN paper [86], the generated Chinese poetry is evaluated with $n - gram$ set to 2 as the word dependencies in Chinese poems consist of at most two characters. However, to evaluate generated Obama speeches which were much longer in length, the $n - gram$ were set to 3 and 4. In the RelGAN paper [89], the generated text for all datasets is evaluated with $n - grams$ ranging between 2 and 5. When the BLEU score is calculated with a specific $n - gram$, it is referred to as $BLEU - n$. Thus, in the RelGAN paper [89], the text generation models are evaluated with $BLEU - 2$ to $BLEU - 5$.

Finally, human evaluation is used to assess the quality of the generated texts. In the SeqGAN paper [86], 20 poems were randomly selected from each model and 70 experts rated each poem as either ‘real’ (1) or ‘generated’ (0). In the RelGAN paper [89], 100 sentences were randomly selected from each model and 10 MTurk participants (see subsection 2.4.1) rated the realness of each sentence between 1 and 5. The average of the scores were then used to compare the performance of the different models.

2.3 Mental Health Screening Surveys

As mentioned, mental illness screening is primarily conducted through surveys administered at medical facilities. There are many different screening surveys used for a variety of mental illnesses. Similar to many related studies, we use the results of these surveys to label our data. In other words, our trained machine learning models are predicting these mental illnesses survey scores. Similar to the screening surveys, these models do not provide diagnosis, only recommendations as to who should be further evaluated by mental health specialists.

2.3.1 Patient Health Questionnaire-9 (PHQ-9)

This common depression screening survey instrument only asks nine multiple-choice questions [92]. The PHQ-9 is highly accurate compared to mental health professional diagnosis [92]. Further, the PHQ-9 has been shown to be effective regardless of mode of administration and participant age group [93]. For all of these reasons, we utilize the PHQ-9 for depression screening.

Each of the nine questions ask participants to reflect on symptoms frequency during the prior two weeks and respond with a value between 0 for *'not at all'* and 3 for *'Nearly every day'*. The total score for the PHQ-9 is the summation of the scores for all of the individual questions, so it ranges from 0 to 27. A higher score is indicative of more severe depression. Table 2.1 contains the interpretation of the PHQ-9 at different score categories [92] and related treatment recommendation [94] if the diagnosis is confirmed by a clinician. At the cutoff score of 10, which is considered to the be threshold for the recommendation of treatment [94], the PHQ-9 has a sensitivity and specificity of 88% [92].

The ninth question of the PHQ-9, referred to as item-9, regards suicidal ideation. Some studies elect to leverage the PHQ-8, which does not include item-9, to screen for depression. The PHQ-8

Table 2.1: PHQ-9 score interpretation [92, 94].

PHQ-9 Score	Depression Severity	Treatment Recommendation
$0 - 4$	None	None
$5 - 9$	Mild	Periodic re-screening, education
$10 - 14$	Moderate	Pharmacotherapy or psychotherapy, education, create treatment plan
$15 - 19$	Moderately Severe	Pharmacotherapy and/or psychotherapy
$20+$	Severe	Pharmacotherapy and psychotherapy

score is interpreted the same as the PHQ-9 score. We elect to ask item-9 of participants as it allows for us to screen for suicidal ideation as well as depression. While the other questions in the PHQ-9 may also be screened for separately, none are as clinically meaningful by themselves as item-9.

2.3.2 General Anxiety Disorder-7 (GAD-7)

The GAD-7 [95] is the anxiety counterpart of the PHQ-9. This survey contains seven multiple-choice questions. Similar to the PHQ-9, participants are asked to reflect on the prior two weeks when responding to questions regarding symptom frequency with a value between 0 for *'not at all'* and 3 for *'Nearly every day'*. The total score for the GAD-7 ranges from 0 to 21. A score of $0 - 4$ indicates mild anxiety, a score of $5 - 9$ indicates mild anxiety, a score of $10 - 14$ indicates moderate anxiety, and a score of $15 - 21$ indicates severe anxiety. A score of 10 is the cutoff for which further evaluation is recommended. At this cutoff, the GAD-7 has a sensitivity of 89% and specificity of 82% [95].

2.3.3 Additional Screening Surveys

There are many other screening surveys with likert scales, especially for depression. Alternative depression screening surveys include the 21 question Beck Depression Inventory (BDI) [96], the 20 question Center for Epidemiological Studies Depression Scale (CES-D) [97], and the 17 question Hamilton Rating Scale for Depression (HRS-D) [98]. However, these surveys all require many questions which is problematic for quick and widespread screening. Even the PHQ-9 and GAD-7 have shortened versions with two questions referred to as the PHQ-2 and GAD-2 which sacrifice performance for speed of completion. Recently, suicide risk rulers [99] have recently been explored as quick and accurate alternatives to the 21 question Beck Scale for Suicidal Ideation (BSS). Alternatively, advances in artificial intelligence has led to the development of computerized adaptive tests which strive for achieve the same accuracy as lengthy questionnaires without sacrificing accuracy. One such example in this domain is the Computerized Adaptive Test Suicide Scale (CAT-SS) [100].

2.4 Related Mental Illness Screening

There are many different modalities that could be used to screen for mental illnesses. I overview a selection of data sources including audio, social media, self-written text, and smartphone sensor data. There is other related literature that leverages sensors that are part of wearable technology [49] or placed in the environment [50] to screen for mental illnesses, but these are less comparable to text messages. Before detailing studies that screen for mental illnesses for each of the aforementioned modalities, I first discuss participant recruitment and data labeling strategies.

2.4.1 Participant Recruitment

There are three main groups that researchers in this domain tend to recruit from: crowd-sourced workers, university students, and social media users. When research focuses on a specific subset of individuals, recruitment may instead occur at an organization catering towards that subset, such as a community group [66] or medical facility [42]. Snowball sampling, the practice of using participants to find more participants, is a strategy that can be used with multiple participant groups.

Crowd-sourced workers. Crowd-sourcing platforms allow businesses and researchers to pay crowd-sourced workers a small fee for completing a task. This is a common strategy to obtain large quantities of labeled data or large quantities of participants. Amazon Mechanical Turk (MTurk) is the most recognized of these platforms and commonly used for research in this domain [59, 54, 101]. While research has shown the collected data is high quality [102], there is nothing preventing participants from repeating the experiments with multiple accounts. Thus, other platforms have emerged targeted towards researchers, such as Prolific [103]. Benefits of Prolific include quality control and the ability to select a specific target audience with over 100 demographic parameters. Recent research in this domain has been leveraging this newer platform [51].

Social media users. By recruiting through social media, researchers are able to easily reach large groups of people and specific subsets of people. In addition to traditional posts, purchased advertisements on these social media platforms can be used to reach the users. In many ways this recruitment approach is similar to that of crowd-sourcing except that a different subset of the population is reached. Social media platforms tend to be favored for research leveraging social media data [24], though other research still uses it for recruitment purposes [66]. There are also social media groups that facilitate survey exchanges.

University students. University students tend to be very plentiful to researchers who conduct research out of universities so they are a convenient group to recruit [47, 43, 26, 104, 66, 48, 44]. The students are primarily motivated to participate through class credit, financial incentives, or raffles. The findings for studies on this population can only be extrapolated to other college students. Further, if students are only recruited from a specific department or school, the students may not be a representative sample even at the university where they were recruited. The student population is at higher risk for mental health illnesses than other age groups [28] so mental health detection research is particularly important for college students. In addition, data collected from students participants is often viewed as more trustworthy than data collected from crowd-sourced participants by medical domain experts.

2.4.2 Mental Illness labeling Strategies

There are four ways related studies collect mental illness labels for their data: clinical assessment, mental illness screening instrument, passive self-declaration, and active self-declaration.

Clinical assessment. Having a clinician assess participants is considered the most reliable ground truth for mental illness [48, 42, 105]. Unfortunately, this labeling strategy is time consuming for both participants and clinicians. As such, it is more challenging and costly to recruit participants if a clinical assessment is required.

Mental illness screening instruments. Screening instruments are the most common way to obtain mental illness labels for participants [54, 59, 56, 58, 106, 101, 22, 42, 51, 26, 104, 107, 66, 44, 41, 108]. It is more simple to obtain than a clinical diagnosis and more trustworthy than self-declaration. There are many different mental illness screening survey instruments that can be deployed to participants. In addition, there are ecological momentary assessment (EMA) measures [26, 44] that have participants report frequently on behavior, feelings, or symptoms. These are often used in conjunction with the less frequently deployed mental illness screening surveys.

Passive self-declaration. Social media data is unique in that messages can be labeled through self-declaration; for instance, Twitter users may post about experiencing depression [55, 60]. Studies that rely on self-declaration are able to build larger datasets as their collection does not require participation, though self-declaration poses is problematic for three main reasons:

1. diagnosed individuals frequently identify with their diagnosis which could influence messaging patterns and message content [59],

2. passive mental illness screening for undiagnosed individuals is the most crucial, and

3. the lack of self-declaration does not indicate a lack of mental illness or symptoms, thus making it impossible to identify a comparative set of individuals without mental illnesses.

Active self-declaration. Asking participants if they are diagnosed with a mental illness is less common. Effectively a mix between the two prior options, it is subject to the limitations of both labeling strategies. Unlike passive-self declaration, active self-declaration is unable to effortlessly collect the same large quantities of data. Further, not all participants with mental illnesses may know they have mental illnesses. As such, studies may alternatively ask participants to actively report on emotional wellbeing, such as stress and sadness [43].

2.4.3 Screening with Text Data

There are many different types of text data. Those featuring in mental illness screening research including transcribed interviews [45, 46], essays [47], and social media posts [24, 25]. While social media messages are most similar to text messages, I first discuss all of these text-based modalities together due to the similarities in how they are analyzed.

Screening Similarities. The research that leverages text modalities for screening purposes focus primarily on detecting depression and depressive symptoms [45, 46, 47, 54, 56, 58, 55, 106, 101, 25]. Many studies use CES-D scores to label the data with a score of at least 22 being indicative of depression [54, 59, 56, 58]. Many studies also use the proprietary Linguistic Inquiry and Word Count (LIWC) software [109] to extract lexical category frequencies [54, 58, 47, 45, 46].

Twitter. Surveys of studies focusing on mental illness screening with social media reveals that Twitter is a particularly popular social media platform for such research [24, 25]. Studies leveraging tweets implement a variety of machine learning models, including support vector classifiers (SVC) [54, 55, 56], regressions [58, 55], random forests (RF) [59], Naive Bayes classifiers [55], and neural networks [60]. The two studies that predicted depression with a binary CES-D score from tweets with the most success achieved an accuracy $= 0.70$ with SVC [54], and F1 $= 0.65$ and AUC $= 0.87$ with RF [59]. Both studies recruited participants from MTurk. De Choudhury et al. [54] collected one year of tweets from 476 participants who reported being diagnosed with depression and performed dimensionality reduction on 188 features involving engagement measures, egocentric network measures, emotion, linguistic style, depression language, and demographics. Reece et al. [59] collected as many recent tweets as possible from 105 depressed participants and 99 healthy participants, extracting features including volume and word category frequencies.

Facebook and Instagram. Facebook and Instagram are also used in some noteworthy depression screening studies though these social media platforms are not as popular as Twitter for such research. For Facebook, a relevant research study collected PHQ-9 scores from 165 new mothers to predict depression from Facebook posts [106]. For Instagram, a relevant research study had 749 crowd-sourced participants complete the PHQ-8 to predict depression with Instagram data [101].

Important features. In these text-based studies, high usage of (singular or plural) self-references were associated with depression [45, 46, 47]. Negative emotion/affect/valence also indicated depression [58, 47, 46]. In one study, depressed Twitter users used statistically significantly more words related to anger [58] which is notable given the link between depression and anger [110].

2.4.4 Screening with Audio Data

While there are a number of audio datasets to predict depression and suicide risk [23], we focus on the most relevant subset. Notably, I focus on datasets that are publicly available, collected recently, and/or collected through a mobile device. The majority of these datasets labeled the audio recordings with PHQ-9 scores. While audio may seem like an unrelated modality to text messages, the transcripts of unscripted voice recordings can be analyzed similarly.

Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ). This dataset contains videos of clinical interviews [22, 111]. 189 sessions conducted with the public around Los Angeles have been publicly released, making it a very popular dataset. These interviews were collected with a virtual agent who asked the same set of main questions in different ways with varied follow-up questions during the clinical interview. The videos provide three different modalities for machine learning purposes: the visual video, the auditory audio, and the audio transcripts. The data is labeled with PHQ-8 scores. This public dataset is popular for depression detection research, especially after featuring in the 2016 Audio/Visual Emotion Challenge and Workshop [112]. In this workshop, openSMILE toolkit [113] was used to extract audio features, though the depression classification model only achieved an F1 of 0.41.

Our Emutivo research team has experimented with applying sub-clip boosting [114] and extract features with topological data analysis [9] to classify depressed participants. We also innovated Audio Assisted BERT (AudiBERT) [11], a multi-modal transfer learning framework, and applied to it the individual interview questions to achieve the highest depression screening results for the paired audio and transcript data in the DAIC-WOZ corpus. For direct and general wellbeing questions, AudiBERT achieved maximum average $F1$ scores of 0.92 and 0.86, respectively [11]. We have since explored the best BERT variant and the impact of ensembling BERT variants for the classification of the transcript data [14]. Further, we assessed the number of follow-up questions required for each core questions with audio [13], text [17], and multimodal models [17].

Interview spoken utterances. Researchers collected an hour of video recordings of 148 adolescents in three different interactions with their families. Half of the adolescents were diagnosed with depression [105]. Speech and textual features were extracted from the video recordings. For speech, the openSMILE toolkit was used. For the transcripts, each word in an utterance with an arousal and a valence rating [115] in order to calculate the per-utterance mean, standard deviation, minimum, and maximum [105]. Support vector classifiers inferred depression with accuracies of 0.65 for the audio features, 0.65 for the text features, and 0.68 for both types of features.

Mobile spoken utterances. Researchers deployed an Android app to collect spoken utterances and PHQ-9 scores from 887 participants [108]. Overall, 5937 short audio recordings spanning 6 distinct tasks were collected from these participants. OpenSMILE and six comparative voice activity detection approaches were adopted in this research. The depression classification models with the sentence level task audio features was most successful with an F1 of 0.41. The researchers mentioned the background noise in the recordings added to the challenge of the task. The small fraction of participants with depression may have also made classification more difficult.

Multi-modal Open Dataset for Mental-disorder Analysis (MODMA). The MODMA dataset currently consists of PHQ-9 labeled electroencephalogram (EEG) and audio recordings clinically assessed hospital patients in China [42]. An EEG test involves attaching electrodes to patients so it is not a passive screening alternative. Therefore, we will focus on the audio component of this dataset. MODMA contains recordings from 52 participants responding to interview questions, reading a fable, and describing a picture. 44% of the participants had depression. Less than a third of the participants were women. No data analysis was conducted in the MODMA dataset paper [42]. So far only one study has been conducted on this audio data [57]. This paper explored the screening potential of a binary fusion tree with SVC and RF as comparison models. The best reported accuracy is 75.8% for male participants and 68.5% for female participants on the gender-dependent models [57]. However, even with 4-fold cross-validation, the size of the partitioned data is concerning and the model may be overfitting.

Environmental audio. In 2019, a research team deployed an Android app to capture the environmental audio of Canadians through the crowd-sourcing platform Prolific [51]. This app was designed to record 15 seconds of audio every 5 minutes. The participants were asked to complete the PHQ-8, GAD-7, the Liebowitz Social Anxiety Scale (LSAS), and the Sheehan Disability Scale (SDS) at the start and end of the two week study. Only 84 of the original 205 participants completed all requirements. While the depression rate in Canada is estimated to be 5%, 37% of the participants screened positive for depression. The authors propose two possibilities [51]: 1) workers on Prolific have higher rates of depression than the general population and 2) self-selection bias may lead depressed individuals to more readily enroll in a study focusing on mental health.

The features extracted from the environmental audio were daily similarity, sleep disturbance on all nights, sleep disturbance on weekends, and speech presence ratio [51]. The analyses were conducted by calculating correlations between these features, mental health surveys, and demographics. Based on the correlation results, the younger participants suffered from worse mental health. The daily similarity and speech presence ratio were both had a statistically significant negative correlation with PHQ-8 scores [51]. None of the environmental audio features had statistically significant correlations with the GAD-7 scores.

2.4.5 Screening with Prospective Smartphone Sensor Data

This section discusses apps that have been deployed to collect datasets with mental health labels. While some of the data collections recruited an insufficient quantity of participants to make any generalizable mental illness conclusions, all of the studies provide useful comparative information. As such, I mention the number of participants and type of data collected for all of the apps. However, I only discuss the results for the papers that screen for mental illnesses with enough participants to construct potentially meaningful machine learning models. Notably, features extracted from prospective smartphone logs are primarily used in conjunction with features extracted from other mobile and wearable sensors.

The apps in this section use a *prospective* design. When data is collected prospectively, the app captures the data as it is being generated by the participant. This means that the data is collected during the time the app is installed on the smartphone. This approach has a few downsides. First, the knowledge of being monitored can alter behavior which leads to bias in the collected data. This common phenomenon is known as the Hawthorne effect. Second, prospective apps continuously uses phone resources which drains phone battery more quickly than normal [66, 43] which can lead to participants being unwilling to continue with the data collection. Overall, participant dropout is a common problem among longitudinal studies.

Reality Commons. In spring 2009, 70 students residing in the same residence hall in New England were recruited to participate in this data collection [43]. Students downloaded an app that collected call logs, SMS logs, Bluetooth co-location sensing, and WLAN-based location sensing. Throughout the study, students reported on any experienced health symptoms, including stress and sadness. Both of these self-reported emotional health labels correspond with less communicative behavior, though the data can not reveal any causal relationships. Notably, the study reported on how the periodic scanning reduced the battery life of the smartphones by $10 - 15\%$. While this dataset does not have mental illness screening labels, it is worthy of mention for being open-source and containing SMS logs.

StudentLife. StudentLife was the first continuous sensing Android app to predict mental health and academic success [26]. This app was deployed to track 48 college students in a computer science class at Dartmouth College over a 10 week term in 2013. In addition to the PHQ-9, the mental health assessments include perceived stress, loneliness, and flourishing scales. These surveys were administered at the beginning and end of the study. Further, ecological momentary assessments (EMA) were administered daily. Students were motivated with raffles throughout the collection process. While SMS text message logs were collected, no message content was collected. The anonymized data from this collection has been publicly released.

In the StudentLife dataset paper, correlation analysis was used to analyze the data. The main findings for depression is that PHQ-9 score has a statistically significant negative correlation with sleep duration and quantity of conversational interactions [26]. The former is expected as one of the questions on the PHQ-9 inquires about sleep quality, though both factors are known to be associated with depression. Further, the perceived stress scale is statistically significantly negatively correlated with sleep duration, conversation frequency, and conversation duration [26]. While the perceived stress scale is not validated to measure clinical anxiety, stress and anxiety are likely correlated with similar behavior. As this is a public dataset, there is much research conducted on this data. Most relevant, an autoencoder-based anomaly detection approach with SVC was applied to the GPS data to achieve an AUC of 0.92 when predicting depression based on PHQ-9 scores [61].

The StudentLife app was leveraged again on 217 undergraduate students [104] during a longitudinal study encompassing the Winter 2020 term (January 5 - March 13) at Dartmouth College, the first term impacted by the Covid-19 pandemic. 178 of the students participated in the study during the Winter 2020 term. Students completed the Patient Health Questionnaire-2 (PHQ-2) and the Generalized Anxiety Disorder-2 (GAD-2) weekly during the term to label the data. The app tracked the students' sedentary time, sleep, location, and phone usage. Depression, anxiety, phone usage, number of locations visited, and week of the term were statistically significantly associated with the quantity of COVID-19 news [104]. Further, depression, anxiety, and sedentary behavior were increased, though it is unclear if this was due to the COVID-19 pandemic or the continuation of an existing trend. While news of the COVID-19 pandemic had begun to surface at the start of Dartmouth College's winter 2020 term, most universities on the East Coast were not directly impacted until the end of the term.

Purple Robot and MoodTraces. In 2013, Purple Robot Android app was deployed for two weeks to collect smartphone sensor data from 40 participants [107]. These participants were recruited through a craigslist advertisement and asked to complete the PHQ-9 at the beginning of the study. 18 and 21 participants had sufficient GPS location and phone useage data for analysis, retrospectively [107]. Another study had users download an android app called MoodTraces to collect daily PHQ-8 and mobility data from GPS tracers [66]. Participants were recruited through academic mailing lists, social media, and charities. Despite raffle incentives, only 28 participants shared sufficient data for the analysis from 2014 to 2015 [66]. Unfortunately, there were insufficient participants in either study to draw meaningful results about depression screening.

LifeRhythm. LifeRhythm [48], a cross-platform app, was first deployed to collect smartphone sensor data from 79 clinically assessed college students at University of Connecticut , 19 of which were assessed as depressed. Over six months in 2015-2016, students were presented with the PHQ-9 every two weeks. In the second deployment in 2017, 39 of the 104 students were assessed as depressed. These students completed the 16 question Quick Inventory of Depressive Symptomatology (QIDS) [116], a tool used to track fine-grained depression symptoms rather than screen for depression. Approximately three-fourths of participants in both data collections were female [48].

LifeRhythm [48] collected a variety of smartphone sensor data including Wifi association data logged based on events. Depending on the platform of the app, the location data was either from periodic GPS collections or event-based. Further Wifi association logs were collected, which differ from the Wifi data collected through the phone. Support vector machines were used to classify the depression symptoms [48], though the paper alludes to this team having performed other analyses on subsets of their collected data. Unfortunately, as the type of location data collected differed based on the smartphone platform, separate SVM models were trained for android and iPhone users. As such, the results were not generalizable as there were only 6 and 13 depressed participants with android and iPhones in the first data collection, respectively [67].

DemonicSalmon. The DemonicSalmon study [44] is the integration of the Depression Monitoring Study (DEMONS) and Social Anxiety Life Monitoring Study (SALMON) studies. For the DemonicSalmon study, the Sensus android app was downloaded by 72 University of Virginia students in 2016 to collect two weeks of smartphone sensor data (including SMS text logs) and ecological momentary assessments (EMA). At the start and end of the study, the participants completed the 20-item Social Interaction Anxiety Scale, the 7-item depression subscale (from the Depression, Anxiety, and Stress Scales), and the Positive and Negative Affect Schedule. Each participant were prompted to complete the EMAs multiple times per day. Overall, 72% of the $3,756$ sent EMAs were completed. Leisure time, physical activity, time spent in town, and number of SMS text messages were positively correlated with positive mood and mental health [44]. While some of these correlations were statistically significant, the correlations were not very strong. The dataset has been made publicly available so further research on this dataset is possible, though it does not share the mental health assessment instruments as similar datasets.

Additional research with log data. Wahle et al. [52] recruited participants to use the Mobile Sensing and Support (MOSS) smartphone app. The number of calls and text messages of the 36 adults who used the app for two weeks were used in random forest and support vector classifiers to screen for depression based on PHQ-9 scores. The best screening model achieved an accuracy of 0.60 [52]. Wang et al. [117, 118] collected smartphone sensor data for months from over 20 outpatients with schizophrenia to predict aggregated scores of mental health indicators of schizophrenia. In this study, the computed features included the daily number and duration of incoming and outgoing calls and SMS text messages. Taylor et al. [119] extracted 50 features from the timing and duration of call and text events over four hour time intervals for 104 users with between 10 and 30 days of data. A total of 343 features were used as input to neural networks to predict self-reported mood, stress, and health. The multitask learning classifiers achieved the highest accuracies [119].

Modeled after the StudentLife collection, Xu et al. [120] recruited two groups of undergraduate students containing 138 and 212 participants respectively. These researchers automatically generated contextually filtered features including the number of daily phone calls from a semester of data to detect depression with an F1 of 0.84. Most recently, Chikersal et al. [121] extracted robust features from a semester of passive sensor data from the 138 aforementioned undergraduates; the features included the number and duration of calls to different contact subsets. Logistic regression and Gradient Boosting classifiers screened for depression with an accuracy of 0.85.

2.5 Datasets Containing Retrospectively Harvested Smartphone Data

2.5.1 Mood Assessment Capable (Moodable) Dataset

The Moodable framework for depression assessment with retrospectively harvested smartphone and social media data [41] was the first to collect text messages with mental health labels. Unlike the StudentLife [26] data collection, Moodable collected text message content as well as the SMS logs. Further, Moodable is innovative among the aforementioned collection apps as the smartphone data was collected retrospectively. This bypasses all of the challenges faced by prospective data collections and allows for instantaneous assessment when deployed to screen for mental illnesses.

The Moodable dataset is internal to WPI and collected prior to my start on the Emutivo research project. The Moodable Android app was deployed to retrospectively harvest data from $300+$ crowd-sourced participants on MTurk in late 2017 through early 2018 under WPI IRB 00007374 File 18-0031 approved 23 October 2017. In addition to the retrospective smartphone and social media data, participants were asked to record themselves reading a sentence as the retrospective app was not able to extract audio files from the phone. All participants were prompted to complete the PHQ-9 through the app to provide a depression screening label.

The phone modalities included text message logs with message content, though only a subset of the participants elected to share text messages. Further, some participants did not have both incoming and outgoing text messages on their phones. After manual data cleaning by the Moodable team [41], 240 participants shared texts messages. For comparison, 335 shared at least one modality with 266 sharing a voice recording and 147 sharing GPS data. Notably, the quantity of text participants may vary depending on the cleaning strategy, window of time, and message direction.

Prior text message screening results. In the Moodable dataset paper [41], feature engineering and machine learning was applied to determine the predictive ability of each of the collected modalities. The text features include incoming text sentiment score moving average, incoming text frequency moving average, and part of speech tag frequencies. The machine learning methods are kNN, SVC, and RF. The results from predicting participant depression based on different PHQ-9 cutoffs are displayed in Figure 2.1. There is no F1 above 0.6 at cutoffs 5, 10, or 15. Even when using features from all modalities, the F1 does not exceed 0.65 for these critical cutoffs [41].

My automated cleaning strategy. From the text content, it became clear some participants repeated the data collection process with multiple MTurk accounts. As the quantity of messages sometimes increased between repetitions of the data collection, our cleaning strategy evolved over time. As such, different subsets of participants may be present in different analyses. The automated cleaning strategy I designed is to order the text messages by date, with the newest at the start. I then remove messages with duplicate meta-data, ignoring the attached session ID. Thus, only the most recent session ID for each MTurk worker should remain in the dataset.

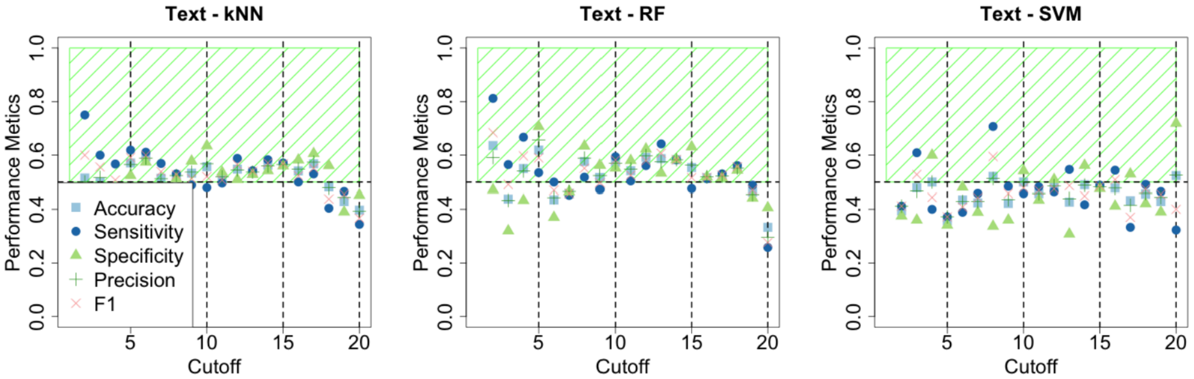


Figure 2.1: Machine learning results with features from Moodable text messages [41].

2.5.2 Early Mental Health Uncovering (EMU) Dataset

The EMU framework leverages passive and active modalities to quickly conduct mental illness screening [12]. We designed and deployed the EMU Android collection app to collect the EMU dataset. In early 2019, 70 unique MTurk participants submitted data. As we stored the smartphones' hardware identifiers, we were able to easily identify unique participants. Our app prompted participants to complete the PHQ-9 for depression labels and the GAD-7 for anxiety labels.

Apart from Moodable, EMU is the only dataset that contains text message content with mental illness labels. While almost half (31) of the participants shared text logs, this number is small for modeling purposes. Likewise, call logs were shared by 25 participants. Thus, EMU logs are primarily used in conjunction with Moodable logs. The most shared EMU modalities were scripted and unscripted voice recordings, which were shared by 63 and 55 participants, respectively.

2.5.3 Student Suicidal and Depression Detection (StudentSADD) Dataset

We further collected of a student dataset that extended from August 2020 to January 2021, referred to as StudentSADD [15]. To appeal to the student audience, we developed website version of the EMU app that was aesthetically similar to the Android app. Understandably, this website version of the app could not collect smartphone logs like the Android version. Overall, 302 unique students from multiple universities contributed data. All students completed the PHQ-9 to label the data.

Unexpectedly, only 11% students elected to share data through the app. Of those students, only a third shared text logs, which we collected without content to preserve student privacy. Instead, text prompt replies were collected for this population, a modality shared by almost all participants. Further, 115 and 110 scripted and unscripted voice recordings were used in screening models [15].

2.5.4 Screening Results Without Text Messages

The Moodable paper [41] used multimodal machine learning models for depression screening, but as mentioned, the F1 scores never exceeded 0.65. As only scripted audio was collected, the transcripts did not vary so there was no purpose in extracting text features. The EMU paper [12] compared the mental illness screening abilities of scripted and unscripted audio recordings using openSMILE features. The best model achieved an F1 of 0.75 by screening for depression with scripted audio features. With only transcript features, models screened for depression, anxiety, and suicidal ideation with F1 scores of 0.54, 0.63, and 0.55, respectively. For depression screening in the StudentSADD paper [15], transfer learning models achieved F1 scores of 0.67, 0.51, 0.63, and 0.64 with text replies, unscripted transcripts, unscripted voice, and scripted voice, respectively.

The voice recordings in these datasets have also been combined in a number of other papers. Our Emutivo research team has experimented with applying sub-clip boosting [114] and extract features with topological data analysis [9] from the Moodable and EMU voice recordings to screen for depression. While Moodable and EMU do not share the same voice prompts, EMU and StudentSADD share the same voice prompts. However, combining the voice recordings from these populations also did not increase depression or suicidal ideation screening capabilities [18].

2.5.5 My Research in Context

While the depression screening results of the Moodable text messages in the original analysis were subpar [41] when compared to other modalities, I hypothesize the screening potential of text messages are much higher. When collected retrospectively, these messages have the potential to provide unbiased, passive, and instantaneous mental health assessment. My research primarily leverages the text messages in the Moodable and EMU datasets. Notably, my research expands upon the text message analysis performed by the original Moodable team in five key ways:

1. Extracting more expansive sets of **features** from the text message data.

2. Identifying the **subset** of text messages that are most useful for screening.

3. Collecting a large **dataset** of text logs labeled with multiple mental illness scores.

4. Experimenting with traditional machine learning, ensemble, and deep learning **methods**.

5. Determining the usefulness of **generated** text for screening purposes.

CHAPTER 3

MENTAL ILLNESS SCREENING WITH TEXT MESSAGE CONTENT

Context: While research is starting to utilize machine learning with social media and smartphone data to replace survey instruments, text message content has not been evaluated as a modality.

Objective: We explore the ability of subsets of text messages to screen for mental illnesses.

Methods: Our approach involves comprehensive feature engineering with 245 features encompassing lexical category frequencies, part of speech tags, sentiment, and volume. We further construct replacement lexicons with categories containing less formal language. We compare the screening ability of subsets texts by training machine learning and ensemble models.

Findings: We screened for depression with $F1 = 0.8$ using two weeks of sent texts and suicidal ideation with $F1 = 0.84$ using one week of sent texts. Our less formal lexicons improved the depression screening capabilities of the models. The received texts from the 25 percent most prolific contacts proved better at depression screening than other subsets of received texts.

This chapter covers material from the following papers:

ML Tlachac, Elke Rundensteiner, “Screening for Depression with Retrospectively Harvested Private versus Public Text”, *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, 2020 [1]

ML Tlachac, Katherine Dixon-Gordon, Elke Rundensteiner, ”Screening for Suicidal Ideation with Longitudinal Text Messages”, *17th IEEE BHI*, pp 1-4, 2021 [2]

ML Tlachac, Avantika Shrestha, Mahum Shah, Benjamin Litterer, and Elke Rundensteiner, “Automated Construction of Lexicons to Improve Depression Screening with Text Messages”, in Submission [3]

ML Tlachac, Ermal Toto, Elke Rundensteiner, “You’re Making Me Depressed: Leveraging Texts from Contact Subsets to Predict Depression”, *16th IEEE BHI*, pp 1-4, 2019 [4]

3.1 Participant Depression Screening with Private Versus Public Text

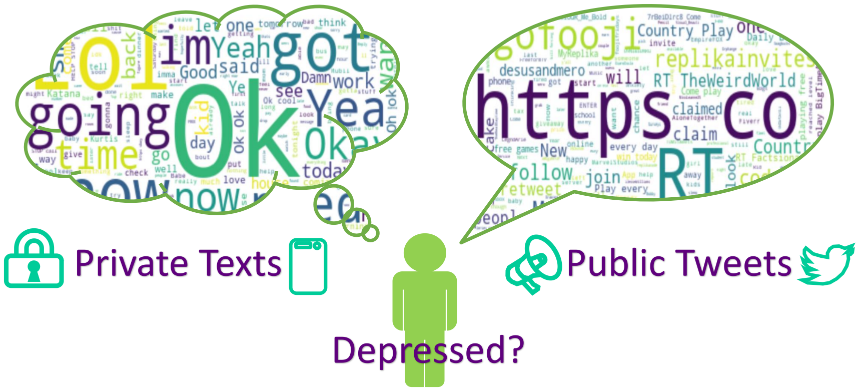


Figure 3.1: Comparison of texts and tweets for Moodable participant 7237 with PHQ-9 = 7.

