Affective State Prediction from Smartphone Touch and Sensor Data in the Wild

Affective State Prediction from Smartphone Touch and Sensor Data in the Wild

Rafael Wampfler
Department of Computer Science
ETH Zurich
Zurich, Switzerland
wrafael@inf.ethz.ch

Severin Klingler
Department of Computer Science
ETH Zurich
Zurich, Switzerland
kseverin@inf.ethz.ch

Barbara Solenthaler
Department of Computer Science
ETH Zurich
Zurich, Switzerland
solenthaler@inf.ethz.ch

Victor R. Schinazi
Department of Psychology
Bond University
Robina, Australia
vschinaz@bond.edu.au

Markus Gross
Department of Computer Science
ETH Zurich
Zurich, Switzerland
grossm@inf.ethz.ch

Christian Holz
Department of Computer Science
ETH Zurich
Zurich, Switzerland
christian.holz@inf.ethz.ch

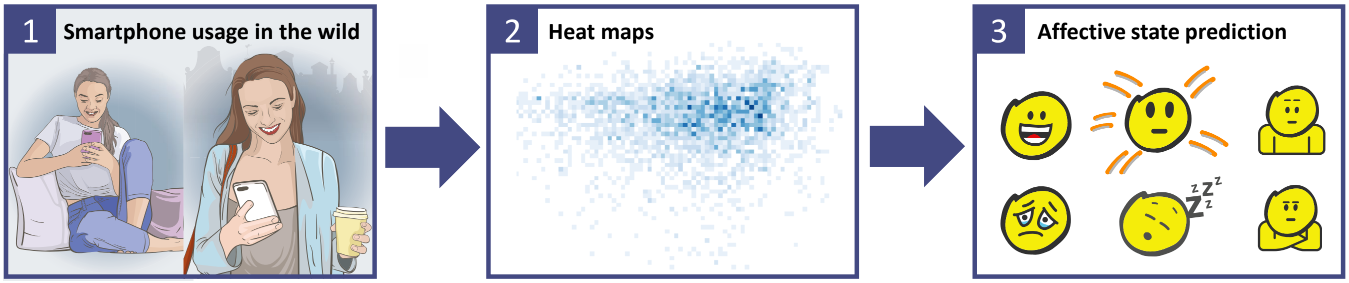


Figure 1: Our system passively records sensor and touch data during smartphone usage in the wild (1). From the recorded data, we extracted two-dimensional heat maps (2) and developed a classification model that predicts users' affective states (3).

ABSTRACT

Knowledge of users' affective states can improve their interaction with smartphones by providing more personalized experiences (e.g., search results and news articles). We present an affective state classification model based on data gathered on smartphones in real-world environments. From touch events during keystrokes and the signals from the inertial sensors, we extracted two-dimensional heat maps as input into a convolutional neural network to predict the affective states of smartphone users. For evaluation, we conducted a data collection in the wild with 82 participants over 10 weeks. Our model accurately predicts three levels (low, medium, high) of valence (AUC up to 0.83), arousal (AUC up to 0.85), and dominance (AUC up to 0.84). We also show that using the inertial sensor data alone, our model achieves a similar performance (AUC up to 0.83), making our approach less privacy-invasive. By personalizing our model to the user, we show that performance increases by an additional 0.07 AUC.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501835

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *User studies*; *Touch screens*; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Classification; Affective Computing; Smartphone; Deep Learning

ACM Reference Format:
Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R. Schinazi, Markus Gross, and Christian Holz. 2022. Affective State Prediction from Smartphone Touch and Sensor Data in the Wild. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3491102.3501835

1 INTRODUCTION

Affective states are psycho-physiological constructs that are used to describe the emotions (short-term) and moods (long-term) of a person exposed to a stimulus [18, 40, 57]. Affective states are typically measured along the valence (positive vs. negative emotions), arousal (intensity of the emotion), and dominance (degree of control of the emotion) dimensions [43]. Affective states can also be grouped into basic emotions (i.e., anger, happiness, sadness, surprise, disgust, and fear) [19]. Awareness of a user's affective state can enhance the quality of interactions, making systems more usable, enjoyable, and effective. Such affect-aware systems are useful in domains such as education and health. For example, a learning application that detects and reacts to frustration can increase motivation and learning gain by adapting task difficulties [61]. Incorporating a user's affective state can also improve personalized recommendations (e.g., for music and news) [1, 44]. Similarly, recognizing a user's affective state can help treating mental health problems such as depression (e.g., as part of a therapeutic chatbot) [52].

The ubiquitous use of smartphones for social interactions (e.g., chat applications and social networks), entertainment (e.g., music and video platforms), and news consumption provides a distinct opportunity for collecting information to recognize the affective states of users. In addition, smartphone use is also highly diverse in context and location (e.g., at home, on the train, or in school), which enables capturing the variability in affective states that may be used for prediction models in real-world environments.

In this paper, we propose a system that accurately predicts affective states in real-world environments. We focus on typing-based applications (e.g., chat and browsing applications), as these are the most used applications [2], as well as smartphone sensor data (i.e., gyroscope and accelerometer sensors). Using data from our in-the-wild user study with 82 participants, we present a learning-based model that accurately predicts three levels of valence, arousal, and dominance from extracted heat maps. We also demonstrate that our model achieves a similar performance when using merely the signals from the inertial sensors on users' phones. We conclude that sensor data is a viable alternative to keyboard data for the prediction of affective states due to the continuous availability of data and the increased protection of privacy. The use of sensor data makes our approach suitable for a large number of devices in the wild and can increase user acceptance.

1.1 Contributions

The contributions of this work are threefold:

- A dataset of smartphone touch and sensor data from a user study with 82 participants (including 30,083 self-reports that captured participants' affective states) conducted in the wild over the course of 10 weeks.
- A deep convolutional affective state classification model that is trained on two-dimensional heat maps of sensor data and typing characteristics from the data we passively captured from the smartphone's on-screen keyboard during everyday use. Due to its small size and low memory consumption, our model can potentially be run on mobile devices, which could improve user experience, accessibility, security, privacy, and energy consumption in future applications.
- An evaluation of our model based on the collected dataset and the self-reports including a comparison of our findings to other work. We show that our model accurately predicts valence (0.83 AUC), arousal (0.85 AUC), and dominance (0.84 AUC) on three levels (low, medium, high). In addition, we show that by processing the two-dimensional heat maps extracted from inertial sensor data alone, our model achieves a similar performance for valence (0.79 AUC), arousal (0.83 AUC), and dominance (0.81 AUC).

2 RELATED WORK

Large-scale labeled datasets are necessary to train and evaluate models for predicting affective states. These datasets are typically collected in laboratory [11, 53] or in-the-wild experiments [4, 59]. In laboratory experiments, emotional states are usually induced by presenting videos [46], pictures [30, 42], or chat conversations [67] with different affective characteristics. In contrast, in-the-wild experiments rely on emotions that are manifested while the user engages with a smartphone (e.g., browsing and chatting [53]) without any explicit emotion induction. Typically, such in-the-wild experiments last for several days [4], weeks [38], or months [24] and include a larger number of participants [35].

To collect the labels during these experiments, researchers often ask participants to complete binary or Likert scale self-reports at specified times (e.g., several times a day) [4] or in response to an event (e.g., user switching an application) [28]. Together with the features extracted from smartphone data during the experiment, the labeled data can then be used to design emotion recognition systems. A comprehensive overview of different experimental designs, data sources, and models is provided by Kołakowska et al. [35].

2.1 Touch Data

The touchscreen of a smartphone allows for the collection of keystroke dynamics and gestures (i.e., tapping, scrolling, and swiping). Keystroke dynamics have been widely investigated for affective state prediction on hardware keyboards [20, 34]. Keystroke parameters consist of timing characteristics such as flight time [12], tap duration [12], and typing speed [67]. Frequency characteristics of keystrokes (i.e., how often selected keys are touched) have also proved to be good predictors of affective states [12, 65]. For example, the usage of backspaces as a source of information on mistakes made while typing has been found to correlate to emotional states [38]. Similarly, Trojahn et al. [65] showed that negative emotions are associated with decreased typing speed and increased error rate. In contrast to hardware keyboards, smartphone keyboards provide additional sources of data including the pressure and size of touch, which can provide further information for predicting affective states [67].

Using such timing and frequency characteristics on smartphone keyboards, Ghosh et al. [28] predicted two levels of happiness, sadness, and stress by employing a personalized random forest classifier. In a previous laboratory experiment, we predicted three levels of valence, arousal, and dominance, and two levels of anger, happiness, sadness, surprise, and stress [67]. Critically, we encoded timing characteristics (i.e., flight time and typing speed) and touch pressure in two-dimensional heat maps and used a semi-supervised model consisting of an autoencoder and a fully connected classification layer. Similarly, Ghosh et al. [27] used an LSTM-based encoder-decoder to learn a low-dimensional feature embedding of time series of typing data (i.e., time and frequency characteristics). These authors predicted happiness, sadness, relaxation, and stress by using a classification network with the first layers shared among users and a final personalized layer.

2.2 Sensor Data

Accelerometers and gyroscopes are common sensors used for affective state prediction on smartphones. For example, Mottelson and Hornbæk [46] showed that positive emotions are accompanied by bigger movements and fewer changes of the orientation of the smartphone while Carneiro et al. [11] found a strong relationship between stress and acceleration. Features from sensor data are calculated either on the aggregated series (i.e., magnitude) or on the three axes (i.e., x, y, and z) separately [30, 46] and belong to either the time (e.g., mean, variance, and interquartile range) [30, 53] or frequency domain [13, 30]. To extract frequency features, a fast Fourier transform (FFT) is typically applied on the sensor data series and a specific number of FFT coefficients are used as feature values [13, 30]. Other computed features from the FFT encompass the peak magnitude, peak magnitude frequency, peak power, and peak power frequency [24]. Finally, researchers have also used computed features such as the deviation of the acceleration from the user's usual behavior [59], the device shaking measured as changes in aggregated acceleration [46], and the activity type (e.g., still, walking, and running) [4].

Using accelerometer data only, Olsen and Torresen [47] predicted valence (using a support vector machine) and arousal (using a multilayer perceptron) on three levels. Similarly, Hashmi et al. [30] predicted the basic emotions using a support vector machine based on timing and frequency features extracted from accelerometer and gyroscope data. In contrast to our work, these previous efforts collected data in a laboratory experiment with induced emotions and used chest-mounted smartphones to track human motion.

