WPI

Dissecting the Paper Creation Process

ML Tlachac

WPI REU 2022

Why am I Presenting About Papers?

My Published Mental Illness Screening Papers

You're Making Me Depressed: Leveraging Texts from Contact Subsets to Predict Depression

ML Tlachac, *Data Science*, *Worcester Polytechnic Institute*, Worcester, MA, USA, mltlachac@wpi.edu
Ermal Toto, *Computer Science*, *Worcester Polytechnic Institute*, Worcester, MA, USA, toto@wpi.edu
Elke Rundensteiner, *Data Science, Computer Science*, *Worcester Polytechnic Institute*, Worcester, MA, USA, rundenst@wpi.edu

***Abstract*—Depression, a prevalent and debilitating mental illness, is frequently undiagnosed. Diagnosis is an important step towards treatment. Currently screening tools, such as the Patient Health Questionnaire-9 (PHQ-9), require patient input. Many studies have used a variety of data types and features to predict depression scores for individuals. In this study, we focus on the predictive ability of a single under-utilized modality indicative of the impact of social interactions: received text messages. Our approach encompasses creating subsets of influential contacts for each participant and engineering features from the text messages of those contact subsets. Overall, our study demonstrates that received text communications are a promising modality when predicting depression scores. Specifically, we found that the F1 score of Gaussian Naive Bayes models leveraging just the text messages from a subset of top contacts performed statistically significantly better by 13.2 percent than the models leveraging text messages from all contacts.**

***Index Terms*—text feature engineering, depression screening, received communications, contact subsets**

I. Introduction

Depression is a debilitating mood disorder characterized by chronic sadness and apathy. Despite being one of the most treatable mental disorders, US Preventative Services Task Force (USPSTF) has identified depression as the leading cause of disability in adults [10], [17]. While the USPSTF recommending depression screening for all adults, an estimated quarter of patients with depression are not diagnosed [7], [17]. This often occurs due to symptoms not being recognized, lack of access to medical resources, and fear of stigma [7]. Currently, questionnaires such as the Patient Health Questionnaire-9 (PHQ-9) are being utilized to screen for depression [10]. However, these screening tools require patient cooperation and unbiased ability to recall symptoms.

A. Related Work

Recent studies have begun to leverage smartphone sensor and social media data for mental health screening [1], [4], [12], [13], [16]. Specifically, location sensor [16], audio [12], and twitter [13] data have been shown to be effective at diagnosing depression. De Choudhury, Counts, and Horvitz conducted comprehensive feature engineering to extract features related

Thanks US Department of Education P200A150306: GAANN Fellowships to Support Data-Driven Computing Research for funding.

Mobile Depression Screening with Time Series of Text Logs and Call Logs

ML Tlachac, *Data Science*, *Worcester Polytechnic Institute*, Worcester, MA, USA, mltlachac@wpi.edu
Veronica Melican, *Computer Science*, *Worcester Polytechnic Institute*, Worcester, MA, USA, vrmelican@wpi.edu
Miranda Reisch, *Data Science*, *Worcester Polytechnic Institute*, Worcester, MA, USA, mhernandezreisch@wpi.edu
Elke Rundensteiner, *Data and Computer Science*, *Worcester Polytechnic Institute*, Worcester, MA, USA, rundenst@wpi.edu

Depression Screening from Text Message Reply Latency*

ML Tlachac[1] and Elke A. Rundensteiner[1]

***Abstract*— Depression is both debilitating and prevalent. While treatable, it is often undiagnosed. Passive depression screening is crucial, but leveraging data from Smartphones and social media has privacy concerns. Inspired by the known relationship between depression and slower information processing speed, we hypothesize the latency of texting replies will contain useful information in screening for depression. Specifically, we extract nine reply latency related features from crowd-sourced text message conversation meta-data. By considering text meta-data instead of content, we mitigate the privacy concerns. To predict binary screening survey scores, we explore a variety of machine learning methods built on principal components of the latency features. Our findings demonstrate that an XGBoost model built with one principal component achieves an F1 score of 0.67, AUC of 0.72, and Accuracy of 0.69. Thus, we confirm that reply latency of texting has promise as a modality for depression screening.**

I. Introduction

Depression is a debilitating mood disorder. While it is one of the most treatable mental disorders [1], it is very prevalent with devastating costs to society as well as individuals. Globally, depression is the leading cause of disability and, when combined with anxiety, costs one trillion per year in lost productivity [2]. Individuals with depression have higher risks for developing other health problems. Left untreated, depression can result in suicide, the tenth leading cause of death in the United States [2].

Unfortunately, misinformation about depression remains rampant. This leads to depression symptoms being unrecognized and fear of stigma. Lack of access to medical resources also contributes to depression being undiagnosed and untreated [3]. On average, it takes 11 years to get treatment after symptoms appear [2]. Additionally, an estimated quarter of individuals with depression are never diagnosed [3]. As the questionnaires typically leveraged to screen for depression are often perceived as intrusive and cumbersome [4], their usefulness is limited by the lack of patient compliance and biased responses.

Thus, there is a need for passive depression screening instruments. Recently, studies have begun to explore the potential of various Smartphone and social media modalities in predicting depression. To mitigate changes in behavior from being monitored, the Moodable approach retrospectively harvests such data [5]. Tweets are a particularly popular modality to screen for depression due to public availability that allows them to be collected retrospectively [6]. These studies primarily predict binary depression with labels collected either from self-declaration or screening surveys. The methods used involve support vector machine (SVM), logistic regression (LR), random forest (RF), and naive bayes (NB) [6].

Our prior research has involved analyzing the screening ability of retrospectively harvested Smartphone data [5], including content from received texts [7], sent texts [8], and tweets [8]. Short message service (SMS) text messages are a promising modality given the popularity of texting- A 2014 study found that texting was the most popular and frequently used Smartphone feature with 97% of American Smartphone owners texting within the past week [9]. As adults under the age of 30 texted the most [9] and suicide is the second leading cause of death in the United States for this age group [2], this further motivates the use of text message data to screen for depression.

However, the primary downside of using text messages to screen for depression is the lack of people willing to share texts. When 202 crowd-sourced participants were surveyed regarding willingness to share Smartphone data with healthcare professionals on a five rating scale from 'completely unwilling' to 'completely willing', browser history was the only modality with a greater percent of participants 'completely unwilling' to share [5]. Specifically, 40.1 percent of participants were 'completely unwilling' to share text messages and only around a third of participants willing to share this modality [5].

Given participant willingness to share other modalities, we surmise that it is the content of the text messages that people wish to remain private. Yet, the text message data contains more than just text message content. The unique aspect of text messages, especially among text-based modalities, is the conversational aspect. The relationship between depression and slower speed of information processing has been documented [10], though not in this context. We hypothesize that this slower processing, which is attributed to less efficient cognitive functioning [10], 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 utilize features engineered from both 'sent' and 'received' text messages. To mitigate privacy concerns from leveraging text messages, our approach considers only meta-data features and no message content.

*This work was supported by US Department of Education P200A150306: GAANN Fellowships
[1]Tlachac and Rundensteiner are with the Department of Data Science, Worcester Polytechnic Institute, Worcester, Massachusetts 10604, USA mtlachac@wpi.edu, rundenst@wpi.edu

Ensembles of BERT for Depression Classification

Saskia Senn[1], ML Tlachac[2], Ricardo Flores[2], and Elke Rundensteiner[2]

***Abstract*— Depression is among the most prevalent mental health disorders with increasing prevalence worldwide. While early detection is critical for the prognosis of depression treatment, detecting depression is challenging. Previous**

clinical interview recordings. While the Bidirectional Encoder Representations from Transformers (BERT) models [11] for text classification were combined with audio classification architectures, the ablation study indicates that the BERT component was most influential to AudiBERT's success. However, RoBERTa, a more robustly trained BERT model, has demonstrated better performance than BERT for mental health applications [12]. Also, previous studies have revealed that ensembles of models can produce more robust classifications than individual models [13], [14].

Thus, within the scope of this research, we investigate the ability of BERT variants and BERT ensembles to classify depression from the transcripts of responses to 12 clinical interview questions. In particular, we hypothesize that ensembling several BERT variants will result in better performance. We compare the depression classification ability of:

1) three different individual BERT variants,
2) two different ensemble method strategies,
3) ensembles containing different BERT models, and
4) three different combinations of architectural layers.

II. Clinical Interview Transcript Data

For this research, we use the transcripts from the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) [15], [16]. The DAIC-WOZ corpus consists of 189 clinical interviews conducted by a virtual agent. Participants were virtually asked a subset of *core* questions with varying amounts of follow-up questions to elicit more details [16].

The interviews are labeled with PHQ-8 depression screening scores. The PHQ-8 contains the first eight Likert scales in the PHQ-9 [17]. The PHQ-8 score ranges from 0 to 24 with a score of 10 being indicative of depression.

We treat each core question in DAIC-WOZ as an individual thematic dataset, as defined in the related literature [8]. Each dataset contains the responses to a single core question and related follow-up questions. In this research, we use the 12 thematic datasets with the most responses, as detailed in Table I. These datasets contain 94 to 105 responses with 21% to 31% of respondents labeled as depressed (PHQ-8$\geq$ 10).

III. Deep Learning Methodology

Our goal is to predict depression with transcripts of responses to clinical interview questions asked by a virtual agent. To accomplish this, we explore the depression screening capabilities of 3 BERT variants and 4 ensembles of BERT variants. Further, we compare two different ensemble method architectures. We focus on BERT classifiers in this research given their demonstrated success with smaller datasets and mental health applications [12], [8].

Screening for Suicidal Ideation with Text Messages

ML Tlachac, *Data Science*, *Worcester Polytechnic Institute*, mltlachac@wpi.edu
Katherine Dixon-Gordon, *Psychological and Brain Sciences*, *University of Massachusetts Amherst*, katiedg@umass.edu
Elke Rundensteiner, *Data Science and Computer Science*, *Worcester Polytechnic Institute*, rundenst@wpi.edu

***Abstract*—Suicide is a leading cause of death in the US, with suicide rates increasing annually. Passive screening of suicidal ideation is vital to provide referrals to at-risk individuals. We study to what degree smartphone-based communication, in particular, text messages, could be leveraged for passively screening for suicidal ideation. We analyze the screening ability of texts sent in different time periods prior to reported ideation, namely, texts from specific weeks only versus accumulative over several weeks. Our approach involves performing comprehensive feature engineering and identifying influential features to train machine learning models. With just the prior week of texts, we were able to predict the existence of suicidal ideation with AUC = 0.88, F1 = 0.84, accuracy = 0.81, sensitivity = 0.94, and specificity = 0.68. The most influential features include word frequencies of words in the car, clothing, affection, confusion, driving, real estate, and journalism categories. This research, demonstrating the potential of text messages to screen for suicidal ideation, will guide the development of screening technologies.**

***Index Terms*—mobile health, suicide ideation, passive screening, machine learning, digital phenotype**

I. Introduction

Suicide is among the top ten causes of death for all age groups in the United States in 2018 [1]. For those aged 10-34, suicide is the second leading cause of death. The suicide rate has increased by 35% since 1999 despite national goals to combat suicide [1]. Psychological autopsies reveal that 90% of individuals who died by suicide suffered mental illness symptoms [2], though 56% were not diagnosed with a mental illnesses [3]. Early identification and referral of at-risk individuals is recommended to prevent suicide [4].

