Automated Construction of Lexicons to Improve Depression Screening with Text Messages

ML Tlachac, Avantika Shrestha, Mahum Shah, Benjamin Litterer, and Elke A. Rundensteiner

***Abstract*—Given that depression is one of the most prevalent mental illnesses, developing effective and unobtrusive diagnosis tools is of great importance. Recent work that screens for depression with text messages leverage models relying on lexical category features. Given the colloquial nature of text messages, the performance of these models may be limited by formal lexicons. We thus propose a strategy to automatically construct alternative lexicons that contain more relevant and colloquial terms. Specifically, we generate 36 lexicons from fiction, forum, and news corpuses. These lexicons are then used to extract lexical category features from the text messages. We utilize machine learning models to compare the depression screening capabilities of these lexical category features. Out of our 36 constructed lexicons, 14 achieved statistically significantly higher average F1 scores over the pre-existing formal lexicon and basic bag-of-words approach. In comparison to the pre-existing lexicon, our best performing lexicon increased the average F1 scores by 10%. We thus confirm our hypothesis that less formal lexicons can improve the performance of classification models that screen for depression with text messages. By providing our automatically constructed lexicons, we aid future machine learning research that leverages less formal text.**

***Index Terms*—mobile health, digital phenotype, supervised machine learning, natural language processing, text classification**

I. Introduction

Depression is a serious mental illness directly affecting around 322 million people globally [1]. Among the top three leading causes of global disability in 2017 [2], [1], depression is costly to individuals and the global economy [3]. Further, depression can lead to suicide [1] which is among the top 20 causes of death in the world [1] and the top 10 causes of death in the US [4]. In particular, untreated and undertreated mental illnesses are present in the large majority of suicides according to psychological autopsies [5]. Further, such mental illnesses increase the likelihood of developing physical health conditions that shorten life expectancy [6].

Fortunately, early depression treatment has proven to be very effective [7]. As such, universal depression screening is recommended by the U.S. Preventative Services Task Force (USPSTF) [8]. However, detecting depression can be challenging as symptoms are not always recognized [9] and remain stigmatized [10]. While depression screening surveys are being increasingly incorporated into routine health care visits [11], these screening surveys require active participation and are thus subject to conscious and unconscious bias in healthcare settings where there may be concern about medical repercussions. To address this problem, researchers have begun to explore using mobile sensor [12], [13], [14], [15], [16] and social media [17], [18] data to passively screen for depression.

Text messages are a particularly promising modality that retain the ubiquitous nature of mobile sensor data and the social interactions captured by social media data. Recent research has extracted features from SMS text messages [19], [20], [21], [22], [16] to screen for depression with machine learning models. Most notably, it was determined that text message content was more effective at screening for depression than tweet content [21]. Lexical category frequency features were proved to be more important when screening for depression and suicidal ideation with text messages than part-of-speech tag frequency, sentiment, and volume features [21], [23]. Yet, the performance of the models were likely limited by the pre-existing lexicons. Empath's emotional and topical lexical categories [24] were used to extract features from the text messages because they contain more modern terms than the proprietary Linguistic Inquiry and Word Count (LIWC) lexical categories [25]. However, these terms are still more formal than would be expected in texts [26]. For example, text messages often contain abbreviations, misspellings, and slang.

In this research we propose an approach to automatically construct alternative lexicons that contain more colloquial terms. In particular, we derive seed words from the 194 existing Empath lexical categories [24] that can be used with Empath's software [27] to identify related words in large text corpuses. Each of the 194 groups of related words form a new lexical category for an alternative lexicon. 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 related words from the three corpuses to create another 9 more robust lexicons.

The purpose of constructing lexicons is to improve the ability of text messages to screen for depression. Therefore, we use our 36 lexicons to extract lexical category frequency features from texts with depression screening labels. We then leverage an assortment of fundamental machine learning methods to screen for depression with these feature sets. To assess the usefulness of our constructed lexicons, we compare their screening results to bag-of-words features and pre-existing Empath lexical features. This study thus serves to investigate the importance of lexicon construction when screening for depression with informal text. Our research contributions include:

1) Strategies for identifying seed words from existing lexical

Tlachac is with Bryant University, Smithfield, RI 02911, USA, email mltlachac@bryant.edu. Research completed while a student at WPI.

Shrestha and Rundensteiner are with Worcester Polytechnic Institute (WPI), Worcester, MA 01609 USA, emails: {ashrestha4, rundenst}@wpi.edu.

Shah is with the University of Rhode Island, Kingston, RI 02881 USA, email: mahum_shah@uri.edu. Research completed at WPI for NSF REU.

Litterer is with the University of Michigan, Ann Arbor, MI 48109 USA, email: blitt@umich.edu. Research completed at WPI for NSF REU.

categories to automatically construct categories,
2) 36 lexicons that were automatically constructed from three linguistically distinct corpuses,
3) A comparison of the usefulness of the constructed lexicons to screen for depression with text messages, and
4) Discussion of important features for depression screening.