2.3 Multimodal Data

One possibility for generalizing and improving performance is to fuse different data modalities for building competent multimodal affective state prediction systems [60]. Several techniques to fuse data have been proposed such as feature-level fusion, decision-level fusion, and data-level fusion [66]. Ruensuk et al. [53] developed a personalized support vector machine model based on touch input, accelerometer data, and gyroscope data. They predicted two levels of valence and arousal during browsing activities and chatting. Wang et al. [68] predicted five levels of valence and arousal. They fused neural network and decision tree classifiers and based their models on data from the accelerometer, gyroscope, GPS (i.e., entropy), light sensor (i.e., indoor vs. outdoor), and network speed. Here, the usage of the GPS signal adds an additional strain on privacy. Other researchers have successfully combined text data with audio data (i.e., speech) [29] and video data [49] for the prediction of emotions. Yang et al. [72] went one step further and combined smartphone data (i.e., front camera, microphone, and keystrokes) with biosensor data (i.e., skin conductance, skin temperature, and blood volume pulse) using an attention-based LSTM system. An extensive comparison of multimodal and unimodal affect classifiers is provided by D'Mello and Kory [17]. While combining different data modalities can improve performance, it can also increase the invasion of privacy (e.g., tracking the user's location or capturing the user's voice and face).

3 DATA COLLECTION

We conducted an experiment in the wild to collect a large-scale dataset of smartphone touch and sensor data. The ethics board of ETH Zurich approved the experiment. During the experiment, we collected keyboard data, sensor data, and context data (e.g., foreground application) while participants used their smartphones in everyday life for approximately 70 days.

3.1 Participants

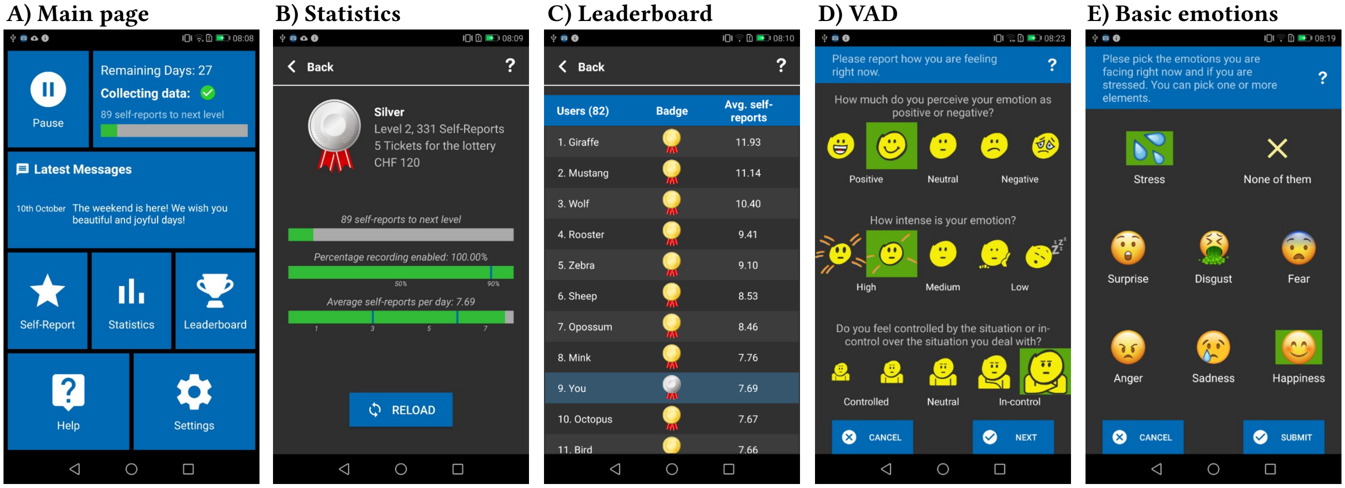
We recruited 82 participants (43 female, 39 male) between the ages of 18 and 43 (mean = 23.0 years, standard deviation SD = 3.64 years). Eleven participants were left-handed and seventy-one participants were right-handed. The majority of participants were students at the bachelor (61 participants), master (13), and Ph.D. (3) levels from ETH Zurich and the University of Zurich. We only considered participants that were German native speakers (due to the keyboard layout) and used typing-based applications (e.g., browsers and chat applications) daily on their smartphones. We recruited only participants using Android devices (Android 7 to 10). The participants used a variety of devices from different manufacturers: Samsung (31 participants), Huawei (21), Xiaomi (7), OnePlus (7), Sony (3), LG (3), Google (2), Nokia (2), Blackberry (1), Fairphone (1), HTC (1), Lenovo (1), Oppo (1), and Wiko (1). Participants actively engaged for an average of 72 days (SD = 2 days) in our experiment.

Compensation. We implemented an incremental reward system and the chance to win an additional price via a lottery similar to other works [32, 64, 69]. Participants were rewarded for their participation depending on their level of contribution and received between CHF 60 and CHF 120 for submitting an average of 3 and 6 self-reports per day, respectively. One participant was awarded an additional CHF 1000 from a lottery draw. See supplemental material for details about the reward system.

3.2 Apparatus

To collect a large-scale dataset in the wild, we developed an Android application consisting of three main components: 1) A user interface providing the participant information, control, and statistics of the experiment, 2) a data logging component for collecting sensor data, context data, and usage logs in the background, and 3) a keyboard that participants had to use during the experiment. In the following, we detail the three components.

Graphical user interface. The main page of the app (Figure 2A) provided information about the number of remaining days of participation and the number of self-reports until the next level is reached. Participants could manually start and pause the data recording. This mechanism enabled privacy when they did not want their data to be recorded. Participants were required to have recording enabled for at least 90% of the time to be eligible for compensation. Furthermore, the experimenter could send messages in the form of notifications to specific or all participants (e.g., information about the experiment or motivating messages). Participants could also access help information (i.e., help text, tutorial videos, and the information sheet) and change the settings (e.g., the storage location of the recorded data, finishing participation, and manual triggering synchronization with the server). The statistics page (Figure 2B)



Remaining Days: 27

Collecting data: ✓

89 self-reports to next level

Latest Messages

10th October The weekend is here! We wish you beautiful and joyful days!

Silver
Level 2, 331 Self-Reports
5 Tickets for the lottery
CHF 120

89 self-reports to next level

Percentage recording enabled: 100.00%

Average self-reports per day: 7.69

Users (82)	Badge	Avg. self-reports
1. Giraffe	●	11.93
2. Mustang	●	11.14
3. Wolf	●	10.40
4. Rooster	●	9.41
5. Zebra	●	9.10
6. Sheep	●	8.53
7. Opossum	●	8.46
8. Mink	●	7.76
9. You	○	7.69
10. Octopus	●	7.67
11. Bird	●	7.66

Please report how you are feeling right now.

How much do you perceive your emotion as positive or negative?

How intense is your emotion?

Do you feel controlled by the situation or in-control over the situation you deal with?

Plese pick the emotions you are facing right now and if you are stressed. You can pick one or more elements.

Figure 2: User interface of the Android application. A) Main page of the application. B) Statistics about self-reports and compensation. C) Leaderboard showing badges (level), average number of self-reports per day, and the rank. Users were assigned animal names to preserve anonymity. Self-reports captured valence, arousal, and dominance (D) and the basic emotions and stress (E). Selected items are highlighted with a green background.

provided information about the number of self-reports, average number of self-reports per day, percentage of enabled recording, and information about compensation and lottery tickets. Finally, in the leaderboard (Figure 2C), participants could track their rank in relation to the other participants in terms of the average number of self-reports per day. To maintain the privacy of the participants, we assigned an animal name to each participant.

Data recording. When the phone was unlocked, the Android application logged the following data in the background: sensor data (i.e., accelerometer, gyroscope, magnetometer, proximity sensor, light sensor, and step counter), device usage logs (e.g., foreground application, charging state, screen orientation, ringer mode, timezone changes, and audio mode), and activity predicted by the activity recognition API of Google (i.e., still, in-vehicle, on a bicycle, running, on foot, tilting, and walking). We did not use all sources of data in this work. For example, we discarded activity from our analysis because on some devices there was a substantial lag in recognition of activities. The data was uploaded in the background to a server several times during the day. The upload was initiated only when the phone was connected to Wi-Fi and all communication was encrypted.

Keyboard. Our application included a keyboard with a layout similar to the default German Android keyboard (Figure 3). Participants were required to use our keyboard during the study. From the keyboard, we recorded touch-related data (i.e., position and timestamps). In the modeling stage, we mapped the touch positions to the keys. Critically, data was not recorded when participants typed passwords, phone numbers, names, postal addresses, and e-mail addresses. The keyboard did not support auto-correction, auto-completion, and swiping. A pre-experiment questionnaire revealed that before the experiment, 79% of the participants had never used swiping, 71% had never used auto-correction, and 75% had never or only rarely used auto-completion.

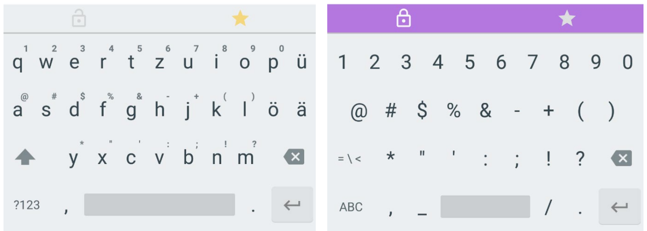


Figure 3: The keyboard included in our application. Two additional buttons in the top bar for enabling private mode (left button) and starting a self-report (right button). The left keyboard has private mode disabled and a self-report available (yellow star) and the right keyboard has private mode enabled (purple top bar) and no self-report available.

We extended the keyboard layout by two additional buttons at the top. The private mode button (left button in the top bar in Figure 3) allowed participants to pause the recording of data directly on the keyboard. By pressing the star button for 2 seconds (right button in the top bar in Figure 3), participants could fill in a self-report. Participants could also start a self-report using the self-report button on the main page of the app (Figure 2). Ninety-three percent of the submitted self-reports were started using the star button on the keyboard.