Smartphones are not only prevalent but also capture large quantities of personal data, making them perfect to passively screen for suicidal ideation and provide referrals to at-risk individuals. 95% of US adults under 35 years of age, the age group most at risk for suicide, owned a personal smartphone in 2018. While research has been conducted on screening for depression with smartphone data [5]–[14], there is a lack of such digital phenotype research for suicidal ideation screening.

While we previously attempted to address this problem [9], the machine learning models yielded an $F1$ score of only 0.58 when predicting the existence of suicidal ideation with features extracted from mobile voice recordings, social media, and phone data. On actively recorded mobile voice, we were able to screen for suicidal ideation with an $F1$ of 0.70 [15]. Yet, we achieved an even higher $F1$ score of 0.81 when screening for depression with the content of sent SMS text messages [10], which have the additional benefit of being passively collected.

We now hypothesize that text messages are a valuable modality to passively screen for suicidal ideation. Yet suicide risk is dynamic [16], [17], 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. Our approach involves comprehensive feature engineering, feature selection, and machine learning. We demonstrate that this approach dramatically increases the suicidal ideation screening ability of data collected from smartphones. Our contributions include:

1) exploring the potential of retrospectively collected crowd-sourced texts for suicidal ideation screening,
2) comparing the screening ability of texts sent in the interval and cumulative weeks prior to reported ideation,
3) identifying the most influential features for screening.

II. Data & Methods

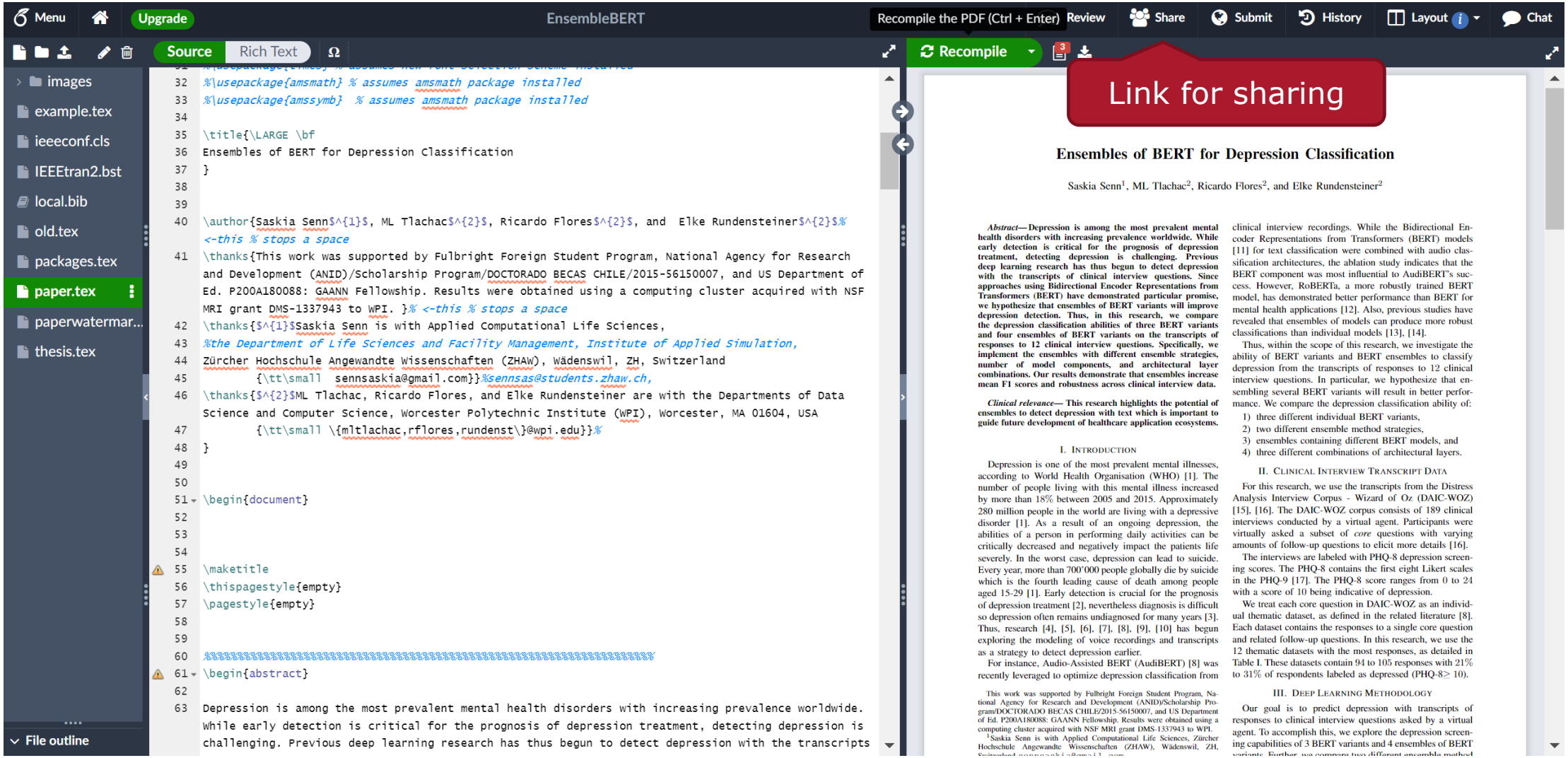
A. Text Message Data

Teams of researchers at Worcester Polytechnic Institute (WPI) collected the Moodable [9] and EMU [18] datasets from 2017 to 2019 under WPI IRB 00007374 File 18-0031. Crowd-sourced participants from Mechanical Turk [19] were asked to share retrospective phone data, record audio, and complete mental illness screening surveys through a mobile collection app [9], [18]. The app administered the Patient Health Questionnaire-9 (PHQ-9) [20]. Each survey item has participants rate the frequency of depression symptoms over the past two weeks on a 4-point scale from 'not at all' to 'nearly every day'. To determine the existence of suicidal ideation, we use the ninth item Q9: "Thoughts that you would be better off dead or of hurting yourself in some way" [20].

Here, we leverage data from the subset of participants 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. In response to the item Q9 of PHQ-9, 39 selected 0, 11 selected 1, 14 selected 2, and 2 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

This work was supported by US Department of Education P200A180088: GAANN grant and the Data Science department at WPI.

Overleaf for Writing Collaboration in LaTeX




```
%\usepackage{amsmath} % assumes amsmath package installed
%\usepackage{amssymb}  % assumes amsmath package installed

\title{\LARGE \bf
Ensembles of BERT for Depression Classification
}


\author{Saskia Senn$^{1}$, ML Tlachac$^{2}$, Ricardo Flores$^{2}$, and  Elke Rundensteiner$^{2}$%
<-this % stops a space
\thanks{This work was supported by Fulbright Foreign Student Program, National Agency for Research and Development (ANID)/Scholarship Program/DOCTORADO BECAS CHILE/2015-56150007, and US Department of Ed. P200A180088: GAANN Fellowship. Results were obtained using a computing cluster acquired with NSF MRI grant DMS-1337943 to WPI. }% <-this % stops a space
\thanks{$^{1}$Saskia Senn is with Applied Computational Life Sciences,
%the Department of Life Sciences and Facility Management, Institute of Applied Simulation,
Zürcher Hochschule Angewandte Wissenschaften (ZHAW), Wädenswil, ZH, Switzerland
        {\tt\small  sennsaskia@gmail.com}}%sennsas@students.zhaw.ch,
\thanks{$^{2}$ML Tlachac, Ricardo Flores, and Elke Rundensteiner are with the Departments of Data Science and Computer Science, Worcester Polytechnic Institute (WPI), Worcester, MA 01604, USA
        {\tt\small \{mltlachac,rflores,rundenst\}@wpi.edu}}%
}


\begin{document}



\maketitle
\thispagestyle{empty}
\pagestyle{empty}


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\begin{abstract}

Depression is among the most prevalent mental health disorders with increasing prevalence worldwide. While early detection is critical for the prognosis of depression treatment, detecting depression is challenging. Previous deep learning research has thus begun to detect depression with the transcripts
```

A Story Starts with a Title

(and a paper also has authors & affiliations)

Examples of Paper Titles

Typically, don't use contractions

You're Making Me Depressed: Leveraging Texts from Contact Subsets to Predict Depression

Depression Screening from Text Message Reply Latency

Screening for Suicidal Ideation with Text Messages

Mobile Depression Screening with Time Series of Text Logs and Call Logs

Ensembles of BERT for Depression Classification

Typically, don't use abbreviations

Author Affiliations

Bottom right corner of first page

Saskia Senn[1], ML Tlachac[2], Ricardo Flores[2], and Elke Rundensteiner[2]

[1]Saskia Senn is with Applied Computational Life Sciences, Zürcher Hochschule Angewandte Wissenschaften (ZHAW), Wädenswil, ZH, Switzerland sennsaskia@gmail.com

[2]ML Tlachac, Ricardo Flores, and Elke Rundensteiner are with the Departments of Data Science and Computer Science, Worcester Polytechnic Institute (WPI), Worcester, MA 01604, USA {mltlachac,rflores,rundenst}@wpi.edu

Underneath title

ML Tlachac
Data Science
Worcester Polytechnic Institute
mltlachac@wpi.edu

Katherine Dixon-Gordon
Psychological and Brain Sciences
University of Massachusetts Amherst
katiedg@umass.edu

Elke Rundensteiner
Data Science and Computer Science
Worcester Polytechnic Institute
rundenst@wpi.edu

Overleaf for Paper Starting Material

```
35  \title{\LARGE \bf
36  Ensembles of BERT for Depression Classification
37  }
38
39
40  \author{Saskia Senn$^{1}$, ML Tlachac$^{2}$, Ricardo Flores$^{2}$, and  Elke Rundensteiner$^{2}$$%
    <-this % stops a space
41  \thanks{This work was supported by Fulbright Foreign Student Program, National Agency for Research
    and Development (ANID)/Scholarship Program/DOCTORADO BECAS CHILE/2015-56150007, and US Department of
    Ed. P200A180088: GAANN Fellowship. Results were obtained using a computing cluster acquired with NSF
    MRI grant DMS-1337943 to WPI. }% <-this % stops a space
42  \thanks{$^{1}$Saskia Senn is with Applied Computational Life Sciences,
43  Zürcher Hochschule Angewandte Wissenschaften (ZHAW), Wädenswil, ZH, Switzerland
44          {\tt\small  sennsaskia@gmail.com}}
45  \thanks{$^{2}$ML Tlachac, Ricardo Flores, and Elke Rundensteiner are with the Departments of Data
    Science and Computer Science, Worcester Polytechnic Institute (WPI), Worcester, MA 01604, USA
46          {\tt\small \{mltlachac,rflores,rundenst\}@wpi.edu}}
47  }
48
49
50  \begin{document}
```

