Learning Language and Multimodal Privacy-Preserving Markers of Mood from Mobile Data

Paul Pu Liang[1⋆], **Terrance Liu**[1⋆], **Anna Cai**[1], **Michal Muszynski**[1], **Ryo Ishii**[1],
Nicholas Allen[2], **Randy Auerbach**[3], **David Brent**[4],
Ruslan Salakhutdinov[1], **Louis-Philippe Morency**[1]
[1]Carnegie Mellon University [2]University of Oregon
[3]Columbia University [4]University of Pittsburgh
{pliang,terrancl,annacai,mmuszyns,rishii,rsalakhu,morency}@cs.cmu.edu
nallen3@uoregon.edu rpa2009@cumc.columbia.edu brentda@upmc.edu

Abstract

Mental health conditions remain underdiagnosed even in countries with common access to advanced medical care. The ability to accurately and efficiently predict mood from easily collectible data has several important implications for the early detection, intervention, and treatment of mental health disorders. One promising data source to help monitor human behavior is daily smartphone usage. However, care must be taken to summarize behaviors without identifying the user through personal (e.g., personally identifiable information) or protected (e.g., race, gender) attributes. In this paper, we study behavioral markers of daily mood using a recent dataset of mobile behaviors from adolescent populations at high risk of suicidal behaviors. Using computational models, we find that language and multimodal representations of mobile *typed text* (spanning typed characters, words, keystroke timings, and app usage) are predictive of daily mood. However, we find that models trained to predict mood often also capture private user identities in their intermediate representations. To tackle this problem, we evaluate approaches that obfuscate user identity while remaining predictive. By combining multimodal representations with privacy-preserving learning, we are able to push forward the performance-privacy frontier.

1 Introduction

Mental illnesses can have a damaging permanent impact on communities, societies, and economies all over the world (World Health Organization, 2003). Individuals often do not realize they are at risk of mental disorders even when they have symptoms. As a result, many are late in seeking professional help and treatment (Thornicroft et al., 2016), particularly among adolescents where suicide is the second leading cause of death (Curtin and Heron, 2019). In addition to deaths, 16% of high school students report having serious suicidal thoughts each year, and 8% of them make one or more suicide attempts (CDC, 2015). This problem is particularly exacerbated as an "echo pandemic" of mental health problems have arisen in the wake of the COVID-19 pandemic (Inkster et al., 2021; Saha et al., 2020).

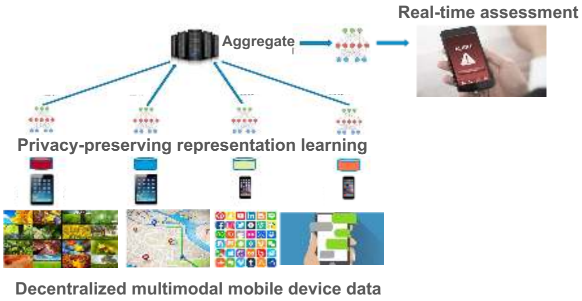

Real-time assessment
Aggregate
Privacy-preserving representation learning
Decentralized multimodal mobile device data

Figure 1: Intensive monitoring of behaviors via adolescents' natural use of smartphones may help identify real-time predictors of mood in high-risk youth as a proxy for suicide risk. While smartphones provide a valuable data source spanning text, keystrokes, app usage, and geolocation, one must take care to summarize behaviors without revealing user identities through personal (e.g., personally identifiable information) or protected attributes (e.g., race, gender) to potentially adversarial third parties.

Intensive monitoring of behaviors via adolescents' natural use of smartphones may help identify real-time predictors of mood in high-risk youth as a *proxy* for suicide risk (Nahum-Shani et al., 2018). While there are inherent limitations in the mismatch between mood prediction and ultimately developing real-time intervention against imminent suicide risk (Coppersmith et al., 2018; Ophir et al., 2020), we believe that the former is a reasonable starting point to tackle similar machine learning problems surrounding affective computing and privacy-preserving learning. Studying mood in this high-risk population is a valuable goal given

[⋆] first two authors contributed equally.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 4170–4187
August 1–6, 2021. ©2021 Association for Computational Linguistics

that suicide attempts are often decided within a short time-lapse and just-in-time assessments of mood changes can be a stepping stone in this direction (Rizk et al., 2019; Oquendo et al., 2020). Technologies for mood prediction can also be a valuable component of decision support for clinicians and healthcare providers during their assessments (Mann et al., 2006; Cho et al., 2019).

Recent work in affective computing has begun to explore the potential in predicting mood from mobile data. Studies have found that typing patterns (Cao et al., 2017; Ghosh et al., 2017a; Huang et al., 2018; Zulueta et al., 2018), self-reporting apps (Suhara et al., 2017), and wearable sensors (Ghosh et al., 2017b; Sano et al., 2018) are particularly predictive. In addition, multimodal modeling of multiple sensors (e.g., wearable sensors and smartphone apps) was shown to further improve performance (Jaques et al., 2017; Taylor et al., 2017). While current work primarily relies on self-report apps for *long-term* mood assessments (Glenn and Nock, 2014), our work investigates mobile behaviors from a high-risk teenage population as a predictive signal for *daily* mood (Franklin et al., 2017; Large et al., 2017).

Prior work has also shown that private information is predictable from digital records of human behavior (Kosinski et al., 2013), which is dangerous especially when sensitive user data is involved. As a result, in parallel to improving predictive performance, a recent focus has been on improving privacy through techniques such as differential privacy (Dankar and El Emam, 2012, 2013; Dankar et al., 2012) and federated learning (McMahan et al., 2016; Geyer et al., 2017; Liang et al., 2020b), especially for healthcare data (e.g., electronic health records (Xu and Wang, 2019)) and wearable devices (Chen et al., 2020).