3.3 Labels from Self-Reports

To gather labeled data for our model, we asked participants to complete self-reports at regular intervals while using their smartphones. To quantify valence, arousal, and dominance, we adapted the Self-Assessment Manikin (SAM) [8] in terms of the dimensions it represents and the number of levels. The SAM is not applicable on smartphones due to its old-fashioned style and the space constraints of smartphone screens. Based on the work by Hayashi et al. [31] and feedback from participants in a pilot study (n = 17), we substituted the figures from the SAM with emojis and reduced the scale to five items (i.e., very low, low, neutral, high, very high). Emojis are commonly used in social networks and other communication applications. We believe that this familiarity made the self-reports more appealing and fostered a fast and accurate understanding of the experimental procedure by the participants.

Figure 2D shows an illustration of the emoji-based self-reports. For the valence dimension, we varied the emojis from a happy face (most positive) to a sad face (most negative). In the arousal dimension, the emojis varied from an awake emoji with large eyes (highest arousal) to a sleepy emoji (lowest arousal). Finally, for the dominance dimension, we increased the size of the emoji to portray control, similar to the SAM. Participants were also asked to select from a series of basic emotions (i.e., happiness, anger, sadness, surprise, disgust, and fear) and stress represented by different emojis (Figure 2E). To track complex emotions, participants were allowed to select all possible combinations of the basic emotions and stress. Participants could also select *None of them* if none of the provided items applied (in that case no other items could be selected).

Following the guidelines by Schmidt et al. [58] and Ghosh et al. [26], we used a combination of time-based and event-based schedules to trigger self-reports. A self-report became available (i.e., the star button on the keyboard turned yellow and started blinking) when four conditions were fulfilled. First, the participant typed at least 80 characters on the keyboard in the current session (we define a session as the period from unlocking the smartphone until it is locked again). Second, the smartphone was unlocked for at least 30 seconds in the current session. Third, between 30 minutes and 60 minutes elapsed since the last self-report was completed. Fourth, data recording was enabled (i.e., private mode on the keyboard was disabled). Depending on the number of average self-reports per day, the minimum amount of time between self-reports (third condition) was set to 30 minutes, 45 minutes, or 60 minutes. This helped to balance the number of self-reports per day and prevented participants from exaggerating submissions of self-reports. Once a self-report became available, participants could start the self-report until the smartphone was locked again (an additional margin of 10 seconds was provided in case participants accidentally locked the phone). We did not enforce a time limit for filling in the self-reports to avoid adding additional pressure on the rating, which could introduce a negative bias. All these parameters were decided based on results from a pilot study (n = 17).

3.4 Procedure

Figure 4A provides an overview of the procedure used in the experiment. After setting up the application (i.e., watching two tutorial videos and granting device permissions), participants conducted a typing test on their default keyboard used before the experiment and on the application keyboard before setting the application keyboard as the new default keyboard. The typing test consisted of six sentences in random order including two well-known pangrams (27–46 characters) [16, 48].

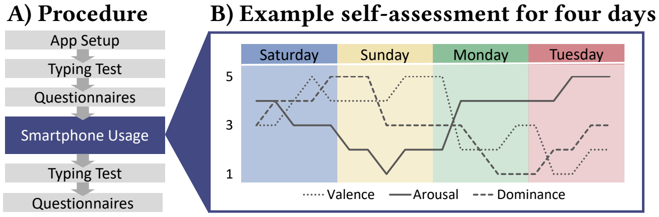


Figure 4: Overview of the different parts of the experiment. A) Overall experimental procedure. B) Changes in valence, arousal, and dominance of a selected participant during four consecutive days.

After the setup was completed, participants used their smartphones for 10 weeks in everyday life, filling in self-reports in regular intervals. We collected a total of 30,083 self-reports covering a large range of the valence-arousal-dominance space. Within the first week, we asked participants to fill in an online questionnaire on demographics and smartphone usage as well as the Patient Health Questionnaire [36] and the Big Five Inventory 2 [15] as measures of mental health and personality traits, respectively. At the end of the experiment, participants again typed the six sentences in random order on our keyboard and their default keyboard used before the experiment. Participants also completed an exit questionnaire on the self-reports regarding their level of understanding and the truthfulness and frequency of their responses. The exit questionnaire also probed their perception of the application's keyboard and smartphone usage. Finally, participants were asked to complete the Patient Health Questionnaire and the Big Five Inventory for a second time. See the supplemental material for more details about the experimental procedure.

Figure 4B depicts the changes in valence, arousal, and dominance during four days of one selected participant. The figure shows that valence and dominance were highest on Saturday and Sunday and decreased on Monday and Tuesday. Arousal showed an opposite pattern with lower values on Saturday and Sunday and higher values on Monday and Tuesday.

3.5 Dataset Validation

Smartphone usage. Figure 5A shows the smartphone usage over the hours of the day and each day of the week aggregated over all participants. Smartphone usage was lowest during the night (1 a.m. to 6 a.m.). During the day, smartphone usage was stable with a peak around 10 p.m. On Saturdays, participants used their smartphones least, whereas on Sundays usage was high throughout the day with peaks in the late afternoon and evening.

We collected data from 13,071 hours of smartphone usage. On average we recorded 3,533 sessions per user (SD = 1,624 sessions, max = 8,861, min = 1,060). On average a session lasted for 183 s (SD = 532 s) and we recorded an average of 47 keystrokes per session (SD = 174 keystrokes). The mean break between sessions was 11 min (SD = 72 min).

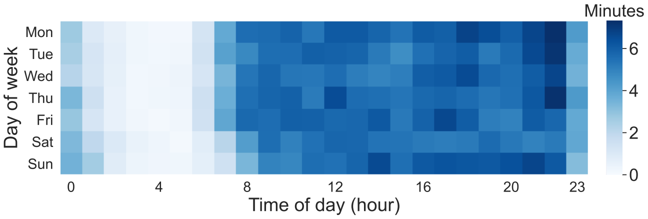


Figure 5: The distribution of average smartphone usage for the days of the week and the times of the day aggregated over all participants.

Self-reports. We collected a total of 30,083 self-reports for valence (669 very low, 2,767 low, 8,071 neutral, 14,642 high, 3,934 very high), arousal (1,643 very low, 5,260 low, 12,572 medium, 7,591 high, 3,017 very high), dominance (1,866 very controlled, 3,256 controlled, 12,823 neutral, 8,089 in-control, 4,049 very in-control) and the basic emotions of anger (selected 1,208 times), happiness (16,425), sadness (1,918), surprise (786), fear (1,628), disgust (515), and stress (4,795). On average, a participant submitted 402 self-reports (SD = 154 self-reports, min = 44, max = 835) totalling 5.64 self-reports per day on average (SD = 2.12 self-reports). Participants also spent an average of 6.76 s (SD = 27.27 s) filling in the self-reports. In addition, an average of 154 keystrokes (SD = 177 keystrokes) and 40 s (SD = 266 s) passed since the start of the session until a self-report was triggered.

We also performed a series of correlations to investigate the relationship between the valence, arousal, and dominance ratings and the basic emotions and stress. Table 1 presents the results for each of these correlations. The effect sizes are largest for valence and smallest for arousal. Notably, these results are a close match to the correlations we found for data collected in a previous laboratory experiment [67]. We found the same directions for the correlations but smaller effects sizes.

Russell and Mehrabian [54] provide a correspondence between valence, arousal, and dominance and the basic emotions based on laboratory experiments. In Table 2, we compare these values to the mean values obtained from the self-reports collected in our experiment. In Russel and Mehrabian's model, the affective dimensions (i.e., valence, arousal, and dominance) spanned the interval $[-1, 1]$. Thus, we mapped the self-reports collected in our experiment to the same interval to obtain a proper measure for comparison. The self-reports collected in our experiment closely match the correspondences found by Russel and Mehrabian. In contrast to Russel and Mehrabian, the mean values for valence, arousal, and dominance are smaller in our data. These differences may be related to the fact that we performed the experiment in the wild without using emotion-eliciting situations as stimuli. Notably, for anger, surprise, and disgust, the mean dominance value shows a reversed sign compared to Russel and Mehrabian's model. For stress, the mean values of all three dimensions (i.e., valence, arousal, and dominance) are around zero. It is known that stress can be positive or negative with different intensity levels [21, 22], thus potentially, positive and negative ratings cancel each other out leading to a mean close to zero.

Table 1: Effect sizes of the Pearson correlations between valence, arousal, and dominance and the basic emotions and stress. Asterisks denote correlations that survived Bonferroni correction (p = 0.0024).

	Anger	Happiness	Sadness	Surprise	Fear	Disgust	Stress
Valence	-0.30^*	$+0.55^*$	-0.37^*	-0.006	-0.22^*	-0.14^*	-0.24^*
Arousal	$+0.05^*$	$+0.27^*$	$+0.01$	$+0.04^*$	$+0.02^*$	-0.008	$+0.004$
Dominance	-0.16^*	$+0.17^*$	-0.18^*	-0.04^*	-0.17^*	-0.10^*	-0.20^*

Table 2: Mean values for valence, arousal, and dominance for the six basic emotions and stress. Results from our study are compared to the correspondences derived by Russell and Mehrabian [54]. All measurements are mapped to the interval $[-1, 1]$. Values in brackets denote standard deviation.

	Valence		Arousal		Dominance	
	Russel	Ours	Russel	Ours	Russel	Ours
Anger	−0.43	−0.36 (0.44)	+0.67	+0.19 (0.52)	+0.34	−0.24 (0.56)
Happiness	+0.76	+0.54 (0.33)	+0.48	+0.20 (0.50)	+0.35	+0.24 (0.51)
Sadness	−0.63	−0.34 (0.50)	+0.27	+0.10 (0.54)	−0.33	−0.21 (0.51)
Surprise	+0.40	+0.29 (0.51)	+0.67	+0.20 (0.54)	−0.13	+0.04 (0.52)
Fear	−0.64	−0.12 (0.51)	+0.60	+0.12 (0.53)	−0.43	−0.21 (0.53)
Disgust	−0.60	−0.18 (0.52)	+0.35	+0.04 (0.55)	+0.11	−0.23 (0.57)
Stress	–	+0.05 (0.49)	–	+0.08 (0.51)	–	−0.08 (0.51)

Keyboard. We recorded an average of 7,669 keyboard sessions per user (SD = 3,341 sessions, min = 2,255, max = 18,445). We define a keyboard session as the time from opening to closing the keyboard. The ten most used keys were the space bar (13.5%), delete key (11.6%), E (8.3%), I (5.5%), A (5.1%), S (4.9%), N (4.7%), H (4.4%), T (4.0%), and R (3.8%).