The goals of this research is to increase the ability to screen for depression of participants from their retrospectively harvested self-written text and compare the screening ability of private versus public texts. Specifically, we leverage text messages and tweets collected retrospectively to when crowd-sourced participants completed the PHQ-9. This is the first study analyzing self-written texts and tweets from the combined Moodable and EMU datasets.

While other aforementioned studies have screened for depression with tweets, we are not aware of any study that has screened for depression with self-written text messages. We explore the question of screening for depression from private versus public text by extracting a rich variety of features involving word category frequencies, part of speech frequencies, sentiment, and volume. We perform feature selection to improve model performance and determine the most important features. Then we explore the performance of various machine learning methods.

This research has four main contributions. First, we examine the potential of crowd-sourced retrospectively harvested self-written text to screen for depression. Second, we compare the depression screening ability of private versus public self-written text. Third, we explore the impact of different temporal quantities of data with diverse machine learning methods. Lastly, we identify the most important features for depression screening from message data.

3.1.1 Dataset

Our study leverages subsets of the Moodable and EMU datasets involving sent text messages and tweets. As the datasets were collected through crowd-sourcing, some issues arose during the data collection process. In all datasets we remove duplicated messages with identical metadata as well as users with only one unique message. To compare the impact of different temporal quantities of data, we extract the messages sent within 14, 28, 42, 56, 182, and 364 days of data collection for each participant. As only 33 participants who provided messages within the last year elected to share both modalities, we consider the text messages and tweet participants separately. Within a year, 162 participants (15 from EMU) generated $43,645$ text messages and 105 participants (10 from EMU) generated $45,908$ tweets. The distribution of PHQ-9 scores for these participants is depicted in Figure 3.2. 55 of the text message participants and 35 of the tweet participants have a PHQ-9≥ 15. Summary statistics for all datasets are in Table 3.1.

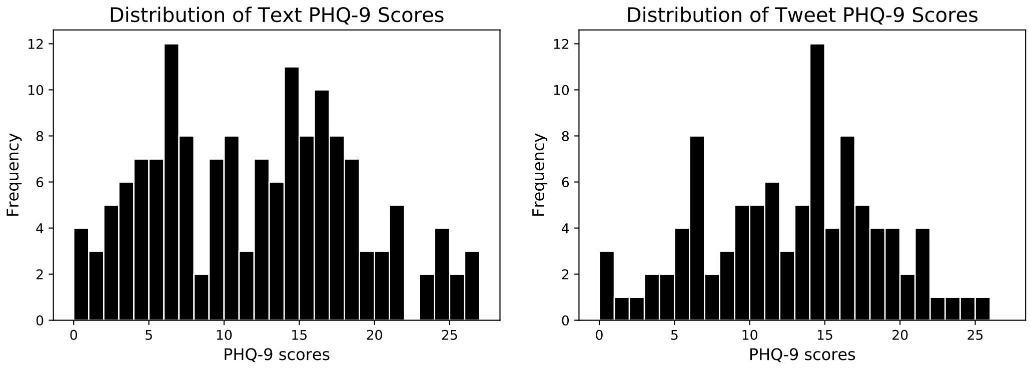


Figure 3.2: PHQ-9 scores of participants with messages within last 364 days.

3.1.2 Methodology

Our approach involves feature engineering, feature selection, and machine learning. As a PHQ-9 score of 15 is the cutoff for moderate depression, we focus on predicting if a PHQ-9 score will be at least 15. Thus, we refer to PHQ-9< 15 as 'not depressed' and PHQ-9≥ 15 as 'depressed'.

Table 3.1: Number of participants and average messages M per participant ± standard deviation (std) for all datasets.

Dataset- days	PHQ < 15	PHQ ≥ 15	PHQ $<$ 15 Avg(M)±std	PHQ $\geq$ 15 Avg(M)±std
Text-14	68	42	76.5 ± 116.9	53.4 ± 85.6
Text-28	79	44	109.9 ± 189.1	86.1 ± 131.7
Text-42	84	44	139.0 ± 270.1	107.7 ± 161.8
Text-56	87	47	162.5 ± 336.4	118.9 ± 188.3
Text-182	92	52	273.5 ± 627.9	174.5 ± 283.4
Text-364	96	55	335.8 ± 856.9	207.5 ± 358.0
Tweet-14	57	32	313.2 ± 551.7	475.6 ± 725.6
Tweet-28	57	32	331.9 ± 592.5	487.5 ± 746.7
Tweet-42	57	32	338.9 ± 603.1	501.3 ± 768.6
Tweet-56	57	32	346.1 ± 610.9	521.4 ± 804.5
Tweet-182	61	34	358.5 ± 360.8	577.4 ± 919.5
Tweet-364	62	35	393.3 ± 650.7	615.0 ± 925.0

Text feature engineering. We only consider features relevant for both modalities. For each dataset, we extract 245 features from the content of each participant's messages involving lexical category frequencies, part of speech (POS) tags, sentiment, and volume.

- **Lexical category features.** While related studies [54, 58, 59] tend to use Linguistic Inquiry and Word Count (LIWC) software to extract use of lexical categories in text corpuses, we instead use the open-source Empath software due to it's distinct advantages. In addition to being able to create custom categories, Empath contains a larger number of preexisting word categories which include more modern terms [69]. Specifically, we use Empath software [69] to calculate the frequency of words in 195 categories for each participant. In addition to 194 preexisting word categories, we leveraged the software to create a new category: *text abbreviation*. We used the social media corpus (reddit) with the seed words 'lol', 'ttyl', and 'brb' to spawn 100 abbreviations.

- **Part of speech tag features.** Similar to the original analysis of the Moodable dataset [41], We also consider part of speech (POS) tags to engineer text features. POS tags have been shown to capture linguistic style [54]. As there are 36 POS tags in the English language, this results in 36 features. Specifically, we generate (word, POS) pairs for every message using Textblob, software designed for textual data processing [122]. For each participant, we calculate the frequency of use or count for each of the POS tags in our text corpus.

- **Sentiment features.** There are two types of sentiment features for text: polarity and subjectivity. For each message, we use Textblob [122] to generate a polarity score $p \in [-1, 1]$ as well as a sentiment score $s \in [0, 1]$ based on the words contained within the tweet. Messages with positive and negative polarity are those with $p > 0$ and p¡0, respectively. Messages with subjective content are those with $s > 0$. For example, exclamation ”wow” has $p = 0.1$ as it is mildly positive and $s = 1.0$ as it is an opinion. We then calculate the percent of positive, negative, and subjective messages for each participant. In addition, we calculate the average and standard deviation of the scores of the messages that are positive, negative, and subjective.

- **Volume features.** As the quantity of social interactions is a known predictors of health [123], we consider 5 volume features for text data including the number of messages as well as the number and standard deviation of words and characters per message.

Feature selection. In order to determine the most important features, we adopt a standard univariate feature selection schema [74]. Specifically, we calculate the chi-squared statistic between each normalized feature and the target variable which consists of two categorical classes. The k features with the largest chi-squared statistic are selected. For comparison purposes, we also perform principal component analysis (PCA) and kernel PCA with a Gaussian kernel (kPCA) on the normalized features.

Machine learning. Other studies in section 2.4 achieved the highest F1 scores for binary depression predictions with logistic regression (LR) [55], random forests (RF) [59], and support vector classifiers (SVC) [54, 56]. Thus, we also experiment with these machine learning methods. Additionally, we include Gaussian Naive Bayes (NB) classifiers known to be effective for document classification [71], XGBoost which provides gradient boosting for decision trees [124], and kNN as it is robust to noise [71]. We consider 3 and 5 neighbors for kNN, linear and Gaussian kernels for SVC, and a depth of 3 for random forests and XGBoost. If not discussed, the default parameter settings were used.

Model evaluation. To mitigate any bias from the unequal distribution of PHQ-9 scores at the cutoff for moderate depression (PHQ-9$\geq$ 15) [125], we apply down-sampling to balance the two classes. For every method, we build a model with $k \in [1, 75]$ features and the top 15 principal components from PCA and kPCA computed on all 245 features. We also build models with PCA features computed from only the top chi-squared selected features. Each model utilizes 5-fold cross-validation. To ensure result robustness, we repeated each experimental procedure 100 times. Our goal is to maximize the $F1$ score, though we also report on the $precision$, $recall$, $specificity$, AUC, and $accuracy$ of the models with the highest $F1$ scores for comparison purposes.

3.1.3 Depression Screening Results

We apply the machine learning methodology with methods NB, LR, SVC, kNN, RF, and XGBoost on two weeks through a year of data for both text and tweet modalities. For tweets, LR models with chi-squared selected features achieved the highest average F1 score for every tweet dataset. Apart from models with half a year of data, the prediction ability of PHQ-9 scores from tweets remains consistent regardless of the temporal quantity of data. As seen in Table 3.2, the average $F1$ scores range from 0.632 to 0.675.

For texts, different methods with chi-squared selected features achieved the highest average F1 scores depending on the text dataset, as seen in Table 3.3 and Figure 3.3. LR achieves the highest average $F1$ score for 14, 182, and 364 days of texts. NB achieves the highest average $F1$ score for 28 days of texts. SVC with a linear kernel achieves the highest average $F1$ score for 42 and 56 days of texts. For 28 and 42 days of data, all three of these methods achieve the highest average

Table 3.2: Models with the highest average F1 scores for each tweet dataset leveraged chi-squared selected features.

Dataset	**Method**	**Features**	**F1 ± std**	**Precision**	**Recall**	**Specificity**	**AUC**	**Accuracy**
Tweets- 14 days	LR	69	0.674 ± 0.065	0.714	0.669	0.723	0.761	0.69
Tweets- 28 days	LR	72	0.662 ± 0.068	0.688	0.665	0.680	0.741	0.67
Tweets- 42 days	LR	57	0.675 ± 0.057	0.719	0.664	0.721	0.765	0.69
Tweets- 56 days	LR	28	0.651 ± 0.055	0.720	0.620	0.746	0.744	0.68
Tweets- 182 days	LR	43	0.632 ± 0.058	0.718	0.591	0.748	0.735	0.67
Tweets- 364 days	LR	73	0.660 ± 0.050	0.691	0.657	0.685	0.748	0.67

Table 3.3: Models with the highest average F1 scores for each text dataset leveraged chi-squared selected features.

Dataset	**Method**	**Features**	**F1 ± std**	**Precision**	**Recall**	**Specificity**	**AUC**	**Accuracy**
Texts- 14 days	LR	10	0.799 ± 0.021	0.721	0.907	0.630	0.809	0.769
Texts- 28 days	NB	27	0.704 ± 0.059	0.648	0.804	0.539	0.751	0.672
Texts- 42 days	SVC	34	0.707 ± 0.058	0.658	0.802	0.553	0.756	0.677
Texts- 56 days	SVC	43	0.714 ± 0.049	0.673	0.794	0.586	0.777	0.690
Texts- 182 days	LR	16	0.781 ± 0.043	0.724	0.869	0.651	0.820	0.760
Texts- 364 days	LR	19	0.766 ± 0.022	0.716	0.837	0.653	0.811	0.745

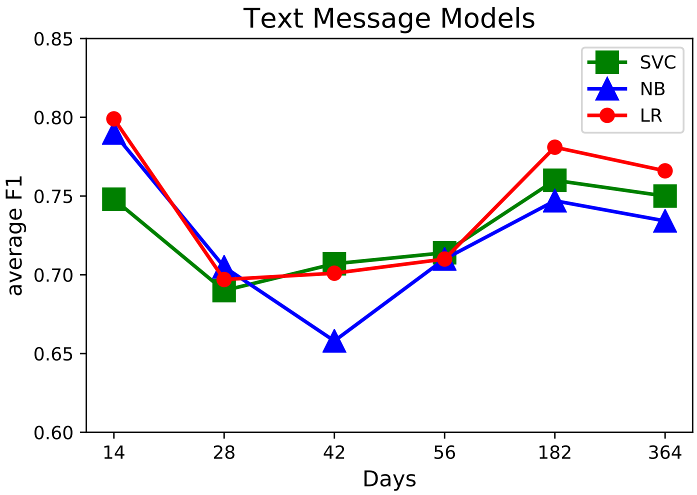


Figure 3.3: Comparing highest average F1 scores for best methods on text datasets.

$F1$ score with 27 and 34 features, respectively.

The best models for texts have a higher average $F1$ score than the best models for tweets, regardless of the temporal quantity of data, as seen in Table 3.2 and Table 3.3. While the $precision$ is higher than the $recall$ for the best tweet models and the $recall$ is higher than the $precision$ for the best text models, the best models for both modalities have comparable $precision$. Despite better performance, the best text models require fewer features than the best tweet models built on features from the same temporal quantity of data (except for 56 days of data).

Overall, the best model is LR with ten chi-squared selected features on 14 weeks of text data. From among the top models, this model has the highest values for all evaluation metrics (except $specificity$), requires the fewest features, and the smallest standard deviation of $F1$ scores. The distribution of $F1$ scores for LR models built on features from 14 weeks of data with around ten features is displayed in Figure 3.4.

3.1.4 Feature Importance Results

The ten features with the largest chi-squared statistics for 14 days of texts are depicted in Figure 3.5 with the normalized frequency of use for each participant. Features list item marker *LS*, *air travel*, *leader*, *pet*, and *politics* remained among the most influential features for all temporal quantities

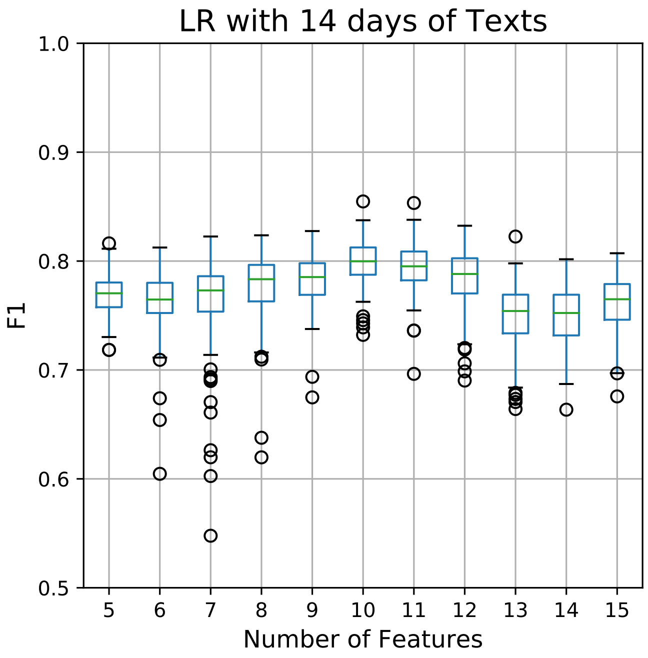


Figure 3.4: Boxplots of 100 LR models on 14 days of text messages.

of texts. None of the sentiment features were among the top features of the best models for any of the text datasets. Volume features are also absent within these models, suggesting the quantity of private messages seem to have little bearing on depression.

For comparison, the ten features with the largest chi-squared statistics for 14 days of tweets include *Subjective Count*, superlative adverb *RBS*, base form verb *VB*, past tense verb *VBD*, personal pronoun *PRP*, comparative adverb *RBR*, *attractive*, *fashion*, *fabric*, and *computer*. However, unlike for texts, many more features contributed to the best performing model for 14 days of tweets. Also, more POS frequencies were influential for tweets. Personal pronouns, a known indicator of depression in public text, is present among the top tweet features but not the top text features.

We can see from Fig. Figure 3.5 that individuals without depression tend to have a higher normalized usage of words in the last 14 days in certain word categories such as *air travel*, *leader*, *ship*, *real estate*, *competing*, and *exercise*. The *pet* category is the only word category where a higher frequency of use is visibly indicative of depression. Thus, it would appear that it is easier to identify individuals who are not depressed through certain high frequency word use than individuals who are depressed.

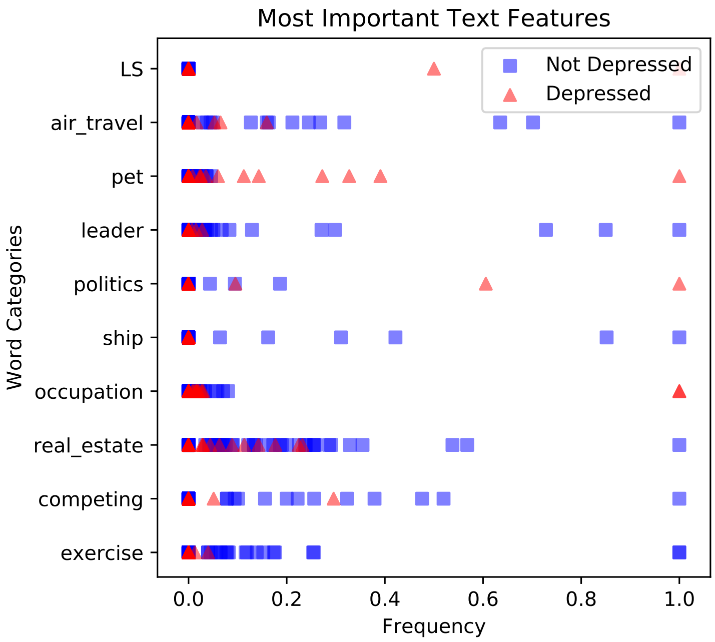


Figure 3.5: Use frequency of top 10 chi-squared features from 14 days of texts.

3.1.5 Discussion

Our study demonstrates that retrospectively harvested text messages have a great potential when screening for depression, more so than publicly posted tweets. Leveraging just text message content, we were able to screen for depression with an average F1 score of 0.8. The popularity of texting further makes it a viable modality to passively screen for depression for most individuals.

Two weeks versus one year of data. We explored the temporal quantity of self-written messages needed to screen for depression ranging from two weeks of messages to a year of data. For texts, the last two weeks of data proved most effective in predicting binary PHQ-9 scores. This is the same amount of time as patients are asked to reflect upon when completing the PHQ-9 [125]. While two weeks may be meaningful, this is the smallest temporal quantity of data we explored. Thus, perhaps the most recent messages are most informative of current mental health. Also, as seen in Table 3.1 not all participants in the datasets with more than two weeks of data had any recent messages. We recommend future studies proceed with two weeks of data or perhaps even less. Limiting the quantity of data reduces the quantity of private information contained within the messages, making it more likely that people will share the data. In addition to achieving the highest average F1 score, the text model with this quantity of data also required the fewest features. The tweet model for two weeks of data resulted in the best metrics of the tweet models in Table 3.2.

Private versus public text. It is not surprising we achieved a higher $F1$ score with private messages versus public messages. This can be explained by tweets being more biased and less personal than text messages as they are shared on a public forum. This may explain why people reported they are much more willing to share tweets than text messages [41]. Despite this finding, more participants in the dataset shared text messages than provided a valid Twitter username. We hypothesize this is due to texting being more popular than tweeting. The lack of tweet participants may have contributed to the lower F1 scores and larger $F1$ standard deviation of the tweet models. Despite this, our best tweet model achieved an average $F1$ score higher than other related works that utilized a screening tool [24], though some models had higher AUC and $accuracy$.

Benefits of machine learning. As mentioned, traditional machine learning has benefits over deep learning in psychiatry: privacy, interpretability, and applicability to smaller datasets. This interpretability allows important features to be easily identified, as seen in Figure 3.5. The $F1$ score of our best text model exceeds the $F1$ score of 0.77 in Alhanai, Ghassemi, and Glass' highly publicized study which predicted a binary PHQ-9 score at the same depression cutoff of 15 using a deep learning model on 142 multi-modal screening interviews [126]. Unlike the deep learning model, the logistic regression model is interpretable and computationally inexpensive. Also, text messages are much less intrusive and quicker to collect than interviews.

Considerations. The majority of participants elected not to share both tweets and texts, thus preventing us from building meaningful models with features engineered from both modalities. Additionally, some participants generated very few messages, perhaps interfering with classification ability. As such, we consider our results a lower threshold regarding the depression screening ability of text messages with machine learning models.

3.2 Suicidal Ideation with Longitudinal Text Messages

We hypothesize that text messages are a valuable modality to passively screen for suicidal ideation. Yet suicide risk is dynamic [127, 128], and it is critical to pinpoint when people are at highest risk for intervention. Thus, we examined the utility of screening based on briefer time intervals. Specifically, we tested if text messages from a particular week versus an aggregated sequence of multiple weeks prior to ideation prediction are most useful for this task. We demonstrate that our approach dramatically increases the suicidal ideation screening ability of data collected from smartphones. This research has three main contributions. First, we explore the potential of retrospectively collected crowd-sourced texts for suicidal ideation screening. Second, we compare the screening ability of texts sent in the interval and cumulative weeks prior to reported ideation. Lastly, we identify the most influential features for suicidal ideation screening.

3.2.1 Text Message Data

In this research we leverage data from the subset of participants in the Moodable and EMU datasets who shared longitudinal SMS text messages. Specifically, participants must have at least one text message in each of the prior two months. 66 participants in the datasets met this criteria. To determine the existence of suicidal ideation, we use item-9 of the PHQ-9. In response to the item-9, 39 participants selected 0, 11 participants selected 1, 14 participants selected 2, and 2 participants selected 3. We consider any positive score to be indicative of suicidal ideation. These participants sent $15,944$ SMS text messages during the 8 past weeks. The number of texts per participant ranged between 2 and 1709 with an average of 242 and median of 66.

We form 15 subsets of messages sent in the weeks prior to completion of the PHQ-9 screening survey (Table 3.4). The *interval* datasets D_i include the texts sent during the week W_i. The *cumulative* datasets C_i contain all texts sent in W_1 through W_i, i.e. all texts in W_i and all more recent weeks preceding the screening day. The data subsets may not contain the same number of participants as not all participants had records of sent SMS text messages on their phone for every

Table 3.4: The number of text messages sent by participants in the interval and cumulative weeks prior to reporting suicidal ideation with the ninth item of the PHQ-9.

	Interval Weeks		Cumulative Weeks	
Week	Participants	Texts	Participants	Texts
1	57	2349	57	2349
2	52	2381	62	4730
3	49	1961	62	6691
4	60	2280	66	8971
5	54	2018	66	10989
6	49	1821	66	12810
7	45	1933	66	14743
8	43	1201	66	15944

week. These interval and cumulative data subsets are identical in week W_1 as both contain the messages sent between day 1 to 7 prior to completion of the PHQ-9. The cumulative dataset for W_8 contains all $15,944$ texts.

3.2.2 Machine Learning Methodology

Our goal is to predict participant suicidal ideation with texts. We accomplish this by training machine learning models on features extracted from texts to predict if item-9> 0.

Feature engineering and selection. We extract 245 features from each data subset, using our feature extraction methodology detailed in subsection 3.1.2. The features include 195 word category frequencies, 36 part-of-speech (POS) tag frequencies, 9 sentiment features, and 5 volume features. We apply chi-squared feature selection [71] to reduce the number of features used when training the machine learning models.

Machine learning methods. After initial exploration, we leverage a selection of popular parametric and non-parametric methods with default parameters [71]: Gaussian Naive Bayes (NB), Logistic Regression (LR), Support Vector Classifier (SVC), and k-Nearest Neighbor (kNN).

Evaluation strategy. To mitigate bias from unbalanced classes, we down sample the data prior to training the machine learning models. We train models with the top 1 to top 20 chi-squared selected

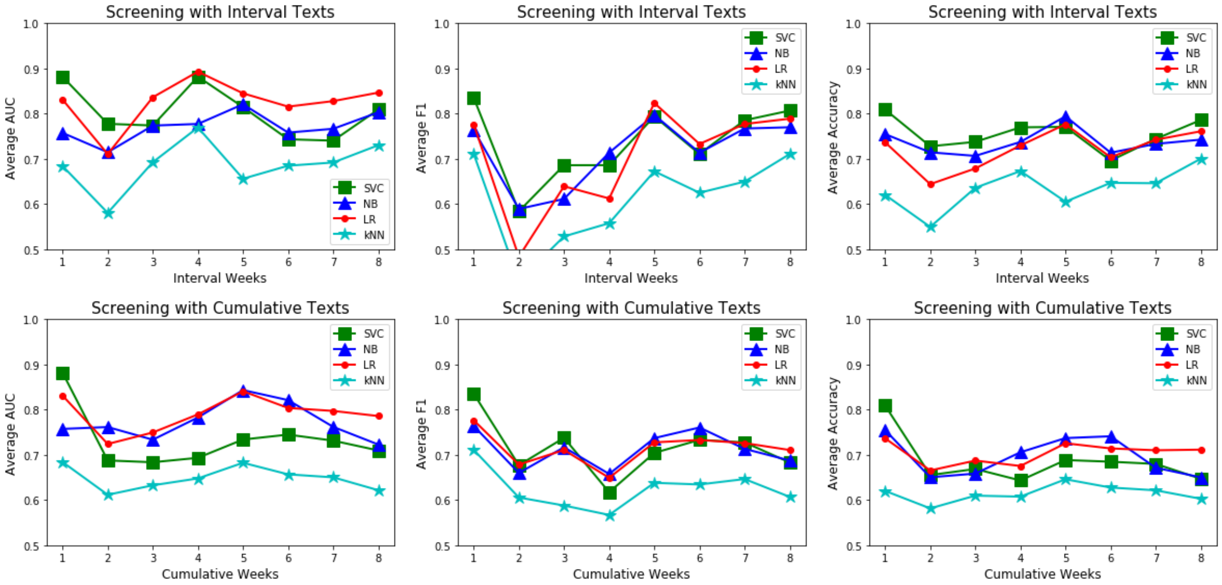


Figure 3.6: Comparison of the suicidal ideation screening ability for the model configurations with the highest average AUC, F1, and accuracy.

features. The models were trained with 5-fold cross-validation. We repeat each experiment 100 times and report the average evaluation metrics to ensure the results are robust.

3.2.3 Suicidal Ideation Screening Results

Figure 3.6 displays the average AUC, $F1$, and $accuracy$ metrics of the model configurations that performed best when screening for suicidal ideation. The highest values for each week are also displayed in Table 3.5. The data and results are the same for the first interval and cumulative week.

Impact of number of weeks. For all of the evaluation metrics, there is a notable decrease in the screening ability between the first and second weeks. For the cumulative weeks, the first week of data performs best across all metrics. While the fourth week of interval data achieved an average AUC of 0.89, this is not statistically significantly different than the average AUC of 0.88 achieved by using the first week of data. Thus we conclude that the best screening model is SVC trained on the one week of texts prior to reporting suicidal ideation.

Table 3.5: The suicidal ideation screening ability of the model configurations with highest average metrics for each subset of texts. For each metric, the average and standard deviation of the 100 scores for the model configuration are displayed.

	Interval Weeks			Cumulative Weeks		
Week	AUC	$F1$	$Accuracy$	AUC	$F1$	$Accuracy$
1	0.88 ±0.06	**0.84** ±0.05	**0.81** ±0.06	**0.88** ±0.06	**0.84** ±0.05	**0.81** ±0.06
2	0.78 ± 0.08	0.58 ± 0.10	0.73 ± 0.07	0.76 ± 0.09	0.68 ± 0.07	0.67 ± 0.08
3	0.84 ± 0.07	0.69 ± 0.12	0.74 ± 0.10	0.75 ± 0.08	0.74 ± 0.06	0.69 ± 0.07
4	**0.89** ±0.05	0.71 ± 0.11	0.77 ± 0.04	0.79 ± 0.07	0.66 ± 0.11	0.71 ± 0.09
5	0.85 ± 0.06	0.82 ± 0.04	0.79 ± 0.07	0.84 ± 0.07	0.74 ± 0.08	0.74 ± 0.06
6	0.82 ± 0.06	0.73 ± 0.08	0.71 ± 0.09	0.82 ± 0.07	0.76 ± 0.04	0.74 ± 0.08
7	0.83 ± 0.08	0.79 ± 0.07	0.74 ± 0.08	0.80 ± 0.07	0.73 ± 0.07	0.71 ± 0.07
8	0.85 ± 0.08	0.81 ± 0.08	0.79 ± 0.08	0.79 ± 0.08	0.71 ± 0.09	0.71 ± 0.08

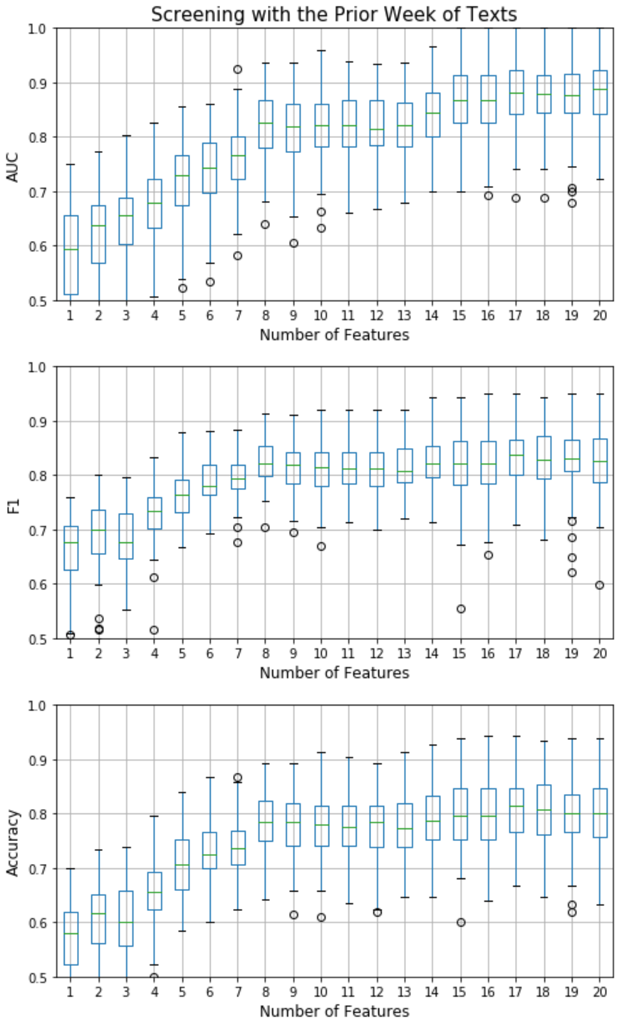


Figure 3.7: Suicidal ideation screening ability for the SVC models using the one prior week of texts with different quantities of chi-squared selected features.

Impact of method. While SVC models performed best for the first week of texts across all metrics, LR and NB models performed better for some other weeks. The kNN models were the worst at screening for suicidal ideation from texts for all metrics and weeks. This suggests a method that assumes a Gaussian or linear data distribution is more appropriate. This parallels screening for depression with texts [1] where SVC, NB, and LR were also the best performing models.

Screening with the prior week of texts. Figure 3.7 displays the distributions of AUC, $F1$, and $accuracy$ suicidal ideation screening scores for the SVC models trained on one week of texts. There is a notable increase at 8 and 15 chi-squared selected features for these metrics, though 17 features achieve the highest averages. Specifically, SVC with 17 features achieves average $AUC = 0.88$, $F1 = 0.84$, $accuracy = 0.81$, $sensitivity = 0.94$, and $specificity = 0.68$. With only 8 features, SVC can achieve $AUC = 0.82$, $F1 = 0.83$, $accuracy = 0.79$, $sensitivity = 0.94$, and $specificity = 0.61$. Even with less features, the $F1$, $accuracy$, and $sensitivity$ are similar. Both model configurations both have high average $sensitivity$.

Important features for screening. The top 8 chi-squared selected features for the prior week of texts are all lexical category frequencies: *car*, *clothing*, *affection*, *tool*, *confusion*, *driving*, *real estate*, and *journalism*. The top 17 features, depicted in Figure 3.8, include another 7 lexical category frequencies, average subjective score, and superlative adverb (RBS). A higher value of these lexical category frequency features is indicative of the participant not reporting suicidal ideation. The lexical categories *ship* and *real estate* were also important when screening for depression [1].

3.2.4 Discussion

Our research demonstrates that text messages are a promising modality to passively screen for suicidal ideation. With just the prior week of texts, the SVC model was able to screen for suicidal ideation with an $AUC = 0.88$, $F1 = 0.84$, $accuracy = 0.81$, $sensitivity = 0.94$, and $specificity = 0.68$. These results represent a dramatic increase in the suicide ideation screening ability of data collected from smartphones.

