# Predicting Well-being Using Short Ecological Momentary Audio Recordings

Yu-Ning Huang
Carnegie Mellon University
Pittsburgh, United States
yuninghu@andrew.cmu.edu

Siyan Zhao
Carnegie Mellon University
Pittsburgh, United States
siyanz@andrew.cmu.edu

Michael L. Rivera
Carnegie Mellon University
Pittsburgh, United States
mlrivera@andrew.cmu.edu

Jason I. Hong
Carnegie Mellon University
Pittsburgh, United States
jasonh@cs.cmu.edu

Robert E. Kraut
Carnegie Mellon University
Pittsburgh, United States
kraut@andrew.cmu.edu

## ABSTRACT

To quickly and accurately measure psychological well-being has been a challenging task. Traditionally, this is done with self-report surveys, which can be time-consuming and burdensome. In this work, we demonstrate the use of short voice recordings on smartphones to automatically predict well-being. In a 5-day study, 35 participants used their smartphones to make short voice recordings of what they were doing throughout the day. Using these recordings, our model can predict the participants' well-being scores with a mean absolute error of 14%, relative to the self-reported well-being ("ground truth"). Both audio and text features from the recordings, especially, MFCC and semantic features, are important for prediction accuracy. Based on the work, we provide suggestions for future research to further improve the prediction result.

## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → **Neural networks**; **Kernel methods**.

## KEYWORDS

datasets, well-being, feature importance

**ACM Reference Format:**
Yu-Ning Huang, Siyan Zhao, Michael L. Rivera, Jason I. Hong, and Robert E. Kraut. 2021. Predicting Well-being Using Short Ecological Momentary Audio Recordings. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3411763.3451629

## 1 INTRODUCTION

Psychological well-being is a key component of one's overall health. To accurately and quickly measure well-being is fundamental to understanding and coping with its fluctuations as a response to changes in life circumstances (e.g., having children, changing jobs, etc.). Traditionally well-being is assessed using self-report surveys [7]. However, surveys impose a burden on users, especially in longitudinal settings where repetitive measures are used. In addition, self-reported surveys can be prone to subjective biases, rendering the measurements less reliable [28, 36].

As an alternative to surveys, researchers have examined how data on ambient surroundings (such as weather), mobile phone and wearable sensor [40], and text (e.g., blog posts [22] and tweets [23]) can be used to predict well-being. Other studies have also explored the possibility of using audio-based data to measure well-being. While these studies have shown success in quantifying well-being without relying on self-report surveys, the setup of these studies are not easily adaptable to everyday situations. For example, some studies require professional audio equipment [4, 14, 19] and others need both audio and video data [9, 11, 20]. Many studies use interview-based data [2, 38], which are difficult to collect in everyday settings and to scale up for large numbers of participants. While a few studies (e.g., [17]) captured ambient noises as an input to calculate social isolation and sleep patterns, which are predicative of well-being, this requires the microphone to be on throughout the day, which can pose computation, energy, and privacy constraints.

In this paper, we explore using smartphones as recording devices to collect short voice recordings for well-being prediction in people's daily settings. In contrast to existing work where professional devices are needed, the current approach uses built-in microphones in smartphones, ubiquitous to the majority of U.S. population [1]. In addition, these recordings are short and in-the-moment snippets of participants' environment, and hence are more natural than interview settings in labs. More importantly, in contrast to always-on audio capture that runs in the background, participants have control over when to make the recordings and the length of the recordings. By using both the audio and text features from the audio recordings, we demonstrate that these short recordings, taken from every-day environments, can provide value in predicting people's well-being.

## 2 DATA COLLECTION

The data in the current paper was collected as part of a larger 5-day study, conducted during the Fall of 2019 to examine the impact of social interaction on well-being. The study was approved by our university's Institutional Review Board. Below, we will describe in

*CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan*
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8095-9/21/05.
https://doi.org/10.1145/3411763.3451629

detail the procedures of the study, variables of our interest, and our analyses methods.

## 2.1 Procedure

We recruited 35 participants (66% female) using a local participant recruiting website. Participants ranged in age from 18 to 45, with 63% younger than 30. Because our delivery of the Ecological Momentary Assessment (EMA) surveys only worked on Android OS, all recruited participants were Android phone users. Twenty-two of the 35 participants were students, and the remaining 13 participants held either part-time or full-time jobs, e.g., social workers, lawyers, and nurses. All participants completed the full data collection.