II. Related Literature

Lexicons contain groupings of similar words that are useful in the analysis of text. LIWC is a popular lexicon for depression detection research [28], [29], [30], [31], though the lexicon was cited as a limitation [31], motivating future research [21] to use the more modern Empath lexicon [24]. There have been a few prior attempts to update and expand lexicons through corpus-based [32], [33] and thesarus-based [34], [35] approaches. In particular, Losada et al. [33] evaluated depression lexica to detect depression with social media posts. However, this research used subsets of more formal lexicons so only a limited number of lexical categories were assessed. Further, Tlachac et al. [21] deduced that different lexical categories were important in the screening of twitter posts and text messages.

Current natural language processing tools are unfortunately inappropriate for text messages. The informal language frequently present in text messages and their short nature poses issues for existing text-based systems which expect longer passages of formal text [26]. For example, the Empath [24] and LIWC [25] lexicons only include correctly spelled words which is not a realistic expectation for text messages. In addition, many existing natural language processing tools are trained on third person narratives [24], [36] instead of first person narratives, which likely further negatively impacts their performance on text messages. For instance, the pre-existing Empath [24] lexical categories were derived from amateur modern fiction and Bidirectional Encoder Representations from Transformers (BERT) [36] was pre-trained on Wikipedia entries.

Prior research has tried to make natural language processing tools more applicable for informal texts. For example, Ek et al. [26] developed a classifier to extract named entities from SMS text messages written in Swedish. Likewise, BERTweet [37] is a language model pre-trained on a corpus of English tweets that can perform named entity recognition, part-of-speech tagging, and text classification. Additionally, knowledge graphs can be manually constructed to aid the encoding of text data for deep learning models, as was done by Xu et al. to detect suicide risk [38]. Despite promising advances in deep learning, machine learning is often more appropriate for many psychopathology tasks [39]. In part, this is due to the small number of participants common in this domain. Further, interpretability is vital for diagnostic applications [39]. Thus, it remains important to improve feature engineering for text classification.

Social media posts are the most common form of text used to screen for mental illness with a recent survey paper [18] highlighting 75 studies published between 2013 and 2018. This overview [18] revealed that Twitter was the most popular platform, English was the most common language, and depression the most frequently modeled mental illness. Reddit posts are also popular in the related tasks of early risk prediction for depression and self-harm [40], [41], [42], [43]. However, private messages are more utilized and likely predictive than public messages. For instance, Tlachac et al. [21] compared the depression screening ability of two weeks of text messages and tweets with 110 and 89 participants, respectively. Logistic regression classifiers achieved F1 scores that were 0.13 higher with the text messages than the tweets.

Transcripts from various sources have also been used in depression screening. On 187 clinical interviews labeled with depression screening scores [44], [45], BERT models [46], [47], [48] achieved the most depression screening success with the transcripts by treating each clinical interview question as a separate dataset. Another study [49] inferred depression from recordings of 148 adolescents in simulated interactions. Each word in the transcripts was tagged with valence and arousal ratings [50] to create features. Three other studies [51], [52], [53] extracted Empath lexical category features to screen for depression with transcripts of mobile voice recordings from 43 adults and 110 students. While both machine learning and deep learning models struggled to classify these transcripts, they performed better on typed responses from 298 students [52]. As they are collected passively, text messages are more challenging to model than typed responses to a given prompt.

III. Data

For this research we leverage the retrospectively harvested text message logs in the Moodable (Mood Assessment Capable Framework) [15] and EMU (Early Mental Health Uncovering) [51] datasets. Collected in 2017-2019 from Amazon Mechanical Turk (mTurk) workers [54], these logs are labeled with Patient Health Questionnaire-9 (PHQ-9) scores. A popular depression screening survey, the PHQ-9 [8] asks participants to reflect on the frequency of nine depression symptoms during the last two weeks with Likert scales ranging from 0 to 3 [55]. Participants with PHQ-9 ≥ 10 screen positive for moderate depression.

For our analysis, we consider the 88 participants in the combined Moodable and EMU datasets who sent at least 5 texts within the two weeks before completing the PHQ-9. In addition to the PHQ-9 [55] asking about the last two weeks, prior research [21] determined this was a sufficient temporal quantity of texts for screening. The 88 participants shared a total of 7914 sent texts during the prior two weeks. As depicted

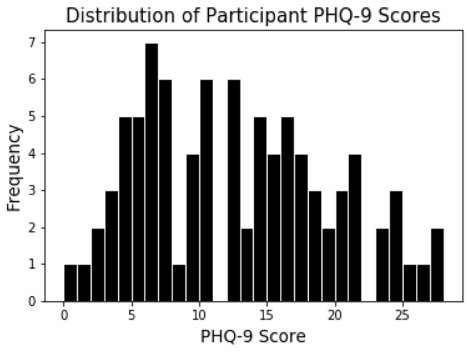

Fig. 1: The participant PHQ-9 scores. PHQ-9 scores range from 0 to 27 with 10 being the cutoff for moderate depression.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3203345

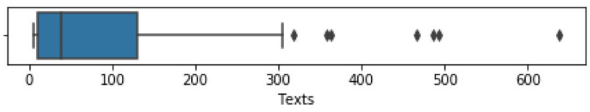

Fig. 2: Distribution of number of texts from each participant.

in Fig. 2, the median number of texts is 39.5. No demographics are available for the 78 Moodable participants [15].