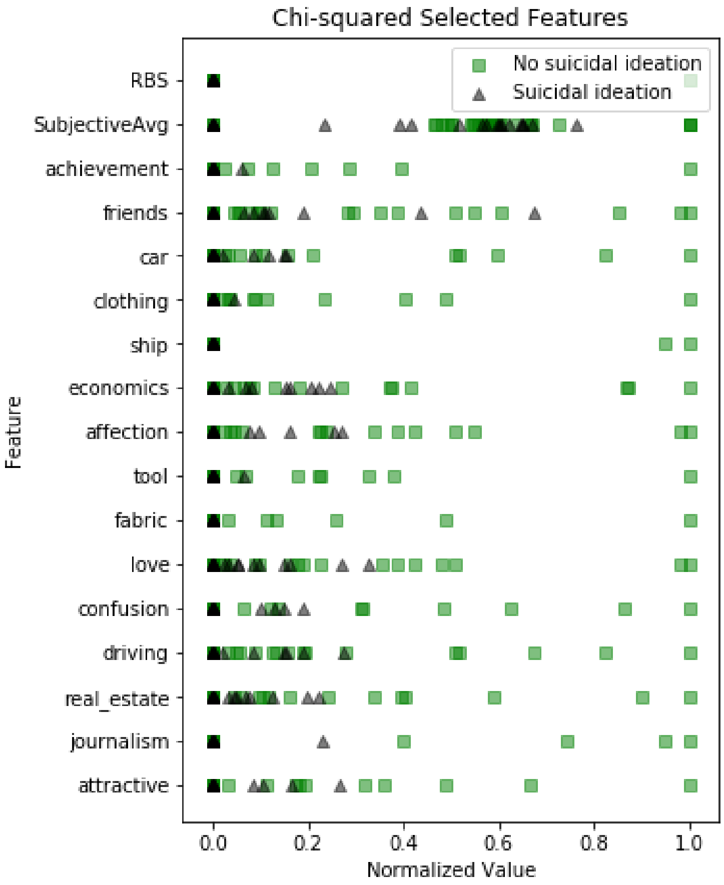


Figure 3.8: The 17 top chi-squared selected features for the prior week of texts; 15 are word category frequencies. For each feature, the normalized values for all 57 participants are displayed.

While we used SMS text messages, this research could apply to any direct messages. We recommend that future research on screening for suicidal ideation with self-written direct messages focuses on the last week of messages, extracts word category frequency features, and uses Gaussian or linear models. The small sample size of crowd-sourced participants with longitudinal texts was the main limitation of this research. Thus, future work could collect larger datasets of labeled direct messages to assess generalizability and test the effectiveness of more advanced models. It is also worth examining more time-sensitive measures of suicide ideation, behaviors, and risk.

3.3 Constructing Lexicons for Depression Screening

Lexical category features have proved useful for screening models [1, 2]. However, we hypothesize that the performance of the models were likely limited by the formal language in the pre-existing lexicons. Thus, in this research we automatically construct alternative lexicons that contain more colloquial terms to improve the depression screening capabilities of text messages. In particular, we derive seed words from the 194 existing Empath lexical categories [129] that can be used to identify related words in large text corpuses [130]. By exploring three strategies to identify seed words, three different quantities of seed words, and three linguistically different corpuses, we construct 27 distinct lexicons. We further combine the categories from the three corpuses for 9 more lexicons. We extract lexical category frequencies from text messages using our 36 lexicons.

To assess the usefulness of our constructed lexicons, we compare their screening results to those of models leveraging bag-of-words features and pre-existing lexical category features. This research thus serves to investigate the importance of lexicon construction when screening for depression with less formal text. There are four main contributions of this research. The first contribution is our proposed strategies for identifying seed words from existing lexical categories to automatically construct new categories. The second contribution is the 36 that we automatically constructed from three linguistically distinct corpuses. The third contribution is a comparison of the usefulness of the constructed lexicons to screen for depression with text messages. The last contribution is a discussion of the important features for depression screening.

3.3.1 Lexicon Construction Methodology

Our goal is to craft alternative lexicons that contain more informal language and thus are more linguistically appropriate for text message classification. The Empath software [129, 130] both provides 194 pre-existing lexical categories and the capability of generating lexical categories from different text corpuses with user specified seed words. We leverage this software in our approach which involves identifying seed words from the pre-existing categories and generating lexicons.

summer hiking cruise rental lake resort location suite traveling holiday boarding tourist coast observatory harbor ticket scenic seaside venue night villa adventure skiing getaway inland traveling hostel camping outside shoreline explore coastline condo luxurious abroad carnival restaurant casino packing tour promotion yearly honeymoon touring seashore limo coastal campground sightseeing spending museum destination visit **vacation** accommodation tropical expressway hangout hotel brochure ride inn condominium nightlife fun ferry excursion secluded lakeside overnight upscale rent spa trip flight travel airport beach surf countryside stay outback journey lax plan waterfront reservation weekend surfing outing yacht drive ocean shore landmark overseas spend nightclub

Table 3.6: The 98 words in the Empath's pre-existing *vacation* category [129].

Identifying seed words. In order to generate lexical categories with the Empath software, we must first specify seed words. These seed words are used by Empath's vector space model to identify the most related words in the text corpus. This is done by calculating the cosine similarity to identify the most similar words to the provided seed word or the vector sum of the seed words [129]. We derive the seed words from the 194 pre-existing Empath lexical categories. Each category has a list of words contained within that category. For example, the 98 words that comprise the category *vacation* are displayed in Table 3.6. We extract 1, 3, and 5 seed words from the words contained in each category to generate alternative categories; a single word may result in an overly broad category but too many seed words may result in an insufficient number of related words. We propose three different strategies to identify seed words:

- Closest (c): we select the words closest to the category name in the category word list. For the category *vacation*, the five closest words are 'destination', 'visit', 'vacation', 'accommodation', and 'tropical'. For the 18 categories where the category name was not in the word list, we manually specified a similar replacement word from within the word list.

- First (f): we select the first words in the category word list. For the category *vacation*, the five first words are 'summer', 'hiking', 'cruise', 'rental', and 'lake'.

- Random (r): we select random words in the category word list. For the category *vacation*, five random words are 'airport', 'carnival', 'location', 'visit', and 'ocean'.

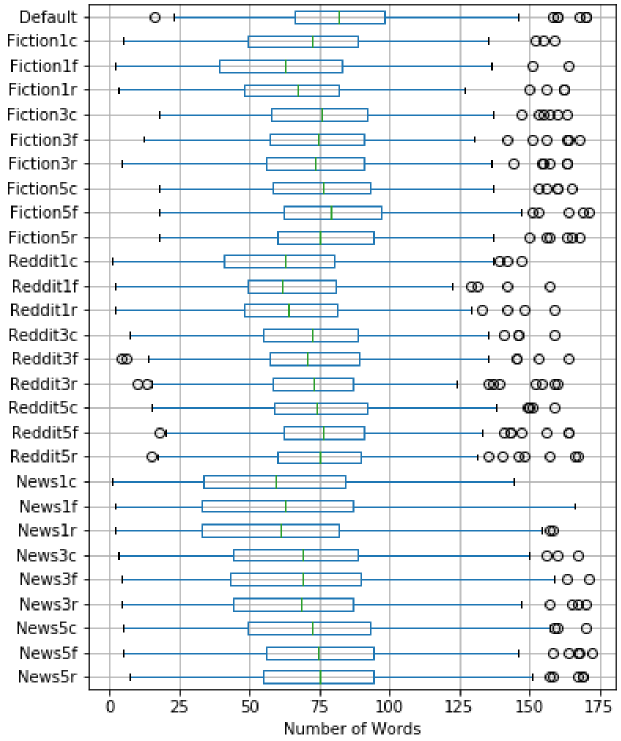


Figure 3.9: The number of words in the categories for each lexicon.

Generating new lexicons. The Empath software [130] can generate new lexical categories from three different text corpuses: amateur modern fiction (Fiction), Reddit posts (Reddit), and The New York Times articles (News). These corpuses are very different linguistically so we generate alternative lexical categories from each of them to determine which is most useful to screen for depression from text messages. We set the size of the alternative categories to be the same size as the pre-existing Empath lexical categories. However, not all seed word combinations resulted in that many words, so some of the generated lexicons contain fewer words. If the seed words are not in the list of most related words, we add the seed words to the list. For each lexicon, we compare the number of words in each of their 194 categories in Figure 3.9. While there is a large variation in the number words in the pre-existing Empath lexical categories, some combinations of seed words produced categories with even fewer words. Lastly, we combine the words in the lexical categories across the corpuses, which we refer to as the combined lexicons. For example, Combined5r contains the union of words in Fiction5r, Reddit5r, and News5r. These combined categories theoretically will provide the linguistical benefits of each corpus.

3.3.2 Feature Extraction Methodology

Data. For this research we leverage the retrospectively harvested text message logs in the Moodable and EMU datasets. For our analysis, we consider the 88 participants in the datasets who sent at least 5 texts within the two weeks prior to completing the PHQ-9 screening survey. The 88 participants shared a total of 7914 sent text messages during the last two weeks. Overall, 53 (60%) of the participants screened positive for depression. Since we are screening for participant depression, we aggregate the text messages sent by each participant. We then extract features from these 88 self-written text passages. Note, we remove all capitalization and punctuation from the text messages prior to feature extraction.

Lexical categories. For each of the 194 categories in the lexicons, we tally the number of instances a participants uses a word in their text passage that are exact matches to words in the category. We then divide this count by the total number of words. In the case of a n-gram word with $n > 1$, the word still counts as a single instance in the text, as defined by the Empath software [129]. We extract lexical category features for each of our 9 Fiction lexicons, 9 Reddit lexicons, 9 News lexicons, and 9 Combined lexicons. This results in 36 unique sets of 194 lexical features.

Default lexicon. To provide a baseline for our lexicons, we use the Empath software [129] to analyze the participant text passages. This results in a set of lexical category features for the preexisting Emapth lexicon, which we further refer to as the Default lexicon in our comparisons.

Bag-of-words. Since the most basic strategy to create features from text data is bag-of-words, we include this strategy as another baseline. In this feature engineering approach, each unique word is a feature. Thus, the number of features is the number of unique words in our entire dataset. The number of times a participant uses a given word is tallied and that count is the value for that feature. Since lexical categories do not capture numerical characters in the text, we remove all words that began with or consisted only of numerical characters. Our participants collectively texted $6,248$ unique words. Thus, the bag-of-words approach resulted in a very sparse feature matrix.

3.3.3 Feature Selection and Machine Learning Methodology

We evaluate the usefulness of the new lexicons by screening for participant depression with the lexical category features. Specifically, the goal of the machine learning models is to screen for moderate depression (PHQ-9$\geq$ 10) from the participants. While this is a standard cutoff [92], is a different cutoff than previous related research [1].

Evaluation strategy. We perform leave-group-out cross validation by creating 100 different test sets with replacement for each lexicon to demonstrate result robustness. The test sets were stratified in respect to the binary depression screening label to ensure the test sets were representative. We normalize the training sets prior to feature selection and apply that transformation to their respective test sets. we consider the best performing models to be those that maximize the average F1 score across all 100 test sets.

Feature selection and data balancing. For each of the crafted lexical feature sets, we experiment with using between 1 and 10 chi-squared selected features for the depression screening models. The feature selection transformation is learned individually for each of the 100 training sets and then applied to their respective test sets. We upsample the training sets to balance the two classes prior to training the machine learning models. No balancing is performed on the test sets.

Machine learning methods. We compare the screening ability of the different sets of lexical category features with support vector classifiers (SVC), logistic regression (LR), Gaussian Naive Bayes (NB), and k-Nearest neighbor (kNN) [71]. We experiment four different SVC kernel functions which we refer to as Gaussian SVC, linear SVC, polynomial SVC, and sigmoid SVC. We also assess the depression screening ability of kNN with 3 and 5 neighbors, which we further refer to as kNN3 and kNN5, respectively.

Figure 3.10 and Table 3.7 showcase the evaluation metrics for the models configurations that maximize the average F1 score for each of the 38 feature sets. The constructed lexicon Fiction5f was the most successful at screening for depression with an average F1 score of 0.79. In com-

Table 3.7: The average ± standard deviation of the metrics for the models configurations with the highest average F1 scores for each lexicon. The p-values (P) are derived from the t-tests that compare the F1 scores from each lexicon with the F1 scores from the bag-of-words (B) and default Empath lexicon (D). The models that are statistically significantly better than the baselines based on these t-tests are marked with asterisks. The lexical categories were generated with seed words that were the 1 to 5 closest (c), first (f), and random (r) words to the names of the categories. The six best models are bolded.

Lexicon	Method	F	F1	Sensitivity	Specificity	Accuracy	F1 P(B)	F1 P(D)
Bag-of-words	Linear SVC	3	0.74±0.04	0.90±0.08	0.23±0.13	0.62±0.06	1.0	0.119
Default	Linear SVC	1	0.72±0.08	0.93±0.13	0.09±0.14	0.59±0.05	0.119	1.0
Fiction1c	Linear SVC	1	0.73±0.04	0.92±0.08	0.13±0.14	0.60±0.05	0.161	0.463
Fiction1f	Polynomial SVC	1	0.75±0.02	0.96±0.04	0.11±0.08	0.62±0.03	0.019*	0.003*
Fiction1r	Polynomial SVC	1	0.74±0.03	0.95±0.06	0.14±0.11	0.62±0.04	0.158	0.015*
Fiction3c	Polynomial SVC	1	0.73±0.03	0.92±0.07	0.14±0.12	0.60±0.04	0.330	0.286
Fiction3f	**LR**	**1**	**0.77±0.06**	**0.88±0.09**	**0.41±0.21**	**0.69±0.09**	**<0.001***	**<0.001***
Fiction3r	Linear SVC	1	0.73±0.03	0.95±0.07	0.05±0.07	0.58±0.03	0.073	0.543
Fiction5c	Polynomial SVC	1	0.73±0.03	0.93±0.07	0.12±0.11	0.60±0.05	0.492	0.226
Fiction5f	**LR**	**1**	**0.79±0.05**	**0.93±0.06**	**0.40±0.18**	**0.71±0.07**	**<0.001***	**<0.001***
Fiction5r	Linear SVC	1	0.73±0.08	0.93±0.12	0.11±0.13	0.60±0.06	0.292	0.700
Reddit1c	**NB**	**1**	**0.76±0.04**	**0.96±0.07**	**0.15±0.10**	**0.63±0.06**	**0.002***	**<0.001***
Reddit1f	kNN3	1	0.71±0.05	0.89±0.09	0.11±0.10	0.57±0.05	<0.001	0.215
Reddit1r	Polynomial SVC	1	0.74±0.03	0.97±0.07	0.05±0.07	0.60±0.04	0.361	0.031*
Reddit3c	Linear SVC	1	0.74±0.03	0.96±0.06	0.07±0.08	0.60±0.04	0.984	0.100
Reddit3f	kNN5	1	0.72±0.05	0.90±0.10	0.15±0.10	0.59±0.06	0.023	0.922
Reddit3r	Linear SVC	1	0.73±0.03	0.93±0.08	0.10±0.11	0.59±0.04	0.171	0.424
Reddit5c	Polynomial SVC	1	0.75±0.03	0.97±0.06	0.10±0.09	0.62±0.04	0.019*	0.003*
Reddit5f	Polynomial SVC	1	0.74±0.05	0.95±0.10	0.10±0.12	0.60±0.05	0.990	0.124
Reddit5r	Sigmoid SVC	3	0.76±0.06	0.88±0.10	0.35±0.15	0.66±0.08	0.010*	0.001*
News1c	**NB**	**1**	**0.77±0.04**	**0.97±0.06**	**0.20±0.11**	**0.66±0.06**	**<0.001***	**<0.001***
News1f	Linear SVC	1	0.72±0.08	0.93±0.01	0.06±0.12	0.58±0.04	0.022	0.604
News1r	kNN3	1	0.75±0.04	0.95±0.07	0.13±0.08	0.61±0.05	0.234	0.021*
News3c	kNN3	1	0.75±0.05	0.96±0.08	0.13±0.10	0.62±0.06	0.036*	0.004*
News3f	Polynomial SVC	1	0.73±0.04	0.94±0.08	0.07±0.11	0.59±0.05	0.138	0.483
News3r	Polynomial SVC	1	0.72±0.03	0.92±0.08	0.07±0.08	0.58±0.04	0.001	0.706
News5c	**NB**	**1**	**0.76±0.04**	**0.97±0.06**	**0.16±0.08**	**0.64±0.06**	**<0.001***	**<0.001***
News5f	Polynomial SVC	1	0.74±0.06	0.96±0.10	0.07±0.11	0.60±0.05	0.964	0.170
News5r	Linear SVC	1	0.72±0.03	0.91±0.07	0.08±0.10	0.57±0.04	<0.001	0.487
Combined1c	Linear SVC	1	0.73±0.04	0.93±0.08	0.12±0.10	0.60±0.05	0.587	0.207
Combined1f	Polynomial SVC	1	0.72±0.08	0.90±0.12	0.16±0.14	0.60±0.07	0.132	0.934
Combined1r	kNN3	1	0.75±0.05	0.88±0.10	0.33±0.16	0.65±0.07	0.066	0.007*
Combined3c	Gaussian SVC	1	0.74±0.05	0.89±0.09	0.26±0.17	0.63±0.06	0.725	0.080
Combined3f	Polynomial SVC	1	0.73±0.04	0.92±0.08	0.15±0.13	0.61±0.05	0.709	0.176
Combined3r	Linear SVC	1	0.72±0.04	0.93±0.09	0.07±0.09	0.58±0.05	0.026	0.911
Combined5c	Linear SVC	1	0.73±0.03	0.94±0.07	0.11±0.09	0.60±0.05	0.735	0.158
Combined5f	LR	1	0.73±0.09	0.85±0.14	0.30±0.18	0.63±0.08	0.302	0.712
Combined5r	**LR**	**1**	**0.77±0.06**	**0.92±0.09**	**0.34±0.14**	**0.68±0.08**	**<0.001***	**<0.001***

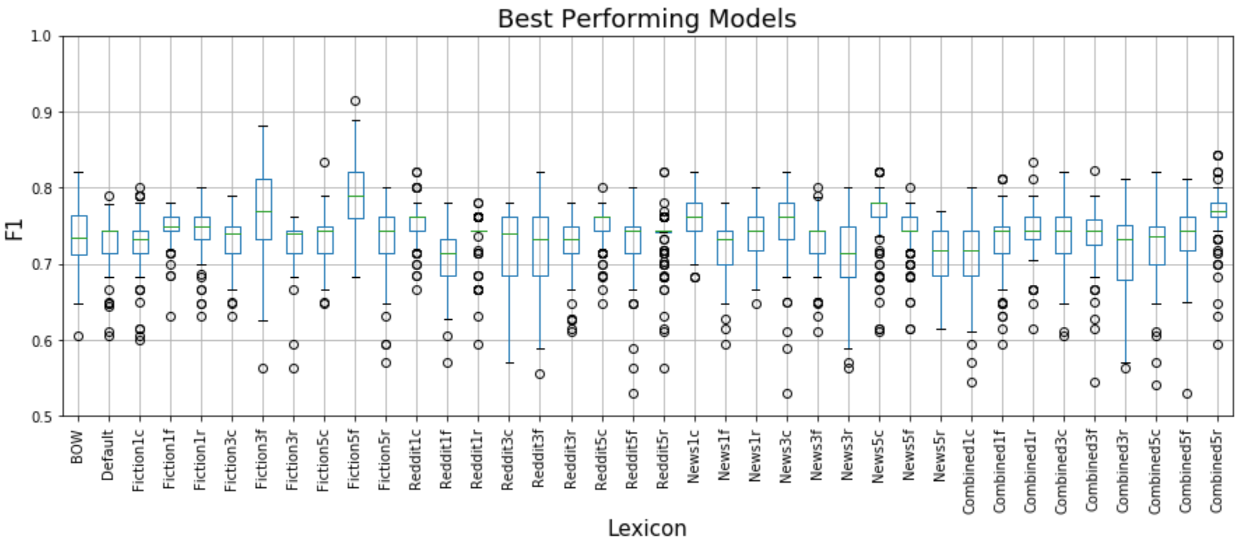


Figure 3.10: The results of the model configurations with the the highest average F1 scores for each lexicon. Bag-of-words (BOW) and Reddit 5c used 3 chi-squared selected features while the other lexicons used only 1 chi-squared selected feature. The lexical categories were generated with seed words that were the 1 to 5 closest (c), first (f), and random (r) words to the category names.

parison our baselines only achieved average F1 scores of 0.74 and 0.72. A t-test revealed that the F1 scores for lexicon Fiction5f were statistically significantly better than the F1 scores for both baselines with p-values < 0.001.

Number of features and method. The best performing model configurations in Table Table 3.7 used all eight machine learning methods. For 23 of the 38 feature sets, linear SVC and polynomial SVC performed the best. In contrast, Gaussian SVC, Sigmoid SVC, and kNN5 were only the best models for a single feature set. Surprisingly, a single chi-squared feature was sufficient to achieve the highest average F1 score for 36 of the 37 lexicons. The exception was Reddit5r whose average F1 score decreases from 0.76 to 0.75 when using only one feature.

Baselines. Our two baselines are bag-of-words and the default Emapth lexicon. The best average F1 scores for both were achieved with Linear SVC models. Unexpectedly, despite the very sparse feature matrix, bag-of-words had a higher average F1 score than the default lexicon, though this difference was not statistically significant according to a t-test (p-value = 0.119). Specifically, three bag-of-word features and one default lexical category feature screened for depression with F1 scores of 0.74 and 0.72, respectively.

Comparison of baselines versus constructed lexicons. 10 of the 36 constructed lexicons had F1 scores that were statistically significantly better than the bag-of-words F1 scores. Likewise, 14 of the 36 constructed lexicons performed statistically significantly better than the Default lexicon. Fiction1f, Reddit1r, News1r, and Combined1r had F1 scores that were statistically significantly better than the Default lexicon but not bag-of-words. The average F1 scores of these constructed lexicons were between 0.74 and 0.75. Reddit1f, Reddit3f, News1f, News3r, News5r, Combined3r are also notable for performing statistically significantly worse than bag-of-words. Thus, it was not advantageous to pair the first seed word identification strategy with the Reddit corpus nor the random seed word identification strategy with the News corpus. It is worth noting that only Reddit1f out of the constructed lexicons had a lower average F1 score than the Default lexicon.

Constructed lexicons. From the results in Table 3.7 and Figure 3.10, we discern certain patterns from the screening ability of our constructed lexicons. For lexicons generated using the Fiction corpus, we notice those that performed the best were created with seed words that were the first words within a category. This was not true for the lexicons that were generated using the Reddit or News corpus. For the News lexicons, those created with seed words closest to the category name performed the best. This is also mostly true for the Reddit lexicons, though the F1 scores are similar for the closest and random seed word identification strategies. Only for Combined lexicons did the random identification strategy perform comparatively well with just one feature.

Combined lexicons. We had expected the combined lexicons to retain the linguistically benefits for all three of the corpuses. However, only one of the Combined lexicons was among the six best models, namely Combined5r. With an average F1 score of 0.77, Combined5r performed better than the three lexicons that contributed to it. The average F1 scores were 0.73, 0.76, and 0.72 for Fiction5r, Reddit5r, and News5r.

Table 3.8: Comparison of the average F1 scores for each method trained on features from the most successful lexicons in Table Table 3.7. The p-values (P) are derived from the t-tests that compare the F1 scores from each of the best lexicons with the F1 scores from the bag-of-words (B) and default Empath lexicon (D).

	Logistic Regression			Naive Bayes			kNN with k=3			kNN with k=5		
Lexicon	F1	P(B)	P(D)	F1	P(B)	P(D)	F1	P(B)	P(D)	F1	P(B)	P(D)
Bag-of-words	0.72	1.0	0.019*	0.72	1.0	0.044*	0.72	1.0	0.238	0.72	1.0	0.014*
Default	0.70	0.019*	1.0	0.70	0.044*	1.0	0.70	0.238	1.0	0.70	0.014*	1.0
Fiction3f	0.77	<0.001*	<0.001*	0.77	<0.001*	<0.001*	0.74	0.005*	<0.001*	0.74	0.013*	<0.001*
Fiction5f	0.79	<0.001*	<0.001*	0.79	<0.001*	<0.001*	0.77	<0.001*	<0.001*	0.77	<0.001*	<0.001*
Reddit1c	0.75	<0.001*	<0.001*	0.76	<0.001*	<0.001*	0.75	<0.001*	<0.001*	0.75	<0.001*	<0.001*
News1c	0.77	<0.001*	<0.001*	0.77	<0.001*	<0.001*	0.77	<0.001*	<0.001*	0.77	<0.001*	<0.001*
Newc5c	0.76	<0.001*	<0.001*	0.76	<0.001*	<0.001*	0.76	<0.001*	<0.001*	0.76	<0.001*	<0.001*
Combined5r	0.77	<0.001*	<0.001*	0.77	<0.001*	<0.001*	0.76	0.001*	<0.001*	0.76	<0.001*	<0.001*

	Gaussian SVC			Linear SVC			Polynomial SVC			Sigmoid SVC		
Lexicon	F1	P(B)	P(D)	F1	P(B)	P(D)	F1	P(B)	P(D)	F1	P(B)	P(D)
Bag-of-words	0.72	1.0	0.064	0.74	1.0	0.119	0.73	1.0	0.120	0.72	1.0	0.178
Default	0.70	0.064	1.0	0.72	0.119	1.0	0.72	0.120	1.0	0.70	0.178	1.0
Fiction3f	0.76	<0.001*	<0.001*	0.76	<0.001*	<0.001*	0.77	<0.001*	<0.001*	0.74	0.032*	<0.001*
Fiction5f	0.79	<0.001*	<0.001*	0.78	<0.001*	<0.001*	0.78	<0.001*	<0.001*	0.75	<0.001*	<0.001*
Reddit1c	0.75	<0.001*	<0.001*	0.75	0.010*	0.002*	0.75	<0.001*	<0.001*	0.75	<0.001*	<0.001*
News1c	0.77	<0.001*	<0.001*	0.76	<0.001*	<0.001*	0.76	<0.001*	<0.001*	0.77	<0.001*	<0.001*
Newc5c	0.76	<0.001*	<0.001*	0.75	<0.001*	<0.001*	0.75	<0.001*	<0.001*	0.76	<0.001*	<0.001*
Combined5r	0.77	<0.001*	<0.001*	0.77	<0.001*	<0.001*	0.77	<0.001*	<0.001*	0.77	<0.001*	<0.001*

The best six lexicons. There were six constructed lexicons that performed notably better than the other lexicons with only one feature. These are Fiction3f, Fiction5f, Reddit1c, News1c, News5c, and combined5r. This leads us to conclude that combining Fiction with the first seed word identification strategy and News with the closest seed word identification strategy were particularly effective for depression screening with text messages. Table 3.8 illustrates the performance of each method for these six best lexicons.

Robustness of the best six lexicons. For every method, the best six lexicons perform statistically significantly better than both baselines according to t-tests with p-values < 0.001 (with the exception of Fiction3f with kNN5 where $P(B) = 0.013$ and Sigmoid SVC where $P(B) = 0.032$). While Gaussian Naive Bayes was not always the best model in Table 3.7 due to higher standard deviation, the average F1 scores notably match that of the best model as displayed in Table 3.8. This makes sense as Naive Bayes algorithms are known to succeed at document classification. However, the impact of the method pales in comparison to the impact of the corpus, number of seed words, and seed word identification strategy.

3.3.4 Important Features for Depression Screening

We apply chi-squared feature selection on the data from all participants to identify the important features. For bag-of-words, the three most important words are ‘looks’, ‘monday’, and ‘likely’. For the default Empath lexicon, the most important category is *negotiate* with the word ‘sell’ used 17 times and the word ‘price’ used 9 times in the text message dataset. Additionally, the words ‘trade’, ‘debt’, ‘guarantee’, ‘compensation’, ’reasonable’, ’negotiation’, ’mortgage’, ’loan’, ’settlement’, and ’barter’ are used three times or less.

For fiction5f, the lexicon that proved most useful for depression classification with text messages, the most important category was *vacation*. The words used in this category are ‘chicago’: 16, ‘town’: 15, ‘downtown’: 10, ‘harbor’: 10, ‘beach’: 8, ‘trip’: 7, ‘vacation’: 5, ‘city’: 5, ‘holiday’: 5, ‘florida’: 5, ‘park’: 5, ‘winter’: 4, ‘california’: 4, ‘summer’: 3, ‘farm’: 3, ‘hiking’: 3, ‘pittsburgh’: 2, ‘fishing’: 2, ‘colorado’: 2, ‘hike’: 1, ‘boats’: 1, ‘countryside’: 1, ‘aquarium’: 1. Unlike the words in the pre-existing Empath category (Table 3.6), the words in the Fiction5f category notably contain names of cities and states.

3.3.5 Discussion

We have improved the mental illness screening capabilities of text messages by introducing lexicons that are better able to classify such informal communications. Our best lexicon increased the average F1 score by 10% over the pre-existing lexicon. Overall, 14 of our 36 constructed lexicons performed statistically significantly better at depression screening with text messages than the pre-existing lexicon. We found that it was particularly advantageous to pair the fiction corpus with the first seed word identification strategy.

Our constructed lexicons could be used to improve classification of other informal text datasets within and outside of the mental health domain. Further, future work may wish to generate customized categories for specific domains. For instance, text messages classifiers may benefit from a category that captures emoticons or numerical terms.

3.4 Depression Screening with Texts from Contact Subsets

The related research has focused on screening with text generated by the participant rather than text generated by contacts of that participant. However, another study discovered that both the quantity and quality of social interactions are known predictors of health [131]. Specifically, negative interactions have been found to be more influential than positive interactions [132]. Additionally, friendships are important to happiness [133] and the number of contacts an individual is comfortable with is associated with better mental health [132], though research indicates people only have three close friends [133]. Likewise, fewer close relationships and lack of social support are associated with depression [134].

Given the influence of social interactions and close relationships on mental health, we hypothesize that received (in contrast to sent) communications will be useful in predicting depression scores. In addition, we hypothesize that communications from a subset of top contacts will be more predictive of depression scores than communications from all contacts. We explore these hypotheses with received smartphone text messages collected in a crowd-sourced study [41]. Our approach consists of creating contact subsets, feature engineering, and machine learning. Our contributions include 1) exploring the potential of received texts, an underutilized modality, in predicting depression screening scores, and 2) comparing the predictive ability of features generated from subsets of top contacts with features from all contacts.

3.4.1 Data

Our study leverages a specific subset of the Moodable dataset [41] consisting of received text messages and PHQ-9 scores, data types which were available for 313 participants. The number of contacts for each of these participants ranges from between 1 and 420 contacts.

3.4.2 Contact Subset Creation

The quantity of social interaction is a known predictor of mental health [131] which fuels our hypothesis that the quantity of text messages sent by a contact relates to the influence exerted on the participant. To test this hypothesis, we explore the impact of utilizing text messages from 11 specific subsets of the top C contacts, including messages from the top one contact and from all contacts. Specifically, we examine three different techniques of calculating the top C contacts, denoted $CP(a)$, $CP(r)$, and $CP(w)$ for participant P.

- For $CP(a)$ contacts, we simply consider the top a contacts for each participant P. As the number of contacts an individual is comfortable with is associated with better mental health [132] and most people only have three close friends [133], we explore $a \in \{2, 3, 4\}$. If a participant does not have a contacts, we use all contacts: $1 \leq CP(2) \leq CP(3) \leq CP(4)$.

- For $CP(r)$ contacts, we consider the top r percent of contacts for each participant P. As the number of contacts for each participant ranges from 1 to 447, the number of top contacts CP with influence may be different for each participant P. We explore $r \in \{25\%, 50\%, 75\%\}$. $CP(r) \geq 1$ for every percent r and participant P.

- For $CP(w)$ contacts, we calculate the number of contacts using Equation 3.1 with weight w for participant P. We explore weight $w \in \{0.25, 0.5, 0.75\}$. Note, $1 \leq CP(0.75) \leq CP(0.5) \leq CP(0.25)$. Similar to $CP(r)$, $CP(w)$ may differ for each participant P. However, the calculation for $CP(w)$ contacts also considers the number of text messages sent by each participant P.

$$C_p(w) = \sum_{i=1}^{n} for F_i = \begin{cases} 1 & \text{if } t_i \geq w(max(T)) \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where w is the user given weight and for Participant P, t_i is the number of texts for the i^{th} contact, T is the set containing the number of texts for all contacts, and n is the total number of contacts.

3.4.3 Text & Emotion Feature Engineering

For each participant, we extract features involving polarity, subjectivity, part of speech (POS) tags, and volume from text messages sent by the aforementioned contact subsets. As there are 11 contact subsets, this forms 11 datasets.

1. *Polarity:* Negative social interactions are associated with depression [132]. As such, we generate features based on the polarity of the text messages for each subset of contacts. We use TextBlob [122] to generate a polarity score p between -1 and 1 for each text message. Text messages with positive and negative polarity are those with $p > 0$ and p¡0, respectively. We calculate the percent and average polarity of positive text messages as well as the percent and average polarity of negative text messages.

2. *Subjectivity:* We also use TextBlob [122] to extract the subjectivity score s of each text message, which ranges between 0 and 1. Text messages with subjective content are those with $s > 0$. We calculate the percent and average subjectivity of subjective text messages.

3. *Part of Speech Tags:* POS tags can capture linguistic style [54]. We use TextBlob [122] to generate $(word, POS)$ pairs. For each of the present 36 types of POS tags, we calculate the percent of that POS tag for each participant.

4. *Volume:* The quantity of social interaction is also important to health [131]. Thus, we calculate the number of contacts, number of text messages, and average number of POS tags per text for each participant P. When considering text messages from the top one contact, we use the total number of contacts.