Participants visited our lab prior to the study to grant informed consent. With the participants' permission, we installed a custom app that collected mobile phone data and delivered surveys to their phones. After completing the installation of the app, participants carried and used their phones as usual for the next 5 days and responded to the surveys when prompted. While participants did not start on the same date, all first initial sessions were scheduled between Wednesday to Friday so that the study spanned across the weekend for all participants. On the 6th day, participants returned to the lab for an end-of-study session where they received compensation based on their survey completion, i.e., $10 for days in which they completed 70% of the EMA surveys and the end-of-day well-being survey.

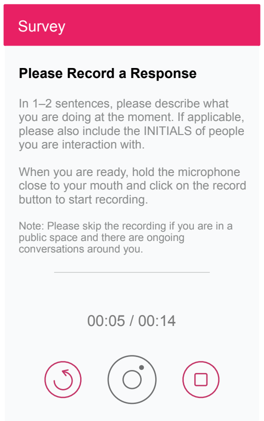
## 2.2 Ecological Momentary Assessments Surveys

During the study, participants received EMA surveys on their phones roughly every 30 minutes between 9:30 AM to 10:30 PM. No survey was delivered before or after so that participants would not be disturbed during their rest.

The EMA surveys were mostly about respondents' social interactions. In this paper we focus on the audio recording that was part of the EMA survey. When participants first opened the survey notifications, they were asked to record an brief audio response describing what they were doing at the moment (Fig. 1). To ensure data privacy, they were asked to use initials instead of personal identifiers when referring to specific individuals in the audio files. Then, the remainder of the text-based survey proceeded to ask about participants' most recent social interaction in the past 10 minutes. Social interactions were defined at the beginning of EMA as "a give-and-take exchange involving two or more people". If the participant had a social interaction, the survey then asked how close the participant felt with the partner involved in the interaction, which medium the interaction was carried out, and their current mood.

## 2.3 End-of-Day Surveys

At the end of each day, participants received a separate survey at 8 PM that assessed their well-being. Participants were asked to complete the survey before they went to bed. The questions relevant to the current paper include 1) a one-item loneliness scale from the Brief Inventory of Thriving [35]; 2) Patient Health Questionnaire-2 for depression [15]; and 3) a four-item version of the Perceived Stress Scale [7]. All these questions are well-established measures.


Survey

Please Record a Response

In 1–2 sentences, please describe what you are doing at the moment. If applicable, please also include the INITIALS of people you are interaction with.

When you are ready, hold the microphone close to your mouth and click on the record button to start recording.

Note: Please skip the recording if you are in a public space and there are ongoing conversations around you.

00:05 / 00:14

**Figure 1: Audio recording screen for the EMA Survey.**

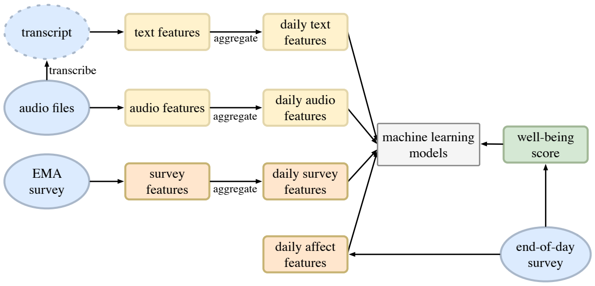


**Figure 2: An overview of the machine learning pipeline.**

While there are other well-being surveys in the literature, these three were available in the current data collection. Because these measures were highly correlated (mean absolute r=0.57), we combined them to calculate an end-of-day well-being score by first reversing the measures (so higher scores mean better well-being), and then standardizing and averaging the measures (Cronbach's alpha of the composite scale was 0.80) [6]. The mean of the final well-being score is 0.08 (SD=0.74, Max=1.06, Min=-2.65, Range=3.71). Higher well-being indicates a better mental health state.

# 3 MACHINE LEARNING

A total of 1247 recordings were collected. The average number of recordings per participant per day is 7.5 (SD=5.2). The average recording length is 4.5 sec. These recordings are used to predict participants' end-of-day well-being scores. A visual representation of the entire machine learning pipeline is shown in Fig. 2.