The distribution of participant PHQ-9 scores are displayed in Fig. 1. Overall, 53 (60%) of the participants screened positive for moderate depression. While this percent is notably higher than in the general population, this phenomenon has also been encountered by other research studies with crowd-sourced workers. In particular, it has been hypothesized [56] that crowd-sourced workers are more likely to be depressed than the general population and self-selection bias makes studies about mental health more appealing to depressed workers.

IV. 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 [24], [27] both provides 194 pre-existing lexical categories as well as the capability of generating new 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 new lexicons. We then evaluate the usefulness of these new lexicons by extracting lexical features from text messages and screening for participant depression with machine learning models.

A. 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 [24].

We derive the seed words from the 194 pre-existing Empath categories. Each category contains a list of words. For example, the 98 words that comprise the category *vacation* are displayed in Table I. While the majority of the words are unigrams, the category *negative emotion* contains four bigrams and one trigram. We extract 1, 3, and 5 seed words from the words of each category to construct 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 strategies to identify seed words:

- Closest (c): we select the words closest to the category name in the category word list [1]. For the category *vacation*, the five closest words are 'destination', 'visit', 'vacation', 'accommodation', and 'tropical'.
- 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'.

[1]There are 18 categories where the category name is not in the word list. For 10, a portion and/or root of the category name is present in the word list. For the remaining 8, we manually selected a replacement word from the word list: *domestic_work* → 'housework', *social_media* → 'forum', *blue_collar_job* → 'worker', *air_travel* → 'airplane', *size_and_shape* → 'sized', *white_collar_job* → 'profession', *negative_emotion* → 'depressed', *positive_emotion* → 'joyful'.

B. Generating New Lexicons

The Empath software [27] 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 linguistically different so we generate lexical categories from each to compare their usefulness for depression screening with text messages.

We set the size of the new 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 generated lexicons contain fewer words. To ensure there are no empty categories, we add the seed words to the list of most related words. For each lexicon, we compare the number of words in each of their 194 categories in Fig. 3.

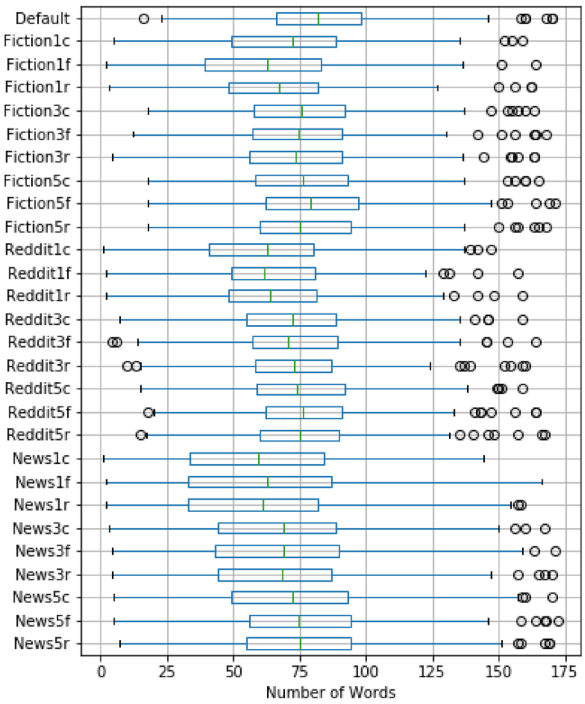

Fig. 3: The number of words in the categories for each lexicon.

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.

C. Extracting Features For Machine Learning

Since we screen for participant depression, we aggregate the text messages for each participant. We then extract features

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 I: The 98 words in the Empath's pre-existing *vacation* category [24].

from these 88 self-written text passages. Note, we remove capitalization and punctuation prior to feature extraction.

Constructed lexicons. For each of the 194 categories in our constructed 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 [24]. We extract lexical category features in this manner 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, we use the Empath software [24] to analyze the text passages. This results in a set of lexical category features for the pre-existing Empath lexicon, which we further refer to as the Default lexicon.

Bag-of-words. Since the most basic strategy to create features from text data is bag-of-words, we also include this strategy as a baseline. In this approach, the number of features is the number of unique words in the dataset. The number of times a participant uses a given word is tallied and that count is the value for that feature. Since the lexical categories do not capture numerical characters, we remove all words that begin with numbers. The participants collectively texted $6,248$ unique words so the resulting feature matrix is very sparse.