See supplemental material for further analysis of the smartphone usage, self-reports, and keyboard usage. Having collected and validated this corpus of 82 participants' smartphone usage and self-reports over 10 weeks, we now describe our method for affective state prediction.

4 METHOD

Our model predicts affective states based on keyboard and inertial sensor data collected during smartphone usage (Figure 6). We encoded these data in two-dimensional heat maps and trained convolutional neural networks to automatically extract meaningful features from the heat maps. For the classification of affective states, we then added a fully connected classification layer. The paragraphs below provide details on every step of our model.

4.1 Heat Maps

From the smartphone data collected in the wild, we extracted two types of two-dimensional heat maps providing intuitive and compact visualizations. First, we created keystroke heat maps that encoded typing characteristics of bigrams (i.e., key combinations) of consecutive keystrokes on the smartphone keyboard. Second, we created sensor heat maps encoding the distribution of the gyroscope and linear acceleration measurements.

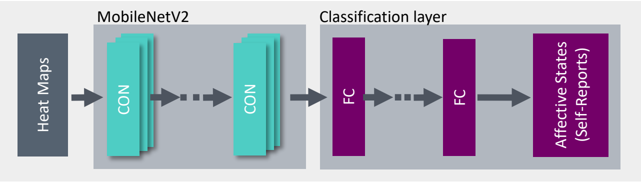


Figure 6: Overview of the main steps of our model. A convolutional neural network (MobileNetV2 [56]) was trained on heat maps created from smartphone keystroke and inertial sensor data. For classification of the affective states, the features learned by MobileNetV2 were used as input to fully connected layers.

Keystroke heat maps. A keystroke $k_i = (x, y, t_{\text{down}}, t_{\text{up}})$ is defined by the coordinates (x, y) on the screen as well as t_{down} and t_{up} providing the timestamp in milliseconds of pressing (touch down) and releasing (touch up) the key, respectively. A text $K = [k_1, \ldots, k_n]$ consists of n keystrokes. Based on the raw input data and motivated by commonly used typing characteristics from other works [12, 67], we extracted three keystroke metrics. First, "up-down" measures the time of moving from one key to the next key (up-down = $t_{i+1,\text{down}} - t_{i,\text{up}}$). Second, "down-down" measures the time of moving between keys as well as the hold time of the first keystroke (down-down = $t_{i+1,\text{down}} - t_{i,\text{down}}$). Third, "down-up" considers the time of moving between keys and the hold time of the first and second keystroke (down-up = $t_{i+1,\text{up}} - t_{i,\text{down}}$). All three keystroke metrics were normalized by the distance between the keys.

To exclude breaks during typing, we chose a threshold of one second between keystrokes. We motivate this threshold by the longest median time per character (400 milliseconds) [10] and the fact that median + 3 $*$ median absolute deviation $= 0.9$ s [39]. By choosing a conservative threshold of one second, we retain delays that are part of natural typing behavior.

Using a window of 80 keystrokes before the self-reports, we aggregated the keystroke metrics into two-dimensional heat maps covering all possible bigrams of characters (a–z including umlauts ä, ö, and ü) and special keys (i.e., delete, space, symbol, shift, return, period, comma, question mark, and exclamation point). In total, we considered 38 keys. We encoded all possible key combinations in a 38×38 heat map H. The rows and columns encode all 38 keys taken into consideration using a centralized alignment of the keys. More frequently used keys in the English language [63] and German language [5, 6] are placed in the middle of the heat map (the space bar is considered to be the most frequent key and the exclamation point and q are the least frequent keys). The first and second keystroke in a bigram is encoded in the row and column, respectively. For example, $H(a, p)$ contains the keystroke metric (i.e., up-down speed, down-down speed, or down-up speed) calculated from the keys a (row) and p (column) of the bigram ap. For multiple occurrences of the same bigram, we averaged all the values of the corresponding cell in the heat map H. In addition, all heat maps were standardized based on the mean heat map during a baseline typing period.

Figure 7A shows an example of an extracted heat map encoding up-down speed. The colors in the heat maps are for visualization purposes only. In our model, we used only one value per pixel. It is visible that the highest up-down speed is concentrated in the bigrams (*space bar, D*), (*E*, *N*), and (*space bar*, *shift key*). See supplemental material for examples of heat maps encoding down-down speed and down-up speed.

Sensor heat maps. For creating two-dimensional sensor heat maps, we extracted the rate of rotation and linear acceleration of smartphones from the inertial sensors. As a preprocessing step, we temporarily aligned the signals and converted the sampling rate to 100 Hz (i.e., downsampling or upsampling), which provided a noise reduction as a positive side effect. We also clipped linear acceleration to $4\,\text{g} = 39.2\,\text{m}\,\text{s}^{-2}$ as 98% of the sensor data were below 4 g. We clipped the gyroscope measurements at $5\,\text{rad}\,\text{s}^{-1}$ because 99% of the measurements were below this threshold. In addition, we only considered linear acceleration and gyroscope measurements greater than $0.02\,\text{m}\,\text{s}^{-2}$ and greater than $0.003\,\text{rad}\,\text{s}^{-1}$, respectively. We chose these thresholds because in a pilot study 95% of the sensor measurements were below these thresholds when the smartphones were lying flat on a table. Thus, we excluded noise inherent to the sensors.

We encoded the three axes combinations into separate heat maps: linear acceleration along the x-axis & rate of rotation around the z-axis, linear acceleration along the y-axis & rate of rotation around the x-axis, and linear acceleration along the z-axis & rate of rotation around the y-axis. We chose these axes combinations because they reflect typical motion sequences. Using a window of 30 seconds before the filled in self-reports, we binned the absolute sensor values into logarithmically spaced bins and counted the number of values in each bin. We chose a logarithmic scale because the absolute sensor measurements are exponentially distributed. Thus, when taking the logarithm, the measurements become approximately normally distributed. Moreover, we believe that the distinction of smaller values is more important than larger values so that also micromotions can be adequately exploited [41]. For the heat maps, we used a resolution of 96 × 96 (i.e., 96 bins in each dimension), because it is divisible by a multiple of two, which is advantageous for the spatial downsampling in a convolutional neural network and it provides a sufficiently high resolution. We standardized all heat maps based on the mean heat map during a baseline period.

Figure 7B shows an example of an extracted heat map for linear acceleration along the x-axis and the rate of rotation around the z-axis (the colors are for visualization purposes only). See supplemental material for examples of heat maps encoding the other axes combinations.

4.2 MobileNetV2

For the keystroke and sensor heat maps, we stacked the three types of heat maps into three channels. To extract meaningful features from the keystroke (38×38×3) and sensor heat maps (96×96×3), we employed a particular type of convolutional neural network called MobileNetV2 [56]. Affective labels are typically sparse and labeled datasets are relatively small for training a network for predicting affective states, making it prone to overfitting. Using a smaller but expressive network such as MobileNetV2 counters this effect.

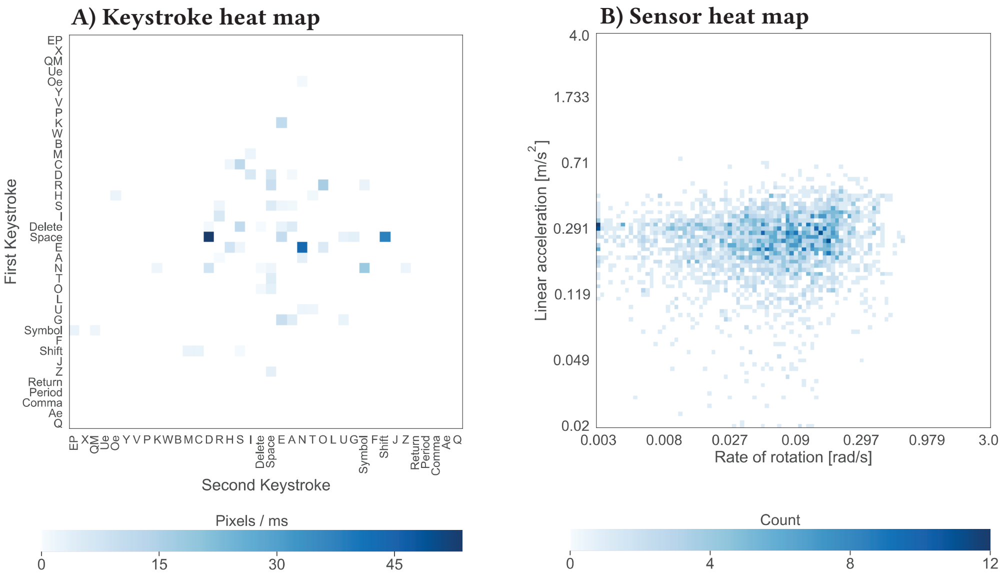


Figure 7: Example of a keystroke heat map extracted from 80 keystrokes (A) and a sensor heat map extracted from 30 seconds of the gyroscope and linear acceleration measurements (B). Abbreviations: exclamation point (EP), question mark (QM), ü (Ue), ö (Oe), and ä (Ae). Color saturation indicates the average up-down speed between consecutive keystrokes (A) and the number of sensor measurements for the combinations of linear acceleration along the x-axis and the rate of rotation around the z-axis (B). The colors are for visualization purposes only.

MobileNetV2 is a network optimized for low memory consumption and high execution speed and is parameterized to meet the resource constraints of mobile devices [55, 71].

The basic building block of MobileNetV2 is the bottleneck depth-separable convolution with residuals which consist of three operations [56]. First, a 1×1 convolution layer expands the number of feature maps. Second, the depthwise convolution applies a single filter to each feature map. Finally, a pointwise convolution with a kernel size of 1×1 is used to combine the outputs of the depth-wise convolutions (i.e., linear combinations of the feature maps) reducing the number of feature maps, and thus the amount of data flowing through the network. The factorization of the convolution into depthwise and pointwise convolutions reduces the computational cost and model size. In addition, the input and output to the basic building block are connected with a residual connection which enables faster training and better accuracy [56].