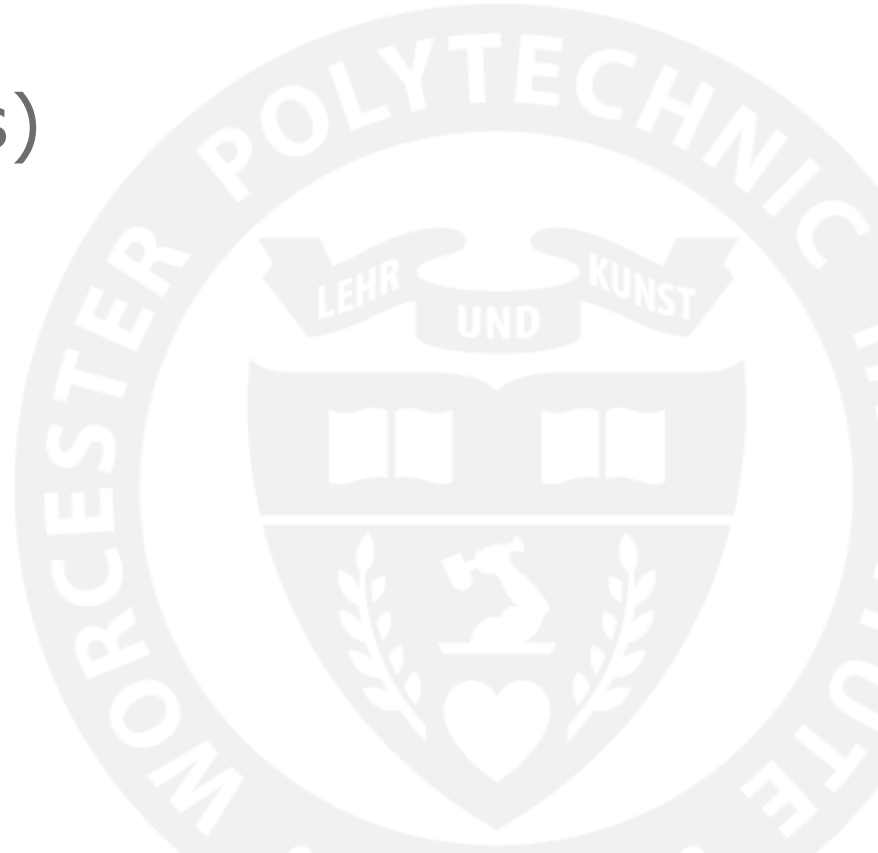
At this venue, funding is here

Separate \thanks{} for each university

Combined email addresses

A Story has a Summary

(A paper has an abstract… and keywords)



Abstract Outline

Abstract Example: Unbalanced

Motivation too long

Abstract—Depression, a prevalent and debilitating mental illness, is frequently undiagnosed. Diagnosis is an important step towards treatment. Currently screening tools, such as the Patient Health Questionnaire-9 (PHQ-9), require patient input. Many studies have used a variety of data types and features to predict depression scores for individuals. In this study, we focus on the predictive ability of a single under-utilized modality indicative of the impact of social interactions: received text messages. Our approach encompasses creating subsets of influential contacts for each participant and engineering features from the text messages of those contact subsets. Overall, our study demonstrates that received text communications are a promising modality when predicting depression scores. Specifically, we found that the F1 score of Gaussian Naive Bayes models leveraging just the text messages from a subset of top contacts performed statistically significantly better by 13.2 percent than the models leveraging text messages from all contacts.

Approach too long

Approach too short

Great Motivation And Problem

Abstract—Depression is among the most prevalent mental health disorders with increasing prevalence worldwide. While early detection is critical for the prognosis of depression treatment, detecting depression is challenging. Previous deep learning research has thus begun to detect depression with the transcripts of clinical interview questions. Since approaches using Bidirectional Encoder Representations from Transformers (BERT) have demonstrated particular promise, we hypothesize that ensembles of BERT variants will improve depression detection. Thus, in this research, we compare the depression classification abilities of three BERT variants and four ensembles of BERT variants on the transcripts of responses to 12 clinical interview questions. Specifically, we implement the ensembles with different ensemble strategies, number of model components, and architectural layer combinations. Our results demonstrate that ensembles increase mean F1 scores and robustness across clinical interview data.

Results Too short

Missing Implication

Abstract Examples: Well Balanced

Abstract—Suicide is a leading cause of death in the US, with suicide rates increasing annually. Passive screening of suicidal ideation is vital to provide referrals to at-risk individuals. We study to what degree smartphone-based communication, in particular, text messages, could be leveraged for passively screening for suicidal ideation. We analyze the screening ability of texts sent in different time periods prior to reported ideation, namely, texts from specific weeks only versus accumulative over several weeks. Our approach involves performing comprehensive feature engineering and identifying influential features to train machine learning models. With just the prior week of texts, we were able to predict the existence of suicidal ideation with AUC = 0.88, F1 = 0.84, accuracy = 0.81, sensitivity = 0.94, and specificity = 0.68. The most influential features include word frequencies of words in the car, clothing, affection, confusion, driving, real estate, and journalism categories. This research, demonstrating the potential of text messages to screen for suicidal ideation, will guide the development of screening technologies.

Motivation

Problem

Approach

Results

Implication

Abstract— Depression is both debilitating and prevalent. While treatable, it is often undiagnosed. Passive depression screening is crucial, but leveraging data from Smartphones and social media has privacy concerns. Inspired by the known relationship between depression and slower information processing speed, we hypothesize the latency of texting replies will contain useful information in screening for depression. Specifically, we extract nine reply latency related features from crowd-sourced text message conversation meta-data. By considering text meta-data instead of content, we mitigate the privacy concerns. To predict binary screening survey scores, we explore a variety of machine learning methods built on principal components of the latency features. Our findings demonstrate that an XGBoost model built with one principal component achieves an F1 score of 0.67, AUC of 0.72, and Accuracy of 0.69. Thus, we confirm that reply latency of texting has promise as a modality for depression screening.

Keywords (5) to make paper easier to find

***Index Terms*—text feature engineering, depression screening, received communications, contact subsets**

***Index Terms*—mobile health, depression screening, time series, feature engineering, machine learning**

***Clinical relevance*— This research highlights the potential of ensembles to detect depression with text which is important to guide future development of healthcare application ecosystems.**

Watch for venue specific requirements

Keywords asked for in submission system

Overleaf: Abstract and Keywords

Starts abstract environment

Ends Keyword environment

```
110 % As a general rule, do not put math, special symbols or citations
111 % in the abstract or keywords.
112 \begin{abstract}
113 Suicide is a leading cause of death in the US, with  suicide rates  increasing annually. Passive
    screening of suicidal ideation is vital to provide referrals to at-risk individuals. We study to what
    degree smartphone-based communication, in particular, text messages,
114 could be leveraged for  passively screening for suicidal ideation. We analyze  the screening ability
    of texts sent in different time periods
115 prior to reported ideation, namely, texts from specific weeks only versus accumulative over several
    weeks.
116 Our approach involves performing comprehensive feature engineering and identifying influential
    features to train machine learning models. With just the prior week of texts, we were able to predict
    the existence of suicidal ideation with AUC = 0.88, F1 = 0.84, accuracy = 0.81, sensitivity = 0.94,
    and specificity = 0.68. The most influential features include word frequencies of words in the car,
    clothing, affection, confusion, driving, real estate, and journalism categories. This research,
    demonstrating the potential of text messages to screen for suicidal ideation,
117 will guide the development of screening technologies.
118
119 \end{abstract}
120
121 \begin{IEEEkeywords}
122 mobile health, suicide ideation, passive screening, machine learning, digital phenotype
    \end{IEEEkeywords}
```

What is in a Story?

(and by extension a paper?)

Story Sections

Intro

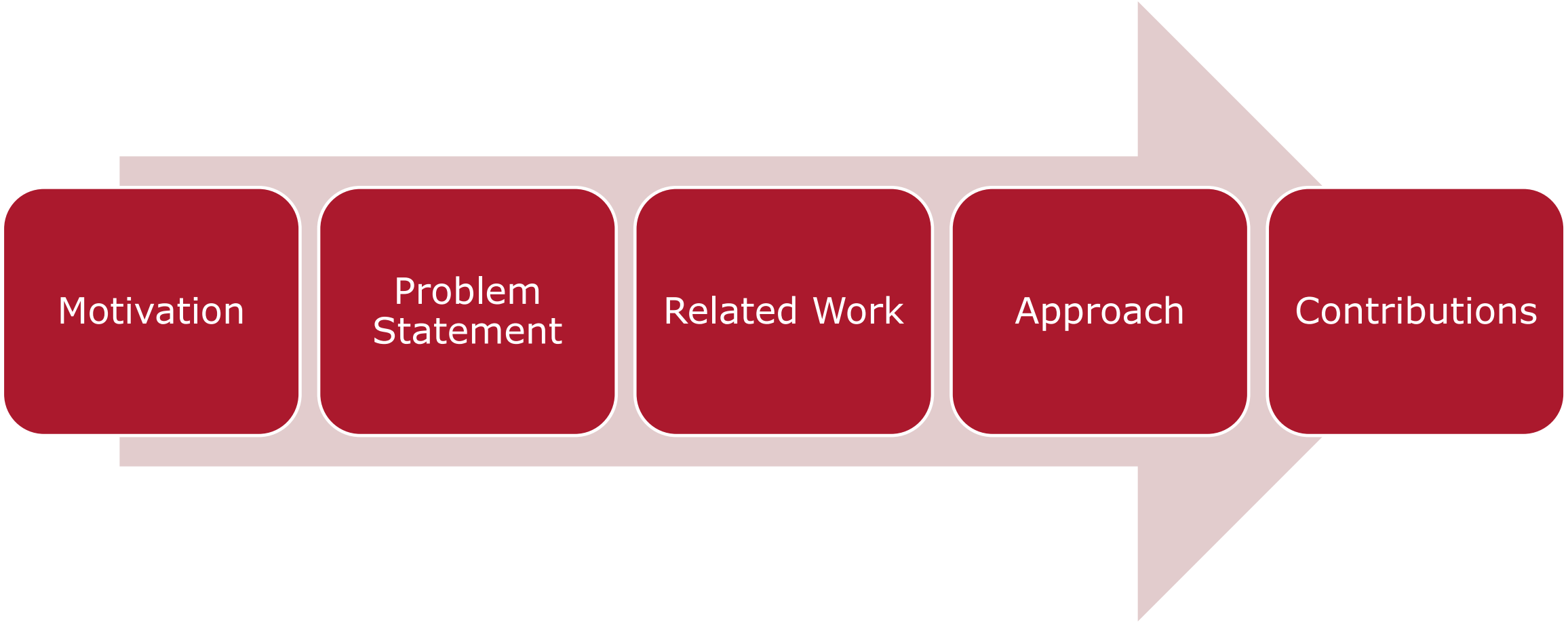
Data

Methods

Results

Discussion and/or Conclusion

What is in an Introduction?



Introduction Examples

I. Introduction

Depression is a debilitating mood disorder characterized by chronic sadness and apathy. Despite being one of the most treatable mental disorders, US Preventative Services Task Force (USPSTF) has identified depression as the leading cause of disability in adults [10], [17]. While the USPSTF recommending depression screening for all adults, an estimated quarter of patients with depression are not diagnosed [7], [17]. This often occurs due to symptoms not being recognized, lack of access to medical resources, and fear of stigma [7]. Currently, questionnaires such as the Patient Health Questionnaire-9 (PHQ-9) are being utilized to screen for depression [10]. However, these screening tools require patient cooperation and unbiased ability to recall symptoms.

Motivation and problem

A. Related Work

Recent studies have begun to leverage smartphone sensor and social media data for mental health screening [1], [4], [12], [13], [16]. Specifically, location sensor [16], audio [12], and twitter [13] data have been shown to be effective at diagnosing depression. De Choudhury, Counts, and Horvitz conducted comprehensive feature engineering to extract features related to engagement, ego-network, emotional expression, linguistic style, and patterns of activity from Twitter data [4].