Example. Participant m9751's five texts contain 24 words; they aggregate to the passage "wait i will call just 5 mins ur aunt here for pray god wait ya they are searching are u free please call call". Table II displays the lexical categories that contain words that match words in the passage. The categories are different for the three lexicons, demonstrating how the seed word identification strategies result in different feature sets. While the word 'aunt' is in the *family* category for all three lexicons, it is only in the *children* category for Fiction5f. Likewise, 'free' is only in the *prison* category for Fiction5f.

D. Feature Selection and Machine Learning Methodology

The goal of the machine learning models is to screen for moderate depression (PHQ-9$\geq$ 10). We perform leave-group-out cross validation [57] 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.

TABLE II: Comparison of the categories captured by participant m9751's texts for three of the constructed fiction lexicons.

Lexicon	Categories
Fiction5c	family, divine, healing, speaking, listen, phone
Fiction5f	family, prison, communication, phone, messaging, children, giving
Fiction5r	family, divine, listen, phone, strength, messaging

Chi-Squared Feature Selection and Upsampling. Feature selection is necessary to reduce the dimensionality of the data and prevent model overfitting. For each of the crafted feature sets, which contain only lexical or bag-of-word features, we experiment with using between 1 and 10 of the features for the depression screening models. Based on related work [21], we perform chi-squared feature selection [57] to determine which features to include in the models. The selected features are those that have the highest chi-squared values in relation to the depression labels in the training set. To prevent overfitting, this 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. Note, no balancing is performed on the test sets.

Machine Learning Methods. We compare the screening ability of the different feature sets with support vector classifiers (SVC), logistic regression (LR), Gaussian Naive Bayes (NB), and k-Nearest neighbor (kNN) [57]. SVC is a versatile method known for being effective at classifying smaller datasets. We experiment four different kernel functions which we refer to as Gaussian SVC, linear SVC, polynomial SVC, and sigmoid SVC. Meanwhile, logistic regression is efficient and Naive Bayes is skilled at document classification [57]. Lastly, the nonparametric k-Nearest Neighbor (kNN) method is robust to noise; we experiment with using kNN with 3 and 5 neighbors, which we further refer to as kNN3 and kNN5, respectively.

Evaluation metrics. Sensitivity and specificity are common metrics used to evaluate diagnostic models. Sensitivity refers to the ratio of depressed participants correctly identified as depressed while specificity refers to the ratio of not depressed participants correctly identified as not depressed. As these are inversely related metrics, we consider the best performing models to be those that maximize the average F1 score across all 100 test sets. F1 places greater value on true positive predictions than accuracy. In addition to F1, we report on the sensitivity, specificity, and accuracy of the best model configurations.

Statistical tests. As we repeat each model configuration 100 times, we obtain 100 F1 scores. We can then use statistical tests to determine if the F1 scores are statistically significantly different between different model configurations and the baselines. We use t-tests and one-way ANOVA tests [58] to compare with one and both baselines, respectively.

E. Software and Lexicon Availability

Upon publication, we will release our code, seed words, lexicons, and feature sets at https://github.com/mltlachac/LexicalCategories. For updates about our research, visit our project website: https://emutivo.wpi.edu.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3203345

