Appendix

A Broader Impact Statement

Learning markers of mood from mobile data presents an opportunity for large-scale adaptive interventions of suicidal ideation. However, there are important concerns regarding its implications to society and policy.

Applications in mental health: Suicide is the second leading cause of death among adolescents. In addition to deaths, 16% of high school students report seriously considering suicide each year, and 8% make one or more suicide attempts (CDC, 2015). Despite these alarming statistics, there is little consensus concerning imminent risk for suicide (Franklin et al., 2017; Large et al., 2017). Current research conducts clinical interviews and patient self-report questionnaires that provide long-term assessments of suicide risk. However, few studies have focused on imminent suicidal risk, which is of critical clinical importance as a step towards adaptive real-time interventions (Glenn and Nock, 2014; Schuck et al., 2019). Given the impact of suicide on society, there is an urgent need to better understand the behavior markers related to suicidal ideation.

"Just-in-time" adaptive interventions delivered via mobile health applications provide a platform of exciting developments in low-intensity, high-impact interventions (Nahum-Shani et al., 2018). The ability to intervene precisely during an acute risk for suicide could dramatically reduce the loss of life. To realize this goal, we need accurate and timely methods that predict when interventions are most needed. Monitoring (with participants' permission) mobile data to assess mental health and provide early interventions is, therefore, a rich opportunity for scalable deployment across high-risk populations. Our data collection, experimental study, and computational approaches provide a step towards data-intensive longitudinal monitoring of human behavior. However, one must take care to summarize behaviors from mobile data without identifying the user through personal (e.g., personally identifiable information) or protected attributes (e.g., race, gender). This form of anonymity is critical when implementing these technologies in real-world scenarios. Our goal is to be highly predictive of mood while remaining as privacy-preserving as possible. We outline some of the potential privacy and security concerns below.

Limitations: While we hope that our research can provide a starting point on the potential of detecting mood unobtrusively throughout the day in a privacy-preserving way, we strongly acknowledge there remain methodological issues where *a lot* more research needs to be done to enable the real-world deployment of such technologies. We emphasize that healthcare providers and mobile app startups **should not** attempt to apply our approach in the real world until the following issues (and many more) can be reliably resolved:

1. We do not make broad claims across teenage populations from only 17 participants in this study. Furthermore, it remains challenging for models to perform person-independent prediction which makes it hard to deploy across large populations.
2. Our current work on predicting daily mood is still a long way from predicting imminent suicide risk. Furthermore, any form of prediction is still significantly far away from integrating methods like this into the actual practice of mental health, which is a challenging problem involving a broad range of medical, ethical, social, and technological researchers (Resnik et al., 2021; Lee et al., 2021).
3. Text and keystrokes can differ for participants who speak multiple languages or non-prestige vernaculars. One will need to ensure that the method works across a broad range of languages to ensure accessibility in its desired outcomes.
4. This study assumes that participants have no restrictions for data/network connections & data plans on their phones, which may leave out vulnerable populations that do not meet this criterion.

Privacy and security: There are privacy risks associated with making predictions from mobile data. To deploy these algorithms across at-risk populations, it is important to keep data private on each device without sending it to other locations. Even if data is kept private, it is possible to decode data from gradients (Zhu and Han, 2020) or pretrained models (Carlini et al., 2020). In addition, sensitive databases with private mobile data could be at-risk to external security attacks from adversaries (Lyu et al., 2020). Therefore, it is crucial to obtain user consent before collecting device data. In our experiments with real-world mobile data, all participants have given consent for their mobile device data to be collected and shared with us for research purposes. All data was anonymized and stripped of all personal (e.g., personally identifiable information) and protected attributes (e.g., race, gender).

Social biases: We acknowledge that there is a risk of exposure bias due to imbalanced datasets, especially when personal mobile data and sensitive health labels (e.g., daily mood, suicidal thoughts and behaviors, suicide risk). Models trained on biased data have been shown to amplify the underlying social biases especially when they correlate with the prediction targets (Lloyd, 2018). This leaves room for future work in exploring methods tailored for specific scenarios such as mitigating social biases in words (Bolukbasi et al., 2016), sentences (Liang et al., 2020a), and images (Otterbacher et al., 2018). Future research should also focus on quantifying the trade-offs between fairness and performance (Zhao and Gordon, 2019).

Overall, we believe that our proposed approach can help quantify the tradeoffs between performance and privacy. We hope that this brings about future opportunities for large-scale real-time analytics in healthcare applications.