MobileNetV2 was developed for images with a resolution of $224 \times 224 \times 3$ and consists of five downsampling layers (i.e., a stride of two). For the keystroke heat maps (38×38×3), we disabled the first three downsampling layers (i.e., setting the stride to one). For the sensor heat maps ($96 \times 96 \times 3$), we disabled the first downsampling layer. This modification of the network was successfully used on the CIFAR10 dataset (containing images with a resolution of $32 \times 32 \times 3$) [3]. Input data is commonly scaled before training. We used Min-Max scaling of the heat maps to the range $[-1, 1]$.

4.3 Classification

To remove noise and foster balanced classes, we simplified the valence, arousal, and dominance measures to three classes (low $\in [1, 2]$, medium $\in [3, 3]$, and high $\in [4, 5]$) of valence (3,436, 8,071, and 18,576 self-reports), arousal (6,903, 12,572, 10,608), and dominance (4,870, 12,066, 11,221).

We took advantage of the learned features from the convolutional neural network by adding a classification network. The final output of the convolutional neural network was passed through a global average pooling layer and a fully connected layer with softmax activation. We aggregated the keystroke and sensor heat maps by stacking the output of the global average pooling layer of the pre-trained networks of the individual heat maps. For the combination of the sensor and keystroke heat maps, we used a fully connected layer with 2048 units between the global average pooling and the softmax layer to foster the learning of mixtures of the extracted features from the heat maps. Due to the prevalent class imbalance, we used balanced class weights to give smaller classes more weight. We trained the whole network on the labeled data (heat maps and corresponding affective states) using backpropagation minimizing the cross-entropy loss using 80 epochs and a batch size of 128. We optimized the networks using stochastic gradient descent with a momentum of 0.9 and a cyclical learning rate using an exponential decay ($\gamma = 0.99994$) with a minimum and maximum learning rate of 10^{-5} and 10^{-2}, respectively [62]. We implemented all networks using the Keras framework with TensorFlow^{TM} back-end.

5 RESULTS

We used the data from our data collection to evaluate our model. We evaluated the performance of our model in terms of accuracy (chance level is 0.33 for three classes and 0.5 for two classes), micro-averaged AUC (chance level is 0.5), and macro-averaged

Table 3: Performance for the prediction of three classes (low, medium, high) of valence, arousal, and dominance. $\mathrm{AUC}_{\mathrm{micro}}$ and $\mathrm{AUC}_{\mathrm{macro}}$ represent micro-averaged AUC and macro-averaged AUC, respectively. The chance level of accuracy and AUC is 0.33 and 0.5, respectively.

Dimension	Heat Map	$\mathrm{AUC}_{\mathrm{micro}}$	$\mathrm{AUC}_{\mathrm{macro}}$	Accuracy
Valence	Keystrokes	0.82	0.76	66%
	Sensors	0.79	0.73	63%
	Combination	0.83	0.78	70%
Arousal	Keystrokes	0.81	0.80	63%
	Sensors	0.83	0.82	64%
	Combination	0.85	0.84	65%
Dominance	Keystrokes	0.82	0.79	67%
	Sensors	0.81	0.79	63%
	Combination	0.84	0.82	68%

AUC (chance level is 0.5). Micro-averaged AUC aggregates the contributions of all classes by considering each element in the label indicator matrix as a label. To account for class imbalance, the macro-averaged AUC averages the class-wise AUCs. We evaluated our model using leave-one-user-out cross-validation to ensure that data of a user in the test set is not used for training. On the training data, we used all available heat maps to compute the baseline heat map. For heat map n of a user in the test set, we used the $n-1$ heat maps to compute the baseline heat map (i.e., the baseline gradually improves the more the user types).

5.1 Affective State Prediction

Table 3 reveals the performance of our model. See supplemental material for additional metrics.

Classification performance. Using the combination of keystroke and sensor heat maps, for valence, arousal, and dominance, the values for micro-averaged AUC (0.83, 0.85, 0.84) are slightly higher than for macro-averaged AUC (0.78, 0.84, 0.82). When considering the percentage of the most frequent class as baseline (valence = 62%, arousal = 42%, and dominance = 40%), the accuracy is above the baseline for valence (70%), arousal (65%), and dominance (68%). Figure 8 shows the confusion matrices for valence, arousal, and dominance evaluated on the combination of the keystroke and sensor heat maps. Often neighboring classes are confused with each other. For all three dimensions, the high class was most often confused with the medium class and vice versa. For arousal (Figure 8B) and dominance (Figure 8C), the low class was often mispredicted as the medium class. In contrast, for valence (Figure 8A) the low class was more often confused as the high class, which may be attributed to the class imbalance.

Heat map comparison. The keystroke heat maps perform slightly better than the sensor heat maps for valence (+0.03 AUC) and dominance (+0.01 AUC). In contrast, for arousal, the sensor heat maps outperform marginally the keystroke heat maps (+0.02 AUC). The combination of the two types of heat maps provides only a marginal improvement in performance (up to 0.04 AUC).

Table 4: F_1-scores for complex emotions formed from two basic emotions and stress. We treat the presence of the complex emotion as the positive class. The number of self-reports for each complex emotion is given in brackets.

	Anger	Happiness	Sadness	Surprise	Fear	Disgust	Stress
Anger	–	0.76 (153)	0.30 (384)	0.24 (101)	0.28 (222)	0.19 (126)	0.46 (429)
Happiness	0.76 (153)	–	0.78 (402)	0.76 (394)	0.77 (518)	0.76 (104)	0.80 (1,563)
Sadness	0.30 (384)	0.78 (402)	–	0.31 (89)	0.37 (418)	0.31 (119)	0.49 (561)
Surprise	0.24 (101)	0.76 (394)	0.31 (89)	–	0.31 (98)	0.23 (53)	0.48 (203)
Fear	0.28 (222)	0.77 (518)	0.37 (418)	0.31 (98)	–	0.28 (105)	0.47 (966)
Disgust	0.19 (126)	0.76 (104)	0.31 (119)	0.23 (53)	0.28 (105)	–	0.46 (214)
Stress	0.46 (429)	0.80 (1,563)	0.49 (561)	0.48 (203)	0.47 (966)	0.46 (214)	–

5.2 Basic Emotion and Stress Prediction

Our model achieved a performance of 90% (0.77 AUC) for anger, 75% (0.81 AUC) for happiness, 93% (0.82 AUC) for sadness, 95% (0.86 AUC) for surprise, 93% (0.85 AUC) for fear, 97% (0.86 AUC) for disgust, and 82% (0.83 AUC) for stress. The differences between AUC and accuracy are due to class imbalance. The basic emotions can also be blended to form complex emotions (e.g., the combination of happiness and sadness results in melancholy) [60]. Table 4 presents the F_1-scores and the number of self-reports (in brackets) for the first-order complex emotions. We evaluated the performance by combining the predictions of the individual models pertaining to the respective basic emotions. Complex emotions formed by happiness were recognized well. Interestingly, the combination of stress and happiness (i.e., positive stress) occurred most often (1,563 times) and was recognized accurately (F_1-score of 0.80). Altogether, we can conclude that our model is also predictive for basic emotions and stress and may even be predictive for complex emotions.

5.3 Window Size Analysis

The results presented in Table 3 are based on keystroke heat maps extracted from 80 characters and sensor heat maps extracted from 30 seconds (i.e., 3,000 sensor values). On average participants typed 74 characters (SD = 43 characters) in the 30-second window before filling in the self-report. Thus, the two types of windows for extracting the keystroke and sensor heat maps are a close match. Nevertheless, considering longer periods can be beneficial for the classification performance, because the model has more data available. As such, we evaluated our model on heat maps extracted on larger windows ranging from 2 minutes to 30 minutes (the minimum time between self-reports was 30 minutes). To analyze different window sizes, we relaxed the constraint of a fixed number of characters (i.e., 80 characters) and sensor measurements (i.e., 3,000 samples). Thus, the heat maps contained a different number of characters and sensor measurements depending on the number and the duration of sessions in the corresponding window. Figure 9A shows the macro-averaged AUC for valence, arousal, and dominance for the different window sizes. Peak performance is reached with a window size of 5 minutes for valence (0.80 AUC), arousal (0.86 AUC), and dominance (0.83 AUC). Further increasing window sizes leads to a substantial drop in the performance for all three dimensions. Overall, performance improvements are only marginal for all three dimensions (up to 0.02 AUC).

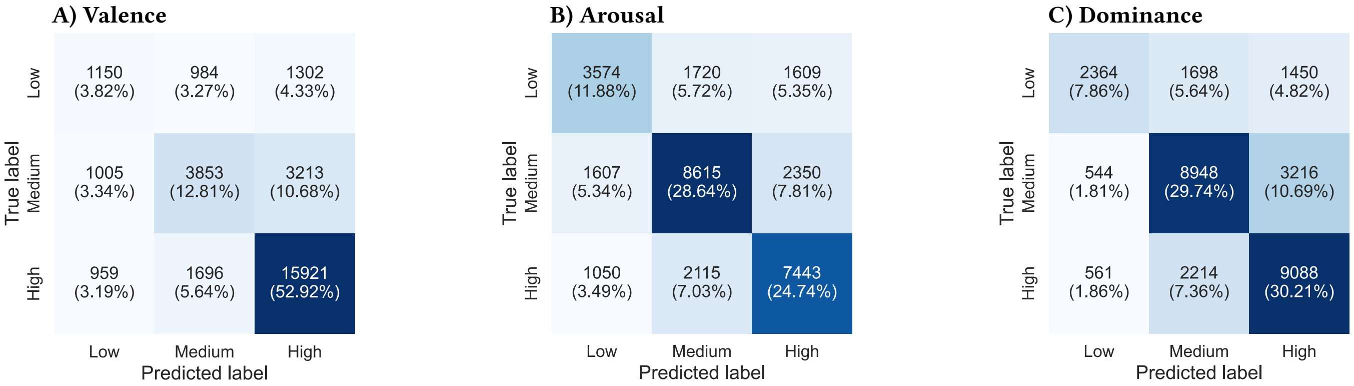


Figure 8: Confusion matrices for classifying three levels (low, medium, high) of A) valence, B) arousal, and C) dominance. Confusion matrices were calculated from predicted self-reports using the combination of keystroke and sensor heat maps.