In this paper, as a step towards using *multimodal privacy-preserving* mood prediction as fine-grained signals to aid in mental health assessment, we analyze a recent dataset of mobile behaviors collected from adolescent populations at high suicidal risk. With consent from participating groups, the dataset collects fine-grained features spanning online communication, keystroke patterns, and application usage. Participants are administered daily questions probing for mood scores. By collecting and working on ground-truth data for this population, we are able to benchmark on a more accurate indicator of mood rather than proxy data such as mood signals inferred from social media content or behavior (Ernala et al., 2019). This unique dataset presents an opportunity to investigate a different medium of natural language processing - *typed text* which presents new challenges beyond conventionally studied written (Marcus et al., 1993) and spoken (Marslen-Wilson and Tyler, 1980) text. We propose multimodal models that *contextualize* text with their typing speeds and app usage. However, these models often capture private user identities in their intermediate representations when predicting mood. As a step towards privacy-preserving learning, we also propose approaches that obfuscate user identity while remaining predictive of daily mood. By combining multimodal contextualization with privacy-preserving learning, we are able to push forward the performance-privacy frontier. Finally, we conclude with several observations regarding the uniqueness of typed text as an opportunity for NLP on mobile data.

2 Multimodal Mobile Dataset

Intensive monitoring of behaviors via adolescents' frequent use of smartphones may shed new light on the early risk of suicidal thoughts and ideations (Nahum-Shani et al., 2018). Smartphones provide a valuable and natural data source with rich behavioral markers spanning online communication, keystroke patterns, and application usage. Learning these markers requires large datasets with diversity in participants, variety in features, and accuracy in annotations. As a step towards this goal, we recently collected a dataset of mobile behaviors from high-risk adolescent populations with consent from participating groups.

We begin with a brief review of the data collection process. This data monitors adolescents spanning (a) recent suicide attempters (past 6 months) with current suicidal ideation, (b) suicide ideators with no past suicide attempts, and (c) psychiatric controls with no history of suicide ideation or attempts. Passive sensing data is collected from each participant's smartphone across a duration of 6 months. Participants are administered clinical interviews probing for suicidal thoughts and behaviors (STBs), and self-report instruments regarding symptoms and acute events (e.g., suicide attempts, psychiatric hospitalizations) are tracked weekly via a questionnaire. All users have given consent for their mobile data to be collected and shared with us for research purposes. This study has been carefully reviewed and approved by an IRB. We follow the NIH guidelines, with a central IRB (single IRB) linked to secondary sites. We have IRB approval for the central institution and all secondary sites.

2.1 Mood Assessment via Self-Report

Every day at 8am, users are asked to respond to the following question - "In general, how have you been feeling over the last day?" - with an integer score between 0 and 100, where 0 means very negative and 100 means very positive. To construct our prediction task, we discretized these scores into the following three bins: *negative* $(0-33)$, *neutral* $(34-66)$, and *positive* $(67-100)$, which follow a class distribution of 12.43%, 43.63%, and 43.94% respectively. For our 3-way classification task, participants with fewer than 50 daily self-reports were removed since these participants do not provide enough data to train an effective model. In total, our dataset consists of 1641 samples, consisting of data coming from 17 unique participants.

2.2 Features

We focused on keyboard data, which includes the time of data capture, the mobile application used, and the text entered by the user. For each daily score response at 8am, we use information collected between 5am on the previous day to 5am on the current day. We chose this 5am-5am window by looking at mobile activity and finding the lowest activity point when most people ended their day: 5am. Since users report the *previous* day's mood (when prompted at 8am), we decided to use this 5am-5am time period to summarize the previous day's activities. Through prototyping, this prompt time and frequency were found to give reliable indicators of the previous day's mood. From this window, we extracted the following features to characterize and contextualize typed text.

Text: After removing stop-words, we collected the top 1000 words (out of approximately 3.2 million) used across all users in our dataset and created a *bag-of-words* feature that contains the daily number of occurrences of each word.

Keystrokes: We also extracted keystroke features that record the exact timing that each character was typed on a mobile keyboard (including alphanumeric characters, special characters, spaces, backspace, enter, and autocorrect). By taking the increase in recorded timing after each keystroke, we obtain the duration that each key was pressed in a sequence of keystrokes during the day. When extracting keystrokes, we removed all small timings under 10^{-2} seconds.

App usage: We count the number of mobile applications used per day, creating a *bag-of-apps* feature for each day. We discard applications that are used by less than 10% of the participants so that our features are generalizable to more than just a single user in the dataset, resulting in 137 total apps (out of the original 640).

In a preliminary analysis, we observed that predictive models performed well when binarizing our feature vectors into boolean vectors, which signify whether a word or app was used on a given day (i.e., mapping values greater than 0 to 1). Our final feature vectors consist of a concatenation of a normalized and a binarized feature vector, resulting in 2000 and 274-dimensional vectors for text and app features respectively. For keystrokes, we found that summarizing the sequence of timings using a histogram (i.e., defining a set of timing buckets and creating a *bag-of-timings* feature) for each day performed well. We chose 100 fine-grained buckets, resulting in a 100-dimensional keystroke vector. Please refer to Appendix B for additional details about the dataset and extracted features.

3 Mood Prediction Methods

In this paper, we focus on studying approaches for learning privacy-preserving representations from mobile data for mood prediction. Our processed data comes in the form of $\{(x_{t,i}, x_{k,i}, x_{a,i}, y_i)\}_{i=1}^{n}$ with $x_t \in \mathbb{N}^{|V_t|=2000}$ denoting the bag-of-words features, $x_k \in \mathbb{N}^{|V_k|=100}$ denoting the bag-of-timings features, and $x_a \in \mathbb{N}^{|V_a|=274}$ denoting the bag-of-apps features. y denotes the label which takes on one of our 3 mood categories: negative, neutral, and positive. In parallel, we also have data representing the corresponding (one-hot) user identity x_{id} which will be useful when learning privacy-preserving representations that do not encode information about user identity x_{id} and evaluating privacy performance.