B Dataset Details

The Mobile Assessment for the Prediction of Suicide (MAPS) dataset was designed to elucidate real-time indicators of suicide risk in adolescents ages $13-18$ years. Current adolescent suicide ideators and recent suicide attempters along with aged-matched psychiatric controls with no lifetime suicidal thoughts and behaviors completed baseline clinical assessments (i.e., lifetime mental disorders, current psychiatric symptoms). Following the baseline clinical characterization, a smartphone app, the Effortless Assessment of Risk States (EARS), was installed onto adolescents' phones, and passive sensor data were acquired for 6-months. Notably, during EARS installation, a keyboard logger is configured on adolescents' phones, which then tracks all words typed into the phone as well as the apps used during this period. Each day during the 6-month follow-up, participants also were asked to rate their mood on the previous day on a scale ranging from $1-100$, with higher scores indicating a better mood. After extracting multimodal features and discretizing the labels (see Section 2), we summarize the final dataset feature and label statistics in Table 9.

C Experimental Setup

We provide additional details on the model implementation and experimental setup.

C.1 Implementation Details

All models and analyses were done in Python. SVM models were implemented with Scikit-learn and MLP/NI-MLP models were implemented with PyTorch. BERT, XLNet, and Longformer models were fine-tuned using Hugging Face (website: https://huggingface.co, GitHub: https://github.com/huggingface).

C.2 Hyperparameters

We performed a small hyperparameter search over the ranges in Table 10. This resulted in a total of 35 hyperparameter configurations for SVM and 12 for MLP (6 for apps only). By choosing the best-performing model on the validation set, we selected the resulting hyperparameters as shown in Table 10.

C.3 Model Parameters

Each model has about two million parameters. See Table 10 for exact hidden dimension sizes.

C.4 Training Resources and Time

All experiments were conducted on a GeForce RTX 2080 Ti GPU with 12 GB memory. See Table 11 for approximate running times.

D Experimental Details

We present several additional analysis of the data and empirical results:

D.1 Details on Mood Prediction

There is often a tradeoff between privacy and prediction performance. To control this tradeoff, we vary the parameter σ, which is the amount of noise added to the identity-dependent subspace across batches and training epochs. In practice, we automatically perform model selection using this performance-privacy ratio R computed on the validation set, where

$$R = \frac{s_{\text{MLP}} - s_{\text{NI-MLP}}}{t_{\text{MLP}} - t_{\text{NI-MLP}}} \tag{4}$$

is defined as the improvement in privacy per unit of performance lost. Here, s is defined as the accuracy in the user prediction task and t is defined as the F1 score on the mood prediction task.

Table 9: Mobile Assessment for the Prediction of Suicide (MAPS) dataset summary statistics.

Users	Datapoints	Modalities	Features	Dimensions	Labels
17	1641	Text Keystrokes App usage	bag-of-words, one-hot bag-of-timings bag-of-apps, one-hot	2000 100 274	Daily mood: negative, neutral, positive

Table 10: Model parameter configurations. *Integer kernel values denote the degree of a polynomial kernel.

Model	Parameter	Value
SVM	C	0.1, 0.5, 1, 2, 3, 5, 10
	Kernel*	RBF, 2, 3, 5, 10
MLP	hidden dim 1 (multimodal & text only)	1024, 512
	hidden dim 2 (multimodal & text only)	128, 64
	hidden dim 1 (keystrokes only)	64, 32
	hidden dim 2 (keystrokes only)	32, 16
	hidden dim 1 (apps only)	128
	hidden dim 2 (apps only)	128, 64
	dropout rate	0, 0.2, 0.5
	learning rate	0.001
	batch size	100
	epochs	200
NI-MLP	λ	0.1, 1, 2, 3, 5, 10
	σ	1, 5, 10, 25, 50, 100, 150

Table 11: Approximate training times (total across 10-fold cross validation and hyperparameter search).

Model	Modality	Time (hours)
SVM	Text + Keystrokes + Apps	10
	Text + Keystrokes	10
	Text + Apps	10
	Text	8
	Keystrokes	1
	Apps	1
MLP (100 epochs, 3 runs)	Text + Keystrokes + Apps	6
	Text + Keystrokes	5
	Text + Apps	6
	Text	5
	Keystrokes	4
	Apps	2
NI-MLP	all	4

In the rare cases where NI-MLP performed better than the original MLP and caused R to become negative, we found this improvement in performance always came at the expense of worse privacy as compared to other settings of λ and σ in NI-MLP. Therefore, models with negative R were not considered for Table 1.

D.2 Details on Preserving Privacy

For Table 5, the model with the best privacy out of those within 5% performance of the original MLP model (or, if no such model existed, the model with the best performance) was selected.

