Affective State Prediction Based on Semi-Supervised Learning from Smartphone Touch Data

Rafael Wampfler[1], Severin Klingler[1], Barbara Solenthaler[1], Victor R. Schinazi[2, 3], Markus Gross[1]

[1]Department of Computer Science, ETH Zurich, Switzerland
[2]Chair of Cognitive Science, ETH Zurich, Switzerland
[3]Institute of Cartography and Geoinformation, ETH Zurich, Switzerland
wrafael@inf.ethz.ch, kseverin@inf.ethz.ch, solenthaler@inf.ethz.ch, scvictor@ethz.ch, grossm@inf.ethz.ch

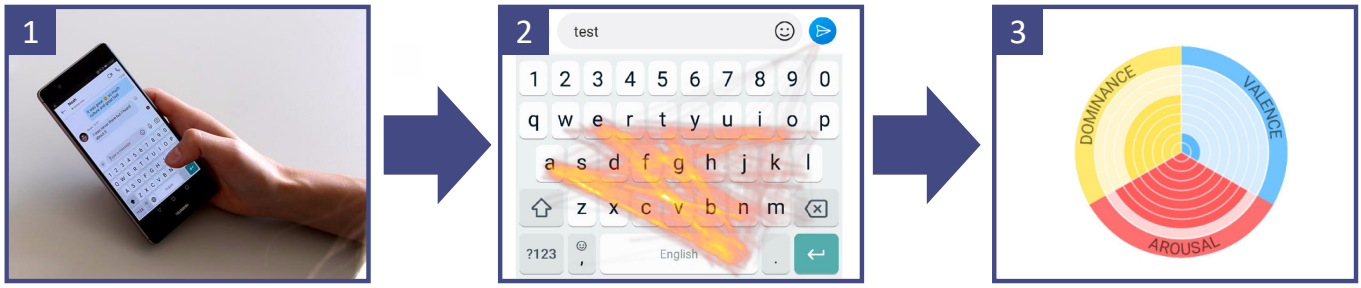

Figure 1. Our system extracts touch input characteristics of users while typing on smartphones (1) and aggregates these metrics into two-dimensional heat maps (2). A semi-supervised classification pipeline dynamically predicts affective states (valence, arousal, and dominance) of the user (3).

ABSTRACT

Gaining awareness of the user's affective states enables smartphones to support enriched interactions that are sensitive to the user's context. To accomplish this on smartphones, we propose a system that analyzes the user's text typing behavior using a semi-supervised deep learning pipeline for predicting affective states measured by valence, arousal, and dominance. Using a data collection study with 70 participants on text conversations designed to trigger different affective responses, we developed a variational auto-encoder to learn efficient feature embeddings of two-dimensional heat maps generated from touch data while participants engaged in these conversations. Using the learned embedding in a cross-validated analysis, our system predicted three levels (low, medium, high) of valence (AUC up to 0.84), arousal (AUC up to 0.82), and dominance (AUC up to 0.82). These results demonstrate the feasibility of our approach to accurately predict affective states based only on touch data.

Author Keywords

Classification; Affective Computing; Smartphone; Deep Learning

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376504

CCS Concepts

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

INTRODUCTION

Interactive systems that include knowledge about the user are capable of providing an optimised user experience. Such systems should be able to dynamically adapt the content, trigger help functions, provide feedback when needed, or show motivational elements that increase user satisfaction and learning gain. Recently, interaction models have been extended to include methods capable of predicting the affective state of the user. Affective states are psycho-physiological constructs used for characterizing emotions (short-term) and moods (long-term) that are experienced by users when engaged with a stimulus [15, 49, 61]. Affective states are often represented along the valence, arousal and dominance dimensions [54] or grouped into basic emotions (i.e., anger, happiness, sadness, surprise, disgust, and fear) [16].

Responding to changes in affective states has been found to be beneficial in educational settings and is associated with increases in learning gain and motivation [29]. Recent work has also suggested that being aware of one's current affective state can be particularly useful in the context of mobile devices as individuals become more dependent on smartphones for social purposes [48]. Here, chat applications are especially relevant as they currently rank as the most used applications on smartphones [2].

The majority of methods to detect affective states rely on biosensor data (e.g., heart rate and skin conductance) [35, 63], body behavior [39], or camera data to infer emotions from facial expressions [59]. However, most of these setups are invasive and potentially costly, which can limit their widespread application. As such, researchers have explored different methods to infer affective states directly from smartphone data, including sensor inputs (e.g., acceleration and gyroscope) [44], application usage patterns [4, 48], and typing speed [21].