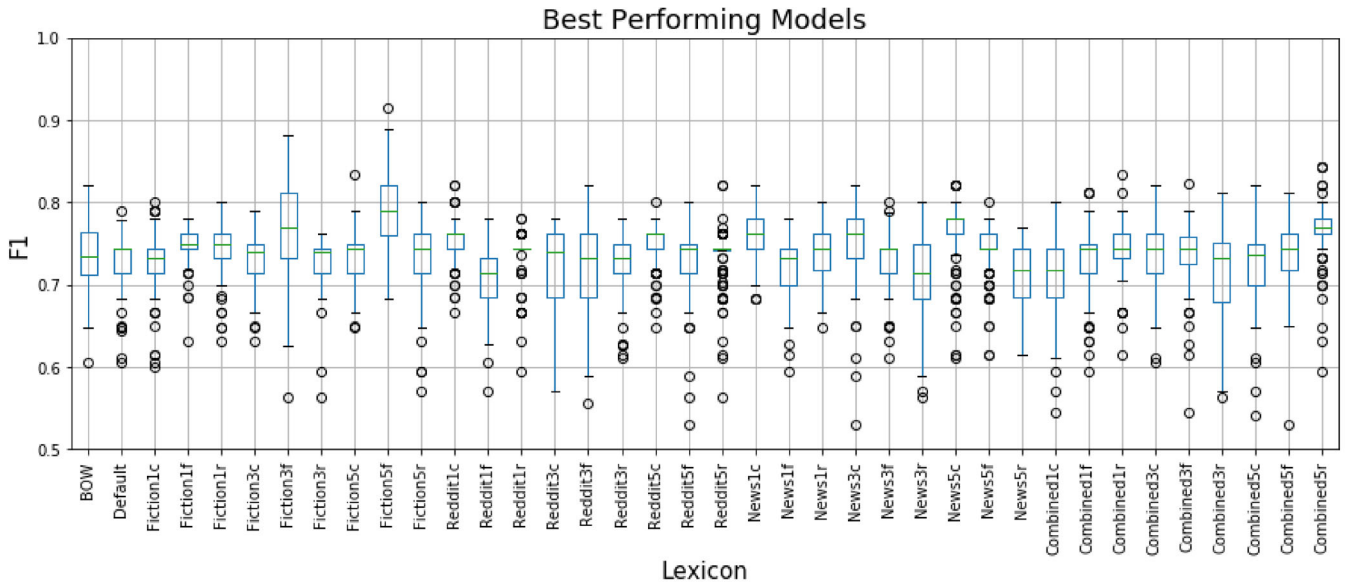

Fig. 4: The results of the model configurations with the the highest average F1 scores for each lexicon. Bag-of-words (BOW) and Reddit 5c used 3 chi-squared selected features while the other lexicons used only 1 chi-squared selected feature. The lexical categories were generated with seed words that were the 1 to 5 closest (c), first (f), and random (r) words to the category names.

V. Results

Fig. 4 and Table III 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 comparison our baselines only achieved average F1 scores of 0.74 and 0.72. The t-tests and the ANOVA 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.

A. Performance of 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 sparser feature matrix, bag-of-words had a higher average F1 than the default lexicon; 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.

10 of the 36 constructed lexicons had F1 scores that were statistically significantly better than both of the baselines based on ANOVA tests. An additional 4 constructed lexicons performed statistically significantly better than the Default lexicon based on t-tests; Fiction1f, Reddit1r, News1r, and Combined1r 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 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.

B. Performance of Constructed Lexicons

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

From the results in Table III and Fig. 4, we discern patterns from the performance 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 seed word identification strategy perform comparatively well with just one feature.

We expected the combined lexicons to retain linguistic benefits for all three corpuses. However, only one combined lexicon was among the six best models, namely Combined5r. With an F1 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.

C. The Six Most Successful Lexicons

There were six constructed lexicons that performed notably better than the other lexicons with only one feature. These

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3203345

TABLE III: The average ± standard deviation of the metrics for the models configurations with the highest average F1 scores for each lexicon. The p-values are derived from the t-tests and ANOVA tests. The t-tests compare the F1 scores from each lexicon with scores from the bag-of-words (B) and default Empath lexicon (D) individually. The ANOVA tests compare the F1 scores from each lexicon with scores from both of the baselines. The models that are statistically significantly better than the baselines are marked with asterisks. The six most successful models are bolded. Fiction5f had the highest average F1 score.