3.1 Unimodal Approaches

We considered two unimodal baselines:

1. Support Vector Machines (SVMS) project training examples to a chosen kernel space and finds the optimal hyperplane that maximally separates each class of instances. We apply an SVM classifier on input data $x_{\text{uni}} \in \{x_t, x_k, x_a\}$ and use supervised

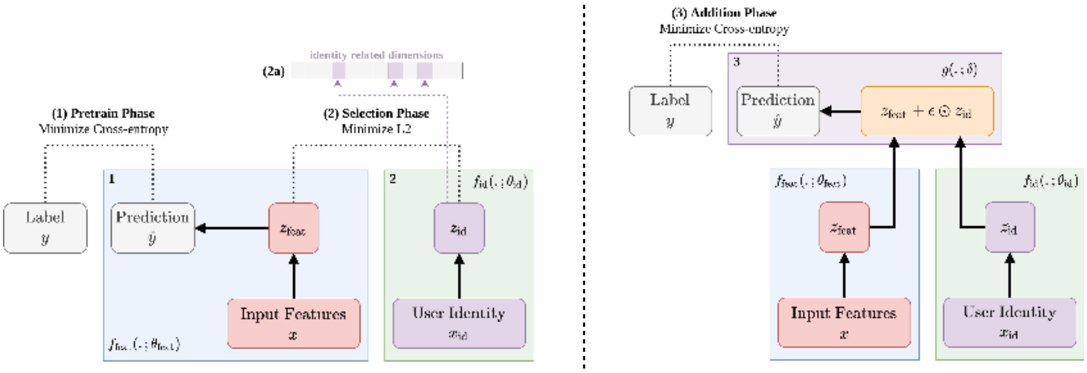


Figure 2: Diagram of the NI-MLP algorithm learned via the (1) *pretrain*, (2) *selection*, and (3) *addition* phases. Boxes with numbers denote which parameters are being optimized in the corresponding step. For example, in the *addition* phase (3), NI-MLP optimizes parameters δ in $g(.;\delta)$. (2a) depicts identity-dependent dimensions z_{id}, which is a sparse vector of size dim(z_{feat}) whose nonzero values (colored purple) signify dimensions of the identity-dependent subspace in z_{feat}.

learning to predict daily mood labels y.

2. Multilayer Perceptrons (MLPs) have seen widespread success in supervised prediction tasks due to their ability in modeling complex nonlinear relationships. Because of the small size of our dataset, we choose a simple multilayer perceptron with two hidden layers. Similarly, we apply an MLP classifier on input data $x_{\text{uni}} \in \{x_t, x_k, x_a\}$ to predict daily mood labels y.

3.2 Multimodal Models

We extend both SVM and MLP classifiers using early fusion (Baltrušaitis et al., 2018) of text and app usage to model multimodal interactions. Specifically, we align the input through concatenating the bag-of-words, bag-of-keystrokes, and bag-of-apps features for each day resulting in an input vector $x_{\text{multi}} = x_t \oplus x_k \oplus x_a$, before using an SVM/MLP classifier for prediction.

3.3 A Step Toward Preserving Privacy

While classifiers trained with traditional supervised learning can learn useful representations for mood prediction, they carry the risk of *memorizing* the identity of the user along with their sensitive mobile usage and baseline mood scores, and possibly *revealing* these identities to adversarial third-parties (Abadi et al., 2016). Therefore, it is crucial to perform mood prediction while also protecting the privacy of personal identities.

We adapt the Selective-Additive Learning (SAL) framework (Wang et al., 2017) for the purpose of privacy-preserving learning. While SAL was originally developed with a very different goal in mind: improving model generalization, we expand SAL to a very important problem in healthcare: preserving privacy. We adapted SAL to learn *disentangled* representations separated into *identity-dependent* private information and *identity-independent* population-level information using three phases:

(1) *Pretrain phase:* The input is a set of (multimodal) features x that are likely to contain both identity-dependent and independent information. The intermediate representation $z_{\text{feat}} = f_{\text{feat}}(x; \theta^*_{\text{feat}})$ is obtained from an MLP classifier pretrained for mood prediction. f_{feat} denotes the classifier with pretrained parameters θ^*_{feat}.

(2) *Selection phase:* Our goal is to now disentangle the identity-dependent and independent information within z_{feat}. We hypothesize that dependent and independent information are encoded in separate subspaces of the feature vector z_{feat}. This allows us to disentangle them by training a separate classifier to predict z_{feat} *as much as possible* given only the user identity:

$$\theta^*_{\text{id}} = \arg\min_{\theta_{\text{id}}} \left(z_{\text{feat}} - f_{\text{id}}(x_{\text{id}}; \theta_{\text{id}})\right)^2 + \lambda ||z_{\text{id}}||_1, \tag{1}$$

where x_{id} denotes a one hot encoding of user identity as input, f_{id} denotes the identity encoder with parameters θ_{id}, and λ denotes a hyperparameter that controls the weight of the ℓ_1 regularizer. f_{id} projects the user identity encodings to the feature space learned by f_{feat}. By minimizing the objective in equation (1) for each (x, x_{id}) pair, f_{id} learns to encode user identity into a sparse vector $z_{\text{id}} = f_{\text{id}}(x_{\text{id}}; \theta^*_{\text{id}})$ representing identity-dependent features: the nonzero values of z_{id} represent dimensions of the identity-dependent subspace in z_{feat}, while the remaining dimensions belong to the

Table 1: Comparison of mood prediction performance across different modalities. Best results in **bold**. For both accuracy and F1 score, models jointly trained on text, keystroke, and apps features outperform models trained using individual modalities. $\star$ denotes that the difference between multimodal and all unimodal models is statistically significant (p-value $<< 0.05$).

	F1 SCORE				ACCURACY			
Modalities	BASELINE	SVM	MLP	NI-MLP	BASELINE	SVM	MLP	NI-MLP
Text + Keystrokes + Apps	19.07	**62.81***	**59.61***	**60.11***	40.18	**67.43***	**63.59***	**64.06***
Text + Keystrokes	19.07	61.19	57.65	58.70	40.18	65.87	61.81	62.61
Text + Apps	19.07	62.08	58.38	52.90	40.18	66.59	62.93	56.76
Text	19.07	61.15	56.27	52.63	40.18	65.83	60.61	56.08
Keystrokes	19.07	57.68	51.43	34.73	40.18	61.03	55.87	39.18
Apps	19.07	58.65	52.29	51.32	40.18	62.65	55.26	55.68

identity-independent subspace.