In this paper, we propose a non-invasive solution that can accurately predict affective states based on sensor data from a mobile device. We achieve this by considering only touch input from the smartphone's on-screen keyboard to generate two-dimensional heat maps of typing characteristics. We train our semi-supervised deep learning architecture on these heat maps to learn a low-dimensional feature embedding. The subsequent classification can predict valence, arousal, and dominance on three levels each (low, medium, high). We demonstrate the effectiveness of predicting the affective states based on the touch characteristics of smartphone users in a data collection study with 70 participants engaged with a chat application and highlight other potential application scenarios.

RELATED WORK

This section provides an overview of previous research in the field of affective state prediction and the related fields of biometrics and stress prediction. We focus our overview on methods that base the prediction on typing characteristics on computer keyboards and the various sensors from smartphones (e.g., touch and gyroscope). The processing of touch data collected from smartphones is at the core of our model. However, typing characteristics on computer keyboards are related to smartphone touch data in so far that typing patterns can resemble the typing patterns on smartphone keyboards. Kanjo et al. [32] provides an overview of other approaches that can be used for measuring affective states (e.g., biosensors).

Biometrics

We use the term biometrics to refer to the measurements and analysis of physical and behavioral characteristics that can be used to identify individuals. We focus specifically on keystroke patterns inferred from data collected through touchscreen and keyboard interactions. Such systems have been previously used for setting more secure passwords [55]. Commonly used features include keystroke dynamics such as down-down, up-down, and down-up timings of keys [3, 14, 51, 55]. Other researchers have also used the ID of the keys [3], the touch positions relative to the center of a touch element [42], and typing difficulty of successive characters [14].

Other researchers have used typing pressure from keyboards [3, 42, 55] and smartphone touch screens [13, 51] for user identification. Here, research has shown that using pressure-sensitive keyboards in addition to traditional keystroke dynamics can significantly reduce the error rate for user identification from 2.04% to 1.41% [51]. Features from pressure data such as the gradient, maximum, and pressure timings have all shown to perform well [42].

In our work, we leverage down-down, up-down, and pressure metrics to predict affective states. In contrast to past work, we consider the spatial distribution of measurements by using two-dimensional heat maps over the keyboard input area.

Stress Prediction

The ubiquity of mobile devices has led to a surge in research focusing on the prediction of stress based on smartphone usage [65]. Researchers have used different modalities for predicting stress based on smartphone data. These include behavioral metrics such as call and text logs and location data stemming from GPS [5, 7], application usage patterns [19], voice recordings [50], and video recordings [11].

Apart from being invasive (e.g., sharing of text logs), relying on these modalities for the prediction of affective states also has the disadvantage of draining the smartphone battery (e.g., the high power consumption of GPS sensors). As such, other work has focused on using sensor-based smartphone data, including touch input and accelerometer data [11, 22]. Carneiro et al. [11] used patterns, accuracy, intensity, and duration of touch events as well as hand gestures to predict stress in real-time while users played a mentally challenging mobile game. In addition, Hernandez et al. [24] showed that typing pressure and the size of the contact area with the mouse tends to increase during stressful situations (i.e., expressive writing, text transcription, and mouse clicking). To measure pressure and contact area, this work relied on pressure-sensitive computer keyboards and capacitive mouses, respectively. Recently, Exposito et al. [18] conducted a similar study on the smartphone and showed that typing pressure increases during stressful situations (i.e., expressive writing). In addition, Sarsenbayeva et al. [60] showed that stress increases tapping frequency but decreases tapping accuracy. These researchers also found that text difficulty had a larger effect on typing performance (measured as the ratio between number of errors and number of entered characters) than the stress level. Finally, other researchers proposed a multi-modal approach jointly measuring accelerometer, microphone data, and social activity data from call and SMS [53].

Most existing approaches for stress prediction used two [7, 50, 5] or three classes [53], and the stress measurement tool of choice were self-reports [24, 53, 19]. Achieved performance ranged from 83% to 100% for two classes (stressed vs. non stressed) [24] and 71% for three classes (low, medium, high) [53].