Lexicon	Method	F	F1	Sensitivity	Specificity	Accuracy	t-test(B)	t-test(D)	ANOVA
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	0.206
Fiction1f	Polynomial SVC	1	0.75±0.02	0.96±0.04	0.11±0.08	0.62±0.03	0.019*	0.003*	0.001*
Fiction1r	Polynomial SVC	1	0.74±0.03	0.95±0.06	0.14±0.11	0.62±0.04	0.158	0.015*	0.007*
Fiction3c	Polynomial SVC	1	0.73±0.03	0.92±0.07	0.14±0.12	0.60±0.04	0.330	0.286	0.137
Fiction3f	**LR**	**1**	**0.77±0.06**	**0.88±0.09**	**0.41±0.21**	**0.69±0.09**	**<0.001***	**<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	0.183
Fiction5c	Polynomial SVC	1	0.73±0.03	0.93±0.07	0.12±0.11	0.60±0.05	0.492	0.226	0.118
Fiction5f	**LR**	**1**	**0.79±0.05**	**0.93±0.06**	**0.40±0.18**	**0.71±0.07**	**<0.001***	**<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	0.352
Reddit1c	**NB**	**1**	**0.76±0.04**	**0.96±0.07**	**0.15±0.10**	**0.63±0.06**	**0.002***	**<0.001***	**<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	0.424
Reddit1r	Polynomial SVC	1	0.74±0.03	0.97±0.07	0.05±0.07	0.60±0.04	0.361	0.031*	0.016*
Reddit3c	Linear SVC	1	0.74±0.03	0.96±0.06	0.07±0.08	0.60±0.04	0.984	0.100	0.110
Reddit3f	kNN5	1	0.72±0.05	0.90±0.10	0.15±0.10	0.59±0.06	0.023	0.922	0.023*
Reddit3r	Linear SVC	1	0.73±0.03	0.93±0.08	0.10±0.11	0.59±0.04	0.171	0.424	0.199
Reddit5c	Polynomial SVC	1	0.75±0.03	0.97±0.06	0.10±0.09	0.62±0.04	0.019*	0.003*	0.001*
Reddit5f	Polynomial SVC	1	0.74±0.05	0.95±0.10	0.10±0.12	0.60±0.05	0.990	0.124	0.076
Reddit5r	Sigmoid SVC	3	0.76±0.06	0.88±0.10	0.35±0.15	0.66±0.08	0.010*	0.001*	<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***	**<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	0.107
News1r	kNN3	1	0.75±0.04	0.95±0.07	0.13±0.08	0.61±0.05	0.234	0.021*	<0.001*
News3c	kNN3	1	0.75±0.05	0.96±0.08	0.13±0.10	0.62±0.06	0.036*	0.004*	<0.001*
News3f	Polynomial SVC	1	0.73±0.04	0.94±0.08	0.07±0.11	0.59±0.05	0.138	0.483	0.191
News3r	Polynomial SVC	1	0.72±0.03	0.92±0.08	0.07±0.08	0.58±0.04	0.001	0.706	0.121
News5c	**NB**	**1**	**0.76±0.04**	**0.97±0.06**	**0.16±0.08**	**0.64±0.06**	**<0.001***	**<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	0.115
News5r	Linear SVC	1	0.72±0.03	0.91±0.07	0.08±0.10	0.57±0.04	<0.001*	0.487	0.031*
Combined1c	Linear SVC	1	0.73±0.04	0.93±0.08	0.12±0.10	0.60±0.05	0.587	0.207	0.173
Combined1f	Polynomial SVC	1	0.72±0.08	0.90±0.12	0.16±0.14	0.60±0.07	0.132	0.934	0.180
Combined1r	kNN3	1	0.75±0.05	0.88±0.10	0.33±0.16	0.65±0.07	0.066	0.007*	<0.001*
Combined3c	Gaussian SVC	1	0.74±0.05	0.89±0.09	0.26±0.17	0.63±0.06	0.725	0.080	<0.001*
Combined3f	Polynomial SVC	1	0.73±0.04	0.92±0.08	0.15±0.13	0.61±0.05	0.709	0.176	0.100
Combined3r	Linear SVC	1	0.72±0.04	0.93±0.09	0.07±0.09	0.58±0.05	0.026	0.911	0.155
Combined5c	Linear SVC	1	0.73±0.03	0.94±0.07	0.11±0.09	0.60±0.05	0.735	0.158	0.148
Combined5f	LR	1	0.73±0.09	0.85±0.14	0.30±0.18	0.63±0.08	0.302	0.712	0.039*
Combined5r	**LR**	**1**	**0.77±0.06**	**0.92±0.09**	**0.34±0.14**	**0.68±0.08**	**<0.001***	**<0.001***	**<0.001***

TABLE IV: Comparison of the average F1 scores for each method trained on features from the most successful lexicons in Table III. The p-values are derived from 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	t-test(B)	t-test(D)	F1	t-test(B)	t-test(D)	F1	t-test(B)	t-test(D)	F1	t-test(B)	t-test(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	t-test(B)	t-test(D)	F1	t-test(B)	t-test(D)	F1	t-test(B)	t-test(D)	F1	t-test(B)	t-test(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*

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 IV displays the performance of each method for these six best 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 exceptions 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 III due to higher standard deviation, the average F1 scores notably match that of the best model as displayed in Table IV. This makes sense as Naive Bayes algorithms are known to succeed at document classification [57]. However, the impact of the method pales in comparison to the impact of the corpus, number of seed words, and seed word identification strategy.

D. Important Features

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 I), the words in the Fiction5f category notably contain names of cities and states.

VI. Discussion

A. Research Implications

As smartphones are ubiquitous, data passively collected from smartphones could greatly increase the rates of screening. In particular, sent text messages are a valuable screening modality as they contain personal information that provide insights into the mental wellbeing of the sender. We improved the mental illness screening capabilities of texts by introducing lexicons that are better able to classify such informal communications. Improving the average F1 by 10%, our Fiction5f lexicon was especially successful for depression screening.

Text messages could be used individually or in conjunction with other smartphone modalities to screen for mental illnesses with machine learning models. These passive screening models could replace active screening surveys in clinical settings, thus reducing the burden on the patients. Alternatively, these passive screening models could be incorporated into smartphones to alert users to changes in their own mental wellbeing.

B. Machine and Deep Learning: Limitations and Future Work

Machine learning is often preferred over deep learning in psychopathology due to effectiveness with limited participants and need for interpretability [39]. Transfer learning models like BERT [36] aim for deep learning to be applicable to smaller datasets and still achieve success for text classification. This was demonstrated by prior research [46], [47], [48] that used BERT to screen for depression with clinical interview transcripts. Yet, BERT classifiers have performed relatively poorly with mobile transcripts [52], [53]. Containing over 100 participants, these datasets are still larger than our dataset.