(3) *Addition phase:* Given two factors z_{feat} and z_{id}, to ensure that our prediction model does not capture identity-related information z_{id}, we add multiplicative Gaussian noise to remove information from the identity-related subspace z_{id} while repeatedly optimizing for mood prediction with a final MLP classification layer $g(z_{\text{feat}}, z_{\text{id}}; \delta)$. This resulting model should only retain identity-independent features for mood prediction:

$$\hat{y} = g\left(z_{\text{feat}} + \epsilon \odot z_{\text{id}}\right) \tag{2}$$

where $\epsilon \sim N(0, \sigma^2)$ is repeatedly sampled across batches and training epochs. We call this approach NOISY IDENTITY MLP, or NI-MLP for short, and summarize the final algorithm in Figure 2.

Controlling the tradeoff between performance and privacy: There is often a tradeoff between privacy and prediction performance. To control this tradeoff, we vary the parameter σ, which is the variance of noise added to the identity-dependent subspace across batches and training epochs. $\sigma = 0$ recovers a standard MLP with good performance but reveals user identities, while large σ effectively protects user identities but at the possible expense of mood prediction performance. In practice, the optimal tradeoff between privacy and performance varies depending on the problem. For our purposes, we automatically perform model selection using this performance-privacy ratio R computed on the validation set, where

$$R = \frac{s_{\text{MLP}} - s_{\text{NI-MLP}}}{t_{\text{MLP}} - t_{\text{NI-MLP}}} \tag{3}$$

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

4 Experiments

We perform experiments to test the utility of text, keystroke, and app features in predicting daily mood while keeping user privacy in mind.

4.1 Experimental Setup

Data splits: Given that our data is longitudinal, we split our data into 10 partitions ordered chronologically by users. We do so in order to maintain independence between the train, validation, and test splits in the case where there is some form of time-level dependency within our labels.

Evaluation: For each model, we run a nested k-fold cross-validation (i.e., we perform 9-fold validation within 10-fold testing). For each test fold, we identify the optimal parameter set as the one that achieves the highest mean validation score over the validation folds. To evaluate NI-MLP, we use the best performing MLP model for each test fold as our base classifier before performing privacy-preserving learning. For all experiments, we report the test accuracy and macro F1 score because our classes are imbalanced. Given the low number of cross-validation folds, we use the Wilcoxon signed-rank test (Wilcoxon, 1992) at 5% significance level for all statistical comparisons (see Appendix C for more experimental details).

4.2 Results on Mood Prediction

We make the following observations regarding the learned language and multimodal representations for mood prediction:

Observation 1: Text, keystroke, and app usage features are individually predictive of mood. To evaluate how predictive our extracted text, keystroke timings, and app usage features are, we first run experiments using SVM, MLP, and NI-MLP on each individual feature separately. Since we have unbalanced classes, we chose a majority classifier (i.e., most common class in the training

Table 2: Mood prediction from text using extended pretrained LM encoders. We find that these models struggle on extremely long contexts of typed text.

Models	F1 SCORE	ACCURACY
BoW	**56.27**	**60.61**
BERT	51.42	58.06
XLNet	19.85	42.40
LongFormer	19.85	42.40

set) as our baseline. From Table 1, we observe that using these three feature types individually outperforms the baseline with respect to accuracy and F1 score. Using the Wilcoxon signed-rank test (Wilcoxon, 1992) at 5% significance level, we found that these improvements over the baseline in both F1 score and accuracy are statistically significant (p-value $<< 0.05$).

Observation 2: Pretrained sentence encoders struggle on this task. We also applied pretrained sentence encoders such as BERT (Devlin et al., 2019) on the language modality for mood prediction. Surprisingly, we found that none of these approaches performed stronger than a simple bag-of-words (see Table 2). We provide two possible explanations for this phenomenon:

1. BERT is suitable for written text on the web (Wikipedia, BookCorpus, carefully human-annotated datasets) which may not generalize to informal typed text that contains emojis, typos, and abbreviations (see Section 4.4 for a qualitative analysis regarding the predictive abilities of emojis and keystrokes for mood prediction).

2. We hypothesize that it is difficult to capture such long sequences of data (>1000 time steps) spread out over a day. Current work has shown that BERT struggles with long sequence lengths (Beltagy et al., 2020). We trained two extensions XLNet (Yang et al., 2019) and LongFormer (Beltagy et al., 2020) specifically designed to take in long-range context but found that they still underperform as compared to a simple bag-of-words approach.

Observation 3: Fusing both text and keystroke timings improves performance. This dataset presents a unique opportunity to study representations of *typed* text as an alternative to conventionally studied written or spoken text. While the latter two use language alone, typed text includes keystroke features providing information about the timings of when each character was typed. In Table 1, we present some of our initial results in learning text and keystroke representations for mood prediction and show consistent improvements over text alone. We further study the uniqueness of typed text by comparing the following baselines:

Table 3: Mood prediction using a MLP from text and keystroke features tallied from (1) all characters, (2) a split between types of characters, as well as (3) aggregated across words.

Modalities	F1 SCORE	ACCURACY
Text	56.27	60.61
Text + Char keystrokes	**57.65**	**61.81**
Text + Split char keystrokes	57.32	61.21
Text + Word keystrokes	56.46	60.68

1. *Text*: bag-of-words only.

2. *Text + char keystrokes*: bag-of-words and bag-of-timings across all characters.

3. *Text + split char keystrokes*: bag-of-words and bag-of-timings subdivided between 6 groups: alphanumeric characters, symbols, spacebar, enter, delete, and use of autocorrect. This baseline presents a more fine-grained decomposition of the typing speeds across different semantically related character groups.