Affective State Prediction

Researches have used different data sources to predict affective states. Most available smartphone systems are complex in terms of the amount and nature of the modalities involved. Systems have been built from smartphone sensor data (e.g., accelerometer, Bluetooth, microphone, and GPS) to grasp user movements and conversational cues [57]. Other systems included the context of the user data (e.g., location and weather) [8, 45], communication data (e.g., call and SMS logs) [8, 48, 56], and interaction data (e.g., web browsing and application usage) [48, 56]. Such systems have provided decent performance with accuracies up to 71% for predicting var-

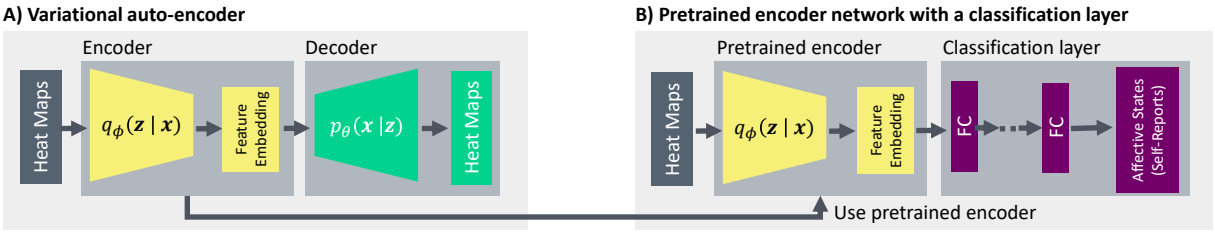

Figure 2. Overview of the main steps of our model. A) A variational auto-encoder is trained on heat maps created from smartphone touch data to learn an efficient low-dimensional feature embedding. B) For classification, the low-dimensional embedding is used as input to fully-connected layers.

ious emotions (i.e., happy, sad, fear, anger, and neutral) [57] and 80% for predicting 3 levels of happiness (i.e., happy, neutral, and unhappy) [8]. Nevertheless, such complex systems are often privacy-invasive and computationally demanding.

Other more light-weight approaches for predicting affective states have exploited touch and typing behavior. Gao et al. [21] predicted four states (i.e., excited, relaxed, bored, and frustrated), each with two levels, with an accuracy between 69% and 77% as well as two levels of arousal and valence with an accuracy of 89%. These researchers used touch pressure and speed of touch features recorded while users were playing a game. Previous work also employed touch data from chat conversations. Lee et al. [45] predicted Ekman's six basic emotions and a neutral state with 67% accuracy using a Bayesian network classifier based on behavior data (i.e., typing speed and touch count) and context data (i.e., location and weather) collected while users used the twitter application. Interestingly, they found that the speed of typing was the most predictive factor. On the other hand, Ghosh et al. [23] predicted four states (i.e., happy, sad, stressed, and relaxed) with a performance of 0.84 AUC using touch statistics (inter-tap durations, number of special characters, and number of deletes). These researchers jointly modeled the typing characteristics and the persistence of emotions by adapting the reported emotions based on a Markov chain.

Other researchers [28] have predicted depression and mania on a regression scale using a personalized deep learning model for bipolar subjects by leveraging temporal dynamics and fusing accelerometer and keyboard metadata (duration of a keypress, time since last keypress, and distance to last keypress). Interestingly, Leow et al. [47] found a positive correlation between higher accelerometer displacements and depression as well as mania.

Key stroke dynamic features such as pressure, latency, and duration have also been used on computer keyboards [52]. Using these features, Epp et al. [17] predicted 15 emotional states on two levels with an accuracy between 77% and 88%. Kołakowska [41] provides an overview of other work on predicting affective states based on computer keyboards.

In contrast to previous work, we are using a light-weight approach by only considering pressure and speed characteristics of touch data and employing a semi-supervised pipeline on heat maps extracted from this data. Moreover, we are using a pressure-sensitive display instead of the contact area [21] to approximate pressure, and we are also considering dominance, which we believe might be necessary for finer-grained distinctions between affective states.

METHOD

We present a semi-supervised classification pipeline for predicting affective states based on touch data collected during typing on smartphones. While touch data is continuously available, ground truth is typically only available in certain intervals (e.g., from self-reports). To make use of the large amount of unlabeled data, we employ variational auto-encoders to infer meaningful low-dimensional embeddings from two-dimensional heat maps (Figure 2A). In a second step, we add a fully connected classification layer to the learned data encoder and fine-tune the entire network for the classification of affective states (Figure 2B). In the following, we provide details on every part of our method.