3.4.4 Machine Learning Methodology

This research predicts whether a PHQ-9 score is greater than 10. To mitigate the unequal distribution of PHQ-9 scores, we apply down-sampling to generate a balanced dataset with 145 participants per class. Our models will be built from the 45 polarity, subjectivity, part of speech, and volume features previously discussed. We have 11 datasets created by generating these features from the texts of the top $1, C(2), C(3), C(4), C(25\%), C(50\%)$, $C(75\%)$, $C(0.25)$, $C(0.5)$, $C(0.75)$, and all contact(s). Every machine learning method is run with all 11 of these datasets. The machine learning methods include kNN, SVC, RF, AdaBoost, XGBoost, LR, and NB. We experiment with 1, 3, 5, 7, and 9 neighbors for kNN; linear, polynomial, Gaussian, and sigmoid kernels for SVC; 1, 2, 3, and 4 maximum tree depth for RF and XGBoost; and Gini impurity and entropy split criterion for RF. Each model uses 5-fold cross-validation. To ensure result robustness and mitigate the effects of random down-sampling, each experimental procedure was repeated 100 times. We evaluate these models with the F1 score.

3.4.5 Results

For every method, we identify the parameter setting and dataset with the highest average F1 score for the 100 trials. The highest average F1 score is above 0.5 for every method. The best performing method is Gaussian Naive Bayes with an F1 score of 0.653. The next best performing method, kNN with k=7, only had an F1 score of 0.596. As such, we proceeded with Gaussian Naive Bayes to analyze the predictive ability of the contact subsets.

Figure 3.11 compares the distribution of F1 scores of 100 Gaussian Naive Bayes models for each of the 11 contact subsets. The models with the highest average F1 scores are those created with features from the top $CP(25\%)$ contacts and $CP(0.25)$ contacts, both with an average F1 score of 0.653. In comparison, the average F1 score for the models created with features from all contacts is 0.577.

By using only features from the top $CP(25\%)$ or $CP(0.25)$ contacts, the F1 score is improved by 13.2 percent in comparison to using features from all contacts. From t-tests, we conclude that models with features derived from the top $CP(25\%)$ contacts and $CP(0.25)$ contacts have statistically significantly higher F1 scores than models with features derived from all contacts with p¡0.0001. Thus, for the task of depression screening, models generated with messages from a subset of contacts outperform models generated with messages from all contacts.

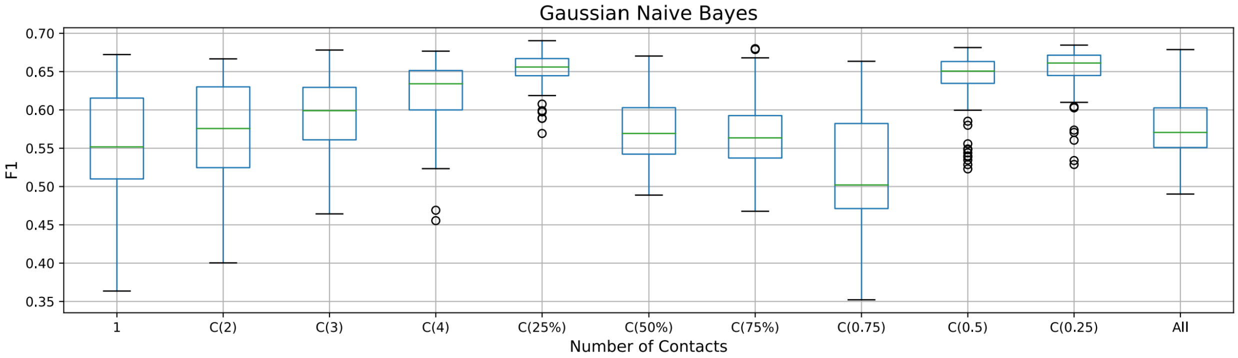


Figure 3.11: F1 scores of 100 Gaussian Naive Bayes models for every contact subset.

3.4.6 Discussion of Implications & Considerations

Leveraging only received text messages, we can predict the binary PHQ-9 score at cutoff 10 with an F1 score of 0.653. The results demonstrate that received communications can be useful in predicting PHQ-9 scores. This suggests received communications are a viable modality, particularity if combined with other modalities. Additionally, we have shown that features derived from the top $CP(25\%)$ and $CP(0.25)$ contacts are statistically significantly better than features derived from all contacts by an improvement in F1 score of 0.076, or 13.2 percent. This research introduces using Equation 3.1 to identify the number of influential contacts. Overall, focusing on a subset of contacts is a promising approach that could be deployed in any study leveraging received communications. Text messages are limited as they do not capture all social interactions. Thus, analyzing text messages is likely most effective for people who rely on texting as a main form of communication. Unfortunately, some of the participants in the Moodable dataset had a minimal number of text messages, likely lowering the screening potential of the received messages.

3.5 Outlook

Ethical considerations and mitigations. The main downside to using text messages to screen for mental illnesses is that they contain private content. This poses some ethical issues in using text messages for screening. Many people are also unwilling to share the content of their text messages [41, 12, 135, 15]. However, there are ways to conduct screening without the message content leaving the phone. For example, features or feature embeddings could be extracted from the texts, and then that extracted information could be transmitted to the database for storage [136]. While this would work in an implementation, it makes research difficult as the features or embedding model must be decided on beforehand. Likewise, in the advent of a trained model, such a model could be deployed on a mobile device [137].

Text messages versus tweets. This research has shown that traditional machine learning models are able to achieve and $F1$ score of 0.8 at PHQ-9 cutoff 15 and 0.74 at PHQ-9 cutoff 10 when leveraging preexisting lexical category features derived from sent text message content [1, 3]. These F1 scores exceed those of models leveraging labeled tweets that were collected in the same manner [1]. Despite tweets being more commonly used in screening research [24, 25], more people text than tweet. In fact, 97% of US adults texted weekly in 2015 [138]. While Twitter remains a relatively popular social media platform [139], only 25% of users in the US produce 97% of the majority of tweets with only 18% of the tweets being original posts [140]. Thus, in more ways than one, text messages seem to be a better mental illness screening modality than tweets.

Text messages versus interview transcripts. We replicated our text message feature engineering strategy [1] to extract features from the transcripts of mobile voice recordings [12, 15, 18]. Using a leave-one-out evaluation strategy, the features from crowdsourced participants' transcripts achieved F1 scores of 0.54, 0.63, and 0.55 to screen for depression, anxiety, and suicidal ideation, respectively [12]. Likewise, the features from student participants' transcripts achieved F1 scores of 0.57 and 0.38 when screening for depression and suicidal ideation, respectively [18]. Combining the two datasets increased the F1 score of the depression screening models to 0.63 [18]. The BERT classifiers did not perform any better when screening for depression or suicidal ideation [18, 15]. While mobile voice recordings certainly have screening merit [12, 15, 19], their transcripts are proving less useful for screening than text message content.

Text messages versus text prompt replies. In addition, we replicated our text message feature engineering strategy [1] to extract features from text prompt replies [15]. Collected similarly to voice recordings, participants were asked to type their response to a prompt instead of recording their voice. Both machine learning models as well as BERT models achieved F1 scores of 0.67 with these text replies. Thus, while text prompt replies were more predictive than mobile voice recording transcripts, text messages are still more predictive than text prompt replies.

Depression versus suicidal ideation Unexpectedly, we were more successful at suicidal ideation screening than depression screening with text messages. With a single week of data, our models screened for suicidal ideation with an F1 score of 0.84 [2]. With two weeks of data, our best models screened for moderate depression (PHQ-9$\geq$ 10) with an F1 score of 0.79 [3] and moderately severe depression (PHQ-9$\geq$ 15) with an F1 score of 0.80 [1]. Unlike for mobile transcripts [12, 19], there is no current sizable dataset containing text message content with anxiety labels. While the EMU collection [12] attempted to collect text message content with PHQ-9 and GAD-7 labels, not enough participants shared sent text messages for modeling purposes. Hence, these texts have only been used to screen for depression in conjunction with the Moodable text messages.

CHAPTER 4

DEPRESSION SCREENING FROM TEXT MESSAGE LOGS

Context: The lack of private data makes SMS logs without text content an attractive passive mental illness screening modality, though communication patterns in these logs have been underutilized.

Objective: We assess the mental illness screening ability of text logs without message content.

Methods: From two weeks of logs, we extract reply latencies and communication time series. We then extract features to screen for depression. We collect a larger dataset of logs with depression and anxiety labels. We thus compare machine learning and deep learning screening models.

Findings: With existing datasets, models achieved F1 scores of 0.67 and 0.72 with latency and time series features, respectively. With our new larger dataset, deep learning models were more successful than machine learning models at lower screening cutoffs while machine learning models were more successful than deep learning models at higher screening cutoffs. Further, depression was easier to screen for than anxiety. SMS text logs without content prove to be useful for depression screening and our new dataset promises to help advance such research.

This chapter covers material from the following papers:

ML Tlachac, Elke Rundensteiner, “Depression Screening from Text Message Reply Latency”, *42nd IEEE EMBC*, pp 5490-5493, 2020 [5]

ML Tlachac, Veronica Melican, Miranda Reisch, Elke Rundensteiner, “Mobile Depression Screening with Contact Timeseries”, *17th IEEE BHI*, pp 1-4, 2021 [6]

ML Tlachac, Ricardo Flores, Miranda Reisch, Katie Houskeeper, Elke Rundensteiner, “DepreST-CAT: Leveraging Smartphone Call and Text Logs Collected During the COVID-19 Pandemic to Screen for Mental Illnesses”, ACM Proceedings on Interactive, Mobile, Wearable and Ubiquitous Technologies, Accepted [7]

4.1 Introduction

4.2 Reply Latencies for Depression Screening

The unique aspect of text messages, especially among text-based modalities, is the conversational aspect. However, the text logs for these conversations contain more than just the content of the text messages. Important screening features may be able to be derived from the communication patterns in these logs.

The relationship between depression and slower speed of information processing has been documented [141], though not in this context. We hypothesize that this slower processing, which is attributed to less efficient cognitive functioning [141], may result in a pattern of slower response times to text messages.

Thus, in this study, we determine whether SMS text message meta-data, namely the latency of response, can be leveraged to screen for depression. This is the first study to capture features engineered using both 'sent' and 'received' text message logs. The contribution of this research is introducing a new novel feature engineering approach for mobile mental illness screening.

4.2.1 Data

Our study focuses on the last two weeks of text message data within the Moodable and EMU datasets. Specifically, we extract the date and direction of the text messages from each participant-contact combination. The direction for the messages could either be 'sent' by the participant or 'received' by the participant. The messages were temporally ordered and we identified all $(received, sent)$ pairs. From each of these pairs, we calculated the latency of the reply, i.e. the seconds between the 'received' and 'sent' messages. Thus, in this manner, we extract a set of reply latencies for each participant.

We restrict the participants to those who replied to at least two messages within the last two weeks, i.e. had at least two latency values within their set of reply latencies. This requirement results in 68 participants and mitigates the quality issues typically stemming from crowd-sourced

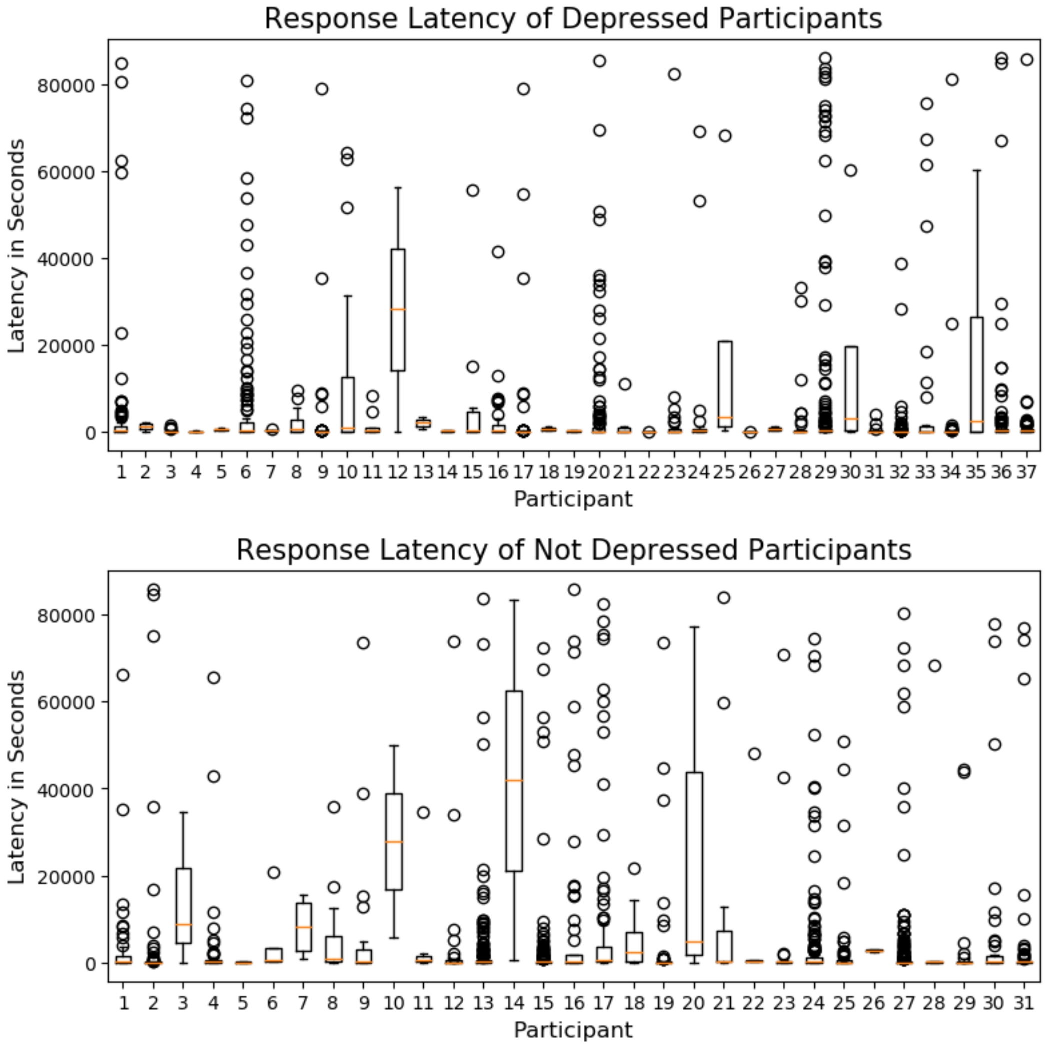


Figure 4.1: Text response latency within the last 14 days for each participant.

data. The latencies for the 37 depressed participants (PHQ-9 ≥ 10) and 31 not depressed participants (PHQ-9 < 10) are depicted in Figure 4.1.

For each participant, we extract nine features from the metadata of their text messages. Seven of these features involve the set of reply latencies. In addition to the minimum and maximum latency, we consider the 10%, 25%, 50%, 75%, and 90% quantiles of the set of reply latencies. We included the 10% and 90% quantiles as these values are less impacted by outliers than the minimum and maximum latencies. In addition to these features, we also record the number of contacts each participant responded to within two weeks and the total number of replies from each participant.

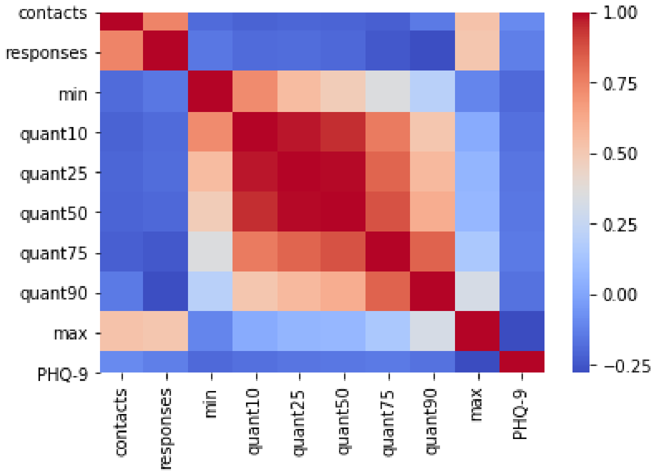


Figure 4.2: Pearson correlation between features.

4.2.2 Principal Component Analysis for Feature Reduction

We suspect there will be a considerable amount of correlation among these features, particularity the quantile features. This is confirmed in Figure 4.2 which shows the correlation among all features. To mitigate this high correlation, we apply principal component analysis (PCA) to our features. When referring to the top p principal components, these are the p principal components that explain the most variance. In addition to traditional PCA which builds linear combination of the original features, we also consider Gaussian kernel PCA which we further refer to as kPCA.

4.2.3 Machine Learning Methodology for Depression Classification

Our goal is to predict if the PHQ-9 score is at least 10. We experiment with a variety of machine learning methods: Gaussian NB, LR, SVM with Gaussian and linear kernels, kNN with $k = 3$ and $k = 5$, RF with $depth \in \{2, 4\}$, XGBoost with $depth \in \{2, 4\}$, and AdaBoost.

We build models with the top one to nine principal components from PCA and kPCA. We down-sample the data to balance the two classes before running the machine learning models, resulting in 31 participants per class. Due to the number of participants, each model employs 5-fold cross-validation. To ensure result robustness, we repeat each experimental procedure 100 times and report on average values. We evaluate our models with F1 score, accuracy, and AUC.

Table 4.1: Best model configuration for each method with PCA.

Method	Parameter	PCs	F1	AUC	Accuracy
NB		8	0.66	0.67	0.60
LR		1	0.55	0.65	0.59
SVM	linear	1	0.55	0.65	0.61
kNN	k=3	8	**0.68**	0.70	**0.68**
RF	3	1	0.64	0.70	**0.69**
XGBoost	Any	1	**0.67**	**0.72**	**0.69**
AdaBoost		1	0.63	0.66	0.55

4.2.4 Screening Results

Table 4.1 lists the best model configurations for each method. Specifically, these model configurations achieved the highest scores for the majority of metrics. As the model configurations involving kPCA were never significantly better than model configurations involving traditional PCA, Table 4.1 contains only models with features derived from traditional PCA. All of the best model configurations leveraged either just the first principal component or eight principal components. The models leveraging just the first principal component are preferred for implementation. As such, we identify XGBoost as the preferred method. Note, the depth parameter was not influential on the XGBoost models. F1, AUC, and Accuracy for the XGBoost models are seen in Figure 4.3.

From Figure 4.2, we can see there is a slight negative correlation between the features and the binary PHQ-9 score, indicating depressed individuals do respond slower. The correlation shows they also have fewer contacts and responses. As previously mentioned, the high correlation among quantile features motivated our use of PCA and building linear combinations of the original latency features was most successful. The coefficients for these principal components are displayed in Figure 4.4. The majority of the models in Table 4.1 leverage just PC1 from Figure 4.4. As seen from these coefficients, the maximum and 90% quantile features barely contribute to the first two principal components. This is likely because these values reflect outlying latencies. The 10%, 25%, and 50% quantile features were most influential to the first principal component, and therefore to the performance of the XGBoost models.

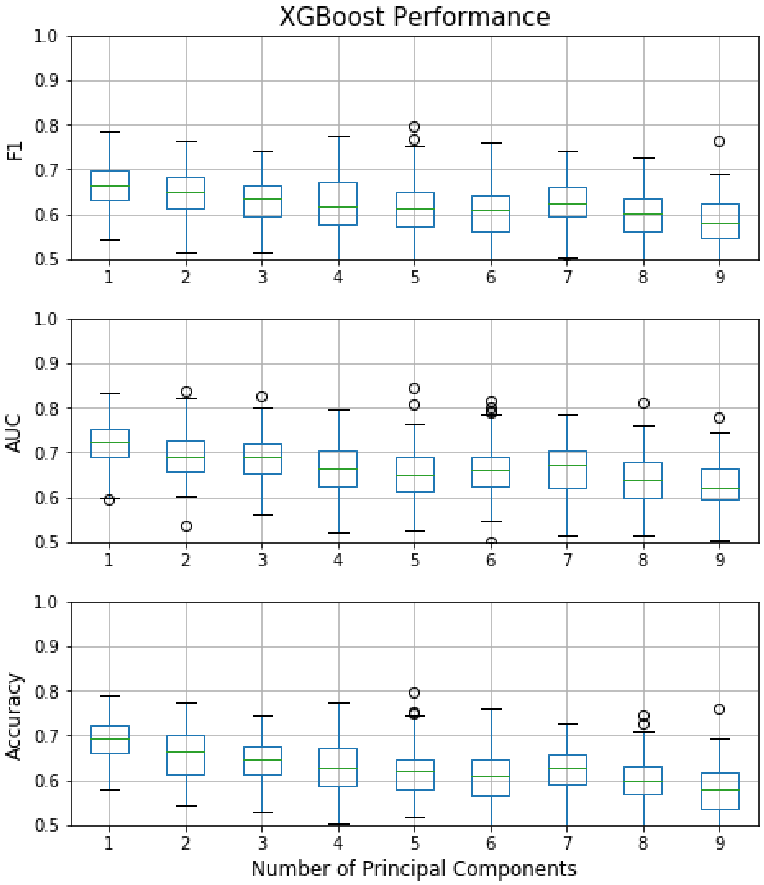
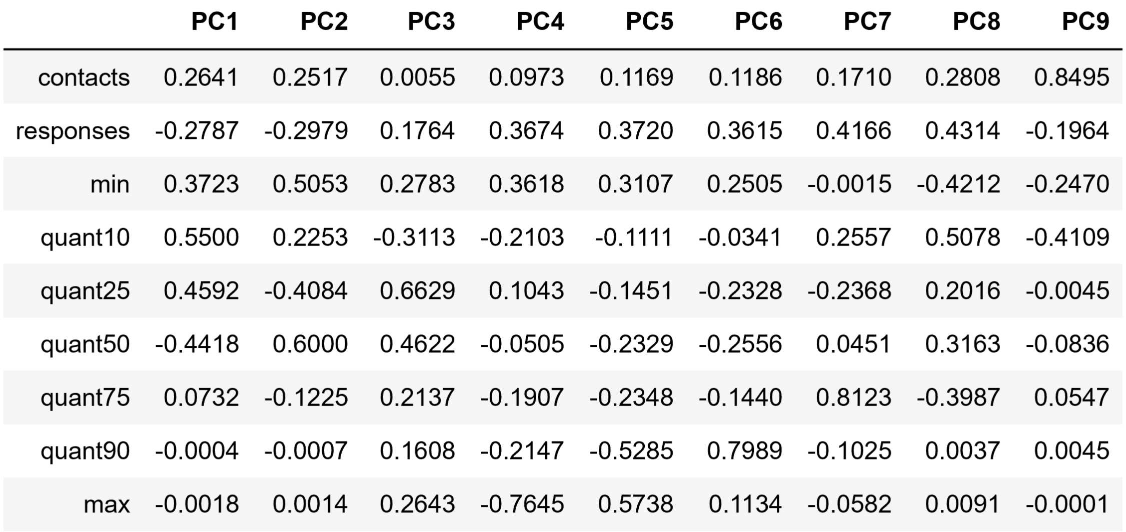


Figure 4.3: 100 trials of XGBoost models with depth 2.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
contacts	0.2641	0.2517	0.0055	0.0973	0.1169	0.1186	0.1710	0.2808	0.8495
responses	-0.2787	-0.2979	0.1764	0.3674	0.3720	0.3615	0.4166	0.4314	-0.1964
min	0.3723	0.5053	0.2783	0.3618	0.3107	0.2505	-0.0015	-0.4212	-0.2470
quant10	0.5500	0.2253	-0.3113	-0.2103	-0.1111	-0.0341	0.2557	0.5078	-0.4109
quant25	0.4592	-0.4084	0.6629	0.1043	-0.1451	-0.2328	-0.2368	0.2016	-0.0045
quant50	-0.4418	0.6000	0.4622	-0.0505	-0.2329	-0.2556	0.0451	0.3163	-0.0836
quant75	0.0732	-0.1225	0.2137	-0.1907	-0.2348	-0.1440	0.8123	-0.3987	0.0547
quant90	-0.0004	-0.0007	0.1608	-0.2147	-0.5285	0.7989	-0.1025	0.0037	0.0045
max	-0.0018	0.0014	0.2643	-0.7645	0.5738	0.1134	-0.0582	0.0091	-0.0001

Figure 4.4: Coefficients for each principal component (PC).

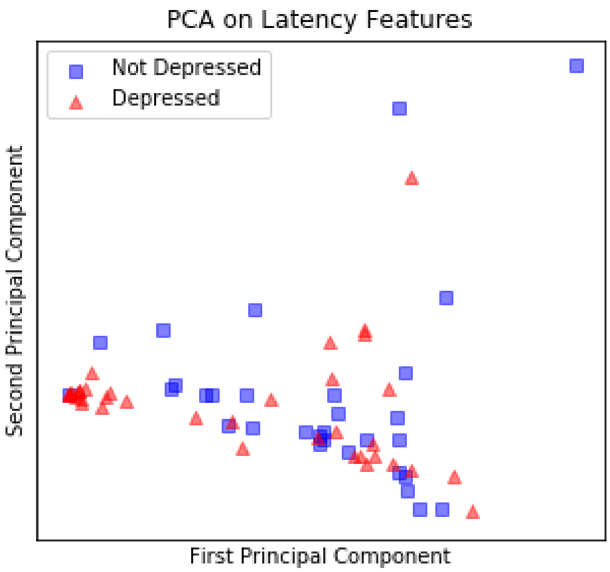


Figure 4.5: Top two principal components.

The first PC covers 42.36 percent of the data variance and the eighth principal component covers less than 0.02 percent of the data variance. Fig. Figure 4.5 displays the two-dimensional separability of the data, which covers 79.4 percent of the data variance. Visually, little separability is added by the second principal component, explaining why the majority of the methods worsened when incorporating more than just the first principal component. While some of the depressed participants on the far left could be easily detected, we can see how other depressed participants are intermixed among not depressed participants and therefore would be difficult to detect regardless of machine learning method deployed.

4.2.5 Discussion

We were able to screen for depression with F1 of 0.67, AUC of 0.72, and Accuracy of 0.69. While these values are lower than some of those achieved with text content [1, 3], we consider text message reply latency to be a promising modality. Reply latency features have the advantage over message content as no private information is required, increasing the number of individuals who would be willing to share this data. In the future, reply latencies could be a part of successful multimodal mobile screening models.

4.3 Communication Time Series for Depression Screening

The depression screening potential of text logs and call logs has not been fully explored. As such, in this research, we capture the temporal aspect of the log data by crafting time series of communications from the logs of 312 participants. We leverage these time series as well as 60 diverse features extracted from each time series as input to train machine learning models for depression screening. Our specific research questions involve whether it is best to screen for depression with:

- Text logs or call logs?
- Incoming, outgoing, or all communications?
- Communication count, average length, or contacts?
- 4, 6, 12 or 24 hour aggregation intervals?
- Time series or features from time series?

4.3.1 Text and Call Log Data

This research leverages the text logs and call logs in the combined Moodable and EMU datasets. Since the PHQ-9 asks about the last two weeks, we consider the last two weeks of text and call logs. Participants with at least two texts or two minutes of calls within the prior two weeks are included in our analysis. Of the 312 participants, 295 and 212 shared text and call logs, respectively. The distributions of PHQ-9 scores for these participants are available in Figure 4.6. The average number of texts was 127 and the average number of calls was 84.

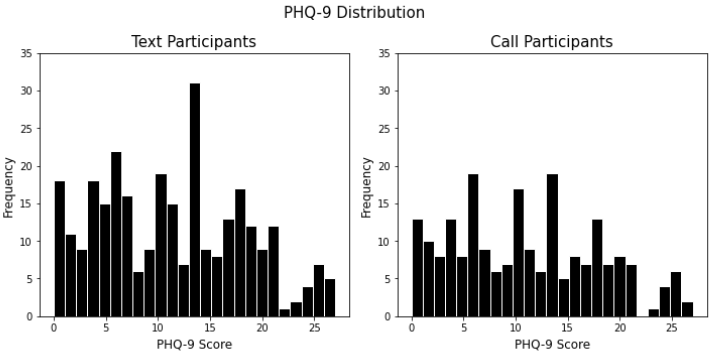


Figure 4.6: PHQ-9 distributions of participants who submitted text and call logs.

4.3.2 Time Series Construction

In order to construct the time series from the logs, we consider the *count* of communications, number of unique *contacts*, and *average length* of communications. While all the call logs contained call duration, not all retrospective text logs contained message size so we approximated this value with the number of characters in the text messages. Further, as calls are more scarce than texts, we consider count to be the summation of seconds of the phone calls. We then calculate these values for every 4, 6, 12, and 24 hours. We refer to these as the aggregation intervals of the time series.

In addition to all communications, we also construct time series with just the incoming and outgoing communications. We require participants to have at least two incoming or outgoing texts or minutes of calls to be included in the analysis. The number of participants who meet these requirements are in Table 4.2. Only a third of participants with text logs submitted outgoing texts.

Table 4.2: Number of participants with each type of data.

Modality	All	Incoming	Outgoing
Text Logs	295	290	99
Call Logs	212	182	197

For each of the six possible data types in Table 4.2, we constructed 12 time series. Thus, each participant has between 12 and 72 time series depending on the data shared. Examples of the 12 time series constructed with number of texts are displayed in Figure 4.7. This participant also has 12 time series for number of text contacts and 12 time series for average text length.

4.3.3 Machine Learning Methodology

Feature engineering. We leverage the Time Series Feature Extraction Library (TSFEL) [142] to extract 16 statistical, 18 temporal, and 26 spectral features from each time series. In addition to considering the 60 count features, 60 contact features, and 60 length features separately, we combine them into a set of 180 features. We have 96 sets of features when considering the 2 log types, 3 directions, 4 aggregation intervals, and 4 time series values. We create 15 principal components (PCs) for each set of features.

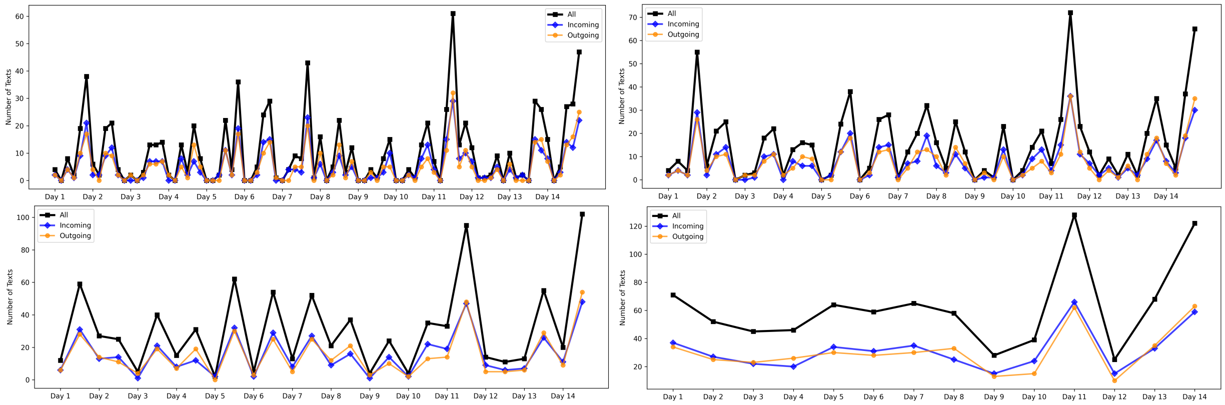


Figure 4.7: Time series representing the count of texts for an example participant with PHQ-9$=7$. The different plots are for different aggregation intervals. For the 4 hour aggregation interval, there are 6 samples per day whereas for the 24 hour aggregation interval, there is only 1 sample per day.

Machine learning methods. As a PHQ-9 score of at least 10 is indicative of depression, our goal is to train classifier to predict if the PHQ-9 score for each participant is at least 10. For the time series, we employ a distance-based approach that uses the k-Nearest Neighbor (kNN) algorithm with dynamic time warping (DTW) distance [143]. Further, we train models with between 1 and 15 PCs from the time series features. We compare the screening ability of kNN with the screening abilities of logistic regression (LR), support vector classifier (SVC) with a Gaussian kernel, and random forest classifier (RF).

Model evaluation. We use stratified sampling to divide the data into train and test sets. We experiment with upsampling and downsampling to balance the training set. For every model configuration, we repeat the experimental procedure 100 times with different train and test sets. We evaluate the depression screening ability of each experimental configuration with the average $F1$ score of the 100 models.