The aforementioned 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 [3]. Specifically, negative interactions have been found to be more influential than positive interactions [15]. Additionally, friendships are important to happiness [5] and the number of contacts an individual is comfortable with is associated with better mental health [15], though research indicates people only have three close friends [5]. Likewise, fewer close relationships and lack of social support are associated with depression [9].

B. Our Approach and Contributions

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 [6]. 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 scores, and
2) comparing the predictive ability of features generated from subsets of top contacts with features from all contacts.

I. Introduction

Suicide is among the top ten causes of death for all age groups in the United States in 2018 [1]. For those aged 10-34, suicide is the second leading cause of death. The suicide rate has increased by 35% since 1999 despite national goals to combat suicide [1]. Psychological autopsies reveal that 90% of individuals who died by suicide suffered mental illness symptoms [2], though 56% were not diagnosed with a mental illnesses [3]. Early identification and referral of at-risk individuals is recommended to prevent suicide [4].

Motivation

Smartphones are not only prevalent but also capture large quantities of personal data, making them perfect to passively screen for suicidal ideation and provide referrals to at-risk individuals. 95% of [illegible] years of age, the age group most at risk for suicide, owned a personal smartphone in 2018. While research has been conducted on screening for depression with smartphone data [5]–[14], there is a lack of such digital phenotype research for suicidal ideation screening.

Problem

While we previously attempted to address this problem [9], the machine learning models yielded an $F1$ score of only 0.58 when predicting the existence of suicidal ideation with features extracted from mobile voice recordings, social media, and phone data. On actively recorded mobile voice, we were able to screen for suicidal ideation with an $F1$ of 0.70 [15]. Yet, we achieved an even higher $F1$ score of 0.81 when screening for depression with the content of sent SMS text messages [10], which have the additional benefit of being passively collected.

Related Work

We now hypothesize that text messages are a valuable modality to passively screen for suicidal ideation. Yet suicide risk is dynamic [16], [17], 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. Our approach involves comprehensive feature engineering, feature selection, and machine learning. We demonstrate that this approach dramatically increases the suicidal ideation screening ability of data collected from smartphones. Our contributions include:

Approach

1) exploring the potential of retrospectively collected crowd-sourced texts for suicidal ideation screening,
2) comparing the screening ability of texts sent in the interval and cumulative weeks prior to reported ideation,
3) identifying the most influential features for screening.

Contributions are bullet points

Overleaf: Bullet Points and Sections

```
151  Our approach involves c                    ering, feature selection, and machine learning. We
      demonstrate
152  that this approach dram                    idal ideation screening ability of data collected
     from smartphones.
153  Our  contributions include:
     \begin{enumerate}
155      \item exploring the potential of retrospectively collected crowd-sourced texts for suicidal
         ideation screening,
156      \item comparing the screening ability of texts sent in the interval and cumulative weeks prior to
         reported ideation,
157      %different weekly subsets of texts from crowd-sourced participants, and
         \item identifying the most influential features for screening.
159  \end{enumerate}
160
161
162
163  \section{Data \& Methods}
164
     \subsection{Text Message Data}
166  Teams of researchers at Worcester Polytechnic Institute (WPI) collected the Moodable
     \cite{MoodableSmartHealth} and EMU \cite{TlachacEMU} datasets from 2017 to 2019 under WPI IRB
```

Remains a single Paragraph

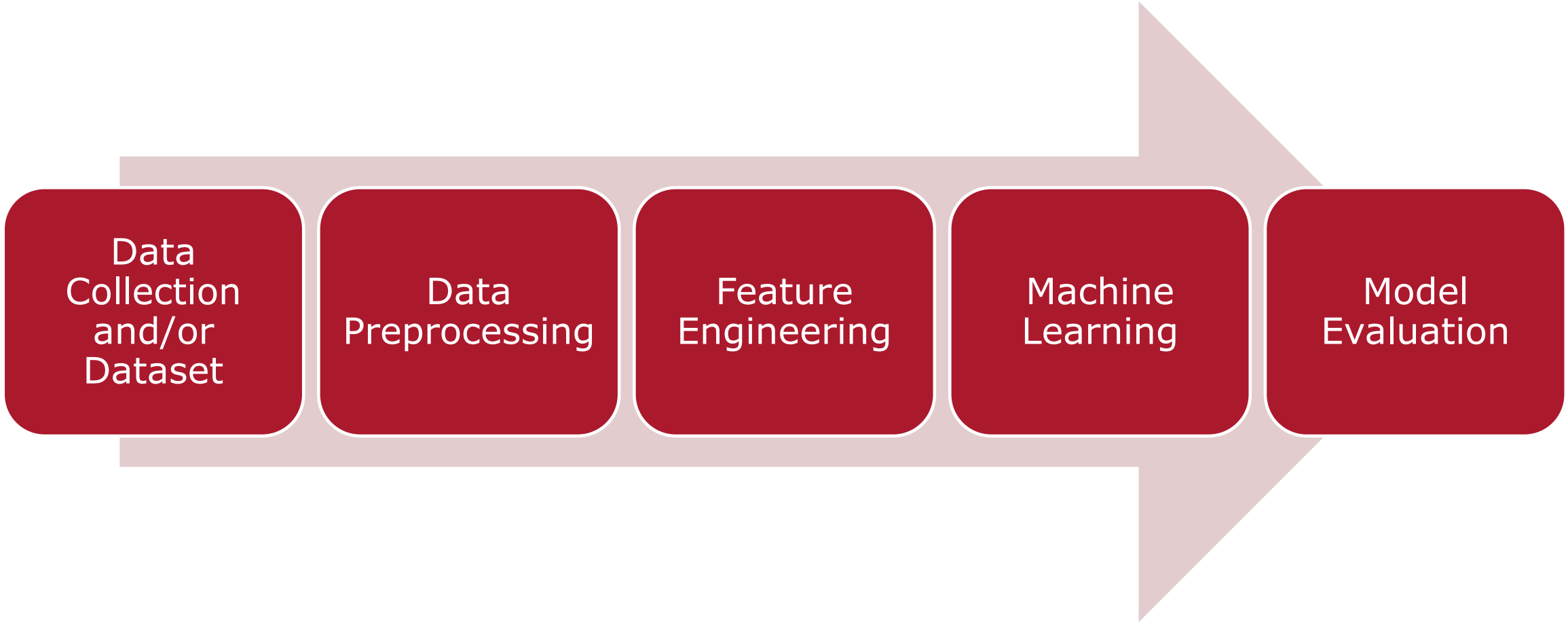
Starts bullet environment

\item for each bullet

May need '\' to keep special characters

For subsection heading

What is in Data & Methodology?



Examples of Paper Data Descriptions

Data as a subsection of Methods

II. Data & Methods

A. Text Message Data

Teams of researchers at Worcester Polytechnic Institute (WPI) collected the Moodable [9] and EMU [18] datasets from 2017 to 2019 under WPI IRB 00007374 File 18-0031. Crowd-sourced participants from Mechanical Turk [19] were asked to share retrospective phone data, record audio, and complete mental illness screening surveys through a mobile collection app [9], [18]. The app administered the Patient Health Questionnaire-9 (PHQ-9) [20]. Each survey item has participants rate the frequency of depression symptoms over the past two weeks on a 4-point scale from 'not at all' to 'nearly every day'. To determine the existence of suicidal ideation, we use the ninth item Q9: "Thoughts that you would be better off dead or of hurting yourself in some way" [20].

Here, we leverage data from the subset of participants 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. In response to the item Q9 of PHQ-9, 39 selected 0, 11 selected 1, 14 selected 2, and 2 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

TABLE I
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

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 I). 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 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.

Tables to describe data

Data as a separate section

Specific section title

II. Clinical Interview Transcript Data

For this research, we use the transcripts from the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) [15], [16]. The DAIC-WOZ corpus consists of 189 clinical interviews conducted by a virtual agent. Participants were virtually asked a subset of *core* questions with varying amounts of follow-up questions to elicit more details [16].

The interviews are labeled with PHQ-8 depression screening scores. The PHQ-8 contains the first eight Likert scales in the PHQ-9 [17]. The PHQ-8 score ranges from 0 to 24 with a score of 10 being indicative of depression.

We treat each core question in DAIC-WOZ as an individual thematic dataset, as defined in the related literature [8]. Each dataset contains the responses to a single core question and related follow-up questions. In this research, we use the 12 thematic datasets with the most responses, as detailed in Table I. These datasets contain 94 to 105 responses with 21% to 31% of respondents labeled as depressed (PHQ-8$\geq$ 10).

TABLE I
THEMATIC DATASET DESCRIPTIONS.

Core Question Description	Count	Depressed
How are you doing today?	105	28.6%
The last time you argued with someone?	103	29.1%
What advice would you give yourself?	102	28.4%
What are you most proud of?	100	28.0%
How are you at controlling your temper?	100	30.0%
When was the last time you felt really happy?	99	28.3%
How easy is it for you to get good sleep?	98	27.6%
How would your best friend describe you?	96	26.0%
What's your dream job?	95	30.5%
What'd you study at school?	95	30.5%
Do you travel a lot?	94	27.7%
Have you been diagnosed with depression?	94	21.3%

Data, Data Preprocessing, & Feature Engineering

II. DATA & METHODOLOGY

A. Data

Worcester Polytechnic Institute (WPI) researchers designed the Moodable approach to collecting retrospective Smartphone sensor data [5]. Under WPI IRB 00007374, teams retrospectively collected data from participants on Mechanical Turk between [illegible] and 2019, resulting in the Moodable and EMU datasets. Collected modalities include sent texts, received texts, tweets, voice samples, etc.

The Patient Health Questionnaire-9 (PHQ-9), a common depression screening survey, was deployed to obtain a depression label for each participant. The PHQ-9 asks participants to reflect upon the prior two weeks when answering 9 questions [1]. Each question has a score between 0 and 3 so aggregate scores are between 0 and 27. A PHQ-9 score of at least 10 is commonly accepted as the threshold for an interim diagnosis of depression [1]. If a clinician were to confirm this interim diagnosis, treatment in the form of psychotherapy and/or antidepressants is typically recommended.

Inspired by the PHQ-9, 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 issues typically stemming from crowd-sourced data. The latencies for the 37 depressed participants (PHQ-9 $\geq$ 10) and 31 not depressed participants (PHQ-9 $<$ 10) are depicted in Fig. 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 latencies, we consider the 10%, 25%, 50%, 75%, and 90% quantiles of the set of reply latencies. We included the [illegible] 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.

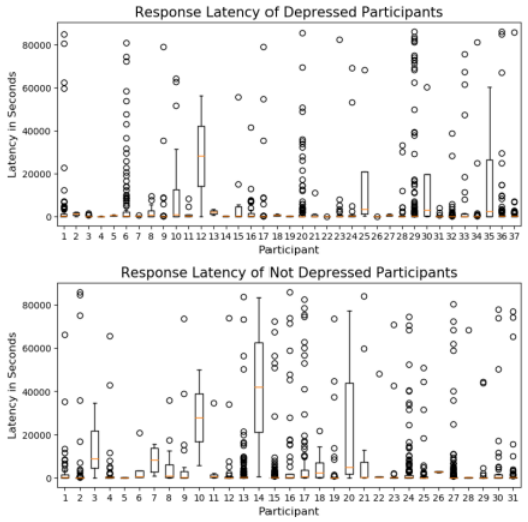


Fig. 1. Text response latency within the last 14 days for each participant.