Heat Maps

Modern smartphones allow for the collection of accurate information about the user's screen inputs. An input $e_i = (x, y, t)$ is defined by the coordinates (x, y) on the screen and the timestamp t in milliseconds. A single touch event $E = [e_1, \ldots, e_n]$ can consist of n touch inputs from the time the user initially touched the screen (e_1, touch down) until he or she releases the screen (e_n, touch up). Based on the raw input data, we can extract several touch event metrics: Down-down speed provides information about the typing speed and is equal to the time difference between two consecutive touch downs normalized by the distance. Up-down speed is equal to the time between a touch up and the subsequent touch down normalized by the distance. Up-down speed provides information about the speed between touch events. In contrast to previous research [3, 13, 55], we do not account for touch duration (down-up speed) since touch events often consist of a single input $E = [e_1]$ for which no duration can be computed. All metrics are standardized based on the mean and standard deviation during a baseline typing period.

Since touch inputs are inherently spatial, we aggregate the touch event metrics into two-dimensional heat maps. These heat maps cover the keyboard region and the send button (see the red dashed line in Figure 4B) as we only include keyboard inputs in this work. We use a sliding window with a window size of 180 seconds shifted by 5 seconds to extract a sequence of heat maps for each user. Since the down-down speed and up-down speed metrics always correspond to two consecutive touch events E_i and E_{i+1} we assign their value to every pixel on a straight line between the events (see Figure 3B and 3C).

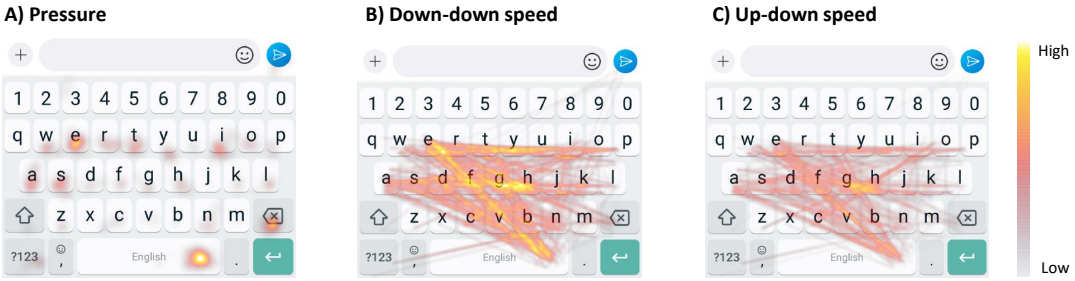

Figure 3. Examples of heat maps extracted from the touch events of a user. A) The color indicates the average pressure applied. B) and C) Consecutive touch events are connected by a line, and the color indicates the down-down and up-down speed between these two events, respectively. The colors are for visualization purposes only.

Finally, we apply Gaussian smoothing to the heat maps to reduce high-frequency noise. We use a kernel of size $k = 31 \times 31$ pixels, which is twice the typical key distance in pixels, and prevents smearing into neighboring keys while keeping inter-key resolution high. In addition, we use $\sigma = 5$ provided by OpenCV [10].

Figure 3 shows examples of extracted heat maps for pressure, down-down speed, and up-down speed. The colors in the heat maps are for visualization purposes only. In our pipeline, we only use one value per pixel.

Variational Auto-encoder

While touch data is available continuously, labels are sparse. We make use of the unlabeled data by learning a low-dimensional representation of the heat maps that capture as much information from the original heat maps as possible. To extract such a low-dimensional representation (also called latent space or embedding), we employ a particular type of neural network called variational auto-encoder [37] (see Figure 2A). Variational auto-encoders have the advantage of providing representations with disentangled factors and allow control over modeling the latent distribution (in our case, multivariate Gaussian) [25, 38]. Previous research has shown that variational auto-encoders are capable of automatically learning meaningful low-dimensional representations in different domains [1, 40].