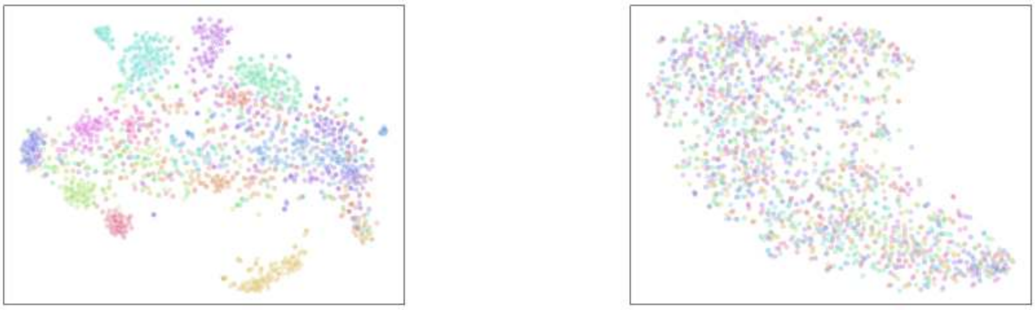
4. *Text + word keystrokes*: bag-of-words and bag-of-timings summed up over the characters in each word. This presents a more interpretable model to analyze the relationships between words and the distribution of their typing speeds.

From Table 3, we observe that keystrokes accurately contextualize text, especially when using fine-grained keystroke distributions across individual characters. Other methods incorporating keystroke features are also all stronger than unimodal models. Different ways of representing keystrokes also provide different levels of interpretability regarding the relationships between words, characters, and keystrokes for mood prediction, which we qualitatively analyze in §4.4.

Observation 4: Multimodal representation learning achieves the best performance. In Table 1, we also compare the performance of our models on combined (text + keystroke + apps) features versus the performance on each individual feature set. For both metrics, combining all features gives better performance over either subset.

4.3 Results on Preserving Privacy

Despite these promising results in mood prediction, we ask an important question: *Does the model capture user identities as an intermediate step towards predicting mood?* To answer this question, we an-



(a) MLP (without privacy-preserving)

(b) NI-MLP (with privacy-preserving)

Figure 3: Visualization of representations learned by (a) MLP and (b) NI-MLP, which have been reduced to two dimensions via t-SNE and colored by participant identity. Representations learned by NI-MLP are no longer separable by users which better preserves privacy.

Table 4: We report user identity prediction performance from raw input data and find that identities are **very easily revealed** from text, keystrokes, and app usage.

	F1 SCORE		ACCURACY	
Modalities	SVM	MLP	SVM	MLP
Text	89.42	92.05	90.60	93.12
Keystrokes	91.36	87.04	90.98	87.15
Apps	85.68	87.49	90.91	92.00

Table 5: Comparison of our privacy-preserving approach (NI-MLP) with the baseline (MLP). We evaluate privacy in predicting user identity from learned **representations** (**lower** accuracy is better), and find that NI-MLP effectively obfuscates user identity while retaining performance. T: text, K: keystrokes, A: apps.

	PERFORMANCE (↑)		PRIVACY (↓)	
Modalities	MLP	NI-MLP	MLP	NI-MLP
T + K + A	59.61	58.48	71.47	**34.49**
T + K	57.65	57.40	64.17	**30.99**
T + A	58.38	57.76	79.04	**65.13**
T	56.27	54.11	76.41	**52.20**
K	51.43	42.48	55.61	**25.71**
A	52.29	49.15	85.94	**66.74**

alyze the privacy of raw mobile data and trained models. We then study our proposed method of learning privacy-preserving features to determine whether it can obfuscate user identity while remaining predictive of daily mood.

How private is the mobile data? We evaluate how much the data reveal user identities by training predictive models with typed text, keystroke timings, and app usage as input and user identity as the prediction target. From Table 4, we observe that all modalities are very predictive of user identity (>87% accuracy), which further motivates the need to learn privacy-preserving features. We further note that identifiable information can be very subtle: while only 28/1000 words were named entities, it was possible to identify the user identity with >87% accuracy, which means that subtle word choice can be identify the user (similarly for apps and keystrokes).

How private are the learned privacy-preserving features? We also study whether our learned features are correlated with user identity through both visualizations and quantitative evaluations.

Visualizations: We use t-SNE (Van der Maaten and Hinton, 2008) to reduce the learned features from trained models to 2 dimensions. After color-coding the points by participant identity, we identify distinct clusters in Figure 3(a), which implies that mood prediction can be strongly linked to identifying the person, therefore coming at the price of losing privacy.

As an attempt to reduce reliance on user identity, we train NI-MLP which is designed to obfuscate user-dependent features. After training NI-MLP, we again visualize the representations learned in Figure 3(b) and we find that they are less visually separable by users, indicating that NI-MLP indeed learns more user-independent features.

Quantitative evaluation: To empirically evaluate how well our models preserve privacy, we extracted the final layer of each trained model and fit a logistic regression model to predict user identity using these final layer representations as input. The more a model preserves privacy, the harder it should be to predict user identity. From Table 5, we observe that we can predict user identity based on the learned MLP representations with high accuracy (>85%) using the most sensitive app usage features. For other modality combinations, user identity can also be decoded with more than 70% accuracy with the exception of keystrokes which are the most private (55%). We achieve significantly more privacy using NI-MLP embeddings - roughly 35%

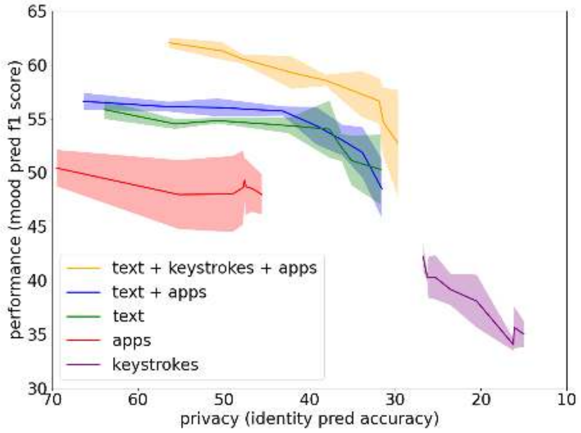


Figure 4: Tradeoff between performance (mood prediction F1 score, **higher** is better) and privacy (identity prediction accuracy, **lower** is better). Shaded regions denote standard deviations from the mean (solid lines). NI-MLP provides a tunable parameter σ to control the tradeoff, which allows us to plot a range of (performance, privacy) points. Using a multimodal model on text, keystroke, and app features obtains better performance and privacy at the same time.

for the best multimodal model, which indicates the possibility of NI-MLP as a means of achieving privacy-preserving mood prediction.