## 3.1 Data Preprocessing

Audio files were first automatically transcribed before textual features were extracted. Any audio files that were empty or not understandable were removed. We tested 3 APIs, i.e., Wit.ai [39], Google Cloud Speech-to-Text [10], and Mozilla DeepSpeech [12], for their transcription accuracy on 7 randomly selected recordings. Google Cloud Speech-to-Text, which had a sample correctness of 100% (Wit.ai: 57%, Mozilla DeepSpeech: 29%), was selected to transcribe the rest of the audio files. In the resulting transcripts, words that were all digits, all punctuation, or stopwords [18] were removed before sentences were tokenized. The resulting text was used to extract the textual features.

## 3.2 Feature Extraction

As the prediction outcome, i.e., well-being, was measured once a day, all input features were aggregated by day as well.

*3.2.1 EMA Survey Features (Baseline1).* To benchmark the performance against using audio recordings, we designed 2 baseline measures with other questions from the surveys. The first baseline was people's reported social interactions from EMA surveys, motivated by literature that has shown consistent associations between social interactions and well-being [3, 30, 34]. Most of the features were summed on a per person per day basis, except for perceived closeness with the interaction partner, which was averaged.

*3.2.2 End-of-Day Survey Features (Baseline2).* We used self-reported affect valence and arousal in the end-of-day surveys as a second baseline. This is supported by [37], which found that Affect Valence, i.e., the positive or negative level of emotion, and Affect Arousal, i.e., the intensity of the emotions, are strong indicators for depression.

*3.2.3 Audio Features.* All audio features were extracted using *pyAudioAnalysis* [8], a python audio analysis library, with a frame size of 50*msecs* and a frame step of 25*msecs*. The means and the variances were then calculated for each window and were further aggregated per person per day.

Three general types of audio features were extracted, i.e., energy, spectral, and length. In [4], energy related features are useful in predicting affect as a higher energy audio infers upbeat mood and a lower energy audio implies calmness. In addition, Zero Crossing Rate (the rate at which the amplitude passes through 0) and Entropy of Energy capture the consistency of the energy. Spectral features are the signal spectrum from a Short Time Fourier Transformation, signaling the timbre of the audio signals. For example, Spectral Centroid is correlated with the brightness of the signal; Spectral Rolloff detects the skewness of the signal; Spectral Flux indicates the spectral change in two successive frames [19]. Mel Frequency Cepstral Coefficients (MFCCs), representing phonemes (distinct units of sound) in the speech signal, are a useful feature verified by previous audio processing work [14, 19]. Apart from the features mentioned, the length of the audio in seconds and the speed (number of words divided by the length of the audio) are also extracted. Literature suggests that the speed, or the tempo, of an audio is correlated to the mood of the audio [4, 19] and the audio length is an unnormalized form of speed.

*3.2.4 Textual Features.* While audio features capture the audio (and vocal) properties of the recordings, textual features contain information about the content of the recordings. We extracted 3 types of text features, i.e., counts, text vectors, and semantic orientation. Text length, i.e., number of words, is a naive way to measure the length of the audio transcript. [21] shows that the count of Part-of-Speech (POS) tags (using *TreeTagger*) is useful in mood classification.

**Table 1: Final Features used in machine learning models. MFCC is Mel Frequency Cepstral Coefficients. TF-IDF is frequency-inverse document frequency. POS is Part-of-Speech. AFINN is an enhanced version of Affective Norms for English Words. LIWC is Linguistic Inquiry and Word Count.**

| Feature Type | | Feature |
|---|---|---|
| Survey | Baseline1: EMA | *{affect_arousal, affect_valence}_{mean, var}, isInteraction, inPerson, closeness* |
| | Baseline2: Affect (End-of-day) | *affect_arousal, affect_valence* |
| Audio | | *length, speed, energy_entropy_{mean, var},<br>spectral_entropy_{mean, var},<br>mfcc_{1, 2, 3, 4}_{mean, var},<br>mfcc_{5, 6, 7, 8, 9, 10, 11, 12, 13}_mean* |
| Text | Text Vector | *TF-IDF* |
| | Others | *POS_{RP, CC, JJ, NP, VB, NNS, TO}, afinn, opinion_positive, LIWC_{Achieve, Affect, Comm, Cogmech, I, Motion, Past, Physical, Posemo, School, See, Self, TV}* |

In addition to simple counts, a text vector is one of the most common ways of representing textual data. We tried 3 text vector options. A Bag-of-word (BOW) model is commonly used in text analysis systems [21, 33]. BOW simply counts the word occurrences. Another text vector utilized is the word lemma frequency [21] acquired using *TreeTagger* [31]. Word lemmas group together words of the same stem (e.g., test, tests, and testing) and can potentially provide a better semantic representation than BOW. Finally, a more sophisticated text vector giving greater weights to important words often used in information retrieval is term frequency-inverse document frequency (TF-IDF) [14].