A variational auto-encoder consists of an encoder and decoder. The encoder network $q_\phi(\mathbf{z}|\mathbf{x})$ learns an efficient compression of the input data $\mathbf{x}$ (heat map) into a low-dimensional space $\mathbf{z}$ using a deep neural network parameterized by ϕ. The decoder network $p_\theta(\mathbf{x}|\mathbf{z})$ reconstructs the input based on sampling from the distribution of the latent space. Here, θ are the parameters of the decoder network. We train the variational auto-encoder using the loss function

$$\mathscr{L}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta \, \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right],$$

where KL denotes the Kullback-Leibler divergence. The left term measures the reconstruction quality, and the right term regularizes the latent space towards the prior $p(\mathbf{z})$. By using the Lagrangian multiplier β, we introduce a trade-off between reconstruction quality and disentanglement of the latent factors fostering a more efficient encoding. This modification of the loss function has been successfully used for training variational auto-encoders [26].

For the auto-encoder, we use two-dimensional convolutions with symmetric encoder and decoder. Depending on the resolution of the input heat maps, it is necessary to down-sample the heat maps to reduce training time. Input data is commonly scaled before training. We use Min-Max scaling of the heat maps per user.

Classification

We take advantage of the learned low-dimensional representation by adding a classification network to the pre-trained encoder network (see Figure 2B). The classification network consists of fully connected layers with rectified linear unit activations except for the last layer, where we use softmax activation for the classification output. The different heat maps are aggregated by stacking the latent space of the individual heat maps.

The fully-connected network is trained on the labeled data (heat maps and corresponding affective states) using backpropagation that minimize the cross-entropy loss. Fine-tuning the classification network has shown good performance in other domains [62].

EXPERIMENT

We conducted a controlled lab experiment to validate our pipeline for the prediction of affective states based on smartphone touch data. The experiment was approved by the ethics board of ETH Zurich. During the experiment, we collected smartphone touch data while participants interacted with a chat application (i.e., Skype) for approximately 70 minutes. We used text-based chat conversations because they are widely used [2] and would be familiar to the participants in the study. In addition, these applications require interaction with the smartphone and can provide the data necessary for testing our prediction model.

Participants

We recruited 70 participants (35 female) between the ages of 18 and 31 (mean = 23.0, standard deviation SD = 2.7) from 20 different departments at the master and bachelor level of ETH Zurich and University of Zurich. We only considered participants that were fluent in English[1] and used smartphone-based

[1] A post-experiment questionnaire revealed that 94% of the participants judged their English level to be "proficient" (C2) or "advanced" (C1) according to the Common European Framework of Reference for Languages.

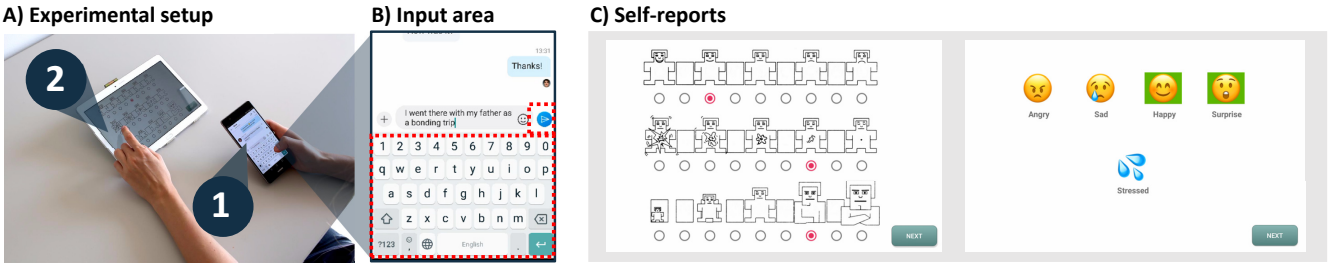

Figure 4. Experimental setup. A) During each session, participants engaged in chat conversations using Skype on a smartphone (1). At regular intervals, participants were asked to complete self-reports on a tablet (2). B) Chat interface and the region that was considered in the prediction model (red-dashed area). C) Self-reports for capturing valence, arousal and dominance (left), basic emotions, and stress level (right).

chat applications on a daily basis. We excluded participants taking any type of medication, tranquilizers, or psychotropic drugs (e.g., anti-depressants) as well as participants affected by any type of the autism spectrum disorders. To control for external environmental factors, we kept the room temperature and the humidity at an average of 23.9°(SD = 0.24°) and 30.1% (SD = 3.6%), respectively. All participants provided written informed consent before the start of the experiment and were rewarded with CHF 45 for their participation. Participants were rewarded with an additional CHF 5 if they missed only one response window when completing the self-report measures.