Interestingly, in Figure 4, we find that the tradeoff curve on a model trained only using app features does not exhibit a Pareto tradeoff curve as expected. We attribute this to randomness in predicting both mood and identities. Furthermore, Wang et al. (2017) found that adding noise to the identity subspace can sometimes improve generalization by reducing reliance on identity-dependent confounding features, which could also explain occasional increased performance at larger σ values.

Note that we do not include privacy results for features learned by SVM, which finds a linear separator in a specified kernel space rather than learning a representation for each sample. Explicitly projecting our features is computationally infeasible due to the high dimensionality of our chosen kernel spaces.

Table 12: Top 5 words associated with positive and negative moods (each row is a different user).

Top 5 positive words	Top 5 negative words
hot, goodnight, ft, give, keep	soon, first, ya, friend, leave
still, y'all, guys, new, come	amazing, see, said, idk, look
mind, days, went, tf, next	tired, hair, stg, snap, anyone
girls, music, happy, mean, getting	omg, people, talking, ask, might

Table 13: Top words associated with positive and negative moods across users. We find that while certain positive words are almost always indicative of mood, others are more idiosyncratic and depend on the user.

Positive words	Positive users	Negative users	Negative words	Negative users	Positive users
make	9	1	i'm/im	10	5
yes	9	1	feel	7	3
got	7	1	yeah	7	5
still	7	1	can't/cant	6	2
wanna	7	1	people	6	4
like	7	2	know	6	4
need	7	2	go	6	5
send	7	2	one	6	6
get	7	2	today	5	1
good	7	3	day	5	2

D.3 Qualitative Analysis

In this section, we provide more empirical analysis on the unimodal and multimodal features in the MAPS dataset.

D.3.1 Understanding the unimodal features

Text: We begin with some basic statistics regarding word distributions. For each user, we tallied the frequencies of each word under each daily mood category (positive, neutral, and negative), as well as the overall number of words in each mood category. We define "positive" words and emojis to be those with a higher relative frequency of positive mood compared to the overall positive mood frequency, and lower than overall negative mood frequency. Likewise, "negative" words and emojis have higher than overall negative mood frequency and lower than overall positive mood frequency. We filtered out words for specific users if the word was used less than 40 times. Finally, we ranked the words by the difference in relative frequency (i.e., a word is "more positive" the larger the difference between its positive mood relative frequency and the user's overall positive mood relative frequency). See Table 12 for examples of top positive and negative words. For each word, we also counted the number of users for which the word was positive or negative. See Table 13 for the words with the highest user counts.

Keystrokes: We show some sample bag-of-timing histograms in Figure 6. It is interesting to find that certain users show a bimodal distribution across their keystroke histograms with one peak representing faster typing and another representing slower typing. Visually, the overall keystroke histograms did not differ that much across users which might explain its lower accuracies in both mood and user prediction when trained with NI-MLP (see Figure 4).

App usage: Similar to "positive" words, we define "positive" apps to be those with higher than overall positive mood relative frequency and lower than overall negative mood relative frequency, and "negative" apps to be the opposite. Apps were also then sorted by difference in relative frequency.

D.3.2 Understanding the multimodal features

Characters with keystrokes: For each user, we plotted histograms of keystroke timings of alphanumeric characters, symbols (punctuation and emojis), spacebar, enter, delete, and use of autocorrect, split across daily mood categories. See Figure 7 for examples across one user. We find particularly interesting patterns in the autocorrect keys and symbols where keystrokes are quite indicative of mood, which attests to the unique nature of typed text.

Words with keystrokes: For each user, we plotted histograms of the word-level keystroke timings of the top 500 words, split across the daily mood categories of positive, neutral, and negative. We also performed Wilcoxon rank-sum tests at 5% signifi-

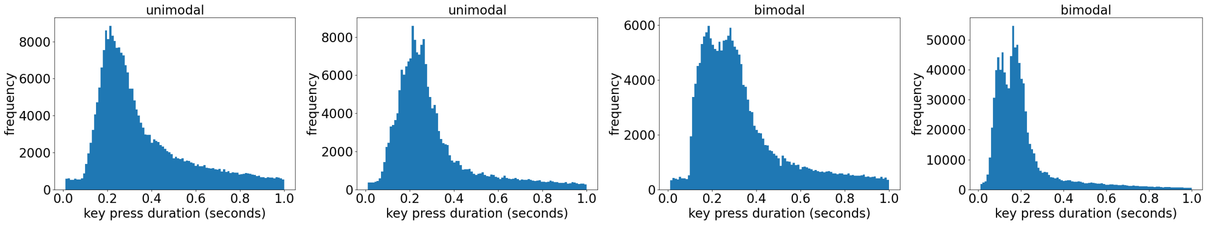

Figure 6: Examples of keystroke timing histograms for different users. We find that the distribution of keystroke timings varies between unimodal and bimodal for different users.