Word semantic orientation is another prevailing method often used to predict mood [14]. We used multiple popular measurements to represent this. The first is the 28 emotional word count [29]. The second is the number of positive words and the number of negative words, established in Opinion Lexicon [13]. We also used AFINN [26], an enhanced version of Affective Norms for English Words (ANEW) [5, 14], to sum up the word valence in a sentence. Lastly, we used Linguistic Inquiry and Word Count (LIWC) [27], which categorizes text to 80 linguistic, psychological and topical categories. This software has shown promising sentiment analysis results in previous works [32].

After extracting these textual features, they were aggregated by summing up on a per person per day basis.

## 3.3 Feature Selection

As we have a large feature space of 3020 dimensions, we needed to prune features that were not important to the prediction. We first did a correlation analysis to understand feature distribution, as well as the correlation between the feature and the target values. From the correlation analysis, we found some interesting features that are worth looking into. For example, *affect_valence_mean* is positively correlated to the well-being score and *affect_valence_var* is negatively correlated to the well-being score (Fig. 3). This indicates that days in which participants use words that reflect more positive affect were associated with better well-being. Days in which

**Table 2: Permutation feature importance of the top 11 most important features and their importance.**

| Feature | Type | RMSE | | MAE | |
|---|---|---|---|---|---|
| | | imp. | rank | imp. | rank |
| *mfcc_1_var* | Audio | 0.060 | 1 | 0.034 | 1 |
| *POS_NP* | Text | 0.015 | 2 | 0.010 | 3 |
| *LIWC_1_Physcal* | Text | 0.014 | 3 | 0.011 | 2 |
| *length* | Audio | 0.009 | 4 | 0.003 | 8 |
| *LIWC_Cogmech* | Text | 0.008 | 5 | 0.004 | 7 |
| *LIWC_Affect* | Text | 0.004 | 6 | 0.006 | 4 |
| *LIWC_I* | Text | 0.003 | 7 | 0.004 | 6 |
| *LIWC_Self* | Text | 0.003 | 8 | 0.004 | 5 |
| *mfcc_1_mean* | Audio | 0.003 | 9 | 0.001 | 11 |
| *afinn* | Text | 0.002 | 10 | 0.003 | 9 |
| *opinion_positive* | Text | 0.002 | 11 | 0.000 | 34 |
| *POS_TO* | Text | 0.000 | 22 | 0.001 | 10 |

participants use words that contain less fluctuations in affect are also positively associated with well-being. This is aligned with [37]'s research conclusion that affect is related to well-being, which strengthens the Baseline2 (Section 3.4) of the model performance evaluation.

Table 1 shows the final features used in the models after all the feature selection process.

*3.3.1 Variance.* During the correlation analysis, we also found that some features, like 28 emotional words, have most of the values being 0. When values in a feature are mostly identical, i.e., having a low variance, this feature is less predicative [16]. To remove these low-variance features, we calculated the variances of all features and kept the half of them with higher variance. In addition, text features with variance smaller than 0.1 were also removed.

*3.3.2 Text Vector Comparison.* Because the text features were many and sparse, we chose the most effective text vector by comparing the effectiveness of BOW, TF-IDF, and word lemma frequency representations using Linear Regression with linear kernel in a 10-fold Cross Validation (CV). The dataset is split into 10 folds. In each of the 10 rounds, 1 of the folds was selected as the testing set and the remaining 9 folds were used as the training set. The test errors of each round are averaged to evaluate model effectiveness. This methodology was also applied to Section 3.3.3 and 3.4 requiring models comparison. The results showed that TF-IDF had a significantly lower mean squared error (MSE) than the other two word vectors (TF-IDF MSE=1.36; BOW MSE=3.58; Lemma MSE=4.41). Therefore, TF-IDF was used in the well-being score prediction models. However, as the TF-IDF still had 970-dimension features, we tested the performance of TF-IDF separately from the remaining textual features in the prediction models.