4.3.4 Results

The plots in Figure 4.8 compare the highest average $F1$ scores of the 100 models that share the same configurations for the log types, communication directions, and aggregation intervals. Details about the best text and call log model configurations for each plot are in Table 4.3 and Table 4.4,

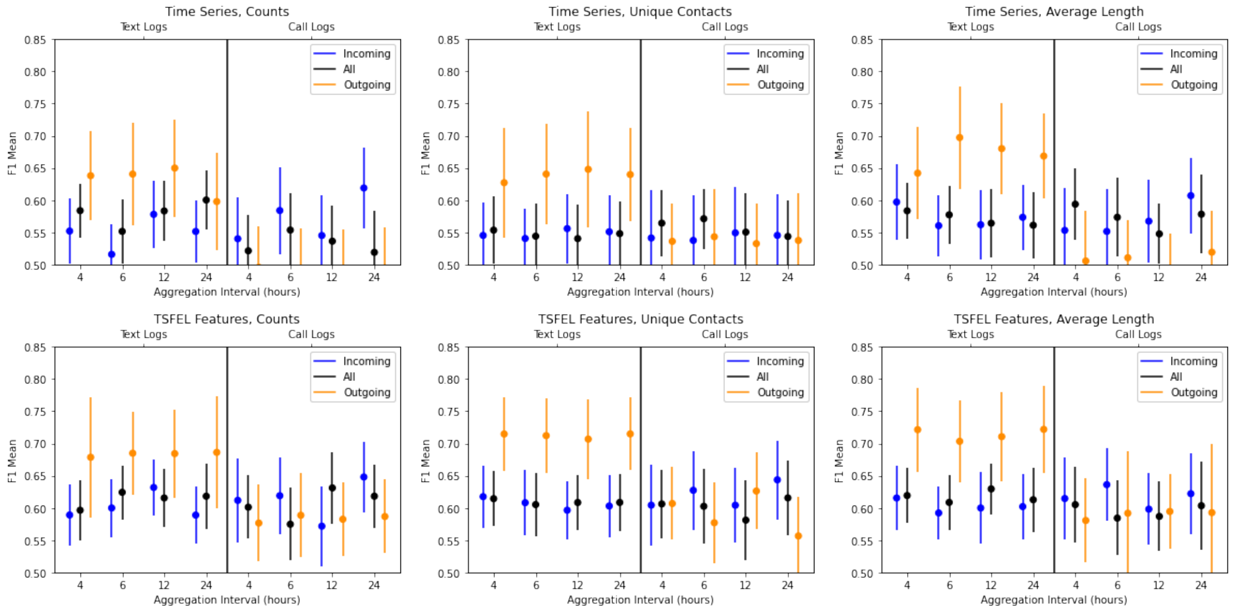


Figure 4.8: Comparison of model configurations with the highest average $F1$ scores.

respectively. These tables also contain the results of combining all TSFEL feature sets. Overall, the answers to our research questions are:

- Text logs are more predictive than call logs.
- Outgoing texts and incoming calls are most predictive.
- Average length was best for text logs while communication count was best for call logs.
- 24 hours was the most predictive aggregation interval.
- TSFEL features were more predictive than time series.

Further, we noticed the model configurations that yielded the highest average $F1$ scores were different for text and call logs. Only 1 principal component was required for text logs while more were required for call logs. LR and downsampling was best for all text log TSFEL feature sets while RF and upsampling was best for all call log TSFEL feature sets.

With text logs we achieved a highest average $F1 = 0.72$. According to a t-test [71], this is statistically significantly ($p < 0.001$) higher than the highest average $F1 = 0.65$ we obtained with call logs. For both time series and TSFEL features, the outgoing texts are statistically significantly ($p < 0.001$) more predictive than either incoming or all texts. TSFEL features are statistically significantly more predictive than time series for text logs with $p < 0.05$ and call logs with $p < 0.001$. The TSFEL features from incoming calls are statistically significantly ($p < 0.05$) more predictive than features from either outgoing or all calls.

Table 4.3: Comparison of model configurations with the highest average $F1$ scores for text logs.

Data	Direction	Interval	Method	PCs	Sampling	F1 ± σ	Precision	Sensitivity	Specificity	AUC	Accuracy
TS - count	Outgoing	6 hrs	kNN		Down	0.65 ± 0.08	0.67	0.64	0.51	0.57	0.59
TS - contacts	Outgoing	12 hrs	kNN		Up	0.65 ± 0.09	0.67	0.64	0.51	0.57	0.59
TS - length	Outgoing	12 hrs	kNN		Up	0.70 ± 0.08	0.71	0.70	0.55	0.62	0.64
TSFEL - count	Outgoing	24 hrs	LR	1	Down	0.68 ± 0.09	0.67	0.71	0.47	0.59	0.61
TSFEL - contacts	Outgoing	24 hrs	LR	1	Down	**0.72** ± 0.06	0.71	**0.72**	0.54	0.63	0.65
TSFEL - length	Outgoing	24 hrs	LR	1	Down	**0.72** ± 0.07	**0.75**	0.71	**0.62**	**0.66**	**0.67**
TSFEL - all	Outgoing	12 hrs	LR	1	Down	0.71 ± 0.06	0.72	0.70	0.58	0.64	0.65

Table 4.4: Comparison of model configurations with the highest average $F1$ scores for call logs.

Data	Direction	Interval	Method	PCs	Sampling	F1 ± σ	Precision	Sensitivity	Specificity	AUC	Accuracy
TS - count	Incoming	24 hrs	kNN		Up	0.62 ± 0.06	0.65	0.60	**0.57**	**0.58**	0.58
TS - contacts	All	6 hrs	kNN		Up	0.57 ± 0.05	0.58	0.57	0.48	0.52	0.53
TS - length	Incoming	24 hrs	kNN		Up	0.61 ± 0.06	0.61	0.61	0.48	0.54	0.55
TSFEL - count	Incoming	24 hrs	RF	15	Up	**0.65** ± 0.06	0.64	**0.67**	0.48	**0.58**	0.59
TSFEL - contacts	Incoming	24 hrs	RF	8	Up	0.64 ± 0.06	0.64	0.66	0.49	**0.58**	0.59
TSFEL - length	Incoming	6 hrs	RF	15	Up	0.64 ± 0.06	0.62	0.66	0.45	0.55	0.57
TSFEL - all	Incoming	24 hrs	RF	12	Up	**0.65** ± 0.06	**0.65**	0.66	0.51	**0.58**	**0.60**

4.3.5 Discussion

We performed detailed experimentation to determine the ability of two weeks of text and call logs to screen for depression. Our best models achieved an average $F1 = 0.72$ with features extracted from time series of outgoing texts and $F1 = 0.65$ with features extracted from time series of incoming calls. Thus, the text logs proved more valuable for depression screening, despite it being more likely that participants deleted text logs than call logs prior to sharing their data. This suggests even a subset of text logs can be valuable for screening purposes.

Despite fewer participants sharing outgoing text logs than any of the other three types of logs, these logs proved most useful in screening for depression. For all four types of logs we compared time series of communication count, communication average length, and unique contacts aggregated every 4, 6, 12, and 24 hours. These results provides valuable insight on the usefulness of logs for mobile depression detection.

4.4 Collecting Text and Call Logs for Mental Illness Screening

Given this promise of logs as screening medium, in this further research, we collect the Depression Stererotype Threat Call And Text log (DepreST-CAT) dataset during the COVID-19 pandemic. Labeled with both depression and anxiety screening scores, this large dataset of retrospective logs presents many new opportunities for developing passive screening tools. We assess the mental illness screening capabilities of the collected logs with unimodal and multimodal machine learning and deep learning methods on the time series of call and text logs. For the machine learning models, we use rich statistical, temporal, and spectral features extracted from the time series. For the deep learning models, we work with state-of-the-art sequential representation learning models augmented with attention to learn which aspects of the time series log most impacts the screening decision. For all models, we also explore the number of weeks of logs prior to the time of the data submission that are most useful for mental illness screening.

While our research was informed by the aforementioned studies [41, 12, 6], our work features several notable advances. Namely, we collect a larger dataset containing both call and text logs from over 365 participants recruited through the Prolific crowdsourcing platform [144]. Screen for multiple mental illnesses is now possible as our data is also labeled with anxiety screening scores. As we have sufficient log data, we also uniquely leverage deep learning methods to perform mental illness screening. Further, as DepreST-CAT was collected during the COVID-19 pandemic, it may offer valuable insights into communication patterns during this unprecedented time.

There are three main contributions of this data collection and analysis research. First, we collected the valuable DepreST-CAT dataset consisting of call logs, texts logs, and mental illness screening survey scores from 369 crowdsourced participants during the COVID-19 pandemic. Second, we conduct an Evaluation study of the depression and anxiety screening potential of machine learning and deep learning models that leverage time series of call and text logs. Lastly, we explore of the impact of the number of weeks of log data on the mental illness screening ability of a rich family of unimodal and multimodal models.

4.4.1 DepreST-CAT Collection Methodology

We collected the Depression Stereotype Threat Call and Text log subset (DepreST-CAT) from December 2020 through April 2021 under the modified IRB. We started our collection a year after COVID-19 emerged in December 2020 and ended our collection in April 2021. This captures the unique period of time between the USA Food and Drug Administration (FDA) issuing emergency authorization for the first COVID-19 vaccines and the US beginning to announce statewide vaccination eligibility for all adults [145]. Call and text logs were collected retrospectively from crowdsourced workers through a mobile app which administers traditional depression and anxiety screening surveys to provide labels for the data.

4.4.2 The Data Collection Application

We modified the Early Mental Health Uncovering (EMU) Android app [12] for the DepreST collection. The app prompted users to share retrospective smartphone data, mental illness screening scores, demographics, and voice recordings. As the app collects passive data retrospectively, the collection process can be completed within a few minutes. Participants were able to delete the app immediately after completing the short collection. All collected data except the voice recordings will be released as part of the DepreST-CAT dataset.

The first page of the app provides the below study overview to which participants must agree to proceed. This page briefly details the study goal, procedure, privacy, and risk. The app then asks participants for permission to collect the passive smartphone modalities: *call logs, text logs, calendar logs,* and *contact logs.* Participants are informed about the bonus payment they will receive for sharing text data, as further described in Section subsection 4.4.3. To ensure informed consent, permission to collect each modality is asked individually. These permissions were asked at the beginning to expedite the collection process as the data could be scraped while the participants completed the remainder of collection. Similar to related work [44], all names and addresses are one-way hashed and sent message content was removed prior to sending the data to our secure server to protect participant privacy.

The app further proceeds to administer the nine-item PHQ-9 for depression screening [92] and the seven-item GAD-7 for anxiety screening [95]. The app next asked participants for demographic information. The multiple choice demographic questions were related to gender, age, student status, history of depression treatment, and racial/ethnic identity. Note, in an attempt to trigger stereotype threat [146, 16] that inspired the name for the dataset, half of the participants randomly received the gender question prior to the screening surveys while the other half received this question after the screening surveys. We also included two COVID-19 related questions on this demographic page to determine how the pandemic impacted the lives of the participants. The first question regarded the quantity of remote work/study to gauge social isolation. The second question was more direct and asked the participants if they knowingly contracted COVID-19.

4.4.3 Participants Recruitment, Eligibility, and Compensation

We recruited participants from Prolific [144], a crowdsourcing platform designed primarily for researchers in need of reliable participants. Our study description on the Prolific platform stated 'Participants will be asked to download an application to their Android phone, answer survey questions, record samples of their voice, and share text logs and messages'. Our participation eligibility requirements included being between 18 and 100 years of age (the maximum range on the Prolific platform), residing in the US, having an Android phone, and being able to answer questions written in English. While initially open to all genders, we temporarily and strategically restricted eligibility to all genders except men for portions of the data collection period to help reduce the initial gender imbalance of the participants.

While the study overview stated that the collection process took four minutes based on our initial testing of the collection app, we allotted for five minutes in our payment estimation. The base pay for completing our study was $\$0.80$ ($\$9.60/$hour), meeting Prolific's recommendation for fair pay [144]. A previously conducted willingness to share study [41] indicated that participants may be less willing to share text logs than the other modalities, we further incentivized the sharing of text logs by offering bonuses ranging from $\$0.05$ to $\$0.45$.

4.4.4 Call and Text Log Screening Methodology

We use the call and text logs to screen for depression and anxiety with machine learning and deep learning methods to provide baselines for DepreST-CAT. Specifically, we screen for mild to moderate depression and anxiety by considering PHQ-9 [92] and GAD-7 [95] cutoffs from 5 to 15. For both the machine learning and deep learning models we construct time series of communications from the call and text logs. Our unimodal models are trained with either the call log time series or text log time series while our multimodal models are constructed with time series from both types of logs. As neither the calendar logs nor the contact logs contain temporal data, we do not use these log modalities in our screening models.

Time series construction. We construct separate time series for the number of incoming texts, number of outgoing texts, seconds of incoming calls, and seconds of outgoing calls by aggregating these values every 24 hours. Further, to explore the longitudinal quantity of data that is best for screening, we end the time series 2, 4, 8, and 16 weeks prior to when the participants submitted the data. As we have a data point every 24 hours, the 2 week time series have 14 values and the 16 week time series have 112 values. Given the four different types of logs and four different time series lengths, we construct 16 time series for each participant. If a participant did not share a particular modality, then the time series values for that modality are set to a default of 0.

Feature engineering. As prior research [6] found that features extracted from the time series were more useful than the raw time series for depression screening, our screening approach involves extracting features from the time series and using the resulting feature vectors to train traditional machine learning models. From each time series, we extract statistical, temporal, and spectral features using the Time Series Feature Extraction Library (TSFEL) [142]. This feature engineering technique yielded 143, 150, 164, and 192 features for the 2, 4, 8, and 16 week time series, respectively. Next, we concatenated the incoming text, outgoing text, incoming call, and outgoing call time series features for each participant into three feature sets for each of the four

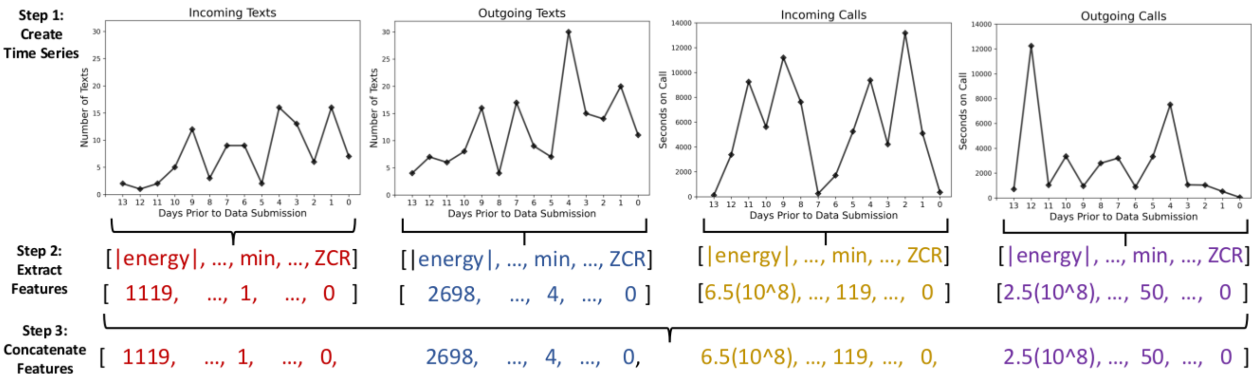


Figure 4.9: An example of the feature engineering process for two weeks of call and text logs for participant Q6N7PT68V with a PHQ-9 score of 12 and a GAD-7 score of 8. We repeat this process for the logs from 4, 8, and 16 weeks prior to data submission. These features are then used in the traditional machine learning models. Not all extracted features are relevant for log time series, such as zero crossing rate (ZCR), so they do not contribute to the principal components.

time series lengths. Specifically, we concatenate the incoming text features with the outgoing text features and the incoming call features with the outgoing call features for the unimodal models. We also concatenate the features from all four time series for the multimodal models. Thus, each participant has a total of 12 feature sets representing three different types of logs in the 2, 4, 8, and 16 weeks prior to data submission. An example of this feature engineering process is depicted in Figure 4.9. Given the large number of features, we use PCA [71] after normalizing the features.

Traditional and ensemble machine learning methods. We assess the screening ability of three standard machine learning methods [71]: k-Nearest Neighbor (kNN) with five neighbors, logistic regression (LR), and support vector classifier (SVC) with a Gaussian kernel. We also consider two tree-based ensemble methods, namely, Random Forest classifier (RF) [71] and eXtreme Gradient Boosting (XGBoost) [77]. The default model parameters are used for these baseline experiments.

Sequence modeling using recurrent neural networks. Recurrent neural networks (RNNs) allow for the analysis of sequential data by generating an embedding for a data sequence [147]. They have been used with great success on problems ranging from natural language processing [148] to time series classification [149]. In time series classification, the goal is to predict the label associated with the overall time series using all the relevant time steps. In general, an RNN model has a hidden state that encodes the information in the time series up to the current time step; typically computed based on the current value of the time step and the previous hidden state. After consuming all the values in the time series, the final hidden state captures the relevant information that can be considered the final sequence embedding. For classification, a final linear layer is added on top of the RNN model that inputs the last hidden state of the RNN and outputs the classification result.

The GRU model for sequence modeling. While RNNs have been shown to be effective for sequential modeling, they are not able to capture long dependencies in time series with many steps. Several variations of the RNN model have been proposed to tackle this issue. Long Short Term Memory (LSTM) are equipped with a special gate designed for storing and forgetting relevant information [150]. Another variant, the Gated Recurrent Unit (GRU) model [151] has become even more prevalent because it can achieve the same performance as the LSTM model with fewer parameters. These GRU models have become particularly popular in the healthcare domain. For instance, they have been used for classifying mortality using time series in electronic health records [152, 153]. We thus adopt this GRU model for depression and anxiety screening. In addition, to address the vanishing problem caused by long sequences where dependencies are lost [154], we extend the GRU model by placing a self-attention layer [155] on top of the model. The attention mechanism creates a context state that is a weighted combination of all hidden to capture longer relationships within the time series of texts and calls. We call this model the GRU-Attention model.

Settings for training the deep sequential model. Unlike the machine learning models, our deep learning sequence models directly consume raw time series to generate a sequence embedding. In our research, the input is two-variant if we use texts (incoming and outgoing) or calls (incoming and outgoing). The input is four-variant when the model consumes all four time series as a multivariate time series. As we produce one value per day, the number of time steps in a time series is equal to number of days, between 14 and 112 days. In other words, the input for each participant varies from 2x14 for 2 weeks, 2x28 for 4 weeks, 2x56 for 8 weeks, and 2x112 for 16 weeks for the unimodal models. Likewise, it is 4x14 for 2 weeks, 4x28 for 4 weeks, 4x56 for 8 weeks, and

4x112 for 16 weeks for the multimodal models. For the experiments, both the GRU and GRU Attention models use a learning rate of 0.001, hidden size of 32, and batch size of 32. To avoid the overfitting problem, we leverage drop out and early stopping strategies. For drop out, we randomly removed 20% of GRU units, and for early stopping, we did not consider epochs greater than 10. With these two techniques, we can ensure a better generalization of deep learning models.

Model training and evaluation strategy. To ensure the robustness of our results, we perform leave-group-out cross-validation with 100 stratified groups (specified with random seeds 0 to 99) for each of the 12 multimodal feature sets. The test sets from the cross-validation comprise 30% of the entire data. All 100 cross-validation groups were used for the machine learning models while only the first five were used for the deep learning models due to computational resources. The training sets are all upsampled for all models with the same random seed (42) to ensure consistency. We train the machine learning models with the top one to four principal components. We also experiment with 2, 4, 8, and 16 weeks of logs for both the machine learning and deep learning models. We repeat the process using only the call logs, only the text logs, and all logs. As we screen for both depression and anxiety, this results in 96 model configurations for each of the five machine learning methods and 24 model configurations for each of the two deep learning methods. We consider the models with the highest average F1 scores to be the most successful at screening.

4.4.5 Collection Results: DepreST-CAT Participants and Smartphone Logs

Out of the 441 total DepreST participants, 361 (81.9%) shared call logs and 348 (78.9%) shared text logs. 369 participants shared at least one of these logs. We designate these 369 participants as the DepreST-CAT participants. These participants shared an impressive total of 143, 280 incoming calls logs, 98, 247 outgoing call logs, 368, 807 incoming call logs, and 76, 118 outgoing text logs. Further, 314 participants shared 5, 186 calendar logs, and 304 participants shared 4, 970 contact logs. Since we collected the logs retrospectively, if a participant elected to share a smartphone log modality, we scraped every log of that type stored on their phone.

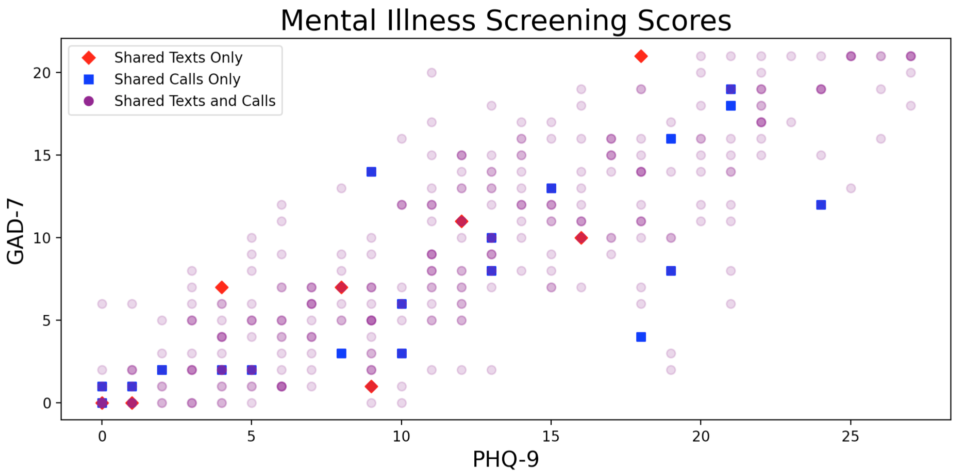


Figure 4.10: The PHQ-9 and GAD-7 scores of the 369 DepreST-CAT participants.

Figure 4.10 plots the PHQ-9 and GAD-7 scores of all of the DepreST-CAT participants based on which of the smartphone log modalities they shared. Only 8 participants shared text logs and not call logs while only 21 participants shared call logs and not text logs. As expected [95], there is visually a high correlation between the PHQ-9 and GAD-7 scores of the DepreST-CAT participants. The distribution of the PHQ-9 and GAD-7 scores for all DepreST-CAT participants are displayed in Figure 4.11. Overall, 56.6% and 43.6% of DepreST-CAT participants screened positive for moderate depression (PHQ-9$\geq$ 10) and moderate anxiety (GAD-7$\geq$ 10), respectively. Notably, 154 (41.7%) participants screened positive for both mental illnesses.

High rates of depression and anxiety in Prolific workers was recently observed in a related study [51]. It was speculated that crowdsourced workers may experience higher rates of mental illness than the general population. Studies regarding mental health may also appeal more to workers who experience mental illness symptoms. As our study description on Prolific did not mention mental health and our study overview in the app only mentioned it briefly, self-selection bias likely played a lesser role in the high mental illness rates in our participant population. In addition to these aforementioned possibilities, we suspect grief, stress, and isolation related to the COVID-19 pandemic contributed to the high rates of mental illness among DepreST-CAT participants.

We asked five demographic and two COVID-19 related questions. As one participant chose not to respond, we report the replies of 368 participants. They included 178 (48.2%) women,

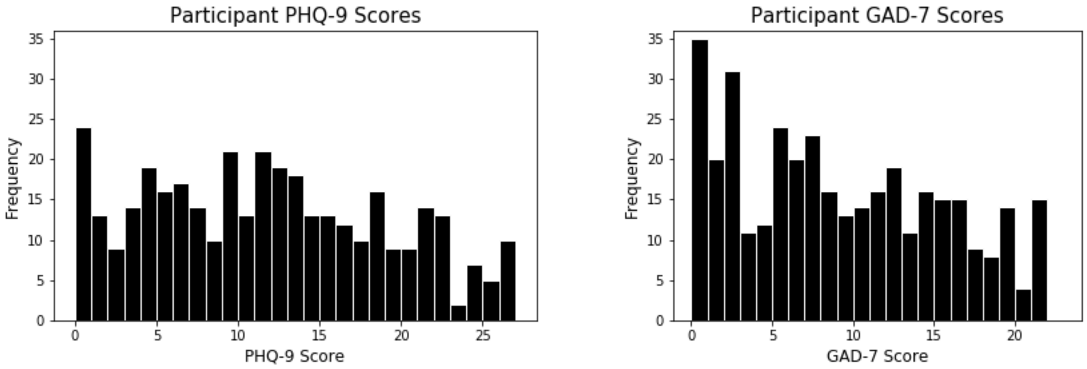


Figure 4.11: The distribution of depression (PHQ-9) and anxiety (GAD-7) screening scores for the 369 DepreST-CAT participants.

Table 4.5: The number of DepreST-CAT participants (N) who screen positive for depression and anxiety at PHQ-9 and GAD-7 cutoffs in the mild (5) to severe (15) score range.

	Cutoff										
	$C5$	$C6$	$C7$	$C8$	$C9$	$C10$	$C11$	$C12$	$C13$	$C14$	$C15$
Depression (N)	288	272	255	241	231	209	196	175	155	136	123
Depression (%)	78.0%	73.7%	69.1%	65.3%	62.6%	56.6%	53.1%	47.4%	42.0%	36.9%	33.3%
Anxiety (N)	257	233	213	190	174	161	145	128	108	97	81
Anxiety (%)	69.6%	63.1%	57.7%	51.5%	47.2%	43.6%	39.3%	34.7%	29.3%	26.3%	22.0%

175 (47.4%) men, and 15 (4.1%) nonbinary. 72 (19.5%) were 18-23 years of age, 202 (54.7%) were 24-39 years of age, 86 (23.3%) were 40-55 years of age, and 8 (2.2%) were 56-100 years of age. 155 (42.0%) reported receiving prior depression treatment. The majority (62.9%) identified as only White. The remaining 136 participants reported 19 different identities including Black (12.5%), Hispanic/Latino (6.8%), and Asian (6.2%). 56 (15.2%) were fully remote, 132 (35.8%) were partially remote, and 146 (39.6%) were not remote. 40 (10.8%) reported having COVID-19.

4.4.6 Machine Learning and Deep Learning Screening Results

We trained classical machine learning as well as deep learning based classification models to screen for depression and anxiety with time series of call and text logs. The percent of the 369 DepreST-CAT participants who screened positive for depression and anxiety at the different PHQ-9 and GAD-7 cutoffs are noted in Table 4.5. At cutoff 5, 78.0% and 69.6% of participants screened positive for depression and anxiety, respectively. Whereas at cutoff 15, 33.3% and 22.0% of par-

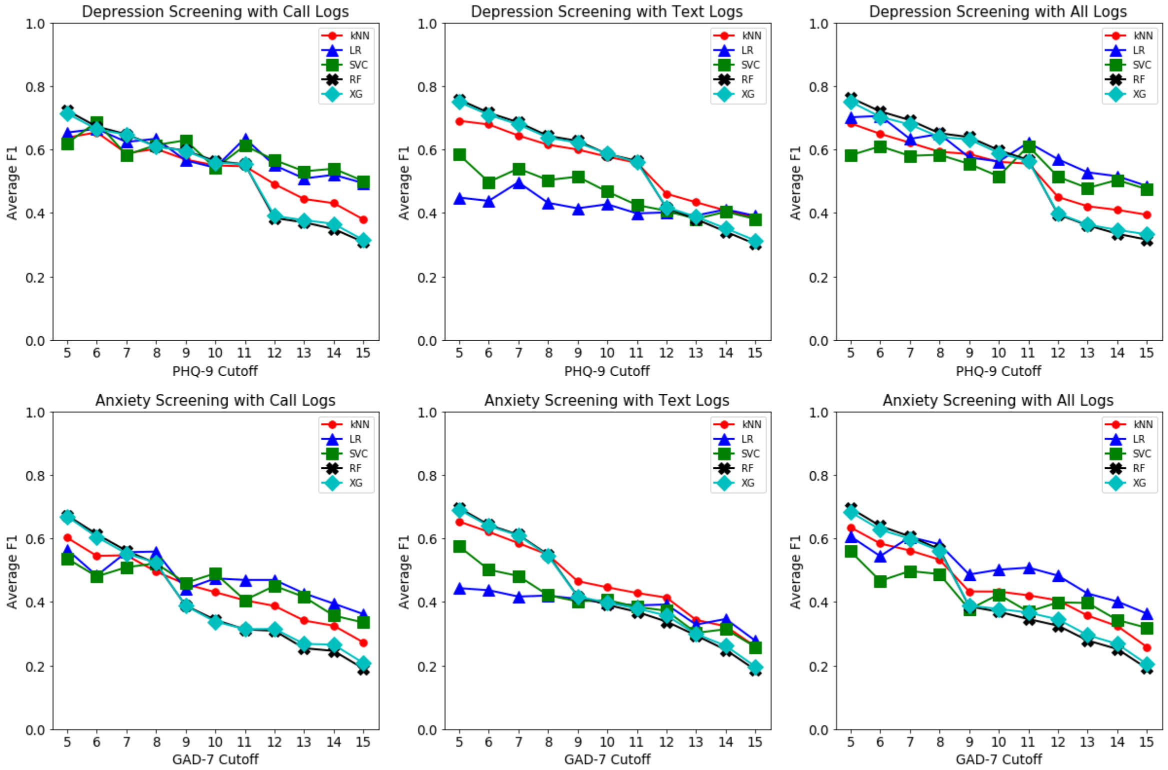


Figure 4.12: The average F1 of the traditional machine learning models screening for depression and anxiety at different screening score cutoffs. The models displayed use two weeks of data and the first principal component.

ticipants screened positive for depression and anxiety, respectively. The data is most naturally balanced at cutoff 10 for depression and at cutoff 8 for anxiety.

We experimented with a variety of parameters in our screening models, including the screening score cutoff, the number of weeks of data prior to submission, the number of principal components, and the machine learning methods. Overall, the number of weeks of data and the number of principal components had relatively little impact on the screening ability of the best performing models. Thus, we focus the majority of our analysis on the simpler models, namely those that use only two weeks of data and only the first principal component.

Impact of cutoffs. Figure 4.12 and Figure 4.13 reveal that the screening models were much more successful at lower cutoffs than at higher cutoffs for both depression and anxiety. As noted in Table 4.8, GRU models trained on all logs were able to achieve average F1 scores of 0.84 and

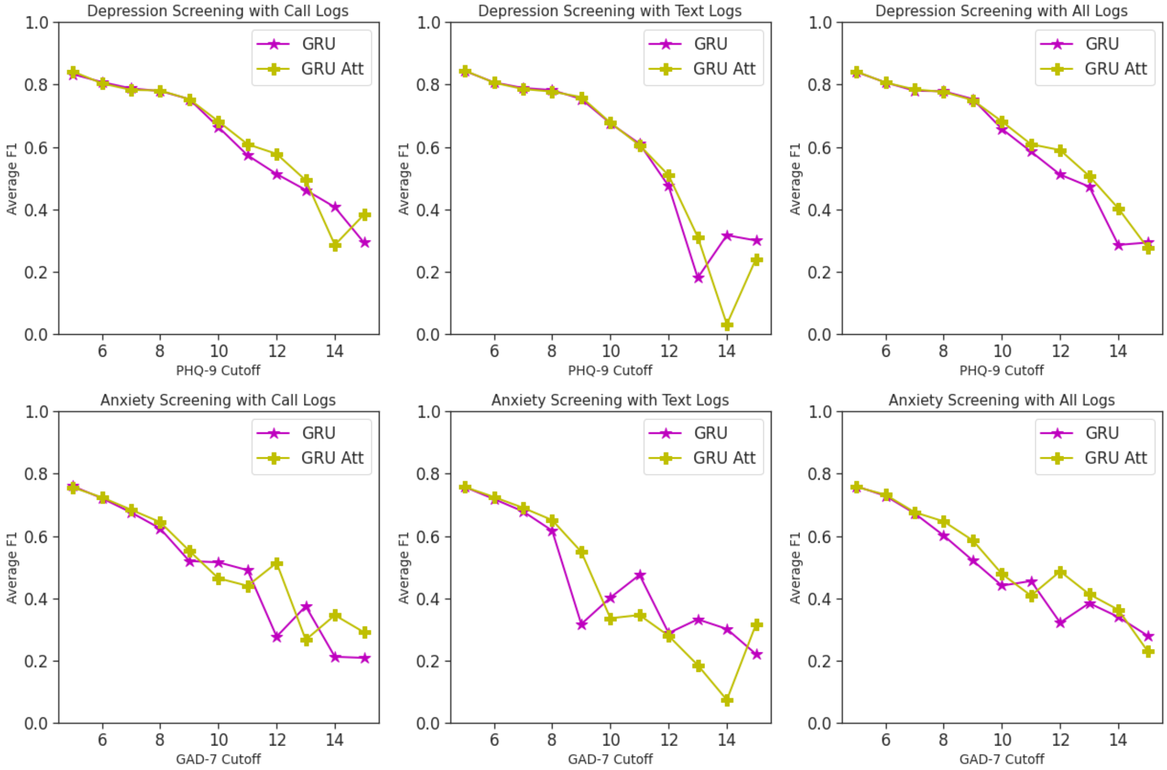


Figure 4.13: The average F1 of the deep learning models screening for depression and anxiety at different screening score cutoffs. The models displayed use two weeks of data.

0.76 at cutoff 5 when screening for depression and anxiety, respectively. For cutoff 15, the most successful depression screening model was SVC on call logs with an average $F1$ of 0.50 and the most successful anxiety screening model was logistic regression on call logs with an average $F1$ of 0.36 (Table 4.6). Thus, our models work well for differentiating participants with no mental illness symptoms from participants with at least mild mental illness symptoms, but they are not as effective at differentiating participants with more mental illness symptoms. This suggests that the communication patterns of mildly, moderately, and severely depressed and anxious participants may not be sufficiently different to be distinguishable. This pattern holds regardless of the number of weeks of data and principal components used in the screening models.

Impact of mental illness. It was easier to screen for depression with the smartphone logs than it was to screen for anxiety. At cutoff 5, the deep learning models that screened for depression achieved an average F1 score that was 0.08 higher than that of the models that screened for anxiety in Table 4.8. Likewise, the machine learning models that screened for depression achieved an average F1 score that was 0.05 higher than the models that screened for anxiety in Table 4.6 and Table 4.7. The pattern holds across all PHQ-9 and GAD-7 cutoffs. These results suggest that depression symptoms have a greater impact on communication patterns than anxiety symptoms.