Apparatus

Participants interacted with five contacts within the Skype application on a Huawei P9 Plus smartphone running Android 7.0. This smartphone provides over 17000 levels of touch pressure sensitivity. The software keyboard used was Gboard with auto-correction and spell-checker features disabled. Throughout the experiment, we recorded their interaction with the device, including sensor (acceleration and orientation) and touch (pressure and position) data. In addition, participants used a Huawei MediaPad M2 tablet to report their emotional state at regular intervals during the experiment. Figure 4A presents the experimental setup.

Self-Reports

To gather ground truth data for our model, we asked participants to complete the Self-Assessment Manikin (SAM) [9] at regular intervals during the experiment. The SAM is a pictorial assessment used to quantify levels of valence, arousal, and dominance on a 9-point scale. Participants were also asked to select from a series of basic emotions (i.e., anger, sadness, happiness, and surprise) represented by different emojis. To ensure that choices were independent, participants were allowed to simultaneously select more than one emoji at a time (e.g., anger and surprise). The basic emotions did not include fear and disgust after a pilot study ($n = 8$) revealed that participants did not experience these emotions during the chat conversations. However, participants had the choice of selecting a "stress" emoji when reporting their emotions. Participants were allowed to select all possible combinations of the basic emotions and stress without any restrictions. Figure 4C shows an illustration of the self-reports.

Procedure

Before the day of the experiment, participants were asked to complete the Patient Health Questionnaire [43] and the Big Five Inventory [30, 31] as measures of mental health and personality traits, respectively. On the day of the experiment, the participants were given an oral overview of the procedure, including an introduction to the self-report questionnaires and an explanation regarding the use of the smartphone. The experimenter then exited the room and used one of the Skype contacts (guiding contact) to start a conversation (5 minutes) with the participants. During this conversation, the experimenter asked 6 predefined questions about well-being, age, living place, work, hobbies, and family. These questions were used to make the participants comfortable with the keyboard and the handling of the smartphone. Next, participants were instructed to watch a nature video for 5 minutes on the smartphone that was used as relaxation and allowed them to acclimate to the room environment. At the end of the nature video, participants were asked to type two well-known pangrams (149 characters) that served as a baseline for touch input during the modeling stage. During the main phase of the experiment, participants chatted with four different Skype users. These Skype users were fake accounts created and controlled by the experimenter sitting in an adjacent room. After finishing all four chat conversations, participants were asked to type once again the two pangrams. Finally, participants completed an exit questionnaire on smartphone use, demographics, and overall mood. Figure 5A provides an overview of the procedure used in the experiment (see supplemental material for additional information about the experiment).

At the beginning of the experiment, participants were given an oral explanation regarding the procedure for answering the self-reports and had a chance to practice with 4 examples. We collected a total of 1893 self-reports covering a large range of the SAM response space.

During the experiment, participants were alerted with an audio notification when it was time to complete the self-report. At this time, the SAM and emojis appeared on the tablet, and participants had 20 seconds to start the self-report. This time buffer, allowed participants to finish the current sentence in the Skype conversation without having to rush to complete the self-report. If participants were slow to respond, the tablet started to vibrate as a final reminder. After completing the self-reports, a delay of 90 seconds was introduced until the

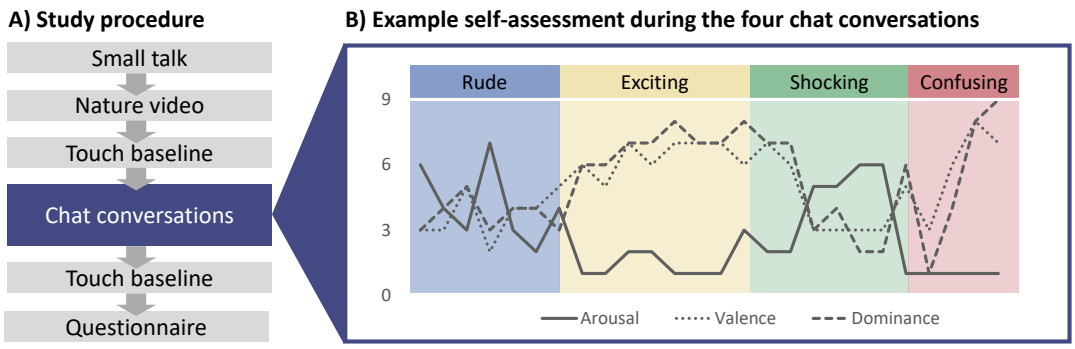

Figure 5. Overview of the different parts of the experiment. A) Overall experimental procedure. B) Changes in valence, arousal, and dominance for one participant during the four chat conversations.

next self-report was presented. This time interval was decided based on feedback from participants in the pilot experiment as it provided the best trade-off between the amount of data collected and the number of interruptions.