*3.3.3 Linear Regression Feature Importance.* After removing sparse features and those with low variance, 44 textual features (not counting TF-IDF) were left. To further reduce the feature dimensionality, we filtered the remaining textual features based on their feature importance using the absolute value t-statistic of the feature [24]: $t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\frac{\sigma(\hat{\beta}_j)}{\sqrt{n}}} \propto \frac{\hat{\beta}_j}{\sigma(\hat{\beta}_j)}$ where $\hat{\beta}_j$ is the jth coefficient, $SE$ is the standard error, $\sigma$ is the standard deviation and $n$ is the number of samples. To obtain the coefficients and the standard deviation, a 10-fold CV Linear Regression with linear kernel was run. We selected the half of the features with the higher importance values.

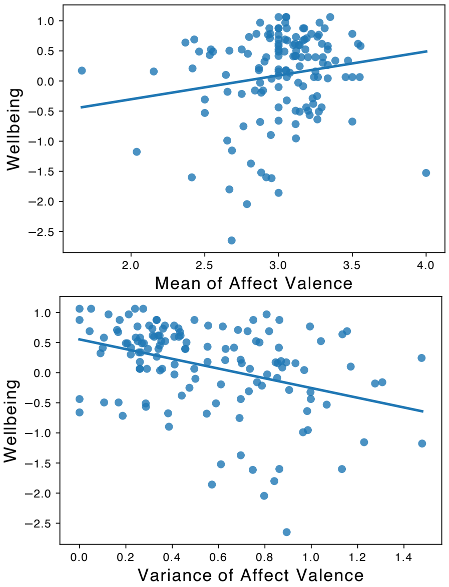


**Figure 3: Affect valence and well-being**

## 3.4 Well-being Score Prediction

Before building the models, we first determined the optimal hyperparameters using Grid Search. Afterwards, the model with the optimal hyperparameters was trained and tested using 10-fold CV. We tested multiple algorithms, i.e., linear regression with non-linear kernels, Multilayer Perceptron (MLP), Decision Tree, Support Vector Machine (SVM), Random Forest and AdaBoost. To understand whether both audio and textual features were necessary in the prediction, we also tested different feature sets. We used both Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to measure model performance, with the goal of minimizing the errors.

*3.4.1 Performance Analysis.* Across all models tested, the two baselines are almost always the least accurate with the highest errors, strongly suggesting that the audio recordings contained important information in predicting well-being.

Of all combinations of machine learning models and feature sets, Multilayer Perceptron with text features except TF-IDF word vectors has the best performance (bolded red values in Table 3). However, when using other feature sets, SVM constantly had the best results (bolded values in Table 3). Furthermore, the hyperparameter searching and training time for SVM was much shorter than MLP. Considering the very small decrease in error and the big gain in computation time, SVM was the best model for the dataset and the problem. With SVM, the best feature set was the audio and text features except TF-IDF word vectors. This suggests that audio recordings and the automatically transcribed text have value in predicting well-being in every-day settings for the general public.

**Table 3: Well-being score prediction 10-fold CV error. Lower RMSE and MAE indicates better results. For each feature set, results for the best performed algorithm is in bold. (Well-being rage=-2.65~1.06)**

| | | Linear Regression | MLP | Decision Tree | SVM | Random Forest | AdaBoost |
|---|---|---|---|---|---|---|---|
| Baseline1: EMA survey features | RMSE | 0.756 | 0.763 | 0.781 | 0.751 | 0.752 | **0.746** |
| | MAE | 0.635 | 0.622 | 0.629 | **0.577** | 0.620 | 0.599 |
| Baseline2: Affect features | RMSE | 0.762 | 0.791 | 0.781 | **0.755** | 0.764 | 0.775 |
| | MAE | 0.630 | 0.658 | 0.648 | **0.601** | 0.632 | 0.632 |
| TF-IDF | RMSE | **0.715** | 0.722 | 0.781 | 0.716 | 0.719 | 0.762 |
| | MAE | 0.603 | 0.595 | 0.600 | **0.581** | 0.589 | 0.612 |
| Text features except TF-IDF | RMSE | 0.686 | **0.647** | 0.801 | 0.692 | 0.734 | 0.728 |
| | MAE | 0.551 | **0.521** | 0.616 | 0.533 | 0.572 | 0.587 |
| Text features including TF-IDF | RMSE | 0.715 | 0.722 | 0.739 | **0.681** | 0.723 | 0.792 |
| | MAE | 0.603 | 0.595 | 0.602 | **0.538** | 0.590 | 0.626 |
| Audio features | RMSE | 0.683 | 0.712 | 0.778 | **0.680** | 0.725 | 0.785 |
| | MAE | 0.551 | 0.545 | 0.609 | **0.542** | 0.599 | 0.610 |
| Audio + text features except TF-IDF | RMSE | 0.684 | 0.715 | 0.822 | **0.665** | 0.718 | 0.740 |
| | MAE | 0.549 | 0.570 | 0.660 | **0.522** | 0.592 | 0.580 |