Impact of log type. The most predictive log type depends on the model type. For SVC, the call logs produce better screening models for almost all PHQ-9 cutoffs and higher GAD-7 cutoffs. While text logs are notably worse in LR models that screen for both depression and anxiety, they perform better at lower GAD-7 cutoffs in kNN models. For ensembles, the call logs perform slightly worse for depression and anxiety screening. The aforementioned drop in screening ability at PHQ-9 cutoff 11 and GAD-7 cutoff 8 for these ensembles exists across all logs types. Call logs, text logs, and all logs perform similarly for the deep learning models at the lower cutoffs. Text logs perform worse after PHQ-9 cutoff 12 and all logs mostly perform better at GAD-7 cutoff 9.

Impact of model type. At the lower PHQ-9 and GAD-7 cutoffs, the deep learning models perform better than the classical machine learning models. However, at PHQ-9 cutoff 9 and GAD-7 cutoff 8 , the performance of the deep learning models starts to decrease drastically (Figure 4.13). Likewise, in Figure 4.12, the performance of the ensemble machine learning models starts out better than the machine learning models but decreases drastically at PHQ-9 cutoff 11 and GAD-7 cutoff 8. The traditional machine learning models in contrast have a more gradual descent in performance. Thus, at the higher PHQ-9 and GAD-7 cutoffs, they are more successful than both the deep learning and ensemble learning models.

Impact of deep learning model. Both of the deep learning models perform the same at the lower PHQ-9 and GAD-7 cutoffs. With a couple exceptions, the GRU model with an attention layer performed better than the GRU model without an attention layer at higher PHQ-9 and GAD-7 cutoffs for all logs. Thus, the attention layer proved beneficial for screening more moderate and severe mental illnesses. In addition to lower average F1 scores, the models screening at

Table 4.6: Comparison of the average ± standard deviation of the F1 scores of the **traditional machine learning** models with different screening cutoffs. These models used the first principal component constructed from the features of two weeks of data.

		Depression			Anxiety		
Cutoff	Method	Call	Text	All	Call	Text	All
5	SVC	0.62 ± 0.20	0.58 ± 0.19	0.58 ± 0.19	0.54 ± 0.15	0.57 ± 0.19	0.56 ± 0.13
6	SVC	0.69 ± 0.12	0.50 ± 0.16	0.61 ± 0.17	0.48 ± 0.16	0.50 ± 0.15	0.47 ± 0.15
7	SVC	0.58 ± 0.17	0.54 ± 0.19	0.58 ± 0.17	0.51 ± 0.13	0.48 ± 0.14	0.50 ± 0.14
8	SVC	0.61 ± 0.14	0.50 ± 0.18	0.58 ± 0.15	0.52 ± 0.14	0.42 ± 0.09	0.49 ± 0.14
9	SVC	0.63 ± 0.09	0.51 ± 0.18	0.55 ± 0.14	0.46 ± 0.14	0.40 ± 0.08	0.38 ± 0.14
10	SVC	0.54 ± 0.15	0.47 ± 0.16	0.52 ± 0.14	0.49 ± 0.10	0.41 ± 0.08	0.42 ± 0.12
11	SVC	0.61 ± 0.11	0.42 ± 0.12	0.61 ± 0.10	0.41 ± 0.15	0.38 ± 0.07	0.37 ± 0.13
12	SVC	0.57 ± 0.08	0.41 ± 0.09	0.51 ± 0.13	0.45 ± 0.09	0.37 ± 0.06	0.40 ± 0.09
13	SVC	0.53 ± 0.09	0.38 ± 0.07	0.48 ± 0.14	0.42 ± 0.04	0.30 ± 0.07	0.40 ± 0.07
14	SVC	0.54 ± 0.03	0.40 ± 0.07	0.50 ± 0.09	0.36 ± 0.08	0.31 ± 0.08	0.34 ± 0.09
15	SVC	0.50 ± 0.04	0.38 ± 0.07	0.47 ± 0.08	0.34 ± 0.04	0.26 ± 0.06	0.32 ± 0.05
5	Logistic Regression	0.65 ± 0.17	0.45 ± 0.05	0.70 ± 0.13	0.56 ± 0.16	0.44 ± 0.06	0.61 ± 0.14
6	Logistic Regression	0.66 ± 0.16	0.44 ± 0.06	0.71 ± 0.11	0.48 ± 0.15	0.44 ± 0.07	0.54 ± 0.14
7	Logistic Regression	0.62 ± 0.15	0.50 ± 0.15	0.63 ± 0.13	0.56 ± 0.14	0.42 ± 0.05	0.60 ± 0.10
8	Logistic Regression	0.63 ± 0.13	0.43 ± 0.09	0.65 ± 0.10	0.56 ± 0.11	0.42 ± 0.06	0.58 ± 0.07
9	Logistic Regression	0.57 ± 0.16	0.41 ± 0.09	0.58 ± 0.14	0.44 ± 0.12	0.41 ± 0.05	0.49 ± 0.10
10	Logistic Regression	0.54 ± 0.15	0.43 ± 0.11	0.56 ± 0.12	0.47 ± 0.11	0.40 ± 0.06	0.50 ± 0.09
11	Logistic Regression	0.63 ± 0.06	0.40 ± 0.06	0.62 ± 0.06	0.47 ± 0.10	0.39 ± 0.06	0.51 ± 0.06
12	Logistic Regression	0.55 ± 0.10	0.40 ± 0.05	0.57 ± 0.06	0.47 ± 0.07	0.39 ± 0.05	0.48 ± 0.04
13	Logistic Regression	0.51 ± 0.12	0.39 ± 0.06	0.53 ± 0.08	0.43 ± 0.03	0.33 ± 0.06	0.43 ± 0.04
14	Logistic Regression	0.52 ± 0.06	0.41 ± 0.06	0.52 ± 0.05	0.39 ± 0.06	0.35 ± 0.07	0.40 ± 0.06
15	Logistic Regression	0.49 ± 0.04	0.39 ± 0.06	0.48 ± 0.04	0.36 ± 0.03	0.28 ± 0.06	0.36 ± 0.03
5	kNN	0.64 ± 0.07	0.69 ± 0.05	0.68 ± 0.04	0.60 ± 0.08	0.65 ± 0.05	0.63 ± 0.05
6	kNN	0.65 ± 0.07	0.68 ± 0.04	0.65 ± 0.05	0.54 ± 0.08	0.62 ± 0.05	0.58 ± 0.05
7	kNN	0.59 ± 0.08	0.64 ± 0.05	0.62 ± 0.04	0.55 ± 0.08	0.59 ± 0.04	0.56 ± 0.05
8	kNN	0.60 ± 0.08	0.61 ± 0.05	0.59 ± 0.05	0.50 ± 0.07	0.55 ± 0.05	0.53 ± 0.05
9	kNN	0.57 ± 0.06	0.60 ± 0.05	0.59 ± 0.05	0.46 ± 0.08	0.47 ± 0.06	0.43 ± 0.05
10	kNN	0.55 ± 0.07	0.58 ± 0.06	0.56 ± 0.04	0.43 ± 0.07	0.45 ± 0.06	0.43 ± 0.06
11	kNN	0.55 ± 0.06	0.56 ± 0.05	0.56 ± 0.05	0.40 ± 0.07	0.43 ± 0.07	0.42 ± 0.06
12	kNN	0.49 ± 0.06	0.46 ± 0.06	0.45 ± 0.06	0.39 ± 0.06	0.41 ± 0.06	0.4 ± 0.06
13	kNN	0.44 ± 0.06	0.43 ± 0.06	0.42 ± 0.05	0.34 ± 0.06	0.34 ± 0.06	0.36 ± 0.05
14	kNN	0.43 ± 0.06	0.41 ± 0.06	0.41 ± 0.06	0.33 ± 0.05	0.32 ± 0.06	0.32 ± 0.07
15	kNN	0.38 ± 0.06	0.39 ± 0.06	0.39 ± 0.07	0.27 ± 0.06	0.26 ± 0.07	0.26 ± 0.06

higher cutoffs had much higher standard deviation. This suggests that the leave-group-out splits had a large impact on the performance of these models.

Impact of machine learning model. The ensemble models in Table 4.7 perform almost identically at all PHQ-9 and GAD-7 cutoffs. These ensembles perform better than the traditional machine learning models at lower cutoffs but worse than the traditional machine learning models at higher cutoffs, as displayed in Figure 4.12. The opposite pattern is observed for SVC and LR

Table 4.7: Comparison of the average ± standard deviation of the F1 scores of the **ensemble machine learning** models with different screening cutoffs. These models used the first principal component constructed from the features of two weeks of data.

		Depression			Anxiety		
Cutoff	Method	Call	Text	All	Call	Text	All
5	Random Forest	0.72 ± 0.05	0.76 ± 0.04	0.76 ± 0.03	0.67 ± 0.07	0.70 ± 0.04	0.69 ± 0.04
6	Random Forest	0.67 ± 0.05	0.72 ± 0.03	0.72 ± 0.03	0.61 ± 0.07	0.64 ± 0.04	0.64 ± 0.04
7	Random Forest	0.65 ± 0.06	0.69 ± 0.04	0.69 ± 0.04	0.56 ± 0.07	0.61 ± 0.04	0.61 ± 0.04
8	Random Forest	0.61 ± 0.05	0.64 ± 0.04	0.65 ± 0.04	0.52 ± 0.05	0.55 ± 0.05	0.57 ± 0.04
9	Random Forest	0.60 ± 0.06	0.63 ± 0.04	0.64 ± 0.04	0.39 ± 0.07	0.42 ± 0.06	0.39 ± 0.07
10	Random Forest	0.56 ± 0.06	0.59 ± 0.04	0.60 ± 0.05	0.34 ± 0.07	0.39 ± 0.06	0.37 ± 0.06
11	Random Forest	0.55 ± 0.05	0.56 ± 0.04	0.57 ± 0.05	0.31 ± 0.07	0.37 ± 0.06	0.35 ± 0.07
12	Random Forest	0.38 ± 0.06	0.42 ± 0.06	0.40 ± 0.05	0.31 ± 0.07	0.34 ± 0.07	0.33 ± 0.07
13	Random Forest	0.37 ± 0.07	0.38 ± 0.07	0.36 ± 0.07	0.25 ± 0.07	0.29 ± 0.07	0.28 ± 0.08
14	Random Forest	0.35 ± 0.07	0.34 ± 0.06	0.33 ± 0.06	0.25 ± 0.08	0.25 ± 0.07	0.25 ± 0.07
15	Random Forest	0.31 ± 0.07	0.30 ± 0.08	0.32 ± 0.06	0.19 ± 0.08	0.19 ± 0.08	0.19 ± 0.07
5	XGBoost	0.71 ± 0.06	0.75 ± 0.04	0.75 ± 0.03	0.67 ± 0.07	0.69 ± 0.04	0.68 ± 0.04
6	XGBoost	0.66 ± 0.05	0.71 ± 0.03	0.70 ± 0.04	0.60 ± 0.07	0.64 ± 0.04	0.63 ± 0.05
7	XGBoost	0.65 ± 0.06	0.68 ± 0.04	0.68 ± 0.04	0.55 ± 0.07	0.61 ± 0.04	0.60 ± 0.05
8	XGBoost	0.61 ± 0.05	0.64 ± 0.04	0.64 ± 0.04	0.52 ± 0.05	0.55 ± 0.05	0.56 ± 0.05
9	XGBoost	0.60 ± 0.05	0.62 ± 0.04	0.63 ± 0.04	0.39 ± 0.07	0.42 ± 0.06	0.39 ± 0.07
10	XGBoost	0.56 ± 0.06	0.59 ± 0.05	0.59 ± 0.05	0.34 ± 0.07	0.40 ± 0.06	0.38 ± 0.06
11	XGBoost	0.55 ± 0.05	0.56 ± 0.04	0.56 ± 0.05	0.31 ± 0.07	0.38 ± 0.06	0.37 ± 0.07
12	XGBoost	0.39 ± 0.07	0.42 ± 0.07	0.4 ± 0.06	0.31 ± 0.07	0.36 ± 0.07	0.34 ± 0.07
13	XGBoost	0.38 ± 0.07	0.39 ± 0.07	0.36 ± 0.07	0.27 ± 0.08	0.30 ± 0.07	0.30 ± 0.07
14	XGBoost	0.37 ± 0.07	0.35 ± 0.06	0.35 ± 0.06	0.27 ± 0.08	0.26 ± 0.07	0.27 ± 0.07
15	XGBoost	0.31 ± 0.07	0.31 ± 0.07	0.33 ± 0.07	0.21 ± 0.07	0.20 ± 0.08	0.20 ± 0.07

models, which perform worse at lower cutoffs and better at higher cutoffs when compared to the other machine learning models. For text logs, kNN is a very strong performer at all PHQ-9 and GAD-7 cutoffs. Yet, kNN is not the best or worst model at any cutoff with call logs and all logs.

Impact of number of weeks of data. The number of weeks of data had very little impact on screening. There was some variation at higher PHQ-9 and GAD-7 cutoffs for the deep learning models, but as mentioned, these models had lower average F1 scores and high standard deviation. Given these results, we suspect even the models that used longer time series likely relied heavily on data from the most recent two weeks. Both the PHQ-9 and GAD-7 screening surveys ask participants to report on symptoms from the last two weeks so it would be logical that the most recent two weeks of data would be most predictive. Likewise, the prevalence of depression and anxiety symptoms can change over time so more historic logs could be misleading. Overall, our findings

Table 4.8: Comparison of the average ± standard deviation of the F1 scores of the **deep learning** models with different screening cutoffs and two weeks of data.

			Depression			Anxiety	
Cutoff	Method	Call	Text	All	Call	Text	All
5	GRU	0.83 ± 0.00	0.84 ± 0.00	0.84 ± 0.00	0.76 ± 0.01	0.76 ± 0.01	0.76 ± 0.01
6	GRU	0.81 ± 0.01	0.81 ± 0.01	0.81 ± 0.00	0.72 ± 0.01	0.72 ± 0.01	0.73 ± 0.01
7	GRU	0.79 ± 0.01	0.79 ± 0.01	0.78 ± 0.01	0.68 ± 0.00	0.68 ± 0.02	0.67 ± 0.02
8	GRU	0.78 ± 0.00	0.78 ± 0.01	0.78 ± 0.01	0.62 ± 0.02	0.62 ± 0.03	0.60 ± 0.04
9	GRU	0.75 ± 0.01	0.75 ± 0.01	0.75 ± 0.01	0.52 ± 0.04	0.32 ± 0.12	0.52 ± 0.05
10	GRU	0.66 ± 0.01	0.68 ± 0.01	0.66 ± 0.02	0.52 ± 0.07	0.40 ± 0.06	0.44 ± 0.09
11	GRU	0.57 ± 0.03	0.61 ± 0.01	0.59 ± 0.02	0.49 ± 0.08	0.48 ± 0.07	0.46 ± 0.07
12	GRU	0.51 ± 0.03	0.47 ± 0.07	0.51 ± 0.03	0.28 ± 0.10	0.29 ± 0.26	0.32 ± 0.17
13	GRU	0.46 ± 0.05	0.18 ± 0.19	0.47 ± 0.08	0.37 ± 0.21	0.33 ± 0.16	0.38 ± 0.06
14	GRU	0.41 ± 0.09	0.32 ± 0.17	0.29 ± 0.10	0.21 ± 0.08	0.30 ± 0.11	0.34 ± 0.03
15	GRU	0.29 ± 0.14	0.20 ± 0.10	0.29 ± 0.10	0.21 ± 0.04	0.22 ± 0.11	0.28 ± 0.04
5	GRU attention	0.84 ± 0.01	0.84 ± 0.00	0.84 ± 0.01	0.76 ± 0.01	0.76 ± 0.00	0.76 ± 0.00
6	GRU attention	0.80 ± 0.01	0.81 ± 0.01	0.81 ± 0.00	0.72 ± 0.00	0.73 ± 0.01	0.73 ± 0.01
7	GRU attention	0.78 ± 0.00	0.79 ± 0.01	0.78 ± 0.00	0.68 ± 0.00	0.69 ± 0.01	0.68 ± 0.00
8	GRU attention	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.00	0.64 ± 0.01	0.65 ± 0.01	0.65 ± 0.01
9	GRU attention	0.75 ± 0.00	0.76 ± 0.00	0.75 ± 0.01	0.55 ± 0.06	0.55 ± 0.03	0.59 ± 0.02
10	GRU attention	0.68 ± 0.01	0.67 ± 0.00	0.68 ± 0.00	0.46 ± 0.13	0.34 ± 0.23	0.48 ± 0.09
11	GRU attention	0.61 ± 0.01	0.61 ± 0.00	0.61 ± 0.01	0.44 ± 0.25	0.35 ± 0.32	0.41 ± 0.23
12	GRU attention	0.58 ± 0.01	0.51 ± 0.06	0.59 ± 0.04	0.51 ± 0.06	0.28 ± 0.21	0.49 ± 0.07
13	GRU attention	0.49 ± 0.03	0.31 ± 0.12	0.51 ± 0.02	0.27 ± 0.22	0.19 ± 0.22	0.41 ± 0.14
14	GRU attention	0.29 ± 0.21	0.03 ± 0.07	0.40 ± 0.11	0.35 ± 0.12	0.07 ± 0.17	0.36 ± 0.11
15	GRU attention	0.38 ± 0.10	0.24 ± 0.20	0.28 ± 0.20	0.29 ± 0.09	0.32 ± 0.11	0.23 ± 0.11

indicate that more than two weeks of log data provides no benefit. While we used a retrospective app, this finding is particularly useful for future screening research that uses prospective apps.

Impact of number of principal components. Including more than the first principal component does not improve the performance of the traditional and ensemble machine learning models. This first principal component covers 48.2%, and 33.6%, 30.0% of the feature variance for call logs, text logs, and all logs, respectively. In comparison, all four principal components respectively cover 70.8%, 68.3%, and 58.6% of the feature variance for these types of logs. Yet, this additional feature variance was not useful for our classification models.

4.4.7 Discussion

Intended use of DepreST-CAT. The DepreST-CAT data is intended to be used by academic researchers to further research alternative mental illness screening technologies and provide insights into communication patterns during COVID-19. The log data could be useful in both supervised and unsupervised machine learning models. DepreST-CAT logs can also be combined with logs in other mental illness datasets to increase data quantity for modeling. In addition to modeling with the DepreST-CAT logs, the collection and screening methodologies detailed in this paper may also provide valuable information for the future of passive mental illness screening.

DepreST-CAT modeling challenges and future opportunities. Our screening results revealed that DepreST-CAT logs were not able to detect either depression or anxiety well at higher mental illness screening scores. We thus urge future research to explore modeling strategies to increase the screening ability of the DepreST-CAT call and text logs for moderate and moderately severe depression. As anxiety was more challenging to predict than depression, we also suggest improving anxiety screening capabilities with DepreST-CAT smartphone logs as future research. For example, such research could employ different strategies to construct the time series, engineer features, and/or concatenate features into more comprehensive feature vectors. Future research could also experiment with different subsets of the logs, machine learning and deep learning methods, and types of supervised learning techniques. In addition to our binary classification approach, it could be useful to develop multinomial classification and regression models for mental illness screening. Lastly, we suggest determining how COVID-19 related factors may have influenced communication patterns in the DepreST-CAT participants.

DepreST-CAT in context. With 369 participants, DepreST-CAT is the largest dataset of smartphone logs with mental illness labels. It is further unique for including anxiety screening score labels and being collected during the COVID-19 pandemic. Unlike related research with smartphone logs [156, 26, 44, 5, 6], we successfully leveraged deep learning models for mental illness classification. Notably, our deep learning models on the DepreST-CAT logs were able to achieve average F1 scores of 0.84 and 0.76 when screening for mild depression (PHQ-9≥ 5) and mild anxiety (GAD-7≥ 5), respectively. Thus, we were able to screen for mild depression (PHQ-9≥ 5) with DepreST-CAT logs with higher F1 scores than related work [1] was able to screen for moderately severe depression (PHQ-9≥ 15) even when given actual text content. Unfortunately, at the cutoffs for moderate depression (PHQ-9≥ 10) and moderate anxiety (GAD-7≥ 10), our deep learning models on the DepreST-CAT logs were only able to achieve average F1 scores of 0.68 and 0.52, respectively. The unimodal traditional machine learning models in the related work [6] achieved a slightly higher F1 score of 0.72 when screening for moderate depression.

Implications of changing phone usage. There are several popular communication apps that present alternatives communication avenues to traditional phone calls and SMS text messages [157]. Thus, DepreST-CAT captures only a subset of mobile communications. Yet, our screening models demonstrated that this communication subset holds promise for mild depression and anxiety screening. We suggest that future work explores the transferability of these screening models to time series of logs from these popular communication apps. Our multivariate modeling approach could also encompass communication time series from any number of sources. As such, a promising future work direction would be to determine how many communication sources are required to construct clinically useful mental illness screening models. Nonetheless, given that smartphones have become ubiquitous [138], it seems very unlikely that phone calls and SMS texts will cease to exist in the future. Therefore, exploring the potential of smartphone communication logs collected during the COVID-19 pandemic to screen for mental illnesses is still valuable.

4.5 Outlook

Contributions. The research in this chapter proposed two new feature sets [5, 6] derived from text logs that can be used in conjunction with other passive sensing features that can be used for detection mental illnesses. However, the biggest contribution of this research is unequivocally the DepreST-CAT dataset [7]. This dataset can be used by researchers to test the screening effectiveness of many different feature sets and methods to screen for both anxiety and depression.

Ethical considerations. Call logs and text message logs without content are a very promising screening modality due to their innocuousness. While there is always a possibility of bad agents using mental illness screening technologies inappropriately, smartphone call and text logs are relatively safe screening modalities for the following two reasons. First, smartphones require explicit permission from users before giving apps access to data stored on the phone. Second, call and text logs can be collected for screening purposes without identifying and/or personal information. For example, we one-way hashed all names and numbers as well as removed message content in DepreST-CAT logs prior to sending the logs to our secure server. Thus, no identifiable information is exposed in the DepreST-CAT logs.

Reply latency. Leveraging an XGBoost model with a single principal component created from text message reply latency features [5], we were able to screen for depression with an F1 score of 0.67, AUC of 0.72, and Accuracy of 0.69. Thus, text message reply latencies are surprisingly adept at screening for depression, though insufficient by themselves. Combining them with other features derived from the text messages is an option. However, only 68 participants in the Moodable and EMU datasets had two reply latencies so the number of participants was smaller than those used for the experiments with text content [1, 3] or time series [6] features. The larger number of participants may have contributed to the greater success of these features at depression screening than reply latencies. Combining the latency features with text content features would remove the privacy benefit of using the logs without text content and the different feature sets prefer different feature reduction techniques. Likewise, combining latency and time series features did not improve screening results over just the latency features or just the time series features [6].

Communication time series. On communication time series from the Moodable and EMU logs, we determined machine learning models were better at depression screening on features derived from the time series than the raw time series [6]. We thus used time series features in the machine learning models when assessing the screening capabilities of DepreST-CAT logs [7]. With these features, the best average F1 scores of the depression screening models was 0.72 for the Moodable/EMU text logs but only 0.59 for the DepreST-CAT text logs at PHQ-9 cutoff 10. The best average F1 score was even lower at 0.45 with these features for anxiety screening models at GAD-7 cutoff 10. While the GRU models on the raw time series performed even worse for anxiety at cutoff 10, they achieved a more respectable F1 score of 0.68 when screening for depression at cutoff 10. While the screening results with the time series features are more impressive at lower PHQ-9 and GAD-7 cutoffs, they are eclipsed by the results of the deep learning models.

Deep learning versus machine learning results. Deep learning models performed well on time series of the DepreST-CAT logs at the cutoffs for mild depression (PHQ-9≥ 5) and anxiety (GAD-7≥ 5); the F1 score was 0.84 for depression screening and 0.76 for anxiety screening. However, the performance of these models quickly deteriorates for higher screening cutoffs. While the GRU models still performed better at the cutoff for moderate depression (PHQ-9≥ 10) than the machine learning models, the machine learning models performed better at the cutoff for moderate anxiety (GAD-7≥ 10) than the GRU models. Neither machine learning nor deep learning models performed well at the cutoff for moderately severe depression (PHQ-9≥ 15) and anxiety (GAD-7≥ 15). Overall, it seems the future of this research would best be served by deep learning models. The lack of intepretability of time series removes much of the benefit of machine learning models.

Text logs versus text content results. We were able to explore the screening potential of the DepreST-CAT logs across multiple screening cutoffs. The results strongly indicate that logs without content are most useful for screening at lower depression screening cutoffs [1]. However, with the existing lexical category features derived from the text content, machine learning models were best able to screen for depression at PHQ-9 of 15 [1, 3], though admittedly not all depression screening cutoffs were explored in this smaller dataset. These results suggests that a model that combines communication patterns and text content from the logs would be able to screen across multiple cutoffs. However, these results were obtained from different datasets, and therefore not directly comparable. Further, this research is still in it's infancy in many ways and the results may be highly subject to the preprocessing of the text logs. Thus, future research may discover a preprocessing strategy and method that can screen for depression at higher cutoffs well without the content of the text logs.

CHAPTER 5

DEPRESSION SCREENING WITH GENERATIVE MODELING

Context: Advances in depression screening research using text messages are limited by the small size of the datasets, weak depression signals in individual messages, and the inability to share private text message content across research teams.

Objective: This research explores the potential of conditional generative models to generate text messages that can be used for depression screening.

Methods: We identify and adapt conditional adversarial approaches for text generation. This involves three weighting strategies and three feedback mechanisms, which can be assembled into nine conditional sequence generative adversarial networks (cSeqGAN). We then conduct a comparative evaluation of these cSeqGAN models to assess the quality of the generated text as well as the usefulness for depression screening. While we implemented our conditioning within a SeqGAN architecture, the weighting strategies and feedback mechanisms are applicable for other generative models.

Findings: The cSeqGAN models produced more realistic text than the unconditioned SeqGAN models, but the unconditioned SeqGAN models produced texts more useful for depression screening than the cSeqGAN models. The comparative study is valuable to further research in the text generation domain.

This chapter covers material from the following papers:

ML Tlachac, Walter Gerych, Kratika Agrawal, Benjamin Litterer, Nicholas Jurovich, Saitheeraj Thatigotla, Jidapa Thadajarassiri, Elke Rundensteiner, “Text Generation to Aid Depression Detection: A Comparative Study of Conditional Sequence Generative Adversarial Networks”, in Revision [8]

5.1 Conditional Generative Adversarial Networks

To advance critical modeling innovations for medical diagnostic applications [20], it is important to generate quality text data to increase the data quantity and amplify the signal for a class label in the data. Further, the generation process can help mitigate privacy concerns associated with identifiable named entities. Since it is challenging to collected large anonymized healthcare datasets [20, 158], data generation promises to improve diagnostic and prognostic modeling. However, these applications require labeled data for training purposes. For example, text data [24, 25, 1] have demonstrated usefulness in screening for mental illness with less burden than traditional screening surveys. Such applications would greatly benefit from *conditional* text generative models [159].

When data quantities are large, a separate unconditioned generative model could be built for each class. However, there are multiple advantages to *conditional* text generation where the generated text can be controlled to match specific classes. One, conditional models are trained with data from all classes which allows for the sharing of parameters, therefore resulting in better language learning and applicability to smaller datasets. The first approaches for conditional text generation relied on *multiple generators* [160] or *discriminators* [161], which negates many benefits of conditional modeling. Also, this does not scale well for use cases with many classes. There are a few other promising text generation methods [162, 158, 163] with a unified architecture containing only one generator and discriminator. However, some fundamental conditional approaches for images such as conditional GANs [159] have yet to be adapted for text.

There is a need for a strategic assessment and comparative study of fundamental conditional text generation approaches. The goal of conditional generation models is to input a small labeled dataset to generate a large quantity of realistic labeled data with data samples for each of the classes. To work well for small datasets, such conditional generative models should have a unified architecture that allows for parameter sharing during training of any of the classes. While there are a few conditional text approaches that leverage a unified architecture [162, 158, 163], there has been no identification nor categorization of fundamental conditional design approaches. Further, it is unknown which designs will be most effective for small datasets.

Our research is the first to conduct an extensive study of unified text generative architectures, i.e., those that have a single generator and a single discriminator. These unified architectures are not only appropriate for small healthcare datasets, but also easily scale for any number of classes – thus increasing their utility. By analyzing existing methods in the literature [159, 160, 161, 162, 158, 164], we identify two core orthogonal design dimensions compatible with unified sequence generation architectures; namely, three weighting strategies and three feedback mechanisms.

To tackle the problem of assessing different conditional text generative approaches, we thus compose each of the alternate weighting strategies and feedback mechanisms into a total of nine fundamental conditional text generation models. While implemented with the popular SeqGAN architecture, these approaches could be integrated with any recurrent generative adversarial model. We then leverage this family of nine models for generating text with depression labels. In addition to datasets with depression screening labels, we also demonstrate our methods on a popular dataset in the related domain of sentiment detection [63]. We performed a comprehensive evaluative study to evaluate the ability of the cSeqGAN and non-conditional SeqGAN models to generate realistic and predictive text. In addition to leveraging standard machine learning metrics, we also designed a user study to obtain human assessment of the generated text.

There are three notable contributions of our research. The first contribution is the identification and adaptation of three weighting strategies and three feedback mechanisms to design conditional text generation models. The second contribution is the assemblage of nine scalable cSeqGAN models that are applicable to small datasets. The third contribution is a comparative evaluation of these nine cSeqGAN models on three real-world datasets.

5.1.1 Existing Conditional Generative Models

There have been a few recent attempts at conditional text generation. The first approach [161] involves introducing adversarial training to a VAE by incorporating a discriminator for each class option. Meanwhile, SentiGAN [160] proposes an architecture with a separate generator for each class option but a single discriminator. These approaches unfortunately do not scale well. While category sentence generative adversarial network (CS-GAN) [162] introduces an auxiliary classifier, only one classifier, one discriminator, and one generator are required regardless of the number of class options. Medical Text Generative Adversarial Network (mtGAN) [158] introduces a conditional constraint to SeqGAN by including features as additional input at every step of the sequence generation process. Most recently, category-aware GAN (CatGAN) [163] uses a hierarchical evolutionary learning algorithm to generate text in an approach that deviates far from the traditional SeqGAN architecture. Some of these text generation approaches are quite promising, but no strategic comparative study has been conducted to compare fundamental conditioning strategies.

5.1.2 Methodological Overview

We compare the three *weighting strategies* and three *feedback mechanisms* we identified for conditional text generation. The *weighting strategies* refer to design choices for how the previously generated word along with the class of the overall text determine the subsequently generated word. The *feedback mechanisms* refer to architectural choices for the critic that determines the realism of the data for the given class. These two design dimensions are orthogonal, in that, each of the possible choices along one dimension can be integrated with all the other choices of the second dimension. That is, a model can be designed to support one of our weighting strategies and one of our feedback mechanisms. In this work, we embed both types of strategies within the popular SeqGAN model architecture [86], thus constructing a total of nine unique conditional sequence generative adversarial networks (cSeqGAN) architectures. Given the unified architecture of these nine cSeqGAN models, we anticipate that the models can be trained to generate text with the relatively small text datasets available for depression screening.

5.1.3 Sequence GANs

As with traditional GANs [165], SeqGANs [86] consist of a generator and a discriminator involved in a minimax game that iteratively improves text quality. However, SeqGAN makes notable changes to the generation process to make it applicable for sequences of discrete tokens. For this paper, the tokens are words.

SeqGAN leverages a recurrent neural network (RNN) with Long Short Term Memory (LSTM) [150] as the generator. The LSTM maps each embedded word $x_t \in x_1, \ldots, x_T$ to a hidden state h_t to create a sequence of hidden states $h_1, \ldots, h_T$. Notably, the LSTM implements the update function g in

$$h_t = g(h_{t-1}, x_t) \tag{5.1}$$

to prevent the vanishing and exploding gradient problem. A softmax output layer then maps these hidden states into an output token distribution.

The SeqGAN generator [86] has the objective of maximizing an expected end reward given the starting state s_0. This expected end reward is calculated using an action-value function of a sequence. Specifically, the REINFORCE algorithm [166] is used to estimate the action-value function. Further, seqGAN [86] evaluates the intermediate state action-value pairs to provide more frequent rewards. This is accomplished using a Monte Carlo search with a roll-out policy that samples the remaining $T - t$ tokens in the sequence. The roll-out policy is repeated multiple times to reduce variance and improve the reward estimations.

As is the case with traditional GANs [165], SeqGAN uses a convolutional neural network (CNN) [167] as discriminator. This CNN is responsible for classifying each input sentence as real or fake. Specifically, a convolution operation is applied to the token embeddings to produce feature maps and the feature maps are pooled. The goal [86] is to minimize the sigmoid cross entropy loss.

5.1.4 Design Dimension: Weighting Strategies

We study three different weighting strategies to condition SeqGAN models for conditional text generation. In particular, we study *sentence weighting* which directly adapts conditional GANs [159] to be applicable for text instead of images, and two different unit weighting strategies inspired by mtGAN [158]. These latter strategies include *single unit weighting* in which the generative model is repeatedly conditioned when generating each word of the sentence, and *dual unit* *weighting* which likewise conditions the generation of each word but learns separate weights for the words and labels respectively. These three weighting strategies are applicable for any model with a recurrent network for a generator.