Data visuals

II. DATA & METHODOLOGY

A. Text and Call Log Data

This research leverages the text logs and call logs in the combined Moodable [9] and EMU [15] datasets. Worcester Polytechnic Institute (WPI) researchers collected retrospective smartphone data from Mechanical Turk [16] participants from 2017 to 2019 under WPI IRB 00007374 File 18-0031. Participants were prompted to complete the popular Patient Health Questionnaire-9 (PHQ-9) [17], the same survey used by the StudentLife app [11] [13]. The PHQ-9 asks participants to reflect on the prior two weeks when completing nine scales regarding frequency of symptoms. The PHQ-9 score can range from 0 to 27 with a score of at least 10 being indicative of depression [17]. We use these PHQ-9 scores to label the logs.

Since the PHQ-9 asks about the last two weeks, we consider the last two weeks of text and call logs. Participants with at

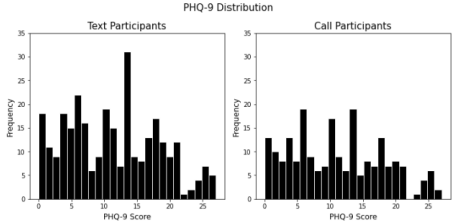


Fig. 1. PHQ-9 distributions of participants who submitted text and call logs.

TABLE I
NUMBER OF PARTICIPANTS WITH EACH TYPE OF DATA.

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

least two texts or two minutes of calls within the prior two weeks are included. [illegible] 212 participants, 295 and 212 shared text and call logs, respectively. The distributions of PHQ-9 scores of these participants are available in Fig. 1. The average number of texts was 127, average number of calls was 84, and the average minutes of calls was 146.

B. Time Series Construction

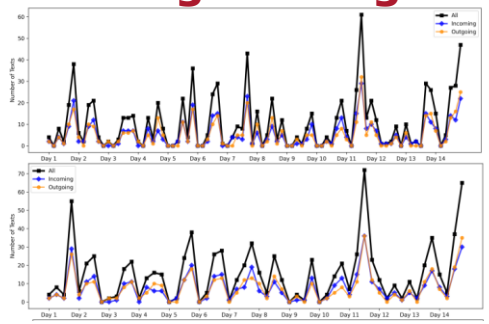
In order to construct the time series from the logs, we consider the *count* of communications, number of unique *contacts*, and *average length* of communications. While all the call logs contained call duration, not all retrospective text logs contained message size so we approximated this value with 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 time series 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. As with all 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 I. Notably, only a third of the participants with text logs submitted outgoing messages.

For each of the six possible data types in Table I, 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 for all aggregation intervals are displayed in Fig. 2. This participant also has 12 time series for number of text contacts and 12 time series for average text length.

C. Feature Engineering

We leverage the Time Series Feature Extraction Library (TSFEL) [18] to extract [illegible] temporal, and 26 spectral features from each time series. In addition to considering the 60 [illegible] and 60 length



Worcester Polytechnic Institute

Overleaf: Tables & Figures

Machine Learning

Machine Learning. Traditional machine learning methods are preferable in psychopathology [24] due to their explainability and effectiveness for small datasets. After initial exploration, we leverage a selection of popular parametric and non-parametric methods with default parameters [23]: Gaussian Naive Bayes (NB), Logistic Regression (LR), Support Vector Classifier (SVC), and k-Nearest Neighbor (kNN).

Justification and Methods

D. Machine Learning Methodology

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 distance [20]. Further, we train models with between 1 and 15 PCs from the time series features. Specifically, we compare the depression prediction ability of the non-parametric kNN method with two parametric methods and a non-parametric ensemble method [21]: Logistic regression (LR), support vector classifier (SVC) with a Gaussian kernel, and random forest classifier (RF).

Goal and Methods

III. Deep Learning Methodology

Our goal is to predict depression with transcripts of responses to clinical interview questions asked by a virtual agent. To accomplish this, we explore the depression screening capabilities of 3 BERT variants and 4 ensembles of BERT variants. Further, we compare two different ensemble method architectures. We focus on BERT classifiers in this research given their demonstrated success with smaller datasets and mental health applications [12], [8].

Goal and Overview

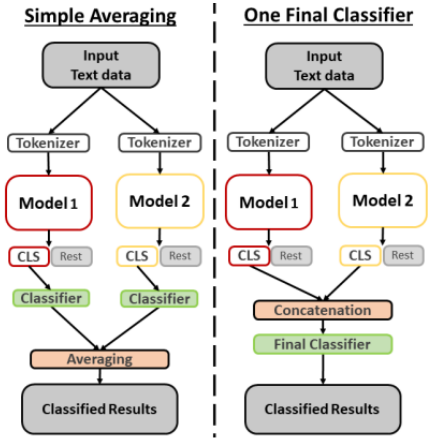


Fig. 1. Comparison of the two ensemble method strategies.

TABLE 2
THE COMPONENTS COMPRISING EACH OF THE DIFFERENT ENSEMBLES.

Ensemble	Method	BERT	RoBERTa	DistilBERT
Ens 1	Simple Averaging	✓	✓	
Ens 2	One Final Classifier	✓	✓	
Ens 3	One Final Classifier		✓	✓
Ens 4	One Final Classifier	✓	✓	✓

A. Individual BERT Variants

BERT. The Bidirectional Encoder Representations of Transformers is a pretrained model for language representation [11]. BERT was revealed to have a superior performance due to its architecture leveraging multi-layer bidirectional Transformer encoder combined with multiple attention heads. The model is pretrained on BooksCorpus and English Wikipedia (16GB) using Masked Language Modeling (*MLM*) and Next Sentence Prediction (*NSP*). This pretraining allows for previously unprecedented success when classifying smaller datasets.

RoBERTa. A previous study [12] suggests that the Robustly Optimized BERT Pretraining Approach (RoBERTa) model [18] is better than BERT for mental health applications. The training approach for RoBERTa is different from BERT base model as there is no *NSP* and instead an extented *MLM* procedure is integrated. Pretrained on a bigger corpus than BERT, RoBERTa includes more informal text data in the pretraining such as a Reddit corpus.

DistilBERT. DistilBERT [19] is a less computationally costly alternative to the BERT base model. Retaining 97% of BERT's language understanding and general-purpose pretraining, DistilBERT reduces the size of the BERT model by 40%. Thus, to reduce computational costs, DistilBERT is more attractive than the BERT base model in ensembles.

B. Ensemble Strategies

Simple Averaging. For the simple averaging ensemble strategy, the final logit result was created by averaging the numeric results in the classification layer of the models, as depicted in Fig.1. As with the individual models, the resulting average was then classified with a threshold of 0.5.

One Final Classifier. For the one final classifier ensemble strategy [8], we concatenated the CLS tokens from all of the individual classifiers in the ensemble, as depicted in Fig.1. The final classification layer was then applied to this concatenated vector to output the final logit for the ensemble.

Model Evaluation

Evaluation Metrics. We evaluate the models with AUC, $F1$. and $accuracu$ metrics. For the models that maximize these scores, we also report on $sensitivity$ and $specificity$ (Eq. 1). These metrics are calculated with the count of true positive tp, false positive fp, false negative fn, and true negative tn predictions. AUC determines the model classification ability at different thresholds by calculating the area under the ROC curve formed by plotting $sensitivity$ and $1 - specificity$. F1 (Eq. 2) is the harmonic mean between the positive predictive value and $sensitivity$. $Accuracy$ denotes the ratio of correctly classified instances.

$$Sensitivity = \frac{tp}{tp + fn}, \; Specificity = \frac{tn}{tn + fp} \quad (1)$$

$$F1 = \frac{2tp}{2tp + fp + fn} \quad (2)$$

Model Training. 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 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.

E. Model Evaluation

Test Set and Data Balancing

Metrics with Justification

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. $F1$ (Eq. 2) is the balance between precision (Eq. 1) and sensitivity (Eq. 2) which are calculated with the number of true positive (tp), false positive (fp), and false negative (fn) predictions. We will also report on the precision, sensitivity, specificity, ROC AUC, and accuracy of the models with the highest average $F1$ scores for comparison purposes.

$$Precision = \frac{tp}{tp + fp}, \; Sensitivity = \frac{tp}{tp + fn} \quad (1)$$

$$F1 = \frac{2(precision)(sensitivity)}{precision + sensitivity} \quad (2)$$

E. Model Implementation and Evaluation

Parameter Settings

Test Set and Data Balancing

Metrics with Justification

Due to our hyperparameter tuning, we ran models with batchsize=8, learning rate=$2x10^{-5}$, and step size=$2x10^{8}$. Our models used 128 tokens, a cross entropy loss function, and an Adam optimizer with weight decay. Each model is trained for 10 epochs. For each thematic dataset, the test sets contain 20% of the responses [8]. To mitigate bias from unequal classes, training sets were upsampled. Experiments are repeated 10 times with randomly initialized weights.

Model performance is evaluated based on metrics calculated with the number of true positive TP, false positive FP, false negative FN, and true negative TN predictions. Our goal is to maximize $F1$ (Eq. 1), a popular metric for diagnostic tasks with unbalanced data. $F1$ is the harmonic mean between the positive predictive value and sensitivity. Also popular for diagnostics, $sensitivity$ is the true positive rate and $specificity$ is the true negative rate. We report the metrics for models with the five highest F1 scores.

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (1)$$

Not needed for accuracy

$$Sensitivity = \frac{TP}{TN + FN}; \; Specificity = \frac{TN}{TN + FP} \quad (2)$$

Methods: Software and Availability

E. Software Tools and Availability

This work was implemented with Python 3.5.2. Visualizations were created with Pandas and Matplotlib. We will release the code at https://github.com/mltlachac/IEEBHI2019.

E. Software and Tool Availability

We will release the code, feature dataset, and additional visualizations at http://github.com/mltlachac/EMBC2020.

C. Availability

We will release the featurized data, code, results, and supplementary figures at *github.com/mltlachac/IEEEBHI2021*. We will post research updates at *emutivo.wpi.edu/*.