Understanding the tradeoff between performance and privacy: NI-MLP provides a tunable parameter σ to control the variance of noise applied on the identity-related dimensions. This parameter σ has the potential to give a tradeoff between privacy and prediction performance. In Figure 4, we plot this tradeoff between performance (mood prediction F1 score, higher is better) and privacy (identity prediction accuracy, lower is better). We find that keystroke features, while themselves not very useful in predicting mood, are highly private features. It is important to note that keystroke features show strong performance when integrated with text and app usage features while also increasing privacy, thereby pushing the Pareto front outwards. It is also interesting to observe that for most models, performance stays level while privacy improves, which is a promising sign for the real-world deployment of such models which requires a balance between both desiderata.

4.4 Qualitative Analysis

To further shed light on the relationships between mood prediction performance and privacy, we performed a more in-depth study of the text, keystroke, and app usage features learned by the model (see Appendix D.3 for more examples).

Table 6: Top emojis associated with positive and negative mood (each row is a different user).

Positive emojis	Negative emojis
🤤 🥰 😘 😍	😟 🤣 😂
:(🐶	🙄 😭 :((❤️ 😟
😟	:(

Table 7: Top 3 apps associated with positive and negative moods (each row is a different user).

Top 3 positive apps	Top 3 negative apps
Photos, Settings, Snapchat	Calendar, Wattpad, SoundCloud
FaceTime, MyFitnessPal, Musically	Notes, App Store, Siri
Weather, Phone, FaceTime	Chrome, App Store, SMS
Weather, Phone, Spotify	Safari, Notes, GroupMe
Spotlight, App Store, Uber	Pinterest, Phone, Yolo
Uber, Netflix, LinkedIn	Phone, Calendar, Safari

Understanding the unimodal features: We first analyze how individual words, keystroke timings, and app usage are indicative of positive or negative mood for different users.

Text: We find that several words are particularly indicative of mood: *can't/cant*, *don't/don't*, and *sorry* are negative for more users than positive, while *yes* is overwhelmingly positive across users (9 pos, 1 neg), but *yeah* is slightly negative (5 pos, 7 neg). We also analyze the use of emojis in typed text and find that while there are certain emojis that lean positive (e.g., 🥰 😘 😍), there are ones (e.g., :(and 😟) that used in both contexts depending on the user (see Table 6).

Apps: In Table 7, we show the top 3 apps associated with positive or negative moods across several users. It is interesting to observe that many outdoor apps (i.e., *Weather, MyFitnessPal, Uber*), photo sharing apps (i.e., *Photos, Snapchat*), and calling apps (i.e., *FaceTime, Phone*) are associated with positive mood, while personal apps such as personal management (i.e., *Calendar, Notes, Siri*), web browsing (i.e., *Chrome, Safari*), and shopping (i.e., *App Store*) are associated with negative mood. However, some of these findings are rather user-specific (e.g., *Phone* can be both positive or negative depending on the user).

Understanding the multimodal features: We also analyze how the *same characters and words* can contribute to *different mood predictions* based on their keystroke patterns. As an example, the distribution of keystrokes for the *enter* character on the keyboard differs according to the daily mood of one user (see Figure 5 and Appendix D.3 for

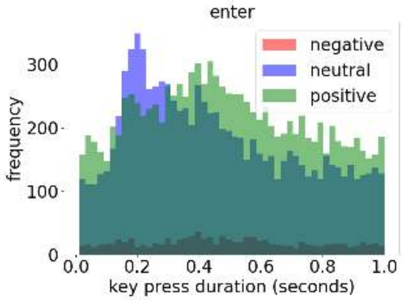


Figure 5: An example where the *'enter'* character keypress is indicative of either positive, neutral, or negative mood depending on the keypress duration.

Table 8: Words with significantly different timings associated with positive and negative moods (each row is a different user).

Slower implies positive	Faster implies positive
just	why, thank, haha
next, was, into, people	making, work, idk
stuff, cute, phone, want, talk, see	they, send, dont, man, going
don't, talk	think, you, all, love

more users). In Table 8, we extend this analysis to entire words. For each of the 500 most common words, we aggregated their accompanying keystroke timings for user-reported positive and negative mood. These two distributions tell us how the same word in different keystroke contexts can indicate different moods. We performed Wilcoxon rank-sum tests at 5% significance level to compare these distributions and recorded the words in which either faster or slower typing was statistically significantly correlated with either mood. Observe how certain semantically positive words like *love*, *thank*, and *haha* become judged as more positive when typed at a faster speed. Therefore, contextualizing text with their keystroke timings offers additional information when learning representations of typed text.

5 Conclusion

In this paper, we investigated the learning of language and multimodal representations of *typed text* collected from mobile data. We studied the challenge of learning markers of daily mood as a step towards early detection and intervention of mental health disorders for social good. Our method also shows promising results in obfuscating user identities for privacy-preserving learning, a direction crucial towards real-world learning from sensitive mobile data and healthcare labels. In addition, our findings illustrate several challenges and opportunities in representation learning from typed text as an understudied area in NLP.