Sentence Weighting

In this first approach, we condition the generator only once while generating each sentence. Specifically, the initial input to the generator is x_0, where x_0 is the concatenation of z and y. $z \in \mathbb{R}^n$ is a draw from a multivariate Gaussian distribution, and y is the class we aim to generate. The generator is an LSTM, such that $x_t = LSTM(x_{t-1})$. Thus, the t word in the sentence is simply the output of the LSTM conditioned on the previous output, such that the initial input contains information of the class to be generated. This means the LSTM generator is only conditioned on the class *once* and is then tasked with generating the entire sentence.

Single Unit Weighting

We modify the previous strategy by conditioning the LSTM on the desired class *at each step* instead of once at the start. Specifically, the Single Unit Weighting takes the form of $x_t = LSTM(W(x_{t-1} \oplus y))$, where W is a weight on the input and $\oplus$ is the concatenation operator. Notably, both the word embedding x_{t-1} and class embedding y share the same weight W.

Dual Unit Weighting

We study another weighting strategy in which the class embedding and word embedding are *separate* inputs into the LSTM generator. In this approach, the generator takes the form of $x_t = LSTM(Wx_{t-1} \oplus Vy)$. W and V are separate weight matrices for the word embedding x_{t-1} and class embedding y. This approach with separate weights is in contrast to the previous approach that used a single weight for both features and labels.

5.1.5 Design Dimension: Feedback Mechanisms

We also study three feedback mechanisms to condition SeqGAN models for conditional text generation, compared in Figure 5.1. These feedback mechanisms can coexist with any of the aforementioned weighting strategies. The first feedback mechanism, which we refer to as *single task feedback*, directly adapts the conditional GANs discriminator [159] for sequence input like text. The other two feedback mechanisms involve two separate tasks: assessing realness and assessing class appropriateness. We refer to the mechanism that only uses one critic as *dual task feedback* and the mechanism that uses two critics as *dual critic feedback*. This latter strategy is modeled after the dual critics in CS-GAN for text generation [162] and GAN-control for image generation [164]. While we use a CNN discriminator like SeqGAN, these three feedback mechanisms are applicable for any generative adversarial model that uses a discriminator.

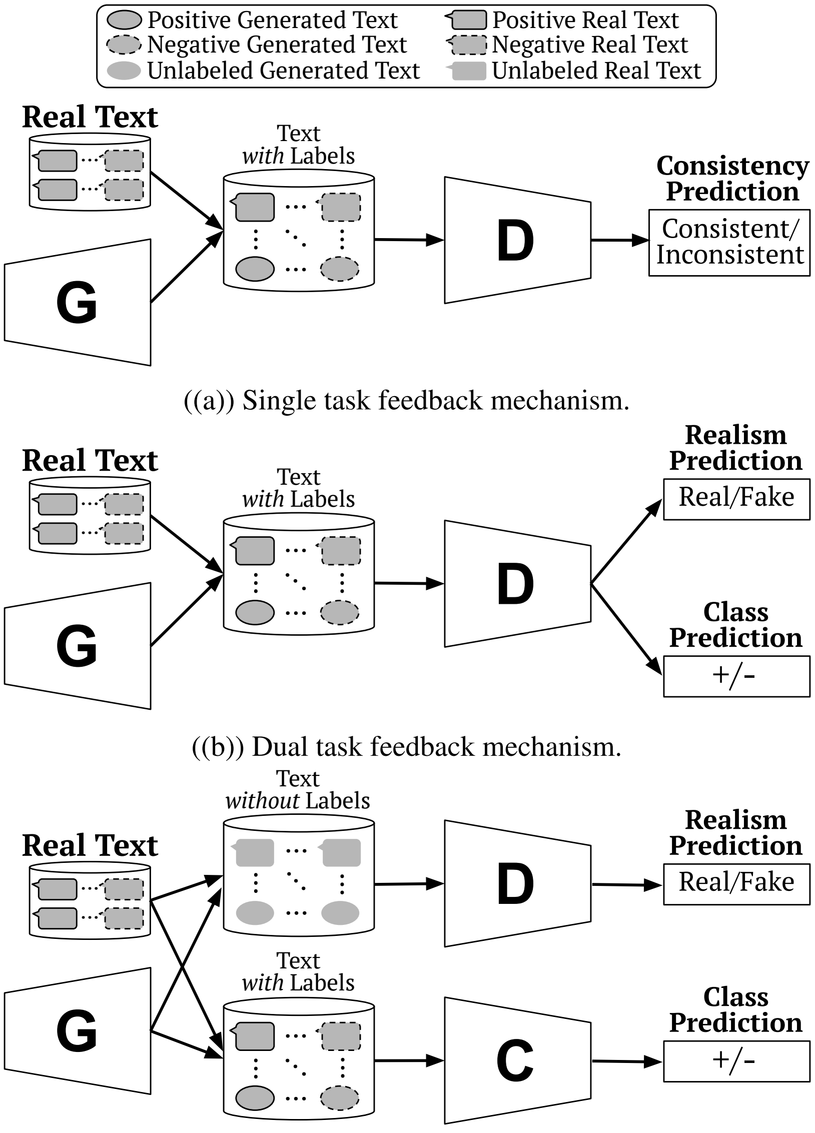
Single Task Feedback

In this feedback mechanism, a single discriminator network decides whether the generated text is realistic *and* whether the generated text matches the condition y as a *single task*. Specifically, the discriminator D is a CNN that is trained with the following cross entropy loss function L_D:

$$\begin{aligned} L_D = \mathbb{E}_{d_g \sim (y,z)}[-log(1 - D(G(y,z),y))] \\ + \mathbb{E}_{d_r \sim data}[log(D(d_r, y))] \end{aligned} \tag{5.2}$$

Thus, the discriminator is trained to distinguish between real and generated data *given knowledge of the class label*, allowing the discriminator to reject text that is in general realistic but does not match the conditioning label. In this case, the generator G is trained with the loss L_G:

$$L_G = \mathbb{E}_{d_g \sim (y,z)}[log(D(G(y,z),y))] \tag{5.3}$$



((a)) Single task feedback mechanism.

((b)) Dual task feedback mechanism.

((c)) Dual critic feedback mechanism.

Figure 5.1: The feedback mechanisms include: (a) a discriminator that performs a single task, (b) a discriminator that performs two tasks, and (c) a discriminator that performs a single task and a classifier that performs a single task. The realism prediction in (b) and (c) only consider the text while the consistency prediction in (a) considers the realness of the concatenated text and label.

Dual Task Feedback

In this feedback mechanism, a single discriminator network performs two separate tasks: it determines whether the generated text is realistic and it determines whether the generated text matches the class label. Let $D^1_i(\tilde{x})$ be D's prediction probability for $y = i$, and let $D^2(\tilde{x})$ be D's prediction for whether $\tilde{x}$ is real or generated. D is then trained with the loss L_D:

$$\begin{aligned} L_D = & \mathbb{E}_{d_g \sim (y,z)}[-log(1 - D^2(G(y, z)))] \\ & + \mathbb{E}_{d_r \sim data}[log(D^2(d_r))] \\ & + \mathbb{E}_{(x,y) \sim data}[log(D^1_y(x))] \end{aligned} \tag{5.4}$$

In this case, the discriminator is trained to distinguish between real and generated text, while also being trained to correctly classify text. The generator is trained with L_G:

$$\begin{aligned} L_G = & \mathbb{E}_{d_g \sim (y,z)}[log(D^2(G(y, z), y))] \\ & + \mathbb{E}_{(x,y) \sim (y,z)}[log(D^1_y(x))] \end{aligned} \tag{5.5}$$

In this setting the generator is trained to fool the discriminator into classifying it as real, while also being given the correct class by the discriminator.

Dual Critic Feedback

This mechanism aims to separate the tasks of determining real from fake and correctly predicting classes. To this end, we utilize a *dual critic* approach in which the discriminator D only determines real text from fake text while a completely separate classifier C performs the classification. In this case, the discriminator D is trained as follows:

$$\begin{aligned} L_D = & \mathbb{E}_{d_g \sim (y,z)}[-log(1 - D(G(y, z)))] \\ & + \mathbb{E}_{d_r \sim data}[log(D(d_r))] \end{aligned} \tag{5.6}$$

Thus, it is trained to determine real text from fake text with no information regarding the class.

The classifier is meanwhile trained to accurately predict classes on the real data with loss L_C:

$$L_C = \mathbb{E}_{(x,y)\sim(y,z)}[log(C(x))] \tag{5.7}$$

Lastly, the generator is trained to generate realistic enough text to fool D while achieving accurate classification of the generated text from C:

$$\begin{aligned} L_G = \mathbb{E}_{d_g\sim(y,z)}[log(D(G(y,z)))] \\ + \mathbb{E}_{(x,y)\sim(y,z)}[log(C(x))] \end{aligned} \tag{5.8}$$

5.1.6 Datasets

We leverage two datasets with depression screening labels in this research. Similar to many healthcare datasets, these datasets are small and would benefit from generative modeling to increase data quantity and improve anonymization. Additionally, we demonstrate our methods on a popular publicly available dataset in the related domain of sentiment detection for replicability purposes.

Clinical interviews. The DAIC-WOZ dataset [22, 111] contains clinical interview transcripts labeled with PHQ-8 scores. Each of the 189 participants were asked a subset of core clinical interview questions with followup questions as needed. We consider each of the 3774 sentences in the transcripts as separate instances.

Text messages. The text messages in the combined Moodable and EMU datasets are labeled with PHQ-9 scores. The PHQ-8 and PHQ-9 share the same moderate depression cutoff of 10 [92]. Since the texts capture real communications and are not in response to clinical prompts, we consider only texts from the participants with more polarizing PHQ-9 scores. Specifically, we only use the 5360 text messages sent within the prior two weeks by participants with PHQ-9≤ 5 and PHQ-9≥ 15.

Movie reviews. The publicly available Stanford's Large Movie Review Dataset [168], commonly referred to as the Internet Movie Database (IMDb) Movie Review Dataset, is popular for binary sentiment classification. The notably brief reviews are highly polarized. While it is a large dataset, we only use 4503 reviews to mimic the size of the other datasets with depression labels.

5.1.7 Experimental Setup

We implement our nine cSeqGAN architectures within the Texygen benchmarking platform [87] for unconditioned text generation. Additionally, We also generate text with unconditioned SeqGAN models as baselines for comparison; a different SeqGAN model is required for each class. To train the generative models, we down sample each dataset to the size of the minority class: 2133 for movie reviews, 1207 interview replies, and 2680 for text messages. The discriminators are pretrained for 50 epochs before the 50 adversarial training epochs. Each model generates 4480 labeled texts. We run each model five times to obtain a confidence interval.

5.1.8 Machine Experimental Evaluation

We evaluate the generated text quality with two established and popular text generation metrics [87]. The first of these metrics is the negative log-likelihood (NLL_{gen}) which is an output of the recurrent generator. A lower NLL_{gen} is indicative of better generated sentence diversity. In contrast, the BLEU score [91] assesses the similarity of the generated sentences with the real sentences. Effectively, a higher score is indicative of more realistic text. Given the short length of our input data, we report on the BLUE-2 scores which specifically assesses 2-gram matches.

We further compare the predictive value of the real text and generated text. BERT [70] is a pretrained language representation model that can be fine-tuned for many tasks. Previously, BERT classifiers have proven effective at classifying short texts, such as sentiment from movie reviews [169] and disaster events from tweets [170]. Thus, we use BERT classifiers with the parameters successful in related work [169]: learning rate of $2x10^{-5}$, 4 training epochs, and batch size of 32.

We only consider texts with at least two words for BERT input. We reserve 300 positive and 300

negative instances from each real dataset to use as testing data. As only four of the 135 conditional generative model runs failed to generate instances of each class with sufficient length, we proceed with 1200 (or the minimum class count) randomly sampled instances from each class as training data. For the real data and SeqGANs output, we also sample 1200 instances of each class.

As we ran each conditional model five times, we have five NLL_{gen}, BLUE-2, and accuracy scores. The average and standard deviation of these scores are reported in Tables Table 5.1-Table 5.3. For the unconditioned SeqGAN models, we average the NLL_{gen} scores for all ten models (5 positive and 5 negative). To calculate BLEU-2 for SeqGAN, we combine the positive and negative output from two models.

5.1.9 Human Experimental Evaluation

We further conduct Turing tests to evaluate the texts by having human rate the text samples. While the related literature [86, 171] only uses Turing tests to assess text quality, we also have humans assess their predictive ability. We recruited 32 STEM students to evaluate the texts. For each dataset, we formed surveys consisting of 66 text samples. Only samples with more than two words were eligible for the surveys. Three positive and three negative texts were randomly selected from the real data, unconditioned model, and nine conditioned models. Since we ran each model five times, we construct five surveys per dataset.

For each sample in the surveys, we ask two binary questions. The first question assesses the predictive quality and the second question assess the realism. For example, the first question for text messages asked *“Is the writer of this text message depressed or not depressed?”* with options “depressed” or “not depressed”. The second question asked *“Was this message created by a human or a computer?”* with options “human” or “computer”.

The questions in each survey were answered by 3 participants. Thus, each of the 5 surveys provides 18 assessments of text sample predictiveness and realness for each type of conditioning. In each survey, we report on the accuracy of responses for each model.

5.1.10 Experimental Results

The results of our experiments are in Tables Table 5.1-Table 5.3. Recall, we assess the generated texts with both a machine evaluation approach and a human evaluation approach. Unlike for the conditioned models, we needed a separate unconditioned SeqGAN model for each class. While the NLL_{gen} and BLEU-2 scores were not applicable for the real text, we did calculate the accuracies of the real text for comparison purposes.

Machine Evaluation. From the results, we observe that the average BLEU-2 scores are higher for all of the cSeqGAN models than the unconditioned SeqGAN models for each dataset. The differences are largest for the clinical interviews in Table Table 5.1 where each of the conditional models have an average BLEU-2 score more than 0.3 higher than the unconditioned models. This indicates that the conditioned models produced more realistic text, likely due to parameter sharing.

Unfortunately, the generated text from the cSeqGAN models were not particularly predictive. For all three datasets, the generated text from the unconditioned models performed better in the BERT classifiers than the text from the conditioned models. Interestingly, the depression screening ability of the generated interview transcripts exceeded that of the real interview transcripts; this indicates that the generative models amplified the signal for the class label. This is not true for the text messages and movie reviews, where real data proved more predictive than generated data.

When considering only the impact of weighting strategies and feedback mechanisms on the evaluation metrics, some patterns emerge. Notably, the lowest NLL_{gen} score was achieved by the models using the sentence weighting strategy. In contrast, the feedback mechanisms that yield the highest NLL_{gen} scores for all weighting strategies are different for each dataset: single task for the movie reviews, dual task for the clinical interviews, and dual critic for text messages.

Comparing the BLEU scores of the nine cSeqGAN architectures also reveal some patterns. Dual unit weighting paired with dual task feedback has very high average BLEU-2 scores for clinical interviews and movie reviews while single unit weighting and single task feedback has very high average BLEU-2 scores for text messages and movie reviews.

Table 5.1: **Clinical Interviews:** Average ± standard deviation of evaluation metrics for the five conditional models with the same weighting strategy and feedback mechanism. For the unconditioned models, these values were calculated for five models trained on depressed sentences and five models trained on not depressed sentences. The machine evaluation accuracy was calculated with BERT classifiers. The accuracy of both human evaluation tasks are also displayed.

		Machine Evaluation			Human Evaluation	
Weighting	Feedback	NLL	BLEU-2	Accuracy	Realness	Predictiveness
Unconditioned	Unconditioned	**0.710 ± 0.084**	0.203 ± 0.008	**0.548 ± 0.004**	0.689 ± 0.143	0.578 ± 0.044
Sentence	Single Task	0.723 ± 0.013	0.544 ± 0.013	0.500 ± 0.008	0.500 ± 0.136	0.467 ± 0.232
Sentence	Dual Task	0.720 ± 0.020	0.550 ± 0.046	0.500 ± 0.008	0.600 ± 0.206	0.567 ± 0.065
Sentence	Dual Critic	0.708 ± 0.024	0.504 ± 0.029	0.504 ± 0.007	**0.745 ± 0.129**	0.589 ± 0.109
Single Unit	Single Task	0.733 ± 0.032	0.527 ± 0.040	0.512 ± 0.006	0.622 ± 0.231	0.544 ± 0.108
Single Unit	Dual Task	0.767 ± 0.032	0.516 ± 0.028	0.512 ± 0.009	0.622 ± 0.065	0.456 ± 0.082
Single Unit	Dual Critic	0.756 ± 0.018	0.539 ± 0.037	0.498 ± 0.004	0.522 ± 0.083	0.444 ± 0.121
Dual Unit	Single Task	0.740 ± 0.026	0.518 ± 0.023	0.506 ± 0.004	0.645 ± 0.156	0.444 ± 0.035
Dual Unit	Dual Task	0.768 ± 0.074	**0.556 ± 0.035**	0.504 ± 0.007	0.622 ± 0.187	0.500 ± 0.117
Dual Unit	Dual Critic	0.757 ± 0.030	0.504 ± 0.028	0.494 ± 0.006	0.711 ± 0.181	**0.600 ± 0.102**
Real Data	Real Data	NA	NA	0.485 ± 0.004	0.800 ± 0.056	0.567 ± 0.089

Table 5.2: **Text Messages:** Average ± standard deviation of evaluation metrics for the five conditional models with the same weighting strategy and feedback mechanism. For the unconditioned models, these values were calculated for five models trained on depressed texts and five models trained on not depressed texts. The machine evaluation accuracy was calculated with BERT classifiers. The accuracy of both human evaluation tasks are also displayed.

		Machine Evaluation			Human Evaluation	
Weighting	Feedback	NLL	BLEU-2	Accuracy	Realness	Predictiveness
Unconditioned	Unconditioned	0.247 ± 0.052	0.223 ± 0.060	**0.674 ± 0.018**	0.500 ± 0.099	0.533 ± 0.125
Sentence	Single Task	0.216 ± 0.005	0.335 ± 0.035	0.487 ± 0.033	0.278 ± 0.182	0.533 ± 0.156
Sentence	Dual Task	**0.214 ± 0.012**	0.315 ± 0.018	0.520 ± 0.029	0.400 ± 0.223	0.433 ± 0.065
Sentence	Dual Critic	0.220 ± 0.004	0.317 ± 0.032	0.478 ± 0.039	0.511 ± 0.177	**0.567 ± 0.042**
Single Unit	Single Task	0.229 ± 0.006	**0.340 ± 0.019**	0.469 ± 0.041	0.433 ± 0.178	0.500 ± 0.070
Single Unit	Dual Task	0.225 ± 0.009	0.326 ± 0.014	0.515 ± 0.039	0.456 ± 0.249	0.522 ± 0.147
Single Unit	Dual Critic	0.234 ± 0.009	0.306 ± 0.034	0.506 ± 0.036	0.511 ± 0.154	0.533 ± 0.075
Dual Unit	Single Task	0.230 ± 0.008	0.325 ± 0.020	0.503 ± 0.036	0.444 ± 0.126	0.500 ± 0.070
Dual Unit	Dual Task	0.225 ± 0.003	0.305 ± 0.033	0.509 ± 0.023	**0.589 ± 0.178**	0.511 ± 0.042
Dual Unit	Dual Critic	0.235 ± 0.004	0.336 ± 0.018	0.506 ± 0.031	0.478 ± 0.120	0.500 ± 0.061
Real Data	Real Data	NA	NA	0.711 ± 0.027	0.856 ± 0.044	0.522 ± 0.097

Table 5.3: **Movie Reviews:** Average ± standard deviation of evaluation metrics for the five conditional models with the same weighting strategy and feedback mechanisms. For the unconditioned models, these values were calculated for five models trained on positive reviews and five models trained on negative reviews. The machine evaluation accuracy was calculated with BERT classifiers. The accuracy of both human evaluation tasks are also displayed.

		Machine Evaluation			Human Evaluation	
Weighting	Feedback	NLL	BLEU-2	Accuracy	Realness	Predictiveness
Unconditioned	Unconditioned	**1.736 ± 0.073**	0.269 ± 0.021	**0.773 ± 0.016**	0.656 ± 0.089	**0.700 ± 0.188**
Sentence	Single Task	2.072 ± 0.055	0.370 ± 0.017	0.490 ± 0.019	0.622 ± 0.022	0.411 ± 0.114
Sentence	Dual Task	2.055 ± 0.065	0.371 ± 0.025	0.519 ± 0.016	0.500 ± 0.149	0.611 ± 0.208
Sentence	Dual Critic	2.016 ± 0.071	0.384 ± 0.019	0.496 ± 0.042	0.500 ± 0.099	0.511 ± 0.133
Single Unit	Single Task	2.286 ± 0.120	0.395 ± 0.009	0.506 ± 0.032	0.589 ± 0.075	0.322 ± 0.155
Single Unit	Dual Task	2.185 ± 0.044	0.387 ± 0.008	0.489 ± 0.043	**0.689 ± 0.155**	0.511 ± 0.074
Single Unit	Dual Critic	2.229 ± 0.031	0.359 ± 0.033	0.516 ± 0.030	0.611 ± 0.157	0.478 ± 0.171
Dual Unit	Single Task	2.266 ± 0.024	0.380 ± 0.025	0.483 ± 0.032	0.467 ± 0.167	0.411 ± 0.264
Dual Unit	Dual Task	2.254 ± 0.053	**0.397 ± 0.011**	0.492 ± 0.033	0.567 ± 0.124	0.378 ± 0.108
Dual Unit	Dual Critic	2.174 ± 0.075	0.374 ± 0.016	0.479 ± 0.019	0.533 ± 0.152	0.567 ± 0.226
Real Data	Real Data	NA	NA	0.831 ± 0.015	0.767 ± 0.108	0.811 ± 0.097

Human Evaluation. As is the standard in unconditioned text generation research [86, 171, 172, 89], we tasked our human evaluators to assess the realness of the generated texts. For each of the datasets, our evaluators understandably achieved the highest average accuracies for the real data. Further, the samples from the unconditioned models were neither rated the most realistic nor the least realistic of the generated text for each dataset. Notably, for the interview transcripts and text messages, the text produced by the sentence weighting strategy paired with the single task feedback mechanism was rated the least realistic by our human evaluators.

We also tasked our human evaluators to assess the predictive value of the generated texts, which we anticipated to be a very difficult for depression detection. This hypothesis was validated, as the highest accuracies were 0.60 and 0.57 for interview transcripts and text messages, respectively. Unexpectedly, these accuracies were achieved with text from conditional models and not the real data. In contrast, our human evaluators achieved greater success at classifying the intentionally polarizing movie reviews and were most successful with the real text. Of the generated text, the movie reviews from the unconditioned models had the highest accuracy of 0.70.

5.1.11 Discussion

Contributions. In this research, we identified three weighting strategies and three feedback mechanisms for conditional adversarial text generation. These approaches combine to create nine unique cSeqGAN architectures. We further conduct a comprehensive comparative evaluation of these conditioning approaches on three small text datasets with depression and sentiment labels. As the first comparative study for fundamental conditional text generation strategies, we provide a valuable resource to inform future text generation applications and research. Our study is particularly useful for the healthcare domain where datasets tend to be small and need augmentation.

Limitations. The BERT classifiers were unfortunately unable to classify the real interviews. As this was not the case for the movie review and text message datasets which achieved accuracies of 0.831 and 0.711 respectively, the transcripts may have been too linguistically different from the BERT pretraining corpus [70] or the movie reviews that informed the BERT parameters [169]. However, unlike prior work that successfully classifies DAIC-WOZ transcripts with BERT [11], we made the task more difficult by combining the responses to all questions in a single corpus and treating each sentence as a separate instance. Despite this preprocessing, the clinical interviews remained the smallest dataset.

It is unfortunate that the machine evaluation metrics indicate there is a trade-off between generating realistic texts and predictive texts. While we selected a BERT classifier based on prior depression detection research with small datasets [11], it is possible that a different classifier would not result in this trade-off, though this is a research direction unto itself. Given our results, neither the unconditioned nor conditioned models currently perform sufficiently for their generated texts to be useful in augmenting the small existing datasets for the purpose of depression detection. Yet, our comprehensive comparative study promises to help further research in this domain and yield a more viable future solution.

Future Opportunities. In this paper, we focused on generating text for depression detection. Thus, we demonstrated the conditional models on the small datasets available in this domain. However, our conditioning strategies are also applicable for larger datasets. While we implemented our nine cSeqGAN architectures to generate data with binary labels, all of the proposed conditional models are also easily scalable for more labels as no extra components are required. Further our weighting strategies can be applied to any generative model with a recurrent network as a generator and our feedback strategies can be applied to any discriminator. Since we implemented the cSeqGAN architectures within the Texygen benchmarking platform [87], it would be easily to apply these conditioning approaches to the other text generation models within the platform.

5.2 Outlook

Overall, our results indicate that the cSeqGAN models do not retain the depression signal well enough for the generated text to be useful in depression screening models. I propose three possible reasons. The first is that the datasets were too small to benefit from the advantages of conditional modeling. Second, it is possible that the real passages of text were too short to contain sufficient depression signal. This could be remedied by concatenating texts from the same users, though this would unfortunately serve to further decrease the size of the datasets, limiting our ability to test this theory. Lastly, it is possible that a conditional adversarial models are simply not the best generative approach for this task.

It is worth noting that I elected for a adversarial approach for a reason. The adversarial approach generated diverse data which is important given the diverse nature of the input messages. Variational autoencoders, the main competition for GANs, attempt to model that data in a fashion that produces less varied output. While this makes sense for some domains, it does not make sense for text message classification as the classifiers would not be able to handle outliers. Further, the conditional approach allowed for the sharing of parameters which should have been advantageous for the small datasets available in the domain. It is worth noting that the comparison of the generative models is somewhat constrained by the metrics available for assessing generated text. This is especially true for text messages which do not tend to follow the linguistic conventions expected by models assessing the text.

Regardless of the mentioned challenges faced by this research, I believe our comparative study is important to share in order to further research in the related domains. For instance, the weighting strategies and feedback mechanisms identified and adapted in our research are applicable to any adversarial text generation model. Our cSeqGAN approaches are also applicable to other forms of text, as demonstrated by the use of movie reviews (Table 5.3).

While the focus of the paper was on our cSeqGAN models, the research in this chapter is also novel for the application of unconditioned seqGAN models and BERT classifiers on a corpus of text messages with depression labels. The BERT classifiers were able to achieve an accuracy of 0.71 on the real text messages and 0.67 on the generated texts from the unconditioned generator (Table 5.2). We consider this accuracy for the generated texts a lower threshold as there have been advancements [88, 89] to seqGAN [86]. The results suggests that the generated texts could potentially replace the real texts to train a classifier. However, such future research is currently limited by the small size of existing datasets and the lack of depression signal contained within a single text message.

We conducted a human evaluation to assess the realness of the generated text, as is common in related text generation literature [86, 88, 89]. However, unlike these related works, we generate texts in two different classes. While we selected our participants based on their expertise on the realness task, we also had them assess the class label. The results of the human evaluation on this second task for text messages are worth discussing, especially in comparison to BERT classifier. We made the task easier for both BERT and the human evaluators by only considering texts from the participants with more polarizing depression screening scores. Further, only longer text messages were eligible to be included in the surveys given to the human evaluators. Yet, on the real text messages, the human evaluators only achieved an accuracy of 0.52 when determining if the text messages were written by a participant who screened positive for depression (Table 5.2). While we expected this task to be difficult for humans, this is notably large difference from the 0.71 accuracy achieved by the BERT classifier. Granted, the BERT classifier did have the advantage of being able to access all of the labeled training data before making the classification decisions. Yet, this discrepancy is still revealing and suggests that individual text messages may not have as weak of a depression signal as presumed.

As mentioned, generated data is typically used to increase the data quantity for modeling purposes or amplify the signal for a class label. Both of these are certainly applicable for text messages where the datasets are small and the signal may be weak. However, there is a third way in which text generation could serve to advance depression screening from text messages. Namely, the generation process could help mitigate privacy issues with text messages. This would allow for the sharing of such text data across research teams which is currently not possible and therefore limiting the advancement of such research.

People share private and identifiable information within text messages, such as addresses, medical information, and passwords. A named entity recognition (NER) model would be able to identify some of this information so it could be redacted, but some private and identifiable information would inevitably not be captured, especially given the linguistic dissimilarities of text messages with the corpuses used to train such models [68]. Thus, pairing a generative model with a named entity model could be an approach to create a corpus of labeled texts that do not contain any identifiable information. The research presented in this chapter was motivated by this goal and represents a concrete step towards determining the appropriateness of a set of generative models to incorporate in such a system.

CHAPTER 6

CONCLUDING THOUGHTS

In this dissertation, I presented my research that explores the ability of text messages to screen for mental illnesses. This research has three main directions: predictive modeling with text message content, predictive modeling with text message logs, and generative modeling with text message content. Text messages are a massively underutilized modality for passive mental illness screening. My research offers important insights into the modeling of this promising screening modality. Additionally, I introduce a unique dataset of smartphone logs to further research in this domain.

Text content. Overall, text content proved to be adept at screening for depression [1] and suicidal ideation [2] with lexical category features. I further automated the construction of alternative lexicons with less formal terms to improve depression screening capabilities [3]; these lexicons will also be useful for other datasets with informal text. Lastly, I explored the screening potential of received text messages by considering messages from different subsets of contacts [4]. The results indicates that text content contains strong depression and suicidal ideation signals, though this research is hampered by small datasets.

Text logs. To mitigate privacy concerns, I also assessed the screening potential of text message logs without content. In particular, I extracted features from text reply latencies [5] and log time series [6]. This data contains useful information for depression screening but in practice would likely have to be combined with features from other smartphone sensors. To further research in this domain, I collected a large dataset of smartphone logs with depression and anxiety labels [7]. Our deep learning models demonstrated impressive screening capabilities with the log time series at lower screening score cutoffs. Given this new dataset, there are many exciting future research avenues for mental illness screening with log data.

Generated Text. While collecting a dataset is one way increase data quantity, another option is generate more data. I thus generate individual texts labeled with screening scores. In particular, I identify and adapt conditional adversarial approaches to generate texts. I pair the three weighting strategies with the three feedback mechanisms for nine generative models. The comparative study reveals that the conditional models are better at generating realistic text while the unconditional models are better at generating predictive text. Notably, the comparative study used BERT to determine the predictive nature of the texts. Thus, this research is novel in that it applies both generative and predictive deep learning models to text messages with depression labels.

6.1 Contributions & Impact

My research has proved both text message logs and text message content have great potential at screening for mental illnesses. The popularity of texting makes it a perfect modality for passively screening for mental health issues. Below, I discuss how my research addresses the biggest challenges in leveraging text messages to screen for mental health.

6.1.1 Small Datasets

Prior to my research, Moodable [41] was the only dataset with text message content labeled with depression scores. Other datasets contained depression labels and text logs without content for 48 students [26], 72 students [44] , and 36 adults [52]. I tackled the limited dataset issue in two key ways: collecting more data and generating more data.

I was involved in designing three data collections [12, 15, 7] that resulted in three datasets with labeled text data. However, two of these datasets [12, 15] contained limited number of logs. Given the population similarities, I used the EMU text logs in conjunction with Moodable text logs to screen for depression in six of the papers that compose this dissertation [1, 2, 3, 5, 6, 8]. While the EMU logs slightly increased the quantity of available data, it is the DepreST-CAT dataset [7] that truly addresses the small dataset challenge. This novel dataset contains the logs, PHQ-9 scores, and GAD-7 scores of 369 crowdsourced participants. As I will be releasing DepreST-CAT, it represents a valuable resource that can truly make an impact on passive mental illness screening research.

Further, I leveraged generative models to create more labeled text content. Given the privacy challenges of collecting more labeled data, this alternative approach utilized data that is already collected. Identifying the best generative model for this task is ongoing research as the conditional models resulted in more realistic texts while the unconditioned models resulted in more predictive texts [8]. Paired with a named entity recognition model that can remove sensitive information, there is the potential that the generated data could be released in the future to advance mental illness screening research with text content.

6.1.2 Text Message Privacy

One of the unique challenges in leveraging text messages to screen for depression is the privacy concerns regarding text message content. The private nature of text messages makes it challenging to college data and share data, thus limiting research. To address this, I have compared the depression screening ability of text message content [1, 2, 3] against similar modalities participants may be more willing to share: text message logs without content [5, 6, 7], tweets [1], text prompts [15], and mobile voice recording transcripts [12, 15, 18].

Text message logs can be collected without content and promise to be a promising screening alternative to text message content. Time series from the DepreST-CAT logs [7] at PHQ-9 cutoff 5 matched the predictive abilities of lexical category features from Moodable and EMU text message content [1] at PHQ-9 cutoff 15. With the release of the DepreST-CAT logs [7], I anticipate that screening models using these logs will soon eclipse models using text message content. Overall, using smartphone logs without content effectively solves the text message privacy concern.

Further, given the screening potential of text message content, my research also involves exploring how the content of these messages can be released while maintaining participant privacy. As mentioned, my strategy for doing so involves generating messages on real data that has undergone some anonymization procedures. The generated texts from the unconditioned models demonstrated promise in screening models [8]. However, this is still ongoing research.

6.1.3 Predictive Data Identification

My research focused on exploring the ability of text messages to predict mental illnesses. This involved identifying the subsets of the text message data that carried the strongest mental illness signals. I compared the depression screening ability of sent text messages [1], received [4] text messages, reply latencies [5], time series of incoming text communications [6], and time series of outgoing text communications [6]. Further, I assessed aggregation interval for the time series of text communications [6]. Additionally, I compared the depression [1] and suicidal ideation [2] screening ability of different longitudinal quantities of self-written text message content. Likewise, for the DepreST-CAT logs [7], I compared different lengths of daily text communication time series for depression and anxiety screening.