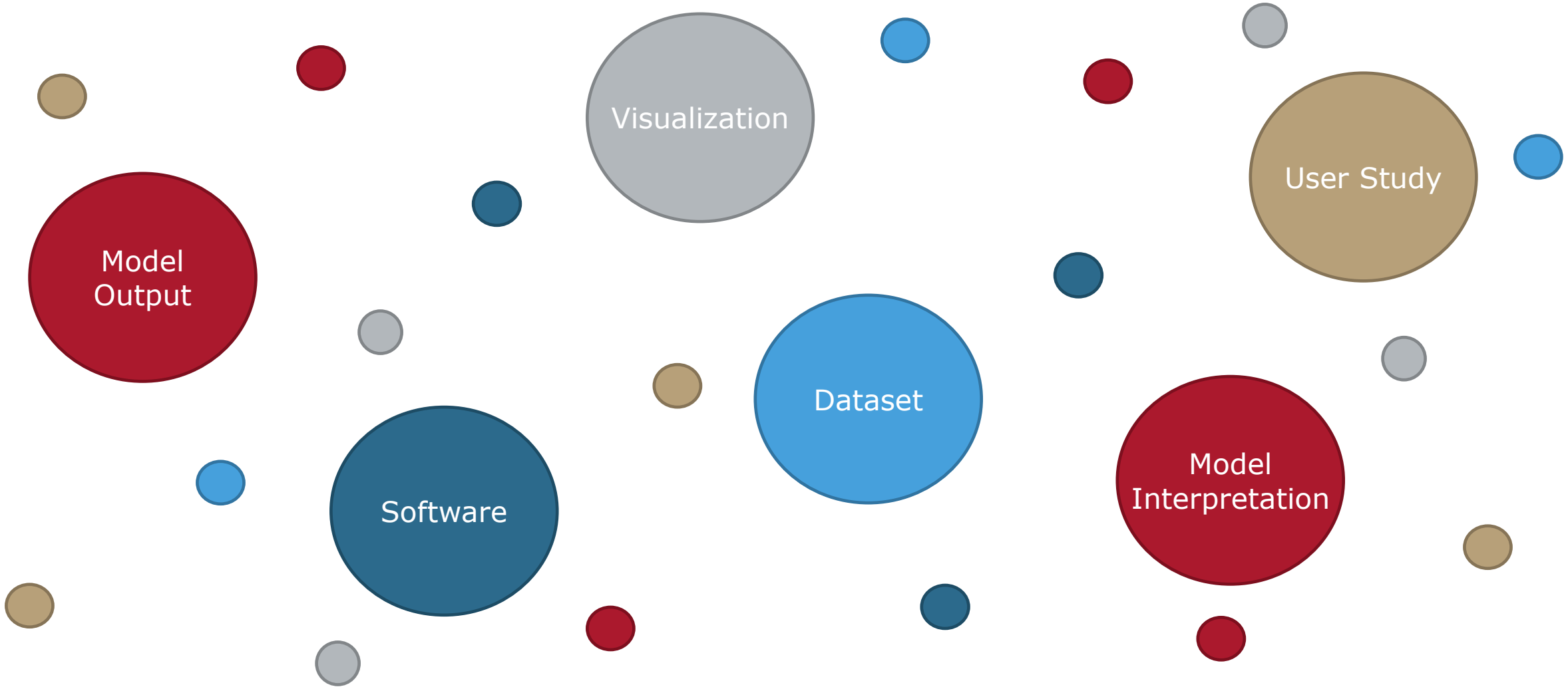
F. Availability

Research updates are available at emutivo.wpi.edu/. We will release the code, featurized data, machine learning results, and more visualizations on github.com/mltlachac/IEEEBHI2021.

F. Availability

Upon publication, we will release the ensemble code at *github.com/sennsaskia/EnsemblesBERT.git* Further research updates will be available at *emutivo.wpi.edu*.

What are Results?



Comparing Different Result Text Formatting

Only text

IV. Results

For every method, we identify the parameter setting and dataset with the highest average F1 score for the 100 trials. These highest average F1 scores with any contact subset for each method are shown in Fig. 2. 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.

Fig. 3 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 $C_P(25\%)$ contacts and $C_P(0.25)$ 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 $C_P(25\%)$ or $C_P(0.25)$ contacts, the F1 score is improved by 13.2 percent in comparison to using features from all contacts. From t-tests, we con-

Subsections

III. RESULTS

Table I 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 I 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 metrics for the XGBoost models. F1, AUC, and Accuracy for the XGBoost models are seen in Fig. 3.

A. Principal Component Analysis Results

From Fig. 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 and have

Bold text

IV. Results

The models' performances aggregated across the 12 thematic datasets are shown in Table 3 and Fig. 2. We observe the highest mean $F1$ of 0.62 for our ensemble models Ens 2 (BLA) and Ens 3 (BLA). These ensembles also have the lowest standard deviations, indicating robustness. Both ensembles have mean sensitivities of 0.64 and specificities of 0.61 which are well balanced for inversely related metrics. Given that DistilBERT in Ens 3 is less computationally costly than BERT in Ens 2, we consider Ens 3 to be the most preferable of our ensembles.

BERT Variants. The individual BERT (BLA), RoBERTa (B), and DistilBERT (BLA) models performs almost as well as best performing ensembles with mean $F1$ scores of 0.60,

Results Should have Many Tables and Figures

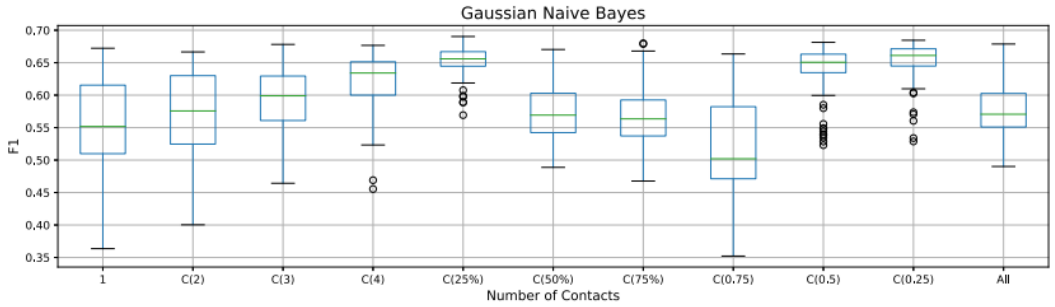


Fig. 3. F1 scores of 100 Gaussian Naive Bayes models for every contact subset.

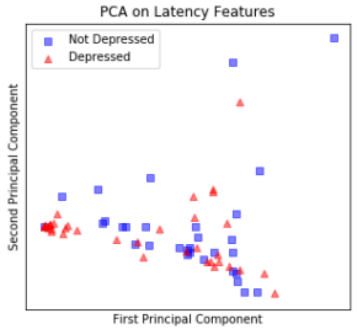


Fig. 5. Top two principal components.

Table caption above

TABLE I

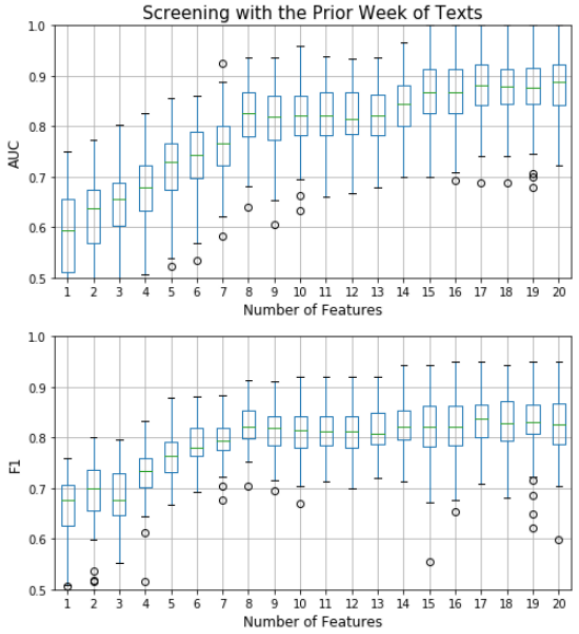
BEST MODEL CONFIGURATION FOR EACH METHOD WITH PCA.

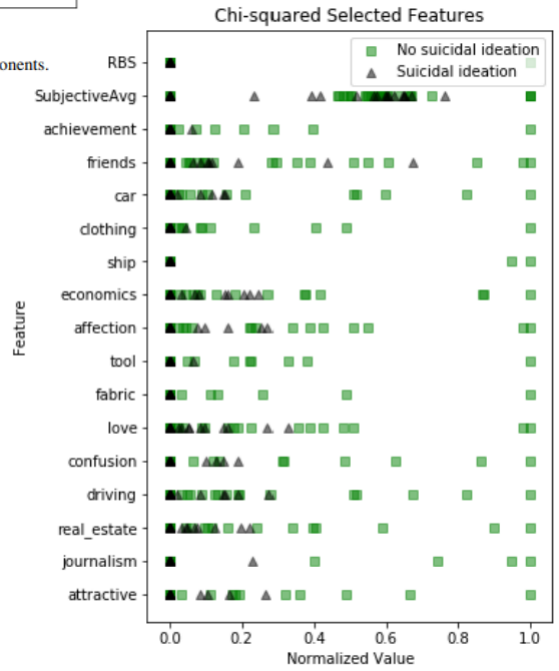
Method	Parameter	PCs	F1	AUC	Accuracy
NB		8	0.66	0.67	0.60
LR		1	0.55	0.65	0.59
SVM	linear	1	0.55	0.65	0.61
kNN	k=3	8	**0.68**	0.70	**0.68**
RF	3	1	0.64	0.70	**0.69**
XGBoost	Any	1	**0.67**	**0.72**	**0.69**
AdaBoost		1	0.63	0.66	0.55

TABLE II

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





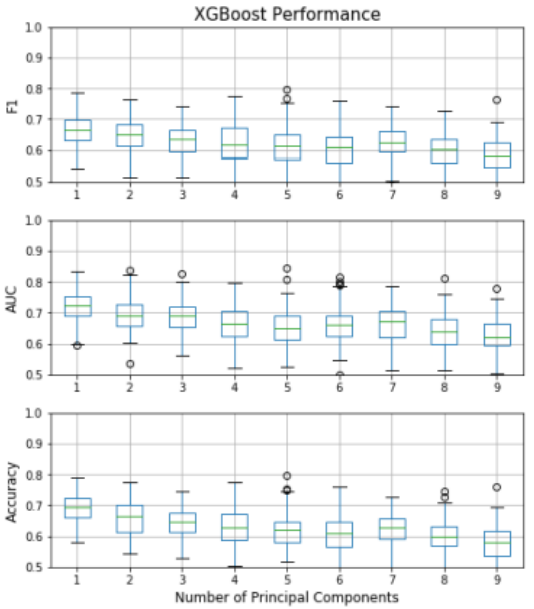


Figure caption below

Fig. 3. 100 trials of XGBoost models with depth 2.

And Even More Tables and Figures

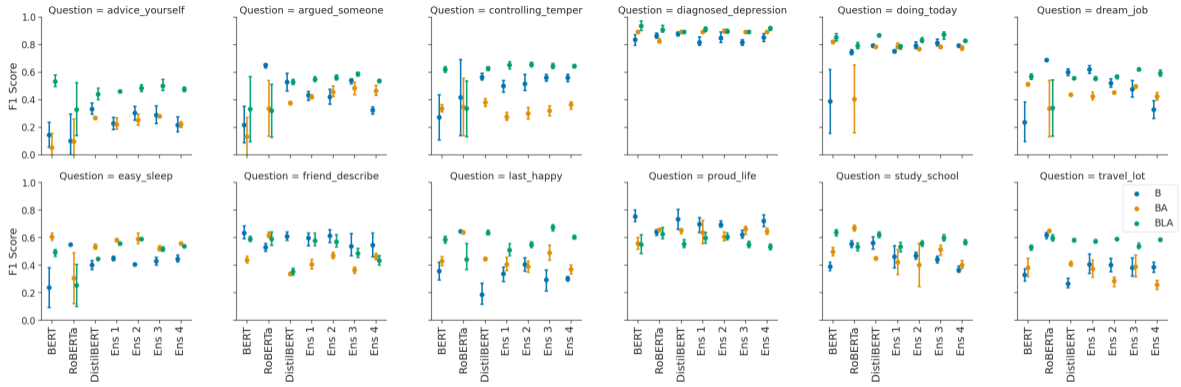


Fig. 3. Mean $F1 \pm$ standard deviation for each dataset, ordered alphabetically. B = base model, A = attention layer, L = LSTM layer.