The primary limitation of this research is thus the small number of participants, which restricted our ability to run deep learning models and compare machine learning models against them. Further, our analysis is limited by the lack of participant sociodemographic attributes. As such, we recommend future work collect a larger dataset of labeled text messages in order to utilize deep transfer learning models like BERTweet [37].

C. Lexicons: Limitations and Future Work

Our feature engineering approach did not consider the ordering of the text messages nor the ordering of the words within the messages, both of which may contain important information for depression screening. The lexicons also do not address the issue of homonyms. Further, since negation semantics are not captured, this potentially renders the emotional lexical categories less useful. While our constructed lexicons contain many bigrams, future work could explore alternative approaches that focus more on context.

In this research, we focused on text messages as they are known to be difficult to model with existing natural language processing tools [26]. However, our constructed lexicons could be used to improve classification of other informal text datasets within and outside of the mental health domain. Our methodology could also be further broadened by considering a different default lexicon, different seed word identification strategies, and different text corpuses. Further, future work could generate customized categories for specific domains and types of text. For instance, text messages may benefit from a category that captures emoticons or numerical terms.

VII. Conclusion

In this research, we constructed 36 alternative lexicons to study their impact on improving the depression screening capabilities of text messages. Our best lexicon, Fiction5f, increased the average F1 score by 10% over the pre-existing lexicon. Overall, 14 of our constructed lexicons performed statistically significantly better in our machine learning models better than the pre-existing lexicon. We conclude that constructing alternative lexicons is useful for classifying less formal text. Therefore, this research provides important insights, especially for mental illness screening with direct messages.

Acknowledgment

We thank US Dep. of Ed. P200A180088, AFRI #1023720, REU site 1560229, and NSF III: Small #1910880. We thank Gerych, Thatigotla, Jurovich, Phillips, Garfinkel, Wu, Halley, O'Neill, Toto, prior Emutivo members, and DAISY lab at WPI.

REFERENCES

[1] World Health Organization, "Depression and other common mental disorders: global health estimates," WHO, Tech. Rep., 2017.

[2] S. James *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," *The Lancet*, vol. 392, pp. 1789–1858, 2018.

[3] D. E. Bloom, E. Cafiero *et al.*, "The global economic burden of noncommunicable diseases," Program on the Global Demography of Aging, Tech. Rep., 2012.

[4] H. Hedegaard and M. Warner, "Suicide mortality in the united states, 1999-2019," *NCHS Data Brief*, no. 398, 2021.

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

[6] J. Firth, N. Siddiqi *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.

[7] G. De Girolamo, J. Dagani *et al.*, "Age of onset of mental disorders and use of mental health services: needs, opportunities and obstacles," *Epidemiology and psychiatric sciences*, vol. 21, no. 1, pp. 47–57, 2012.

[8] A. L. Siu, K. Bibbins-Domingo *et al.*, "Screening for depression in adults: US preventive services task force recommendation statement," *JAMA*, vol. 315, no. 4, pp. 380–387, 2016.

[9] R. Epstein, P. Duberstein *et al.*, ""i didn't know what was wrong:" how people with undiagnosed depression recognize, name and explain their distress," *J. of gen. internal medicine*, vol. 25, no. 9, pp. 954–61, 2010.

[10] J. H. Wirth and G. V. Bodenhausen, "The role of gender in mental-illness stigma: A national experiment," *Psychological Science*, vol. 20, no. 2, pp. 169–173, 2009.

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

[12] R. Wang, F. Chen *et al.*, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *ACM Ubicomp*, 2014, pp. 3–14.

[13] S. Ware, C. Yue *et al.*, "Predicting depressive symptoms using smartphone data," *Smart Health*, vol. 15, pp. 1–16, 2020.

[14] M. Boukhechba, A. R. Daros *et al.*, "Demonicsalmon: Monitoring mental health and social interactions of college students using smartphones," *Smart Health*, vol. 9, pp. 192–203, 2018.

[15] A. Dogrucu, A. Perucic *et al.*, "Moodable: On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data," *Smart Health*, vol. 17, pp. 1–17, 2020.

[16] M. L. Tlachac, R. Flores *et al.*, "DepreST-CAT: Retrospective smartphone call and text logs collected during the covid-19 pandemic to screen for mental illnesses," *ACM IMWUT*, vol. 6, no. 2, 2022.

[17] S. C. Guntuku, D. B. Yaden *et al.*, "Detecting depression and mental illness on social media: an integrative review," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 2017.

[18] 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.

[19] M. L. Tlachac, E. Toto, and E. Rundensteiner, "You're making me depressed: Leveraging texts from contact subsets to predict depression," in *IEEE BHI*, 2019, pp. 1–4.

[20] M. Tlachac and E. A. Rundensteiner, "Depression screening from text message reply latency," in *IEEE EMBC*, 2020, pp. 5490–5493.

[21] 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, pp. 3326–3332, 2020.