5.4 Personalization

Affective states can be individual and can reflect idiosyncrasies in users. While there may be similar typing and sensor patterns between users characterizing similar affective states, leveraging user-specific data can improve the performance of the model. To investigate the extent of performance gain for a participant with N filled in self-reports, we used the first n self-reports to fine-tune the whole model using five epochs and predicted then the $N - n$ remaining self-reports. Figure 9B reveals the macro-averaged AUC in terms of n (i.e., the number of self-reports used to fine-tune the model). Fine-tuning on only 10 self-reports provides already a slight performance improvement (up to 0.02 AUC). The performance improvement plateaus at around 40 to 60 self-reports used for fine-tuning. The performance improvement is large for valence (+0.07 AUC), reaching a performance of 0.85 AUC. The performance improvements for arousal (+0.05 AUC) and dominance (+0.04 AUC) are slightly smaller, leading to a performance of 0.89 AUC and 0.86 AUC, respectively.

5.5 Ablation Study

A model can only be as good as the data that supports it. If the data (i.e., the heat maps) show clear patterns, we can achieve a well-performing model with only a little amount of data. On the other hand, if the data is noisy, a much larger dataset is needed to achieve the same performance. In our experiment, we collected a homogeneous dataset consisting of mostly bachelor and master students around the age of 23. Thus, we hypothesize that typing and smartphone usage behavior were similar among participants and less training data is needed to achieve a good performance for the classification of affective states. To test our hypothesis, we conducted an ablation study by training the model on data from a subset of the participants. To accomplish this, we selected a percentage of participants at random in each of the 82 training sets of the leave-one-user-out cross-validation. Figure 9C shows the macro-averaged AUC for different percentages of users in the training data. Performance plateaus at around 60% (49) of participants for valence (0.77 AUC) and around 80% (66) of participants for arousal (0.83 AUC) and dominance (0.81 AUC). Thus, a subset of the users (i.e., between 60% and 80%) is enough to achieve a performance close to the performance reached when using data from all the users. By linear extrapolation, we can roughly predict that with the double amount of participants (i.e., 164 participants), we could come close to a performance of around 0.83 AUC for valence, 0.89 AUC for arousal, and 0.87 AUC for dominance.

5.6 Runtime Analysis

We conducted a runtime analysis of the different parts of our model. Our computing environment consisted of an Intel® Xeon® CPU E5-2630 v4 @ 2.20GHz and an NVIDIA GeForce® GTX 1080 Ti. The prediction of a new data point consisted of extracting keystroke heat maps (mean = 0.84 s, SD = 0.03 s) and sensor heat maps (mean = 0.23 s, SD = 0.18 s), followed by the convolutional neural network and the fully connected layer for the classification of the affective states (mean = 0.0036 s, SD = 0.004 s). Summing up these values leads to a prediction time of 1.07 seconds if considering both types of heat maps. If only the keystroke heat maps and sensor heat maps are used, the prediction time amounts to 0.84 seconds and 0.23 seconds, respectively. The higher runtime for creating the keystroke heat maps compared to the sensor heat maps is due to the preprocessing (i.e., mapping touch positions to keys, calculating the metrics between the key pairs, and sanity checks).

6 DISCUSSION

We presented a model that can be used on mobile devices for predicting valence, arousal, and dominance, the basic emotions, and stress. We believe that the ability to run our model on smartphones can improve user experience, provide ubiquitous access to affective state predictions, and can be beneficial for security and privacy. To that end, we used the MobileNetV2 network that can also counter possible overfitting effects due to its smaller size. More complex networks (e.g., VGG or ResNet) could increase the model performance on a PC but might not run on mobile devices due to the increased memory footprint [7].

Heat maps. The predictions of our model were based on heat maps extracted from keystroke data and sensor data collected during smartphone usage in real-world environments. We found that

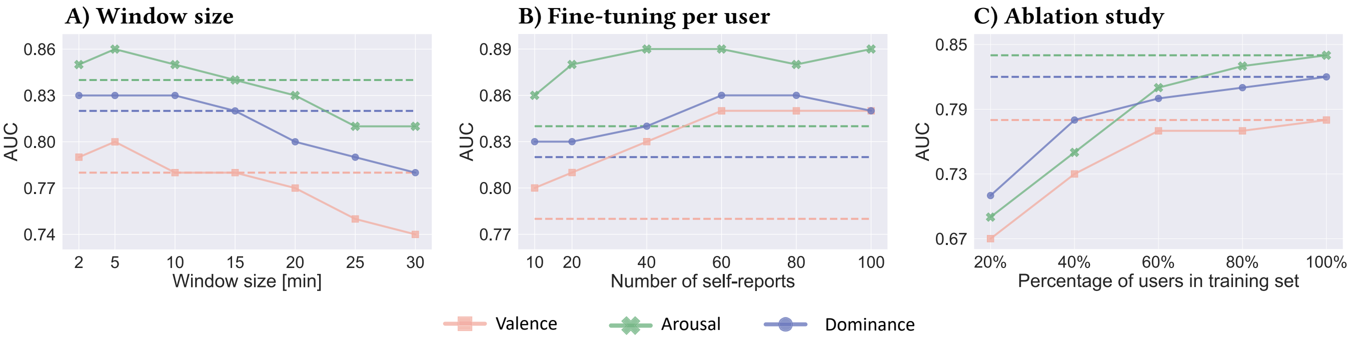


Figure 9: Macro-averaged AUC for the classification of three levels (low, medium, high) of valence, arousal, and dominance using A) different window sizes for the heat map extraction, B) fine-tuning the network per user on varying number of self-reports, and C) different number of users in the training set. The dashed lines represent the baseline performance (Table 3).

both types of heat maps are capable of accurately predicting valence, arousal, and dominance. In particular, the sensor heat maps showed the best performance for predicting arousal (0.83 AUC), while the keystroke heat maps were most predictive for valence (0.82 AUC) and dominance (0.82 AUC). These results are in line with the findings by Olsen and Torresen [47], reporting that accelerometer data is more predictive for arousal than valence.

The keystroke heat maps provide an intuitive and compact visualization of typing patterns compared to the heat maps presented in our previous work [67]. From the keystroke heat maps, typed text can be partially recovered. Nevertheless, we only encoded information about consecutive keystrokes but ignored time-series information. Thus, a full reconstruction of the typed text from the heat map alone is difficult, especially when considering a sequence of words with multiple occurrences of the same bigrams. On the other hand, the recording of sensor data is less privacy-invasive. Sensor data is also less prone to bias than typing data (i.e., users might be more aware of their typing behavior than of their smartphone holding behavior). In addition, the runtime to extract sensor heat maps is substantially lower than that of keystroke heat maps. In conclusion, we suggest the sensor heat maps as most appropriate for the prediction of affective states in real-world applications.

We also found that typing speed significantly differs depending on the location of the key pairs on the keyboard (i.e., key combinations typed with both thumbs were significantly faster). By subtracting the baseline heat map per participant, we correct for these differences in typing speed. In addition, the convolutional neural network is capable of learning potential keyboard layout-based bias in typing speed.

Window size. We used 80 characters and 30 seconds of accelerometer and gyroscope data to extract the keystroke and sensor heat maps, respectively. In practice, 30 seconds of sensor data can be stored continuously in the background of the smartphone until the user has typed 80 characters. If only sensor data is used for the prediction, the restriction does not apply anymore and predictions are possible more often (i.e., also when users did not type). For larger window sizes it takes longer until a prediction is possible. We showed that peak performance is reached with a window size of 5 minutes (+0.02 AUC). A potential explanation for the performance improvement is that with larger window sizes the model can implicitly gauge the total time spent on the smartphone from the sparseness of the heat maps (i.e., a sparser heat map implies a less active user). On the other hand, if the window size becomes too large, the heat maps become too dense and noisy which degrades the performance.

Personalization of the model. We also showed that fine-tuning our model per participant can substantially improve the performance for valence (+0.07 AUC), arousal (+0.05 AUC), and dominance (+0.04 AUC). Peak performance for personalizing the model was reached using the first 40 to 60 self-reports of the participants. After the start of the experiment, it took some time until the participants got used to filling in the self-reports (i.e., the variance tended to be larger for the first self-reports). Thus, 40 to 60 self-reports were necessary for fine-tuning the model to learn the stable self-report pattern of the participants and reaching peak performance. A reason for this performance improvement is that the network can learn keystroke and sensor patterns typical for a specific participant. Moreover, for participants that reported one class (e.g., high valence) more often than other classes, the model can shift towards predicting this class with a higher probability. Other researchers reported performance improvements for personalized models of up to 6.3% for predicting valence [14] and 17.6% for predicting arousal [25].

Comparison with prior work. A direct comparison of the performance of our model is difficult due to differences in the measurement of affective states and experimental setups. Olsen and Torresen [47] reported slightly higher accuracy for the prediction of arousal (+10%) but lower accuracy for valence (−19%). In contrast to our work, they captured accelerometer data during sequences of walking from only 10 participants. In comparison to Ruensuk et al. [53], our model performed similarly for valence (+1%) and arousal (−7%), although these authors predicted only two levels of valence and arousal.

In comparison to our previous work [67], the performance of our model in terms of macro-averaged AUC was superior for arousal (+0.04 AUC) and dominance (+0.02 AUC) but inferior for valence

(−0.05 AUC). The inferior performance for valence may be attributed to class imbalance (11% low, 27% medium, 62% high) or the data collection in the wild. With regard to the basic emotions, our model performed better for surprise (+0.1 AUC) and stress (+0.03 AUC) but was inferior for anger (−0.07 AUC), happiness (−0.07 AUC), and sadness (−0.05 AUC). In our previous work, we used data from a laboratory experiment and three minutes of data to extract heat maps. In contrast, in this work, we used only 30 seconds of data. Laboratory experiments provide more control while in the wild experiments provide more ecological validity (i.e., less control) and offer the possibility of collecting larger datasets. In addition to most other works, we also considered dominance, which we believe is an important dimension of affective states.

6.1 Implications and Potential Applications

The ability to predict affective states has a broad range of applications. In the following, we detail two applications that could take advantage of affective state predictions from our model.