TABLE II
COMPARISON OF THE MODEL CONFIGURATIONS WITH THE HIGHEST AVERAGE $F1$ SCORES FOR TEXT LOGS.

Data	Direction	Interval	Method	PCs	Sampling	F1 $\pm \sigma$	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 III
COMPARISON OF THE MODEL CONFIGURATIONS WITH THE HIGHEST AVERAGE $F1$ SCORES FOR CALL LOGS.

Data	Direction	Interval	Method	PCs	Sampling	F1 $\pm \sigma$	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**

TABLE 3
MEAN ± STANDARD DEVIATION ACROSS ALL 12 THEMATIC DATASETS. B = BASE MODEL, A = ATTENTION LAYER, L = LSTM LAYER.

Model Architecture	F1			Sensitivity			Specificity		
	B	BA	BLA	B	BA	BLA	B	BA	BLA
BERT	0.35±0.29	0.47±0.24	0.60±0.18	0.31±0.27	0.45±0.25	0.61±0.19	0.52±0.39	0.62±0.30	0.60±0.25
RoBERTa	0.58±0.22	0.43±0.33	0.47±0.31	**0.73**±0.29	0.56±0.44	0.54±0.36	0.30±0.32	0.17±0.29	0.34±0.32
DistilBERT	0.54±0.21	0.50±0.18	0.59±0.16	0.49±0.22	0.44±0.20	0.59±0.18	0.73±0.22	0.74±0.19	0.62±0.23
Ens 1	0.52±0.18	0.49±0.20	0.60±0.13	0.47±0.19	0.43±0.22	0.62±0.15	0.75±0.18	**0.76**±0.16	0.61±0.21
Ens 2	0.53±0.17	0.49±0.20	**0.62**±0.12	0.49±0.17	0.44±0.21	0.64±0.14	0.74±0.18	**0.76**±0.18	0.61±0.20
Ens 3	0.51±0.18	0.51±0.18	**0.62**±0.13	0.47±0.19	0.48±0.19	0.64±0.14	0.73±0.19	0.72±0.18	0.61±0.26
Ens 4	0.49±0.21	0.48±0.20	0.60±0.14	0.46±0.21	0.44±0.22	0.63±0.15	0.75±0.19	0.75±0.18	0.57±0.20

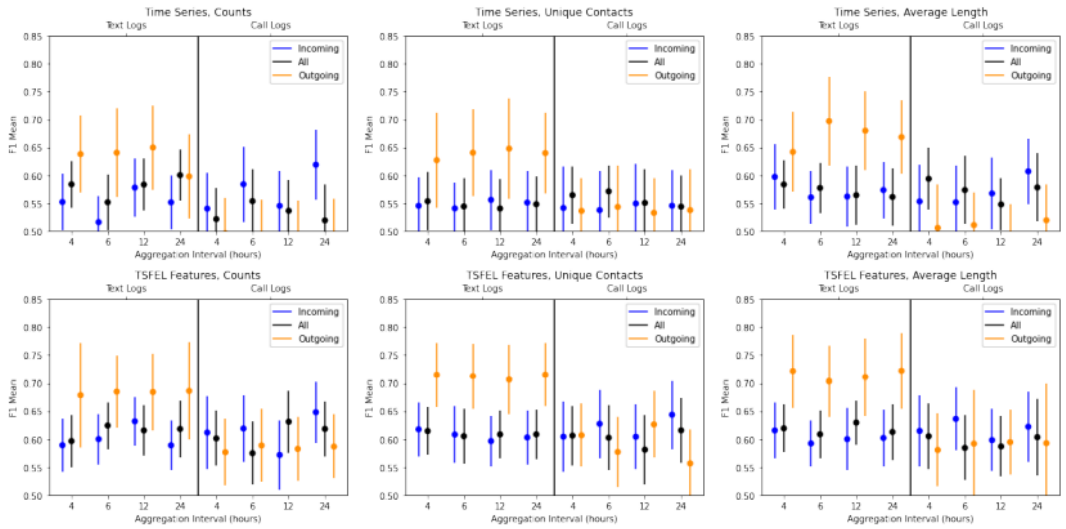


Fig. 3. Comparison of the model configurations with the highest average $F1$ scores for text and call logs.

Figures can span two columns

Figures can span one column

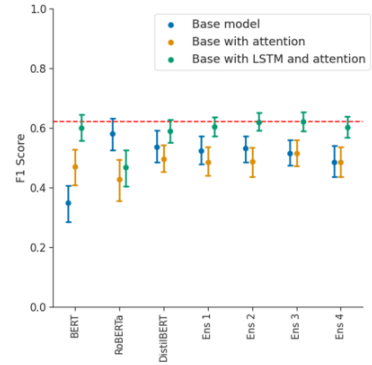
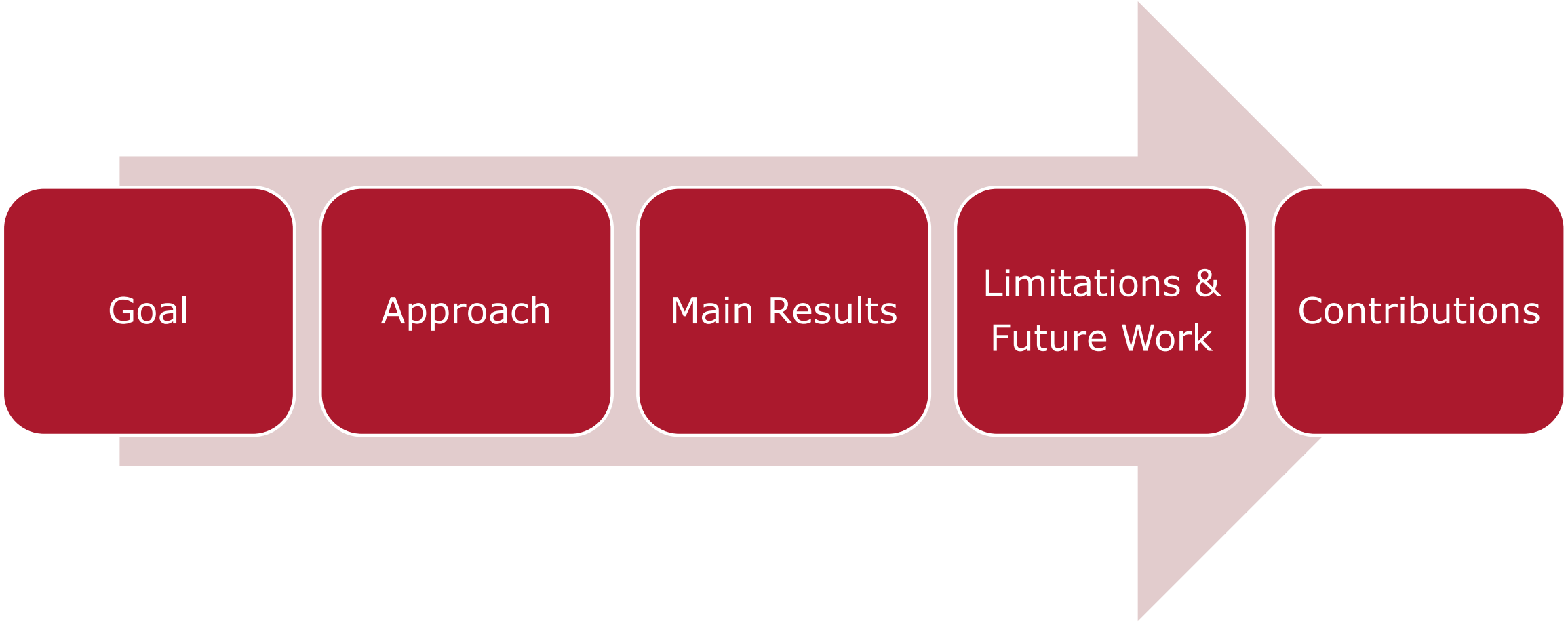


Fig. 2. Mean $F1 \pm$ standard deviation for all thematic datasets. Red line indicates highest mean value of 0.62.

Bold the best values in tables

What is in a Discussion/Conclusion?



Examples of Conclusions

VI. Conclusion

Our study confirms that received communications could be useful in predicting mental health. We predicted a binary — Conclusion

PHQ-9 score at cutoff 10 with an average F1 score of 0.653 using features extracted from retrospectively gathered text messages from a subset of contacts. In addition, we found that generating features from a subset of contacts rather than all contacts statistically significantly improved the F1 score by 13.2 percent. — Best Results

Future work involves including other modalities, predicting for other PHQ-9 cutoffs, including additional features, and performing feature selection/reduction. In addition, this idea of screening for depression with received communications from a subset of contacts could be expanded to other datasets. — Future Work

IV. Conclusions

Leveraging an XGBoost model with text message reply latency features, we were able to predict a binary PHQ-9 score with an F1 score of 0.67, AUC of 0.72, and Accuracy of 0.69. While these values are lower than some of those achieved with tweet content [6], we consider text message — Best Results

reply latency to be a promising modality that warrants further research. Texting popularity makes it a valuable modality to screen for depression. Reply latency features have the advantage over message content as no private information is required, increasing the number of individuals who would be — Justification

willing to share this data. The main limitation of this research is the limited number of participants. As reply latency — Limitation

features contain valuable information, future work involves collecting more data and combining latency features with other features extracted from various Smartphone modalities. — Future Work

IV. Conclusion

We performed detailed experimentation to determine the ability of two weeks of text and call logs to screen for — Goal

depression. We 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. We — Best Results

compared time series of communication count, communication average length, and unique contacts aggregated every 4, 6, — Approach

12, and 24 hours. These results provides valuable insight on the usefulness of logs for mobile depression detection. Future — Contribution

work involves integrating the predictions of text and call logs with the predictions of other passively collected modalities in a multi-modal model to improve mental illness screening. — Future Work

V. Conclusion

Our research demonstrates that text messages are a promising modality to passively screen for suicidal ideation. With — Conclusion

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. — Best Results

Further, this research could guide the future development of successful screening technologies. — Contribution

Examples of Discussion with Conclusion

V. Discussion, Limitations, & Future Work

The results support our hypothesis and previous research [13], [14] that ensembles of BERT variants perform better than individual models for depression classification. While the mean $F1$ scores of our ensembles (BLA) were only slightly higher than those of the best performing individual models, our ensembles were notably more robust, as evidenced by their small standard deviations. As suggested by related literature [12], the base RoBERTa model performed decently for the mental health datasets. RoBERTa did not benefit from fine tuning so BERT (BLA) surpassed it's performance. Yet, RoBERTa proved to be a critical component of the successful ensembles (BLA).

AudiBERT [8] improved depression detection ability by combining BERT with an audio model. We discovered that ensembling BERT models can also improve classification. Thus, integrating multiple text models with multiple audio models could further enhance the performance of AudiBERT.

While DAIC-WOZ [15] is the best available dataset for this research, the small number of participants limits model performance. Our ensemble models were quite stable on the small datasets but collecting larger datasets will help further establish the robustness of our ensembles in this domain. While we used transcripts, our ensembles are applicable to any text data such as social media posts.