[22] M. L. Tlachac, V. Melican *et al.*, "Mobile depression screening with time series of text logs and call logs," in *IEEE BHI*, 2021, pp. 1–4.

[23] M. L. Tlachac, K. Dixon-Gordon, and E. Rundensteiner, "Screening for suicidal ideation with text messages," in *IEEE BHI*, 2021, pp. 1–4.

[24] E. Fast, B. Chen, and M. Bernstein, "Empath: Understanding topic signals in large-scale text," in *ACM CHI*, 2016, pp. 4647–57.

[25] J. W. Pennebaker, R. J. Booth *et al.*, "Linguistic inquiry and word count: Liwc2015 operator's manual," 2015.

[26] T. Ek, C. Kirkegaard *et al.*, "Named entity recognition for short text messages," *Procedia Soc.*, vol. 27, pp. 178–187, 2011.

[27] E. Fast, B. Chen, and M. Bernstein, "Lexicons on demand: Neural word embeddings for large-scale text analysis." in *IJCAI*, 2017, pp. 4836–40.

[28] M. De Choudhury, M. Gamon *et al.*, "Predicting depression via social media," in *AAAI conf. on weblogs and social media*, 2013.

[29] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in *SIGCHI conf. on human factors in computing systems*, 2013, pp. 3267–3276.

[30] M. Park, C. Cha, and M. Cha, "Depressive moods of users portrayed in twitter," in *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD 2012*, 2012, pp. 1–8.

[31] A. Reece, A. Reagan *et al.*, "Forecasting the onset and course of mental illness with twitter data," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[32] L. Wang and R. Xia, "Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision," in *EMNLP*, 2017, pp. 502–510.

[33] D. E. Losada and P. Gamallo, "Evaluating and improving lexical resources for detecting signs of depression in text," *Language Resources and Evaluation*, vol. 54, no. 1, pp. 1–24, 2020.

[34] C. Fellbaum, "A semantic network of english: the mother of all wordnets," in *EuroWordNet*. Springer, 1998, pp. 137–148.

[35] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *LREC*, 2010, pp. 2200–2204.

[36] J. Devlin, M. Chang *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL*, pp. 4171–86, 2018.

[37] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," *EMNLP Demos*, pp. 9–14, 2020.

[38] Z. Xu, Y. Xu *et al.*, "Detecting suicide risk using knowledge-aware natural language processing and counseling service data," *Social Science & Medicine*, vol. 283, p. 114176, 2021.

[39] 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.

[40] R. Martınez-Castano, A. Htait *et al.*, "Early risk detection of self-harm and depression severity using bert-based transformers," in *CLEF*, 2020, pp. 1–16.

[41] H. Hosseinabad, E. Ersi, and A. Vahedian, "Detection of early sign of self-harm on reddit using multi-level machine," in *CLEF*, 2020, pp. 1–10.

[42] L. Achilles, M. Kisselew *et al.*, "Using surface and semantic features for detecting early signs of self-harm in social media postings," in *CLEF*, 2020, pp. 1–14.

[43] T. Basu and G. V. Gkoutos, "Exploring the performance of baseline text mining frameworks for early prediction of self harm over social media," in *CLEF*, 2021, pp. 928–937.

[44] J. Gratch, R. Artstein *et al.*, "The distress analysis interview corpus of human and computer interviews," in *LREC*, 2014, pp. 3123–3128.

[45] D. DeVault, R. Artstein *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *AAMAS*, 2014, pp. 1061–8.

[46] E. Toto, M. L. 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.

[47] R. Flores, M. L. Tlachac *et al.*, "Transfer learning for depression screening from follow-up clinical interview questions," *Deep Learning Applications*, vol. 4, 2022, in Press.

[48] S. Senn, M. L. Tlachac *et al.*, "Ensembles of bert for depression classification," in *44th EMBC*, 2022, in press.

[49] 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*. IEEE, 2014, pp. 1–5.

[50] 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.

[51] M. L. Tlachac, E. Toto *et al.*, "Emu: Early mental health uncovering framework and dataset," in *IEEE ICMLA: Health*, 2021, pp. 1311–8.

[52] M. L. Tlachac, R. Flores *et al.*, "Studentsadd: Mobile depression and suicidal ideation screening of college students during the coronavirus pandemic," *IMWUT*, vol. 6, no. 2.

[53] ——, "Early mental health uncovering with short scripted and unscripted voice recordings," *Deep Learning Applications*, vol. 4, 2022, in Press.

[54] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[55] 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.

[56] D. Di Matteo, K. Fotinos *et al.*, "The relationship between smartphone-recorded environmental audio and symptomatology of anxiety and depression: exploratory study," *JMIR Form. Res.*, vol. 4, no. 8, pp. 1–13, 2020.

[57] F. Pedregosa, G. Varoquaux *et al.*, "Scikit-learn: Machine learning in python," *J. of machine learning research*, vol. 12, pp. 2825–2830, 2011.

[58] P. Virtanen *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.