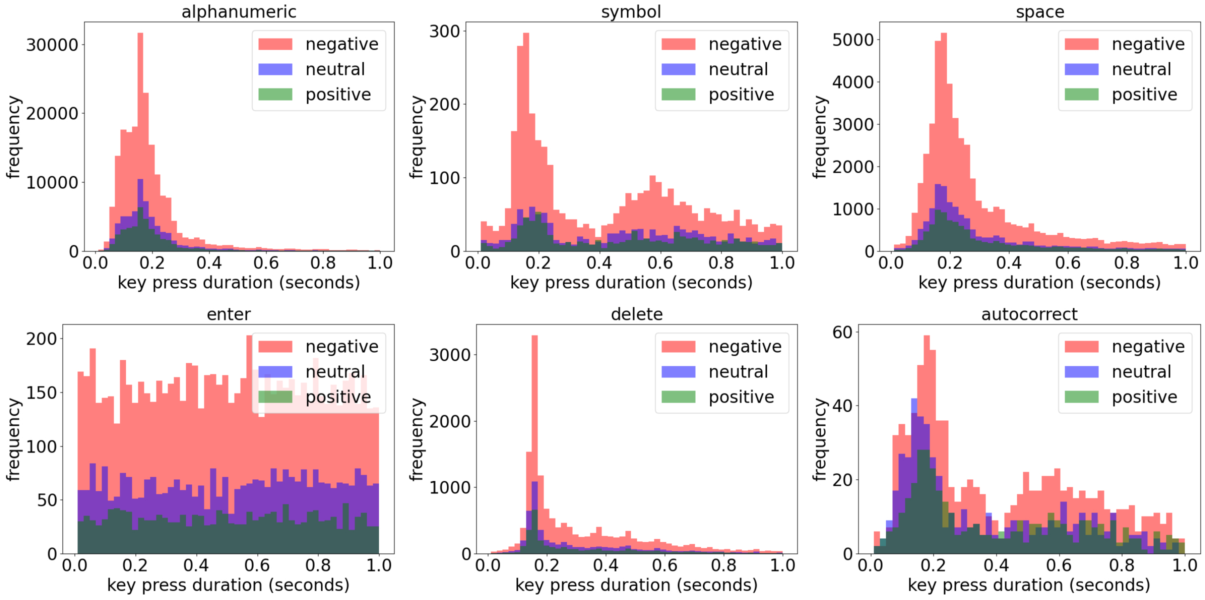

Figure 7: Example of more character key-presses and how their keystroke patterns can be indicative of either positive, neutral, or negative mood. We find particularly interesting patterns in the autocorrect keys and symbols where keystrokes are quite indicative of mood.

cance level (Wilcoxon, 1992) between the timings of positive and negative mood for each user/word combination to determine which words had significantly different timings between positive and negative mood.

E Negative Results and Future Directions

Since this is a new dataset, we explored several more methods throughout the research process. In this section we describe some of the approaches that yielded initial negative results despite them working well for standard datasets:

1. **User specific models:** We also explored the setting of training a separate model per user but we found that there was too little data per user to train a good model. As part of future work, we believe that if NI-MLP can learn a user-independent classifier, these representations can then be used for further finetuning or few-shot learning on each specific user. Previous work in federated learning (Smith et al., 2017; Liang et al., 2020b) offers ways of learning a user-specific model that leverages other users' data during training, which could help to alleviate the lack of data per user.

2. **User-independent data splits:** We have shown that text, keystrokes, and app usage features are highly dependent on participant identities. Consequently, models trained on these features would perform poorly when evaluated on a user not found in the training set. We would like to evaluate if better learning of user-independent features can improve generalization to new users (e.g., split the data such that the first 10 users are used for training, next 3 for validation, and final 4 for testing). Our initial results for these were negative, but we believe that combining better privacy-preserving methods that learn user-independent features could help in this regard.

3. **Fine-grained multimodal fusion:** Our approach of combining modalities was only at the input level (i.e., early fusion (Baltrušaitis et al., 2018)) which can be improved upon by leveraging recent work in more fine-grained fusion (Liang et al., 2018). One such example could be to align each keystroke feature and app data to the exact text that was entered in, which provides more fine-grained contextualization of text in keystroke and app usage context.