6.1.4 Text Message Linguistics

Additionally, text messages are a unique modality in that they contain informal language which poses challenges to existing models [68]. I originally elected to use Empath lexical categories to derive features from the text content as it contained more modern terms than other lexicons [1]. However, given the informal nature of texts, I further constructed alternative lexical categories to improve the depression screening capabilities of the machine learning models [3].

6.2 Implementation of Mobile Screening

Implementation within clinical settings. As smartphones are ubiquitous [138], smartphone logs present the opportunity to automate mental illness screening in clinical settings. Leveraging such data would assist in making depression screening universal which is recommended by the US Preventative Services Task Force (USPSTF) [40]. Instead of patients actively completing PHQ-9 and GAD-7 screening scores in primary care practices [173], they could instead passively submit their smartphone log history. Many mental illnesses could potentially be screened for without any additional effort or time from the patients. Also, unlike screening surveys, the retrospective log data would not be influenced by biases regarding mental illnesses. Alternatively, screening models trained on two weeks of retrospective logs would also be effective on logs collected prospectively over two weeks, thus allowing for the screening models to passively monitor patient mental health over time. This could be particularly useful in monitoring the effects of new medication and alerting care teams to unexpected reactions.

Implementation outside of clinical settings. There has been recent research indicating that mobile interventions can improve mental health [174, 175]. Thus, screening models leveraging smartphone logs could be seamlessly used in conjunction with mobile intervention apps. Other research has found that tracking mood increases emotional self-awareness and mental wellbeing in participants who are depressed and anxious [176]. This indicates that simply providing individuals with their screening scores over time may be beneficial. As mental illness symptoms may not be recognized [37], the passive mobile screening models could also simply serve to alert individuals when to seek care from mental health services.

6.3 Current and Future Research Directions

6.3.1 Current Screening with Text Content Research

Absent from chapter 3 was the use of more advanced models such as BERT classifiers for participant depression screening with text messages. Applying BERT for this task is a challenge due to the small available dataset of text messages labeled with depression screening scores, weak depression signals spread across many messages, and linguistic differences of text messages to the training corpus. I am currently in the process of determining how to overcome these challenges to make BERT classifiers applicable for screening with text messages.

BERT classifiers were able to achieve 0.71 when screening for depression with text messages [8]. However, these BERT classifiers in chapter 5 were advantaged by the test set being artificially balanced and only containing the text messages from participants with the most polarizing PHQ-9 depression screening scores. Further, the predictions were conducted at the message level which, while a seemingly more difficult task than using larger text passages, is not as diagnostically useful.

6.3.2 Current Screening with Text Logs Research

The DepreST-CAT dataset [7] is newly collected and offers many possible research directions. In particular, I am analyzing the results of deep learning models that used the call and text log communication time series to screen for self reported suicidal ideation. Further, I am in the process of exploring the usefulness of different feature sets extracted from the text and call logs to screen for depression and anxiety at higher screening score cutoffs.

6.3.3 Current Generated Text Research

Since the generated text from the unconditioned models proved to be the most useful generated text for the BERT classifiers [8], I will be comparing the depression screening performance of text generated from a variety of existing unconditioned generative models. This research will understandably be informed by the results of the aforementioned research regarding the applicability of BERT classifiers for text messages.

Further, we have labeled named entities in a subset of the text messages. We are using these to test the effectiveness of existing named entity recognition models and train our own named entity recognition models. As mentioned in chapter 5, the end goal is to pair a named entity recognition model with a generative model to generate a corpus of anonymized texts for depression screening. This research will thus also involve assessing the anonymity of the resulting texts.

6.3.4 Available Future Research Directions

My research has focused on binary classification of text. However, mental illness screening models could also be categorical or numerical. Thus, this is a possibility for future research. Further, my research has primarily analyzed the screening capabilities of single modalities. The only multimodal models involving text were those used for the DepreST-CAT call logs and texts [7]. In my related research, the multimodal models have involved voice recordings and the transcripts of those voice recordings [11, 12, 13, 17, 18, 19]. As such, there is an opportunity to use the text content and log features proposed in my research [5, 6, 1] in conjunction with features from other smartphone sensors to screen for mental illnesses. The screening potential of these features can also be assessed on other datasets. Further, other feature engineering and methods can be applied to the DepreST-CAT logs [7] to further the goal of passive universal mental illnesses screening.

REFERENCES

[1] M. L. Tlachac and E. Rundensteiner, "Screening for depression with retrospectively harvested private versus public text," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, 2020.

[2] M. Tlachac, K. Dixon-Gordon, and E. Rundensteiner, "Screening for suicidal ideation with text messages," pp. 1–4, 2021.

[3] M. Tlachac, A. Shrestha, M. Shah, B. Litterer, and E. Rundensteiner, "Automated construction of lexicons to improve depression screening with text messages," In submission.

[4] M. L. Tlachac, E. Toto, and E. Rundensteiner, "You're making me depressed: Leveraging texts from contact subsets to predict depression," *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 1–4, 2019.

[5] M. L. Tlachac and E. Rundensteiner, "Depression screening from text message reply latency," in *International Conferences of the IEEE Engineering in Medicine and Biology Society*, 2020, pp. 5490–5493.

[6] M. Tlachac, V. Melican, M. Reisch, and E. Rundensteiner, "Mobile depression screening with time series of text logs and call logs," pp. 1–4, 2021.

[7] M. L. Tlachac, R. Flores, M. Reisch, K. Houskeeper, and E. Rundensteiner, "Studentsadd: Mobile depression and suicidal ideation screening of college students during the coronavirus pandemic," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Accepted.

[8] M. Tlachac, W. Gerych, K. Agrawal, B. Litterer, N. Jurovich, S. Thatigotla, J. Thadajarassiri, and E. Rundensteiner, "Text generation to aid depression detection: A comparative study of conditional sequence generative adversarial networks," In Revision.

[9] M. L. Tlachac, A. Sargent, E. Toto, R. Paffenroth, and E. Rundensteiner, "Topological data analysis to engineer features from audio signals for depression detection," in *19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2020.

[10] E. Toto, M. L. Tlachac, F. Stevens, and E. Rundensteiner, "Audio-based depression screening using sliding window sub-clippooling," in *19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2020.

[11] E. Toto, M. Tlachac, and E. A. Rundensteiner, "Audibert: A deep transfer learning multimodal classification framework for depression screening," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 4145–4154.

[12] M. Tlachac, E. Toto, J. Lovering, R. Kayastha, N. Taurich, and E. Rundensteiner, "Emu: Early mental health uncovering framework and dataset," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2021, pp. 1311–1318.

[13] R. Flores, M. Tlachac, E. Toto, and E. A. Rundensteiner, "Depression screening using deep learning on follow-up questions in clinical interviews," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2021, pp. 595–600.

[14] S. Senn, M. L. Tlachac, R. Flores, and E. Rundensteiner, "Ensembles of bert for depression classification," in *44nd International Conference of IEEE Engineering in Medicine and Biology Society (EMBC)*, Accepted.

[15] M. L. Tlachac, R. Flores, M. Reisch, R. Kayastha, N. Taurich, V. Melican, C. Bruneau, H. Caouette, J. Lovering, E. Toto, and E. Rundensteiner, "Studentsadd: Mobile depression and suicidal ideation screening of college students during the coronavirus pandemic," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Accepted.

[16] M. L. Tlachac, M. Reisch, B. Lewis, R. Flores, L. Harrison, and E. Rundensteiner, "Impact of stereotype threat on mobile depression screening," In revision.

[17] R. Flores, M. L. Tlachac, E. Toto, and E. Rundensteiner, "Transfer learning for depression screening from follow-up clinical interview questions," *Deep Learning Applications*, vol. 4, In Submission.

[18] M. L. Tlachac, R. Flores, E. Toto, and E. Rundensteiner, "Early mental health uncovering with short scripted and unscripted voice recordings," *Deep Learning Applications*, vol. 4, In Submission.

[19] M. Reisch, M. Tlachac, R. Flores, E. Toto, and E. Rundensteiner, "Mental health classification utilizing multimodal deep learning with mobile speech recordings," In Preparation.

[20] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual review of clinical psychology*, vol. 14, pp. 91–118, 2018.

[21] R. Redlich, J. R. Almeida, D. Grotegerd, N. Opel, H. Kugel, W. Heindel, V. Arolt, M. L. Phillips, and U. Dannlowski, "Brain morphometric biomarkers distinguishing unipolar and bipolar depression: A voxel-based morphometry–pattern classification approach," *JAMA psychiatry*, vol. 71, no. 11, pp. 1222–1230, 2014.

[22] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, *et al.*, "The distress analysis interview corpus of human and computer interviews.," in *Language Resources and Evaluation*, CiteSeer, 2014, pp. 3123–3128.

[23] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Communication*, vol. 71, pp. 10–49, 2015.

[24] S. Guntuku, D. Yaden, M. Kern, L. Ungar, and J. Eichstaedt, “Detecting depression and mental illness on social media: An integrative review,” *Current Opinion in Behavioral Sciences*, vol. 18, 2017.

[25] S. Chancellor and M. De Choudhury, “Methods in predictive techniques for mental health status on social media: A critical review,” *NPJ digital medicine*, vol. 3, no. 1, pp. 1–11, 2020.

[26] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, “Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones,” in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 2014, pp. 3–14.

[27] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “Mimic-iii, a freely accessible critical care database,” *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[28] Substance Abuse and Mental Health Services Administration, “Key substance use and mental health indicators in the united states: Results from the 2019 national survey on drug use and health,” 2020.

[29] A. Woodall, C. Morgan, C. Sloan, and L. Howard, “Barriers to participation in mental health research: Are there specific gender, ethnicity and age related barriers?” *BMC psychiatry*, vol. 10, no. 1, pp. 1–10, 2010.

[30] D. E. Bloom, E. Cafiero, E. Jané-Llopis, S. Abrahams-Gessel, L. R. Bloom, S. Fathima, A. B. Feigl, T. Gaziano, A. Hamandi, M. Mowafi, *et al.*, “The global economic burden of noncommunicable diseases,” Program on the Global Demography of Aging, Tech. Rep., 2012.

[31] S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, *et al.*, “Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the global burden of disease study 2017,” *The Lancet*, vol. 392, no. 10159, pp. 1789–1858, 2018.

[32] J. Firth, N. Siddiqi, A. Koyanagi, D. Siskind, S. Rosenbaum, C. Galletly, *et al.*, “A blueprint for protecting physical health in people with mental illness: Directions for health promotion, clinical services and future research,” *Lancet Psychiatry*, 2019.

[33] E. Isometsä, “Psychological autopsy studies–a review,” *European psychiatry*, vol. 16, no. 7, pp. 379–385, 2001.

[34] H. Hedegaard, S. Curtin, and M. Warner, “Increase in suicide mortality in the united states, 1999–2018,” *NCHS Data Brief*, vol. No. 366, 2020.

[35] M. Weist, M. Rubin, E. Moore, S. Adelsheim, and G. Wrobel, “Mental health screening in schools,” *Journal of School Health*, vol. 77(2), 2007.

[36] P. S. Wang, P. A. Berglund, M. Olfson, and R. C. Kessler, “Delays in initial treatment contact after first onset of a mental disorder,” *Health services research*, vol. 39, no. 2, pp. 393–416, 2004.

[37] R. M. Epstein, P. R. Duberstein, M. D. Feldman, A. B. Rochlen, R. A. Bell, R. L. Kravitz, C. Cipri, J. D. Becker, P. M. Bamonti, and D. A. Paterniti, “” i didn’t know what was wrong:” how people with undiagnosed depression recognize, name and explain their distress,” *Journal of General Internal Medicine*, vol. 25, no. 9, pp. 954–961, 2010.

[38] K. Demyttenaere, A. Bonnewyn, R. Bruffaerts, T. Brugha, R. De Graaf, and J. Alonso, “Comorbid painful physical symptoms and depression: Prevalence, work loss, and help seeking,” *Journal of affective disorders*, vol. 92, pp. 185–193, 2006.

[39] A. Halfin, “Depression: The benefits of early and appropriate treatment,” *American Journal of Managed Care*, vol. 13, no. 4, 2007.

[40] A. L. Siu, K. Bibbins-Domingo, D. C. Grossman, L. C. Baumann, K. W. Davidson, M. Ebell, F. A. Garcıa, M. Gillman, J. Herzstein, A. R. Kemper, *et al.*, “Screening for depression in adults: Us preventive services task force recommendation statement,” *Jama*, vol. 315, no. 4, pp. 380–387, 2016.

[41] A. Dogrucu, A. Perucic, A. Isaro, D. Ball, E. Toto, E. A. Rundensteiner, E. Agu, R. Davis-Martin, and E. Boudreaux, “Moodable: On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data,” *Smart Health*, pp. 100–118, 2020.

[42] H. Cai, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, Q. Zhao, *et al.*, “Modma dataset: A multi-model open dataset for mental-disorder analysis,” *arXiv*, arXiv–2002, 2020.

[43] A. Madan, M. Cebrian, D. Lazer, and A. Pentland, “Social sensing for epidemiological behavior change,” in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, 2010, pp. 291–300.

[44] M. Boukhechba, A. R. Daros, K. Fua, P. I. Chow, B. A. Teachman, and L. E. Barnes, "Demonicsalmon: Monitoring mental health and social interactions of college students using smartphones," *Smart Health*, vol. 9, pp. 192–203, 2018.

[45] S. Vanheule, M. Desmet, R. Meganck, and R. Hogenraad, "The tongue ever turns to the aching tooth: A pilot study of depressed patients' self-preoccupation," *Journal of the American Psychoanalytic Association*, vol. 61, no. 6, 2013.

[46] E. Newell, S. McCoy, M. Newman, J. Wellman, and S. Gardner, "You sound so down: Capturing depressed affect through depressed language," *Journal of Language and Social Psychology*, vol. 37, no. 4, pp. 451–474, 2018.

[47] S. Rude, E. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[48] S. Ware, C. Yue, R. Morillo, J. Lu, C. Shang, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang, "Predicting depressive symptoms using smartphone data," *Smart Health*, vol. 15, pp. 1–16, 2020.

[49] R. S. McGinnis, E. W. McGinnis, J. Hruschak, N. L. Lopez-Duran, K. Fitzgerald, K. L. Rosenblum, and M. Muzik, "Wearable sensors and machine learning diagnose anxiety and depression in young children," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, IEEE, 2018, pp. 410–413.

[50] R. F. K. Martin, P. Leppink-Shands, M. Tlachac, M. DuBois, C. Conelea, S. Jacob, V. Morellas, T. Morris, and N. Papanikolopoulos, "The use of immersive environments for the early detection and treatment of neuropsychiatric disorders," *Frontiers in Digital Health*, vol. 2, p. 40, 2021.

[51] D. Di Matteo, K. Fotinos, S. Lokuge, J. Yu, T. Sternat, M. A. Katzman, and J. Rose, "The relationship between smartphone-recorded environmental audio and symptomatology of anxiety and depression: Exploratory study," *JMIR Form Res*, vol. 4, no. 8, 2020.

[52] F. Wahle, T. Kowatsch, E. Fleisch, M. Rufer, S. Weidt, *et al.*, "Mobile sensing and support for people with depression: A pilot trial in the wild," *JMIR mHealth and uHealth*, vol. 4, no. 3, e5960, 2016.

[53] A. Smith, "How americans use text messaging," *Pew Research Center: Internet & Technology*, 2011.

[54] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[55] M. Nadeem, G. Coppersmith, and S. Sen, "Identifying depression on twitter," *arXiv preprint arXiv:1607.07384*, 2016.

[56] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, “Recognizing depression from twitter activity,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 2015.

[57] Z. Liu, D. Wang, L. Zhang, and B. Hu, “A novel decision tree for depression recognition in speech,” *arXiv preprint arXiv:2002.12759*, 2020.

[58] M. Park, C. Cha, and M. Cha, “Depressive moods of users portrayed in twitter,” in *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*, ACM New York, NY, vol. 2012, 2012.

[59] A. Reece, A. Reagan, K. Lix, P. Dodds, C. Danforth, and E. Langer, “Forecasting the onset and course of mental illness with twitter data,” *Scientific Reports*, vol. 7, no. 1, 2017.

[60] A. Benton, M. Mitchell, and D. Hovy, “Multi-task learning for mental health using social media text,” *arXiv preprint arXiv:1712.03538*, 2017.

[61] W. Gerych, E. Agu, and E. Rundensteiner, “Classifying depression in imbalanced datasets using an autoencoder-based anomaly detection approach,” in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, IEEE, 2019, pp. 124–127.

[62] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, “Multimodal fusion of bert-cnn and gated cnn representations for depression detection,” in *AVEC*, 2019, pp. 55–63.

[63] J. Han, Z. Zhang, N. Cummins, and B. Schuller, “Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives,” *IEEE Computational Intelligence Magazine*, vol. 14, no. 2, pp. 68–81, 2019.

[64] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[65] C. Sen, T. Hartvigsen, B. Yin, X. Kong, and E. Rundensteiner, “Human attention maps for text classification: Do humans and neural networks focus on the same words?” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4596–4608.

[66] L. Canzian and M. Musolesi, “Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis,” in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 1293–1304.

[67] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang, “Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data,” in *2016 IEEE Wireless Health (WH)*, IEEE, 2016, pp. 1–8.

[68] T. Ek, C. Kirkegaard, H. Jonsson, and P. Nugues, “Named entity recognition for short text messages,” *Procedia-Social and Behavioral Sciences*, vol. 27, pp. 178–187, 2011.

[69] E. Fast, B. Chen, and M. Bernstein, “Empath: Understanding topic signals in large-scale text,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, 2016.

[70] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[71] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, 2011.

[72] Scikit-learn Developers, “Principal component analysis,” 2018.

[73] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[74] Scikit-learn Developers, “Feature selection,” 2019.

[75] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, “Avec 2016: Depression, mood, and emotion recognition workshop and challenge,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC ’16, Amsterdam, The Netherlands: ACM, 2016, pp. 3–10, ISBN: 978-1-4503-4516-3.

[76] D. B. Dwyer, P. Falkai, and N. Koutsouleris, “Machine learning approaches for clinical psychology and psychiatry,” *Annual Review of Clinical Psychology*, vol. 14, pp. 91–118, 2018.

[77] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACD Sigkdd International conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[78] V. Morde, “Xgboost algorithm: Long may she reign!” *Towards Data Science*, 2019.

[79] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[80] A. Murarka, B. Radhakrishnan, and S. Ravichandran, “Detection and classification of mental illnesses on social media using roberta,” *arXiv preprint arXiv:2011.11226*, 2020.

[81] D. Q. Nguyen, T. Vu, and A. T. Nguyen, “Bertweet: A pre-trained language model for english tweets,” *arXiv preprint arXiv:2005.10200*, 2020.

[82] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, “Huggingface’s transformers: State-of-the-art natural language processing,” *arXiv preprint arXiv:1910.03771*, 2019.

[83] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, pp. 2672–2680, 2014.

[84] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.

[85] C. Donahue, J. McAuley, and M. Puckette, “Adversarial audio synthesis,” *arXiv preprint arXiv:1802.04208*, 2018.

[86] L. Yu, W. Zhang, J. Wang, and Y. Yu, “Seqgan: Sequence generative adversarial nets with policy gradient,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[87] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu, “Texygen: A benchmarking platform for text generation models,” in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 1097–1100.

[88] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang, “Long text generation via adversarial training with leaked information,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[89] W. Nie, N. Narodytska, and A. Patel, “Relgan: Relational generative adversarial networks for text generation,” in *International conference on learning representations*, 2018.

[90] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv: 1312.6114*, 2013.

[91] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[92] K. Kroenke, R. L. Spitzer, and J. B. Williams, “The phq-9: Validity of a brief depression severity measure,” *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[93] N. F. BinDhim, A. M. Shaman, L. Trevena, M. H. Basyouni, L. G. Pont, and T. M. Alhawassi, “Depression screening via a smartphone app: cross-country user characteristics and feasibility,” *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. 29–34, 2014.

[94] R. S. DeJesus, K. S. Vickers, G. J. Melin, and M. D. Williams, "A system-based approach to depression management in primary care using the patient health questionnaire-9," in *Mayo Clinic Proceedings*, Elsevier, vol. 82, 2007, pp. 1395–1402.

[95] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: The gad-7," *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.

[96] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An inventory for measuring depression," *Archives of general psychiatry*, vol. 4, no. 6, pp. 561–571, 1961.

[97] L. S. Radloff, "The ces-d scale: A self-report depression scale for research in the general population," *Applied psychological measurement*, vol. 1, no. 3, pp. 385–401, 1977.

[98] M. Hamilton, "The hamilton rating scale for depression," in *Assessment of depression*, Springer, 1986, pp. 143–152.

[99] L. O'Connor, C. Larkin, A. F. Ibrahim, M. Allen, B. Wang, and E. D. Boudreaux, "Development and pilot study of simple suicide risk rulers for use in the emergency department," *General hospital psychiatry*, vol. 63, pp. 97–102, 2020.

[100] R. D. Gibbons, D. Kupfer, E. Frank, T. Moore, D. G. Beiser, and E. D. Boudreaux, "Development of a computerized adaptive test suicide scale—the cat-ss," *The Journal of clinical psychiatry*, vol. 78, no. 9, pp. 1376–1382, 2017.

[101] B. J. Ricard, L. A. Marsch, B. Crosier, and S. Hassanpour, "Exploring the utility of community-generated social media content for detecting depression: An analytical study on instagram," *Journal of Medical Internet Research*, 2018.

[102] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?," 2016.

[103] S. Palanab and C. Schittera, "Prolific.ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.

[104] J. F. Huckins, A. W. DaSilva, W. Wang, E. Hedlund, C. Rogers, S. K. Nepal, J. Wu, M. Obuchi, E. I. Murphy, M. L. Meyer, *et al.*, "Mental health and behavior of college students during the early phases of the covid-19 pandemic: Longitudinal smartphone and ecological momentary assessment study," *Journal of medical Internet research*, vol. 22, no. 6, e20185, 2020.

[105] M. Asgari, I. Shafran, and L. B. Sheeber, "Inferring clinical depression from speech and spoken utterances," in *2014 IEEE international workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, 2014, pp. 1–5.

[106] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, “Characterizing and predicting postpartum depression from shared facebook data,” in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work Social Computing*, ACM, 2014, pp. 626–638.

[107] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr, “Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study,” *Journal of Medical Internet Research*, vol. 17, no. 7, e175, 2015.

[108] Z. Huang, J. Epps, D. Joachim, and M. Chen, “Depression detection from short utterances via diverse smartphones in natural environmental conditions.,” in *INTERSPEECH*, 2018, pp. 3393–3397.

[109] J. W. Pennebaker, R. J. Booth, R. L. Boyd, and M. E. Francis, “Linguistic inquiry and word count: Liwc2015 operator’s manual,” 2015.

[110] A. Sahu, P. Gupta, and B. Chatterjee, “Depression is more than just sadness: A case of excessive anger and its management in depression,” *Indian journal of psychological medicine*, vol. 36, no. 1, pp. 77–79, 2014.

[111] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, *et al.*, “Simsensei kiosk: A virtual human interviewer for healthcare decision support,” in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, 2014, pp. 1061–1068.

[112] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, “Avec 2016: Depression, mood, and emotion recognition workshop and challenge,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ACM, 2016, pp. 3–10.

[113] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, ser. Springer Theses, Recognizing Outstanding Ph.D. Research. Springer International Publishing, 2016.

[114] E. Toto, M. Tlachac, F. L. Stevens, and E. A. Rundensteiner, “Audio-based depression screening using sliding window sub-clip pooling,” in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2020, pp. 791–796.

[115] A. B. Warriner, V. Kuperman, and M. Brysbaert, “Norms of valence, arousal, and dominance for 13,915 english lemmas,” *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[116] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber, *et al.*, “The 16-item quick inventory

of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): A psychometric evaluation in patients with chronic major depression," *Biological psychiatry*, vol. 54, no. 5, pp. 573–583, 2003.

[117] R. Wang, M. S. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, J. Kane, M. Merrill, E. A. Scherer, V. W. Tsent, and D. Ben-Zeev, "Crosscheck: Toward passive sensing and detection of mental health changes in people with schizophrenia," in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, 2016, pp. 886–897.

[118] R. Wang, W. Wang, M. S. Aung, D. Ben-Zeev, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, J. Kane, E. A. Scherer, and M. Walsh, "Predicting symptom trajectories of schizophrenia using mobile sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–24, 2017.

[119] S. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 200–213, 2017.

[120] X. Xu, P. Chikersal, A. Doryab, D. K. Villalba, J. M. Dutcher, M. J. Tumminia, T. Althoff, S. Cohen, K. G. Creswell, J. D. Creswell, J. Mankoff, and A. K. Dey, "Leveraging routine behavior and contextually-filtered features for depression detection among college students," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–33, 2019.

[121] P. Chikersal, A. Doryab, M. Tumminia, D. K. Villalba, J. M. Dutcher, X. Liu, S. Cohen, K. G. Creswell, J. Mankoff, J. D. Creswell, M. Goel, and A. K. Dey, "Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 28, no. 1, pp. 1–41, 2021.

[122] S. Loria, "Textblob: Simplified text processing," 2018.

[123] S. Cohen, "Social relationships and health.," *American psychologist*, vol. 59, no. 8, p. 676, 2004.

[124] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[125] K. Kroenke, R. Spitzer, and J. Williams, "The phq-9: Validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, 2001.

[126] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech*, 2018.

[127] E. M. Kleiman, B. J. Turner, S. Fedor, E. E. Beale, R. W. Picard, J. C. Huffman, and M. K. Nock, “Digital phenotyping of suicidal thoughts,” *Depression and Anxiety*, vol. 35, no. 7, pp. 601–608, 2018.

[128] E. K. Czyz, C. A. King, and I. Nahum-Shani, “Ecological assessment of daily suicidal thoughts and attempts among suicidal teens after psychiatric hospitalization: Lessons about feasibility and acceptability,” *Psychiatry Research*, vol. 267, pp. 566–574, 2018.

[129] E. Fast, B. Chen, and M. S. Bernstein, “Empath: Understanding topic signals in large-scale text,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 4647–4657.

[130] ——, “Lexicons on demand: Neural word embeddings for large-scale text analysis.,” in *IJCAI*, 2017, pp. 4836–4840.

[131] S. Cohen, “Social relationships and health.,” *American psychologist*, vol. 59, no. 8, p. 676, 2004.

[132] K. Rook, “The negative side of social interaction: Impact on psychological well-being.,” *Journal of personality and social psychology*, vol. 46, no. 5, p. 1097, 1984.

[133] M. Demir and M. Özdemir, “Friendship, need satisfaction and happiness,” *Journal of Happiness Studies*, vol. 11, no. 2, pp. 243–259, 2010.

[134] I. Kawachi and L. Berkman, “Social ties and mental health,” *Journal of Urban health*, vol. 78, no. 3, pp. 458–467, 2001.

[135] J. Rooksby, A. Morrison, and D. Murray-Rust, “Student perspectives on digital phenotyping: The acceptability of using smartphone data to assess mental health,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[136] X. Dai, X. Kong, T. Guo, and Y. Huang, “Cinet: Redesigning deep neural networks for efficient mobile-cloud collaborative inference,” in *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, SIAM, 2021, pp. 459–467.

[137] S. S. Ogden and T. Guo, “MODI: Mobile deep inference made efficient by edge computing,” in *USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18)*, 2018.

[138] Pew Research Center. “Smartphone ownership is growing rapidly around the world but not always equally.” (2019).

[139] M. Alkhathlan, M. Tlachac, L. Harrison, and E. Rundensteiner, ““honestly i never really thought about adding a description”: Why highly engaged tweets are inaccessible,” in *IFIP Conference on Human-Computer Interaction*, Springer, 2021, pp. 373–395.

[140] C. McClain, R. Widjaya, G. Rivero, and A. Smith, "The behaviors and attitudes of us adults on twitter," *Pew Research Center*, 2021.

[141] C. R. Gale, G. D. Batty, S.-A. Cooper, I. J. Deary, G. Der, B. S. McEwen, and J. Cavanagh, "Reaction time in adolescence, cumulative allostatic load, and symptoms of anxiety and depression in adulthood: The west of scotland twenty-07 study," *Psychosomatic medicine*, vol. 77, no. 5, p. 493, 2015.

[142] "Tsfel: Time series feature extraction library," *SoftwareX*, vol. 11, p. 100 456, 2020.

[143] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," vol. 11, Jan. 2004, pp. 70–80.

[144] S. Palan and C. Schitter, "Prolific. ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.

[145] J. Howard, "All 50 states now have expanded or will expand covid vaccine eligibility to everyone 16 and up," *Cable News Network (CNN) Health*, 2021.

[146] S. J. Spencer, C. Logel, and P. G. Davies, "Stereotype threat," *Annual review of psychology*, vol. 67, pp. 415–437, 2016.

[147] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[148] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[149] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: A review," *Data mining and knowledge discovery*, vol. 33, no. 4, pp. 917–963, 2019.

[150] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[151] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[152] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[153] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li, "Brits: Bidirectional recurrent imputation for time series," *arXiv preprint arXiv:1805.10572*, 2018.

[154] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *International conference on machine learning*, PMLR, 2013, pp. 1310–1318.

[155] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding,” *arXiv preprint arXiv:1703.03130*, 2017.

[156] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, *et al.*, “Sensing the” health state” of a community,” *IEEE Pervasive Computing*, vol. 11, no. 4, pp. 36–45, 2011.

[157] B. Wetzel, R. Pryss, H. Baumeister, J.-S. Edler, A. S. O. Gonçalves, and C. Cohrdes, ““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,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 12, p. 6212, 2021.

[158] J. Guan, R. Li, S. Yu, and X. Zhang, “Generation of synthetic electronic medical record text,” in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2018, pp. 374–380.

[159] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv: 1411.1784*, 2014.

[160] K. Wang and X. Wan, “Sentigan: Generating sentimental texts via mixture adversarial networks.,” in *IJCAI*, 2018, pp. 4446–4452.

[161] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, “Toward controlled generation of text,” in *International Conference on Machine Learning*, PMLR, 2017, pp. 1587–1596.

[162] Y. Li, Q. Pan, S. Wang, T. Yang, and E. Cambria, “A generative model for category text generation,” *Information Sciences*, vol. 450, pp. 301–315, 2018.

[163] Z. Liu, J. Wang, and Z. Liang, “Catgan: Category-aware generative adversarial networks with hierarchical evolutionary learning for category text generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 8425–8432.

[164] A. Shoshan, N. Bhonker, I. Kviatkovsky, and G. Medioni, “Gan-control: Explicitly controllable gans,” *arXiv preprint arXiv:2101.02477*, 2021.

[165] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *arXiv preprint arXiv:1406.2661*, 2014.

[166] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.

[167] X. Zhang and Y. LeCun, "Text understanding from scratch," *arXiv preprint arXiv: 1502. 01710*, 2015.

[168] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 142–150.

[169] S. A. Rauf, Y. Qiang, S. B. Ali, and W. Ahmad, "Using bert for checking the polarity of movie reviews," *International Journal of Computer Applications*, vol. 975, p. 8887,

[170] H. M. Zahera, I. A. Elgendy, R. Jalota, M. A. Sherif, E. Voorhees, and A. Ellis, "Fine-tuned bert model for multi-label tweets classification.," in *TREC*, 2019, pp. 1–7.

[171] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang, "Long text generation via adversarial training with leaked information," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[172] K. Lin, D. Li, X. He, Z. Zhang, and M.-T. Sun, "Adversarial ranking for language generation," *arXiv preprint arXiv:1705.11001*, 2017.

[173] M. L. Savoy and D. T. O'Gurek, "Screening your adult patients for depression," *Family practice management*, vol. 23, no. 2, pp. 16–20, 2016.

[174] D. Bakker, N. Kazantzis, D. Rickwood, and N. Rickard, "A randomized controlled trial of three smartphone apps for enhancing public mental health," *Behaviour Research and Therapy*, vol. 109, pp. 75–83, 2018.

[175] J. A. Flett, H. Hayne, B. C. Riordan, L. M. Thompson, and T. S. Conner, "Mobile mindfulness meditation: A randomised controlled trial of the effect of two popular apps on mental health," *Mindfulness*, vol. 10, no. 5, pp. 863–876, 2019.

[176] D. Bakker and N. Rickard, "Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: Moodprism," *Journal of affective disorders*, vol. 227, pp. 432–442, 2018.

VITA

ML Tlachac started as a Data Science graduate student at Worcester Polytechnic Institute in 2016. After graduating with an applied mathematics major and sociology minor from University of Wisconsin- Eau Claire, the interdisciplinary nature of data science appealed to Tlachac. Since joining the Data Science program at WPI, Tlachac has gravitated towards health informatics research. Tlachac began research on mental health detection with machine learning after completing a Master's thesis on longitudinal antibiogram modeling. An internship at NextEra Analytics inspired Tlachac to also explore generative modeling. Tlachac has mentored teams of undergraduate, graduate, and postgraduate students on the Emutivo project for four years. Upon graduating, Tlachac will continue educating students as a data science professor at Bryant University. Tlachac is a mental health advocate who enjoys writing, solving puzzles, and walking with dog Bumper.

mltlachac@wpi.edu

mltlachac.github.io/

github.com/mltlachac

linkedin.com/in/mltlachac

emutivo.wpi.edu