VI. Conclusion

Our research demonstrates that ensembling text classification models indeed improves performance for depression screening. In particular, we recommend ensembling AudiBERT and DistilBERT with One Final Classifier strategy for this task. While not always advantageous for individual models, fine-tuning with additional SUM and attention layers before classification benefited the ensembles. This research could guide future development of successful depression classification models for healthcare application ecosystems.

Summary of Results

Compare to Literature

Limitation and Future Work

Goal

Approach

Main Result

Contribution

A Story Ends with Acknowledgments

(and references if a paper)

Acknowledgments After Conclusion

May need to mention funding

ACKNOWLEDGMENT

We thank Dogrucu, Perucic, Isaro, Ball, Toto, and Prof Agu for innovating the retrospective collection approach and collected the Moodable dataset. We thank prior Emutivo teams and the DAISY research community at WPI for their support.

ACKNOWLEDGMENT

We thank Prof. Agu, Resom, Dogrucu, Peruic, Isaro, and Damon for innovating the Moodable approach and for sharing the resulting Moodable dataset.

ACKNOWLEDGMENT

We thank Caouette, Kayastha, Bruneau, Hartvigsen, Kakar, Toto, Flores, prior Emutivo teams, and the DSRG lab at WPI.

Always 'we'

ACKNOWLEDGMENT

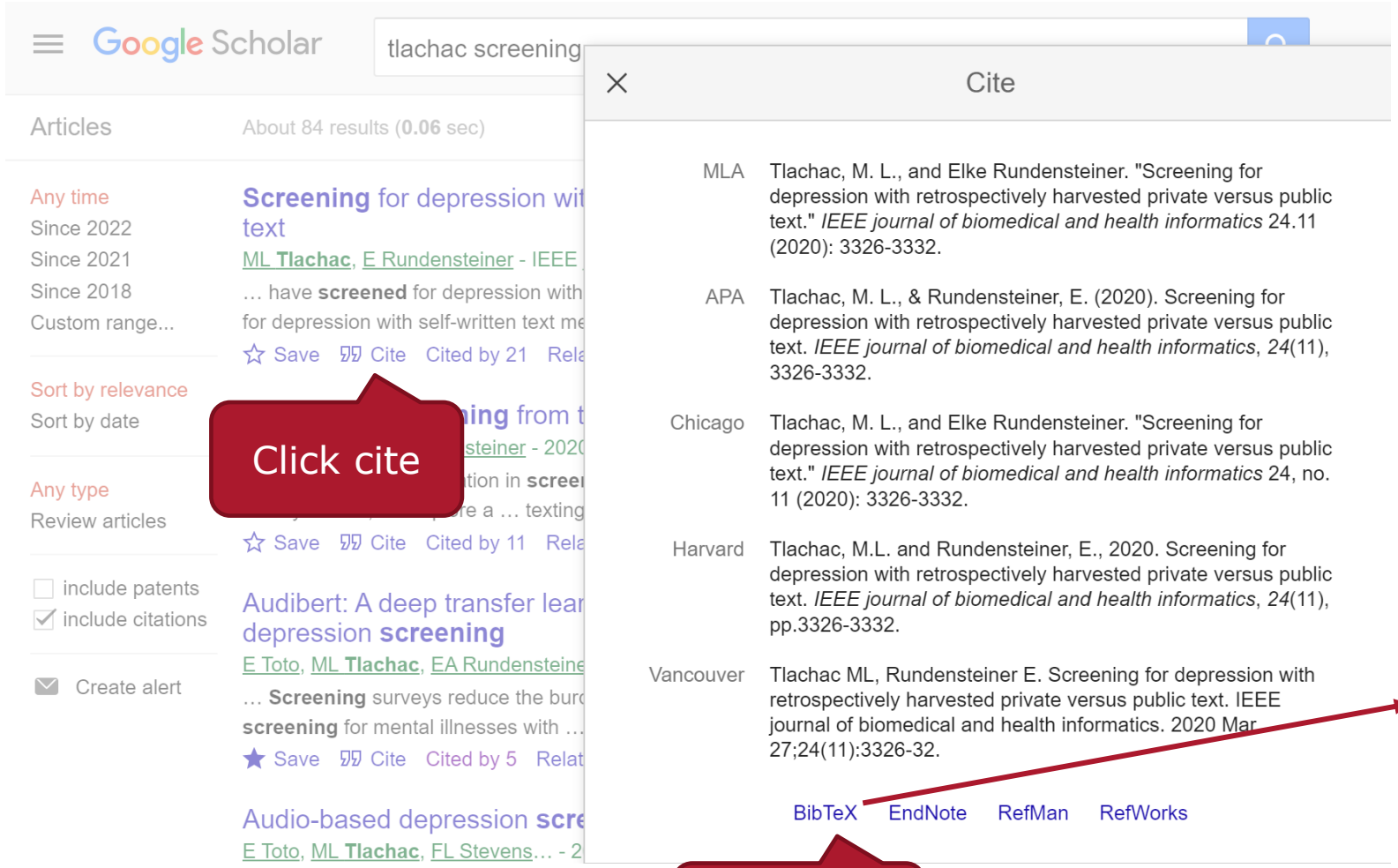
We thank Ermal Toto, Samuel S. Ogden, Marissa Bennett, and the DSRG community at WPI for their support. We thank Gao, Flannery, Resom, Assan, and Wu for collecting the EMU dataset. We thank Prof. Agu, Dogrucu, Peruic, Isaro, and Ball for innovating the Moodable approach and collecting the Moodable dataset.

ACKNOWLEDGMENT

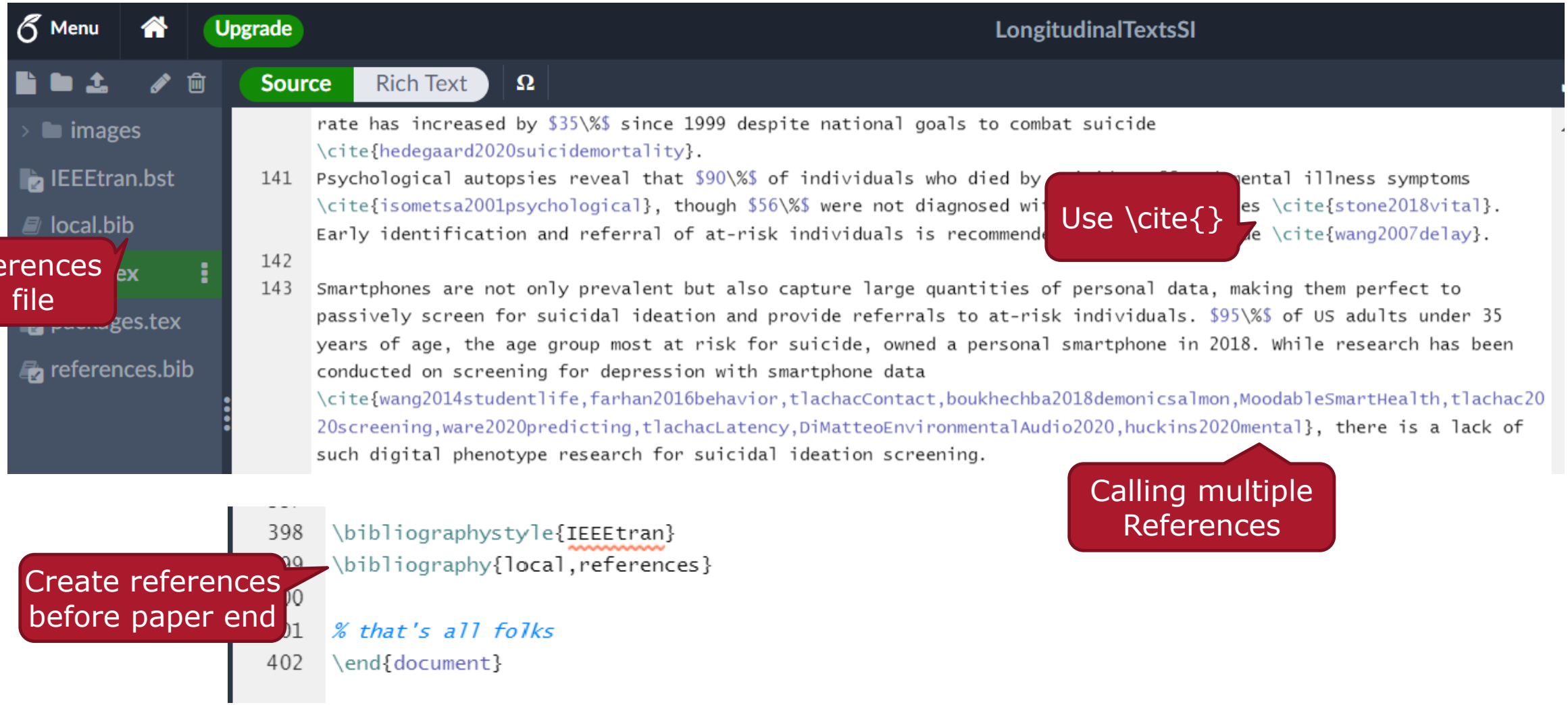
We thank Ermal Toto, Souyma Joshi, and the DAISY lab at WPI. We thank Claus Horn at ZHAW.

No 'also'

Google Scholar to Find Bibliography Entries



References in Overleaf



Half to Full Column of References

REFERENCES

[1] World Health Organization, "Depression and other common mental disorders: global health estimates," Tech. Rep., 2017. [Online]. Available: https://apps.who.int/iris/handle/10665/254610

[2] A. Halfin, "Depression: the benefits of early and appropriate treatment," *American Journal of Managed Care*, vol. 13, no. 4, 2007.

[3] R. M. Epstein *et al.*, ""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(9), 2010.

[4] N. Cummins *et al.*, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, 2015.

[5] M. Valstar *et al.*, "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.

[6] M. Rodrigues Makiuchi *et al.*, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 55–63.

[7] E. Toto *et al.*, "Audio-based depression screening using sliding window sub-clip pooling," in *IEEE ICMLA*, 2020.

[8] 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.

[9] R. Flores *et al.*, "Depression screening using deep learning on follow-up questions in clinical interviews," in *IEEE ICMLA*, 2021.

[10] M. L. Tlachac *et al.*, "Emu: Early mental health uncovering framework and dataset," in *IEEE ICMLA Special Session Machine Learning in Health*, 2021.

[11] J. Devlin *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, 2019.

[12] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Detection and Classification of mental illnesses on social media using RoBERTa," *arXiv*, 2020.

[13] H. Dang *et al.*, "Ensemble BERT for Classifying Medication-mentioning Tweets," in *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, 2020.

[14] S. Ghosh and A. Chopra, "Using Transformer based Ensemble Learning to classify Scientific Articles," *arXiv*, 2021.

[15] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews." in *Proceedings of the 9th International Conference on Language Resources and Evaluation*. CiteSeer, 2014.

[16] D. DeVault *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.

[17] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9," *Journal of General Internal Medicine*, vol. 16, no. 9, 2001.

[18] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv*, 2019.

[19] V. Sanh *et al.*, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv*, 2020.

Extra Tips

For papers

Extra Paper Tips

- Find the formatting instructions through venue website
- Find three example papers from venue past years
- Make the writing accessible and formal
- Write more than four pages and then reduce to four page
- Add tables, plots, and bullet points to improve readability
- Make sure plots are greyscale and colorblind friendly