*3.4.2 Permutation Feature Importance.* We used the permutation feature method to understand the relative importance of the feature in predicting well-being, by removing a single feature and observing the increase in error [24]. Permutation feature importance was calculated using both the RMSE and MAE metrics, the same metrics used in Section 3.4. The error was obtained using the best SVM model in Section 3.4 and 10-fold CV.

The top most important features were mostly in agreement across RMSE and MAE. Note that the 11th most important feature ranked by RMSE (*opinion_positive*), surprising, has a close to zero negative importance when measured by MAE. This means that removing positive opinion word features will introduce more extreme errors. More generally, more than half of the most important features were semantic textual features. This confirms the importance of text in predicting one's well-being and audio features alone would not be sufficient. LIWC features were especially informative as 6 of the 11 features were from LIWC categories. The importance of I and self-related words is consistent with existing work on the association between self-focus, observed as the frequency of first-person pronouns, and many well-being measure, such as depression, anxiety, and negative mood [25, 41].

To summarize, our model produces a prediction result within ±0.52 (14.1% of the total well-being score range) of the "ground truth" or survey based values of well-being scores. This suggests that short audio recordings are valuable in predicting one's well-being in daily-settings. Both audio and textual feature are critical to the prediction performance. Next, we conducted an error analysis to understand reasons for this error.

*3.4.3 Error Analysis.* To better understand entries with high errors, predictions made by the best-performed model-feature combination (i.e., SVM model using audio and text features except TF-IDF) are ranked by the squared error of the predicted well-being score. We examined the top 10 and 20 entries with the highest errors. As both sets of high-error entries lead to similar conclusions, we will only show top 10 entries for reasons of space (Table 4).

**Table 4: Top 10 error entries by participant.**

| Participant | Date (YY-MM-DD) | Error |
|---|---|---|
| SAS2002 | 2019-07-19 | 2.80 |
| SAS2004 | 2019-07-29 | 3.23 |
| SAS2008 | 2019-07-28 | 2.40 |
| SAS2015 | 2019-08-30 | 3.08 |
| SAS2015 | 2019-08-31 | 3.86 |
| SAS2015 | 2019-09-01 | 3.10 |
| SAS2015 | 2019-09-02 | 7.60 |
| SAS2016 | 2019-08-30 | 2.19 |
| SAS2027 | 2019-11-02 | 2.28 |
| SAS2027 | 2019-11-03 | 5.29 |

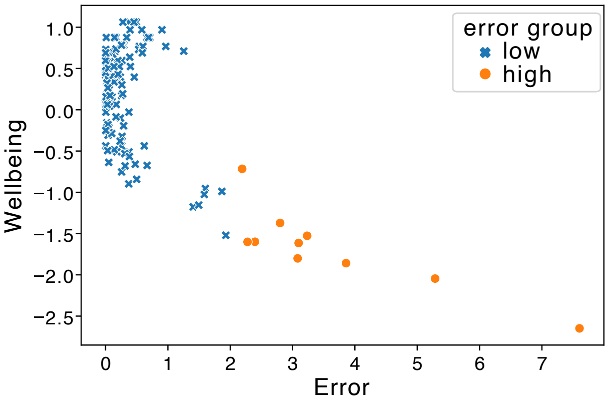


**Figure 4: High error entries (in orange) distribution. High error entries mostly have negative well-being.**

Fig. 4 shows that all high error entries have a well-being score lower than -1, with all but one being less than -1.5. As the well-being score was standardized (with the SD being 0.74), these points were effectively 2 SD below the sample average, meaning that these high-error cases were the low 5% of the data. As there are only a few of these cases, the models did not effectively learn the representation of these data points, which lead to high errors. These high error entries also mostly came from the same participants (Table 4). Four of the 10 high error entries were from participant *SAS2015* and 2 from *SAS2027*. A commonality between these participants was that they usually only had negative well-being entries. This again confirms the insight that the unbalanced positive and negative well-being scores may be the reason for high error.