Limitations & future work: While our approach shows promises in learning representations for mood prediction, several future directions on the modeling and NLP side include: 1) better models and pre-training algorithms for NLP on typed text, 2) algorithms that provide formal guarantees of privacy (Dwork, 2008), and 3) federated training from decentralized data (McMahan et al., 2016) to improve privacy (Geyer et al., 2017) and fairness (Liang et al., 2020a) of sensitive data. We describe more limitations and future social implications of our work in our broader impact statement in Appendix A.

Acknowledgements

This material was based upon work partially supported by the National Science Foundation (Awards #1750439 and #1734868) and the National Institutes of Health (Award #U01MH116923). MM was supported by the Swiss National Science Foundation (#P2GEP2_184518). RS was supported by NSF IIS1763562 and ONR Grant N000141812861. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, National Institutes of Health, or Office of Naval Research, and no official endorsement should be inferred. We would also like to acknowledge NVIDIA's GPU support and the anonymous reviewers for their extremely helpful comments.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NIPS*.

Bokai Cao, Lei Zheng, Chenwei Zhang, Philip S Yu, Andrea Piscitello, John Zulueta, Olu Ajilore, Kelly

Ryan, and Alex D Leow. 2017. Deepmood: modeling mobile phone typing dynamics for mood detection. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 747–755.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

CDC. 2015. *Suicide Facts at a Glance 2015*.

Yiqiang Chen, Xin Qin, Jindong Wang, Chaohui Yu, and Wen Gao. 2020. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*.

Chul-Hyun Cho, Taek Lee, Min-Gwan Kim, Hoh Peter In, Leen Kim, and Heon-Jeong Lee. 2019. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: prospective observational cohort study. *Journal of medical Internet research*, 21(4):e11029.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Sally C Curtin and Melanie P Heron. 2019. Death rates due to suicide and homicide among persons aged 10–24: United states, 2000–2017.

Fida Kamal Dankar and Khaled El Emam. 2012. The application of differential privacy to health data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 158–166.

Fida Kamal Dankar and Khaled El Emam. 2013. Practicing differential privacy in health care: A review. *Trans. Data Priv.*, 6(1):35–67.

Fida Kamal Dankar, Khaled El Emam, Angelica Neisa, and Tyson Roffey. 2012. Estimating the re-identification risk of clinical data sets. *BMC medical informatics and decision making*, 12(1):66.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16.

Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Katherine M Musacchio, Adam C Jaroszewski, Bernard P Chang, and Matthew K Nock. 2017. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychological bulletin*, 143(2):187.

Robin C Geyer, Tassilo Klein, and Moin Nabi. 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.

Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2017a. Evaluating effectiveness of smartphone typing as an indicator of user emotion. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 146–151. IEEE.

Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2017b. Tapsense: Combining self-report patterns and typing characteristics for smartphone based emotion detection. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–12.

Catherine R Glenn and Matthew K Nock. 2014. Improving the short-term prediction of suicidal behavior. *American journal of preventive medicine*, 47(3):S176–S180.

He Huang, Bokai Cao, S Yu Phillip, Chang-Dong Wang, and Alex D Leow. 2018. Dpmood: Exploiting local and periodic typing dynamics for personalized mood prediction. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 157–166. IEEE.

Becky Inkster et al. 2021. Early warning signs of a mental health tsunami: A coordinated response to gather initial data insights from multiple digital services providers. *Frontiers in Digital Health*, 2:64.

Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. 2017. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 202–208. IEEE.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805.

Matthew Michael Large, Daniel Thomas Chung, Michael Davidson, Mark Weiser, and Christopher James Ryan. 2017. In-patient suicide: selection of people at risk, failure of protection and the possibility of causation. *BJPsych Open*, 3(3):102–105.

Ellen E Lee, John Torous, Munmun De Choudhury, Colin A Depp, Sarah A Graham, Ho-Cheol Kim, Martin P Paulus, John H Krystal, and Dilip V Jeste. 2021.

Artificial intelligence for mental healthcare: Clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020a. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020b. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*.

Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 150–161, Brussels, Belgium. Association for Computational Linguistics.

Kirsten Lloyd. 2018. Bias amplification in artificial intelligence systems. *CoRR*, abs/1809.07842.

Lingjuan Lyu, Han Yu, and Qiang Yang. 2020. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

J John Mann, Dianne Currier, Barbara Stanley, Maria A Oquendo, Lawrence V Amsel, and Steven P Ellis. 2006. Can biological tests assist prediction of suicide in mood disorders? *International Journal of Neuropsychopharmacology*, 9(4):465–474.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

William Marslen-Wilson and Lorraine Komisarjevsky Tyler. 1980. The temporal structure of spoken language understanding. *Cognition*, 8(1):1–71.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2016. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*.

Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462.

Yaakov Ophir, Refael Tikochinski, Christa SC Asterhan, Itay Sisso, and Roi Reichart. 2020. Deep neural networks detect suicide risk from textual facebook posts. *Scientific reports*, 10(1):1–10.

Maria A Oquendo, Hanga C Galfalvy, Tse-Hwei Choo, Raksha Kandlur, Ainsley K Burke, M Elizabeth Sublette, Jeffrey M Miller, J John Mann, and Barbara H Stanley. 2020. Highly variable suicidal ideation: a phenotypic marker for stress induced suicide risk. *Molecular psychiatry*, pages 1–8.

Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating user perception of gender bias in image search: The role of sexism. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, SIGIR '18, page 933–936, New York, NY, USA. Association for Computing Machinery.

Philip Resnik, April Foreman, Michelle Kuchuk, Katherine Musacchio Schafer, and Beau Pinkham. 2021. Naturally occurring language as a source of evidence in suicide prevention. *Suicide and Life-Threatening Behavior*, 51(1):88–96.

Mina M Rizk, Tse-Hwei Choo, Hanga Galfalvy, Emily Biggs, Beth S Brodsky, Maria A Oquendo, J John Mann, and Barbara Stanley. 2019. Variability in suicidal ideation is associated with affective instability in suicide attempters with borderline personality disorder. *Psychiatry*, 82(2):173–178.