Mental health. Affective states are connected to physiological and mental health [9]. As such, recognizing a person's affective state can assist with the treatment of health problems by either calling a psychologist or by the intervention of the system with the user itself [52]. For example, Woebot [70] is a smartphone-based therapeutic chatbot and tracks the mood of the user from chat conversations with the user. The chatbot then tries to increase the mood of the user by adapting the conversation based on the inferred mood. On the other hand, MoodPrism [51] presents a colorful summary of emotional health based on daily self-reports capturing the mood of the users. The feedback from MoodPrism can help users identify patterns in their feelings, which in turn can improve wellbeing [50]. Similarly, Mood Meter [45] is a smartphone application helping to identify moods throughout the day using self-reports. These apps can benefit from affective state predictions from our model to help improve emotion recognition accuracy and replacing self-reports.

Personalized recommendations. Personalized recommendations have become ubiquitous on smartphones providing users a personalized experience that can increase motivation and commitment [44]. Personalized recommendations can be explicit (e.g., music and video recommendations) or implicit (e.g., personalized search results and news article recommendations). Considering the affective states of users for personalized recommendations can improve the recommendations qualitatively for different application domains. For example, Mizgajski et al. [44] developed personalized news article recommendations based on self-reported emotions of the users during news browsing and reading activities. In particular, they showed that incorporating pleasant emotions improved the recommendation quality substantially. A similar system was proposed for music recommendations on Spotify tailored to the emotion of the user [1]. Both systems can benefit from emotion predictions from our model, eliminating the need for user interaction to gauge the emotional state.

6.2 Limitations

We acknowledge potential limitations of the approach presented in this paper. We analyzed the runtime of our model on a computer. On a mobile device, the runtime of extracting the heat maps and the inference time of the network might be slightly higher. To keep runtime low, the model could be deployed on a server. Another limitation is the number of data required until a prediction can be made. In our experiment, we ensured that we have enough sensor and keystroke data available by unlocking a self-report when the user typed at least 80 characters and used the smartphone for at least 30 seconds. In practice, using only the sensor heat maps for the prediction of affective states relaxes the constraint of typing 80 characters while providing a similar performance. In addition, by allowing the participants to fill in self-reports only every 30 minutes, we could have missed finer-grained changes in affective states. Furthermore, due to the requirement of having typed at least 80 characters for unlocking a self-report, the windows used for creating the sensor heat maps always contained keystrokes. As such, predicting affective states using sensor heat maps during periods with no keyboard input (e.g., watching videos on YouTube) requires further research. Finally, the study was restricted to a population consisting of German-speaking bachelor and master students, and future studies are required for a proper generalization of our model to the wider population.

6.3 Future work

Future work could improve the model by leveraging time-series information of the self-reports and the collected data using an LSTM architecture. Although such an approach showed promising results for predicting mood [33], our initial experiments did not improve the performance of our model. Furthermore, context data (e.g., ambient light, application type, daytime, and weekday) could be exploited to improve the model. Models relying on context data showed to be promising for affective state prediction [35] but at the cost of an increased invasion of privacy. Here again, our first results did not reveal performance improvements when enriching our model with context data. To improve the input representation of our model, the heat maps could be expanded to the three-dimensional space to capture movement in space. Similarly, typical smartphone moving patterns can be analyzed based on the gyroscope and accelerometer readings and then used to improve our heat maps. Finally, we also captured the personality traits and levels of depression of the participants in the study. In a next step, we will build a predictive model for personality traits and depression level based on our collected smartphone data. Recently, personality was successfully predicted based on touchscreen-based interactions [37] and smartphone accelerometer data [23].

7 CONCLUSION

In this paper, we presented a model for predicting affective states (up to 0.85 AUC), basic emotions (up to 0.86 AUC), and stress (0.83 AUC) based on two-dimensional heat maps extracted from users' touch events on smartphone keyboards and the signals from the inertial sensors. We evaluated our model with data collected in the wild from 82 participants over 10 weeks. By fine-tuning the network per participant, we achieved substantial performance improvements (up to +0.07 AUC). We also showed that we achieve a similar performance using sensor heat maps alone without any keystroke heat maps, which is beneficial for privacy and runtime efficiency (0.23 seconds vs. 0.84 seconds).

The novelty of our contribution consists of our model that processes heat maps of touch and sensor data to provide accurate assessments of affective states on different types of mobile devices. The keystroke heat maps extracted from touch events, provide an intuitive and compact visualization of the typing characteristics. In addition, the keystroke heat maps allowed us to investigate the distribution of keystroke pairs in relation to the measured affective states (e.g., more frequent occurrence of keystroke pairs with a backspace when experiencing negative emotions). By extracting heat maps from sensor signals instead of directly processing raw sensor data, our approach takes into account the relationship between acceleration and rotation and provides a less privacy-invasive way for affective state prediction compared to keystroke heat maps. The findings of our work are important because they demonstrate the applicability of affective state prediction beyond laboratory settings with a minimal amount of privacy invasion.

ACKNOWLEDGMENTS

We thank Tobias Günther for insightful discussions and all the participants for their time and effort. This work was supported by Tamedia, Ringier, NZZ, SRG, VSM, viscom, and the ETH Zurich Foundation.

REFERENCES

[1] Pedro Álvarez, Francisco Javier Zarazaga-Soria, and Sandra Baldassarri. 2020. Mobile music recommendations for runners based on location and emotions: The DJ-Running system. *Pervasive and Mobile Computing* 67 (2020), 101242. https://doi.org/10.1016/j.pmcj.2020.101242

[2] Androidrank. 2021. Website. Retrieved December 20, 2021 from https://www.androidrank.org.

[3] Maneesh Ayi and Mohamed El-Sharkawy. 2020. RMNv2: Reduced Mobilenet V2 for CIFAR10. In *10th Annual Computing and Communication Workshop and Conference (CCWC)*. 287–292. https://doi.org/10.1109/CCWC47524.2020.9031131

[4] Anja Bachmann, Christoph Klebsattel, Matthias Budde, Till Riedel, Michael Beigl, Markus Reichert, Philip Santangelo, and Ulrich Ebner-Priemer. 2015. How to Use Smartphones for Less Obtrusive Ambulatory Mood Assessment and Mood Recognition. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (Osaka, Japan). Association for Computing Machinery, New York, NY, USA, 693–702. https://doi.org/10.1145/2800835.2804394

[5] Karl-Heinz Best. 2005. Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics* 11 (2005), 9–31.

[6] Albrecht Beutelspacher. 1996. *Kryptologie*. Vol. 7. Springer. 10 pages.

[7] Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. 2018. Benchmark Analysis of Representative Deep Neural Network Architectures. *IEEE Access* 6 (2018), 64270–64277. https://doi.org/10.1109/ACCESS.2018.2877890

[8] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25 (1994), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

[9] Cynthia Breazeal. 2011. Social robots for health applications. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 5368–5371. https://doi.org/10.1109/IEMBS.2011.6091328

[10] Daniel Buschek, Benjamin Bisinger, and Florian Alt. 2018. ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173829

[11] Davide Carneiro, José Carlos Castillo, Paulo Novais, Antonio Fernández-Caballero, and José Neves. 2012. Multimodal behavioral analysis for non-invasive stress detection. *Expert Systems with Applications* 39 (2012), 13376–13389. https://doi.org/10.1016/j.eswa.2012.05.065

[12] Matteo Ciman and Katarzyna Wac. 2018. Individuals' Stress Assessment Using Human-Smartphone Interaction Analysis. *IEEE Transactions on Affective Computing* 9 (2018), 51–65. https://doi.org/10.1109/TAFFC.2016.2592504

[13] Liqing Cui, Shun Li, and Tingshao Zhu. 2016. Emotion Detection from Natural Walking. In *International Conference on Human Centered Computing*. Springer International Publishing, Cham, 23–33.

[14] Daxiang Dai, Qun Liu, and Hongying Meng. 2016. Can your smartphone detect your emotion?. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. 1704–1709. https://doi.org/10.1109/FSKD.2016.7603434

[15] Daniel Danner, Beatrice Rammstedt, Matthias Bluemke, Lisa Treiber, Sabrina Berres, Christopher J. Soto, and Oliver P. John. 2016. *Die deutsche Version des Big Five Inventory 2 (BFI-2)*. GESIS - Leibniz-Institut für Sozialwissenschaften, Mannheim. 20 pages. https://doi.org/10.6102/zis247

[16] Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. Observations on Typing from 136 Million Keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174220

[17] Sidney D'Mello and Jacqueline Kory. 2012. Consistent but Modest: A Meta-Analysis on Unimodal and Multimodal Affect Detection Accuracies from 30 Studies. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, 31–38. https://doi.org/10.1145/2388676.2388686

[18] Panteleimon Ekkekakis. 2012. Affect, mood, and emotion. *Measurement in Sport and Exercise Psychology* (2012), 321–332.

[19] Paul Ekman. 1999. Basic Emotions. *Handbook of cognition and emotion* 98 (1999).

[20] Clayton Epp, Michael Lippold, and Regan L. Mandryk. 2011. Identifying Emotional States Using Keystroke Dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 715–724. https://doi.org/10.1145/1978942.1979046

[21] Susan Folkman. 2008. The case for positive emotions in the stress process. *Anxiety, Stress, and Coping* 21 (2008), 3–14. https://doi.org/10.1080/10615800701740457

[22] Susan Folkman and Judith Tedlie Moskowitz. 2000. Stress, Positive Emotion, and Coping. *Current Directions in Psychological Science* 9 (2000), 115–118. https://doi.org/10.1111/1467-8721.00073

[23] Nan Gao, Wei Shao, and Flora D. Salim. 2019. Predicting Personality Traits From Physical Activity Intensity. *Computer* 52 (2019), 47–56. https://doi.org/10.1109/MC.2019.2913751

[24] Enrique Garcia-Ceja, Venet Osmani, and Oscar Mayora. 2016. Automatic Stress Detection in Working Environments From Smartphones' Accelerometer Data: A First Step. *IEEE Journal of Biomedical and Health Informatics* 20 (2016), 1053–1060. https://doi.org/10.1109/JBHI.2015.2446195

[25] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. EMMA: An Emotion-Aware Wellbeing Chatbot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–7. https://doi.org/10.1109/ACII.2019.8925455

[26] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2021. Designing an Experience Sampling Method for Smartphone Based Emotion Detection. *IEEE Transactions on Affective Computing* 12 (2021), 913–927. https://doi.org/10.1109/TAFFC.2019.2905561