## 4 DISCUSSION

The current work demonstrates the possibility of using short audio recordings in every-day settings to predict one's well-being. Our models produce prediction results within ±0.52 (14.1% of the total well-being score range) of the reported well-being scores using SVM with both text and audio features from the recordings (excluding TF-IDF features). This result outperformed the 2 baselines, i.e., predictions based on self-report data describing respondents' social interactions and their mood. This suggests that short voice recordings, made on smartphones in every-day settings and in contrast to always-on audio capture, can be used as inputs to predict well-being. Among the features used, MFCC and affect-related features are valuable for predicting well-being. In addition to the performance evaluation, this work opens up opportunities to further improve prediction results.

First, from the Error Analysis in Section 3.4.3, the lack of negative well-being representation is a main reason why the machine learning models cannot work more effectively. A more balanced positive and negative well-being dataset will help with this issue. In addition, the current study contains a small dataset. While this demonstrates the possibility of getting good prediction results with a small dataset, more data points will not only help improve the prediction results but also gives more room to balance the outcome class. This can be done by having more participants, conducting longer studies, making the audio recording a required task.

There are other areas that future work can improve on. First, the current study did not provide much guidance on what people should record (Fig. 1), As a result, most of the participants left very brief recordings. Having a more structured instruction will help collect longer audio samples, which provides more robust features for machine learning. Moreover, The current work only predicts people's well-being at the end of the day. The research could be improved by collecting a more continuous and instantaneous prediction of well-being, e.g., multiple times a day. Not only would more data improve prediction accuracy, it could also be used to better understand the events that influence well-being. Finally, future work is needed to test the generalizability of the models with other samples.

It is also worth-noting the trade-offs of using audio-based prediction. These recordings are much faster to collect compared to multi-item survey questions. Also, the results are less prone to subjective bias that may be present when filling out surveys. However, audio recordings are difficult to do in situations such as in a library or during a lecture. Under these circumstances, text-based descriptions can be used as an alternative as our results suggest that models using only text-based features can yield decent performance.

## 5 CONCLUSION

Using text and audio features from short audio recordings in everyday setting can predict one's well-being. In particular, MFCC, affect-related, and semantic-related measurements are valuable for the prediction. We also identify directions for future studies to further improve the performance.

## REFERENCES

[1] 2020. Demographics of Mobile Device Ownership and Adoption in the United States. https://www.pewresearch.org/internet/fact-sheet/mobile/

[2] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting Depression with Audio/Text Sequence Modeling of Interviews.. In *Interspeech*. 1716–1720.

[3] Michael J Bernstein, Matthew J Zawadzki, Vanessa Juth, Jacob A Benfield, and Joshua M Smyth. 2018. Social interactions in daily life: Within-person associations between momentary social experiences and psychological and physical health indicators. *Journal of Social and Personal Relationships* 35, 3 (2018), 372–394.

[4] Aathreya S Bhat, VS Amith, Namrata S Prasad, and D Murali Mohan. 2014. An efficient classification algorithm for music mood detection in western and hindi music using audio feature extraction. In *2014 Fifth International Conference on Signal and Image Processing*. IEEE, 359–364.

[5] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings.* Technical Report. Technical report C-1, the center for research in psychophysiology ....

[6] Moira Burke and Robert E Kraut. 2016. The relationship between Facebook use and well-being depends on communication type and tie strength. *Journal of computer-mediated communication* 21, 4 (2016), 265–281.

[7] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of health and social behavior* (1983), 385–396.

[8] Theodoros Giannakopoulos. 2015. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PloS one* 10, 12 (2015).

[9] Yuan Gong and Christian Poellabauer. 2017. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 69–76.