Koustuv Saha, John Torous, Eric D Caine, and Munmun De Choudhury. 2020. Psychosocial effects of the covid-19 pandemic: Large-scale quasi-experimental study on social media. *Journal of medical Internet research*, 22(11):e22600.

Akane Sano, Sara Taylor, Andrew W McHill, Andrew JK Phillips, Laura K Barger, Elizabeth Klerman, and Rosalind Picard. 2018. Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: observational study. *Journal of medical Internet research*, 20(6):e210.

Allison Schuck, Raffaella Calati, Shira Barzilay, Sarah Bloch-Elkouby, and Igor Galynker. 2019. Suicide crisis syndrome: A review of supporting evidence for a new suicide-specific diagnosis. *Behavioral sciences &amp; the law*, 37(3):223–239.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. 2017. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pages 4424–4434.

Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. 2017. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 715–724.

Sara Ann Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2017. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*.

Graham Thornicroft, Nisha Mehta, Sarah Clement, Sara Evans-Lacko, Mary Doherty, Diana Rose, Mirja Koschorke, Rahul Shidhaye, Claire O'Reilly, and Claire Henderson. 2016. Evidence for effective interventions to reduce mental-health-related stigma and discrimination. *The Lancet*, 387(10023):1123–1132.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 949–954. IEEE.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer.

World Health Organization. 2003. Investing in mental health.

Jie Xu and Fei Wang. 2019. Federated learning for healthcare informatics. *arXiv preprint arXiv:1911.06270*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763.

Han Zhao and Geoff Gordon. 2019. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems*, volume 32, pages 15675–15685. Curran Associates, Inc.

Ligeng Zhu and Song Han. 2020. Deep leakage from gradients. In *Federated Learning*, pages 17–31. Springer.

John Zulueta, Andrea Piscitello, Mladen Rasic, Rebecca Easter, Pallavi Babu, Scott A Langenecker, Melvin McInnis, Olusola Ajilore, Peter C Nelson, Kelly Ryan, et al. 2018. Predicting mood disturbance severity with mobile phone keystroke metadata: a biaffect digital phenotyping study. *Journal of medical Internet research*, 20(7):e241.

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	bag-of-words, one-hot	2000	Daily mood: negative, neutral, positive
		Keystrokes	bag-of-timings	100	
		App usage	bag-of-apps, one-hot	274	

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-

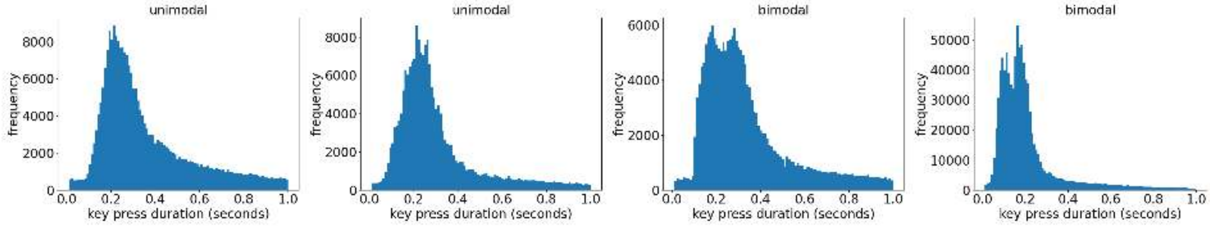


Figure 6: Examples of keystroke timing histograms for different users. We find that the distribution of keystroke timings varies between unimodal and bimodal for different users.

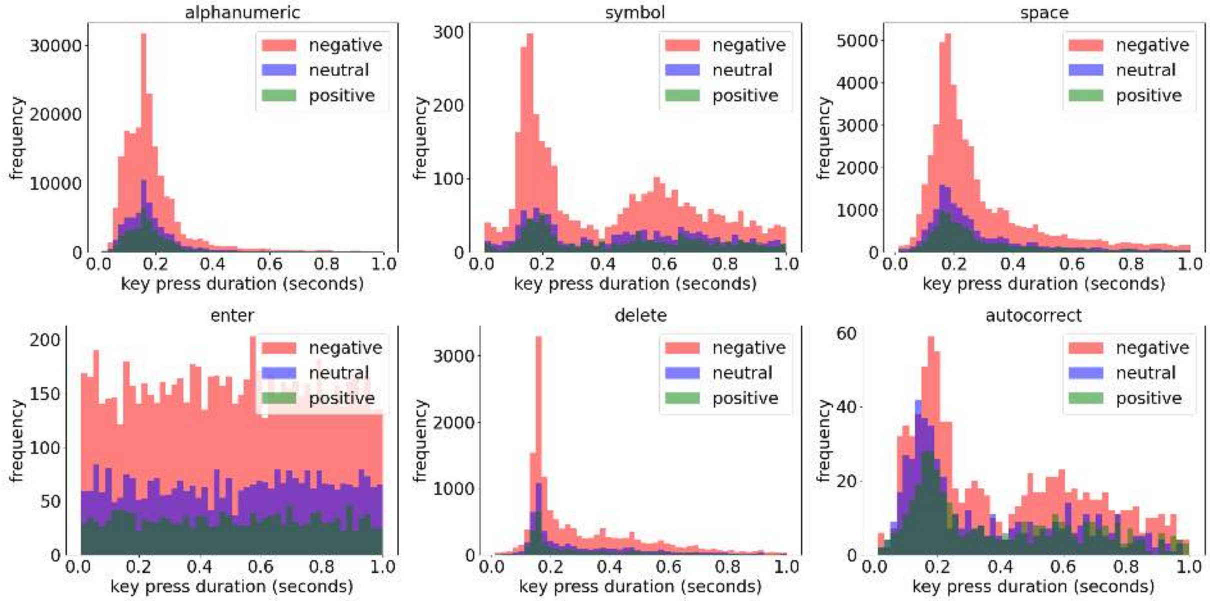


Figure 7: Example of more character key-presses and how their keystroke patterns can be indicative of either positive, neutral, or negative mood. We find particularly interesting patterns in the autocorrect keys and symbols where keystrokes are quite indicative of mood.

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.