[27] Surjya Ghosh, Shivam Goenka, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Representation Learning for Emotion Recognition from Smartphone Keyboard Interactions. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 704–710. https://doi.org/10.1109/ACII.2019.8925518

[28] Surjya Ghosh, Sumit Sahu, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. EmoKey: An Emotion-aware Smartphone Keyboard for Mental Health Monitoring. In *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*. 496–499. https://doi.org/10.1109/COMSNETS.2019.8711078

[29] Yue Gu, Shuhong Chen, and Ivan Marsic. 2018. Deep Multimodal Learning for Emotion Recognition in Spoken Language. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5079–5083. https://doi.org/10.1109/ICASSP.2018.8462440

[30] Muhammad Arslan Hashmi, Qaiser Riaz, Muhammad Zeeshan, Muhammad Shahzad, and Muhammad Moazam Fraz. 2020. Motion Reveal Emotions: Identifying Emotions From Human Walk Using Chest Mounted Smartphone. *IEEE Sensors Journal* 20 (2020), 13511–13522. https://doi.org/10.1109/JSEN.2020.3004399

[31] Elaine C. S. Hayashi, Julián E. Gutiérrez Posada, Vanessa R. M. L. Maike, and M. Cecília C. Baranauskas. 2016. Exploring New Formats of the Self-Assessment Manikin in the Design with Children. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3033701.3033728

[32] Jennifer Healey, Lama Nachman, Sushmita Subramanian, Junaith Shahabdeen, and Margaret Morris. 2010. Out of the Lab and into the Fray: Towards Modeling Emotion in Everyday Life. In *International Conference on Pervasive Computing*. Springer Berlin Heidelberg, 156–173.

[33] He Huang, Bokai Cao, Philip S. Yu, 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)*. 157–166. https://doi.org/10.1109/ICDM.2018.00031

[34] Agata Kołakowska. 2013. A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *2013 6th International Conference on Human System Interactions (HSI)*. 548–555. https://doi.org/10.1109/HSI.2013.6577879

[35] Agata Kołakowska, Wioleta Szwoch, and Mariusz Szwoch. 2020. A Review of Emotion Recognition Methods Based on Data Acquired via Smartphone Sensors. *Sensors* 20 (2020). https://doi.org/10.3390/s20216367

[36] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders* 114 (2009), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

[37] Ludwig Küster, Carola Trahms, and Jan-Niklas Voigt-Antons. 2018. Predicting personality traits from touchscreen based interactions. In *2018 10th International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX.2018.8463375

[38] Hosub Lee, Young Sang Choi, Sunjae Lee, and I. P. Park. 2012. Towards unobtrusive emotion recognition for affective social communication. In *2012 IEEE Consumer Communications and Networking Conference (CCNC)*. 260–264. https://doi.org/10.1109/CCNC.2012.6181098

[39] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* 49 (2013), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

[40] Katharina Lochner. 2016. *Successful Emotions*. Springer Fachmedien, Wiesbaden. https://doi.org/10.1007/978-3-658-12231-7

[41] Yuanchao Ma, Bin Xu, Yin Bai, Guodong Sun, and Run Zhu. 2012. Daily Mood Assessment Based on Mobile Phone Sensing. In *2012 9th International Conference on Wearable and Implantable Body Sensor Networks*. 142–147. https://doi.org/10.1109/BSN.2012.3

[42] C. Maramis, L. Stefanopoulos, I. Chouvarda, and N. Maglaveras. 2017. Emotion Recognition from Haptic Touch on Android Device Screens. In *Precision Medicine Powered by pHealth and Connected Health*. Springer Singapore, 205–209.

[43] Albert Mehrabian and James A. Russell. 1974. *An approach to environmental psychology*. The MIT Press, Cambridge.

[44] Jan Mizgajski and Mikołaj Morzy. 2019. Affective Recommender Systems in Online News Industry: How Emotions Influence Reading Choices. *User Modeling and User-Adapted Interaction* 29 (2019), 345–379. https://doi.org/10.1007/s11257-018-9213-x

[45] MoodMeter 2021. Website. Retrieved December 20, 2021 from https://moodmeterapp.com/.

[46] Aske Mottelson and Kasper Hornbæk. 2016. An Affect Detection Technique Using Mobile Commodity Sensors in the Wild. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, 781–792. https://doi.org/10.1145/2971648.2971654

[47] Andreas Fsrøvig Olsen and Jim Torresen. 2016. Smartphone accelerometer data used for detecting human emotions. In *2016 3rd International Conference on Systems and Informatics (ICSAI)*. IEEE, 410–415. https://doi.org/10.1109/ICSAI.2016.7810990

[48] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How Do People Type on Mobile Devices? Observations from a Study with 37,000 Volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. Association for Computing Machinery. https://doi.org/10.1145/3338286.3340120

[49] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 439–448. https://doi.org/10.1109/ICDM.2016.0055

[50] Sophie C. Reid, Sylvia D. Kauer, Stephen J. C. Hearps, Alexander H. D. Crooke, Angela S. Khor, Lena A. Sanci, and George C. Patton. 2011. A mobile phone application for the assessment and management of youth mental health problems in primary care: a randomised controlled trial. *BMC Family Practice* 12 (2011), 1–14.

[51] Nikki Rickard, Hussain-Abdulah Arjmand, David Bakker, and Elizabeth Seabrook. 2016. Development of a Mobile Phone App to Support Self-Monitoring of Emotional Well-Being: A Mental Health Digital Innovation. *JMIR Mental Health* 3 (2016). https://doi.org/10.2196/mental.6202

[52] Giuseppe Riva, Rafael A. Calvo, and Christine Lisetti. 2015. Cyberpsychology and affective computing. *The Oxford Handbook of Affective Computing* (2015), 547–558.

[53] Mintra Ruensuk, Hyunmi Oh, Eunyong Cheon, Ian Oakley, and Hwajung Hong. 2019. Detecting Negative Emotions during Social Media Use on Smartphones. In *Proceedings of Asian CHI Symposium: Emerging HCI Research Collection*. Association for Computing Machinery, 73–79. https://doi.org/10.1145/3309700.3338442

[54] James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality* 11 (1977), 273–294.

[55] Ahmed T. Sahlol, Mohamed Abd Elaziz, Amani Tariq Jamal, Robertas Damaševičius, and Osama Farouk Hassan. 2020. A Novel Method for Detection of Tuberculosis in Chest Radiographs Using Artificial Ecosystem-Based Optimisation of Deep Neural Network Features. *Symmetry* 12 (2020), 1146. https://doi.org/10.3390/sym12071146

[56] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4510–4520.

[57] Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information* 44 (2005), 695–729. https://doi.org/10.1177/0539018405058216

[58] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. 2018. Labelling Affective States "in the Wild": Practical Guidelines and Lessons Learned. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. Association for Computing Machinery, 654–659. https://doi.org/10.1145/3267305.3267551

[59] Sandra Servia-Rodríguez, Kiran K. Rachuri, Cecilia Mascolo, Peter J. Rentfrow, Neal Lathia, and Gillian M. Sandstrom. 2017. Mobile Sensing at the Service of Mental Well-Being: A Large-Scale Longitudinal Study. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 103–112. https://doi.org/10.1145/3038912.3052618

[60] Nusrat J. Shoumy, Li Minn Ang, Kah Phooi Seng, D. M. Rahaman, Motiur, and Tanveer Zia. 2020. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications* 149 (2020), 1–26. https://doi.org/10.1016/j.jnca.2019.102447

[61] K. Dmello Sidney, Scotty D. Craig, Barry Gholson, Stan Franklin, Rosalind Picard, and Arthur C. Graesser. 2005. Integrating affect sensors in an intelligent tutoring system. In *Affective Interactions: The Computer in the Affective Loop Workshop at Intl. Conf. on Intelligent User Interfaces*. AMC Press, 7–13.

[62] Leslie N. Smith. 2017. Cyclical Learning Rates for Training Neural Networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 464–472. https://doi.org/10.1109/WACV.2017.58

[63] Robert L. Solso and Joseph F. King. 1976. Frequency and versatility of letters in the English language. *Behavior Research Methods & Instrumentation* 8 (1976), 283–286. https://doi.org/10.3758/BF03201714

[64] Mirjam Stieger, Marcia Nißen, Dominik Rüegger, Tobias Kowatsch, Christoph Flückiger, and Mathias Allemand. 2018. PEACH, a smartphone-and conversational agent-based coaching intervention for intentional personality change: study protocol of a randomized, wait-list controlled trial. *BMC Psychology* 6 (2018), 1–15. https://doi.org/10.1186/s40359-018-0257-9

[65] Matthias Trojahn, Florian Arndt, Markus Weinmann, and Frank Ortmeier. 2013. Emotion Recognition through Keystroke Dynamics on Touchscreen Keyboards. In *Proceedings of the 15th International Conference on Enterprise Information Systems - Volume 3: ICEIS,*. INSTICC, SciTePress, 31–37. https://doi.org/10.5220/0004415500310037

[66] Johannes Wagner, Elisabeth André, and Frank Jung. 2009. Smart sensor integration: A framework for multimodal emotion recognition in real-time. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–8. https://doi.org/10.1109/ACII.2009.5349571

[67] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R. Schinazi, and Markus Gross. 2020. Affective State Prediction Based on Semi-Supervised Learning from Smartphone Touch Data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–13. https://doi.org/10.1145/3313831.3376504

[68] Ping Wang, Luobing Dong, Wei Liu, Ningning Jing, et al. 2020. Clustering-Based Emotion Recognition Micro-Service Cloud Framework for Mobile Computing. *IEEE Access* 8 (2020), 49695–49704.

[69] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, 3–14. https://doi.org/10.1145/2632048.2632054

[70] Woebot 2021. Website. Retrieved December 20, 2021 from https://woebothealth.com.

[71] Qian Xiang, Xiaodan Wang, Rui Li, Guoling Zhang, Jie Lai, and Qingshuang Hu. 2019. Fruit Image Classification Based on MobileNetV2 with Transfer Learning Technique. In *Proceedings of the 3rd International Conference on Computer Science and Application Engineering*. Association for Computing Machinery. https://doi.org/10.1145/3331453.3361658

[72] Kangning Yang, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2021. Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition. *IEEE Transactions on Affective Computing* (2021). https://doi.org/10.1109/TAFFC.2021.3100868