[10] Google, LLC. 2020. *Google Cloud Speech-to-Text.* https://cloud.google.com/speech-to-text

[11] Rahul Gupta, Nikolaos Malandrakis, Bo Xiao, Tanaya Guha, Maarten Van Segbroeck, Matthew Black, Alexandros Potamianos, and Shrikanth Narayanan. 2014. Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. 33–40.

[12] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567* (2014).

[13] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 168–177.

[14] Xiao Hu and J Stephen Downie. 2010. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th annual joint conference on Digital libraries*. 159–168.

[15] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2003. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Medical care* (2003), 1284–1292.

[16] Max Kuhn, Kjell Johnson, et al. 2013. *Applied predictive modeling.* Vol. 26. Springer.

[17] Nicholas D Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. 2011. Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th international ICST conference on pervasive computing technologies for healthcare*. 23–26.

[18] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028* (2002).

[19] Lie Lu, Dan Liu, and Hong-Jiang Zhang. 2005. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing* 14, 1 (2005), 5–18.

[20] Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed Ai-Shuraifi, and Yunhong Wang. 2013. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. 21–30.

[21] Gilad Mishne et al. 2005. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, Vol. 19. 321–327.

[22] Gilad Mishne, Maarten De Rijke, et al. 2006. Capturing Global Mood Levels using Blog Posts.. In *AAAI spring symposium: computational approaches to analyzing weblogs*, Vol. 6. 145–152.

[23] Saif M Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. *arXiv preprint arXiv:1708.03696* (2017).

[24] Christoph Molnar. 2019. *Interpretable Machine Learning.* https://christophm.github.io/interpretable-ml-book/.

[25] Nilly Mor and Jennifer Winquist. 2002. Self-focused attention and negative affect: a meta-analysis. *Psychological bulletin* 128, 4 (2002), 638.

[26] Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903* (2011).

[27] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[28] Donald A Redelmeier, Joel Katz, and Daniel Kahneman. 2003. Memories of colonoscopy: a randomized trial. *Pain* 104, 1-2 (2003), 187–194.

[29] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[30] Gillian M Sandstrom and Elizabeth W Dunn. 2014. Social interactions and well-being: The surprising power of weak ties. *Personality and Social Psychology Bulletin* 40, 7 (2014), 910–922.

[31] Helmut Schmid. 2013. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*. 154.

[32] H Andrew Schwartz, Maarten Sap, Margaret L Kern, Johannes C Eichstaedt, Adam Kapelner, Megha Agrawal, Eduardo Blanco, Lukasz Dziurzynski, Gregory Park, David Stillwell, et al. 2016. Predicting individual well-being through the language of social media. In *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific, 516–527.

[33] Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)* 34, 1 (2002), 1–47.

[34] Michael F Steger and Todd B Kashdan. 2009. Depression and everyday social activity, belonging, and well-being. *Journal of counseling psychology* 56, 2 (2009),

289.

[35] Rong Su, Louis Tay, and Ed Diener. 2014. The development and validation of the Comprehensive Inventory of Thriving (CIT) and the Brief Inventory of Thriving (BIT). *Applied Psychology: Health and Well-Being* 6, 3 (2014), 251–279.

[36] Seymour Sudman and Norman M Bradburn. 1973. Effects of time and memory factors on response in surveys. *J. Amer. Statist. Assoc.* 68, 344 (1973), 805–815.

[37] David Watson, Lee Anna Clark, and Greg Carey. 1988. Positive and negative affectivity and their relation to anxiety and depressive disorders. *Journal of abnormal psychology* 97, 3 (1988), 346.

[38] James R Williamson, Diana Young, Andrew A Nierenberg, James Niemi, Brian S Helfer, and Thomas F Quatieri. 2019. Tracking depression severity from audio and video based on speech articulatory coordination. *Computer Speech & Language* 55 (2019), 40–56.

[39] Wit.ai. 2020. *Wit.ai.* https://wit.ai

[40] H. Yu, E. B. Klerman, R. W. Picard, and A. Sano. 2019. Personalized Wellbeing Prediction using Behavioral, Physiological and Weather Data. In *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*. 1–4. https://doi.org/10.1109/BHI.2019.8834456

[41] Johannes Zimmermann, Timo Brockmeyer, Matthias Hunn, Henning Schauenburg, and Markus Wolf. 2017. First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients. *Clinical psychology & psychotherapy* 24, 2 (2017), 384–391.