Tasks

To trigger different affective states, we created four different types of chat conversations (i.e., exciting, shocking, rude, and confusing) by varying the content and context of the text messages sent to the participants. Participants saw a list of five contacts in the Skype application on the smartphone that was provided to them. Each contact was associated with one of the conversation types. A fifth contact representing the experimenter was created to guide participants through the experiment and to provide help in case of questions.

To make the chat task more credible, we employed NVIDIA's face generator [33] to create fake profile pictures for each of the four contacts. We used the image of the experimenter for the fifth contact. In addition, participants were told that the four contacts were real people sitting next door. All conversations followed a predetermined script to keep them consistent across participants. The paragraphs below describe in more detail each of these conversations.

Exciting conversation. During this conversation, the participants were chatting about their most beautiful holiday experience. This conversation was designed to make participants remember and reminisce, leading to positive feelings (e.g., enjoyment).

Shocking conversation. This conversation focused on the topic of the Rohingya refugee crisis, which is an ongoing persecution of Muslim Rohingya people in Myanmar by the government. This conversation was intended to sadden the participants leading to negative feelings (e.g., anger).

Rude conversation. In this conversation, we asked participants to provide help with a malfunctioning smartphone. Independent of the help participants provided, they could not resolve the issue at any point during the conversation. Here, the Skype contact chatting with participants became increasingly rude and was intended to trigger negative feelings (e.g., anger) and surprise.

Confusing conversation. For this conversation, we used Cleverbot [12]. Cleverbot is a well-known chatbot that learns from past conversations. We have found this chatbot to be a good way to trigger confusion, anger, and surprise. We have reset the chatbot engine for every participant to avoid introducing potential bias from conversations with previous participants. A post-experiment questionnaire revealed that 63% of the participants did not recognize that this conversation was with a chatbot.

The order of conversations was randomized across participants with the exception that the confusing conversation was always last to prevent participants from behaving differently should they recognize that they were chatting with a chatbot [27]. This lead to the counterbalancing of three conditions and a total of six orders. With our randomization approach, we achieved an almost complete counter balanced distribution (12, 11, 11, 14, 10, 12). In general, the average duration of the rude and confusing conversations (836 seconds and 650 seconds) was shorter compared to the exciting and shocking conversations (1212 seconds and 1272 seconds). These shorter durations may be related to the fact that participants became tired of engaging in the conversations.

Figure 5B depicts the changes in valence, arousal, and dominance during the four chat conversations for one participant. The figure shows that valence increases during the exciting and confusing conversations and decreases for the other two conversations (we see the opposite pattern for arousal). The rude and shocking conversations seemed to be more intense than the exciting and confusing conversations. We also see that dominance is following a similar pattern than valence with the participant feeling more in control during the exciting and confusing conversations.

RESULTS

We evaluated our classification pipeline based on the data we collected during the experiment. We collected 1893 self-reports on the affective and emotional state of participants that were used as the ground truth to our model. Because the SAM is scored on a 9-point scale, we evaluated the performance of the classifier for three classes (low, medium, high) of valence, arousal, and dominance. We also recorded 3720 minutes of touch data from which we extracted 44625 heat maps for each

of the three types of heat maps (i.e., pressure, down-down speed, and up-down speed). We also reveal the runtime of our method to analyze the real-time applicability of our method. To measure the performance of our model, we calculated the accuracy (chance level = 0.33 for three classes and 0.5 for two classes) and the micro-averaged area under curve (AUC) of the receiver operating characteristic (ROC) curve (chance level = 0.5). The micro-averaged AUC aggregates the contributions of all classes by considering each element of the label indicator matrix as a label. Because these two metrics are both affected by class imbalance, we also calculated the macro-averaged AUC (chance level = 0.5) by taking the mean of the class-wise AUCs. We have evaluated our model using leave-one-user-out cross-validation to ensure that data from a user is not used for training and testing at the same time.

Network Parameters

Variational auto-encoder. For each of the three types of heat maps, we have trained a variational auto-encoder to learn a low-dimensional representation. We used a resolution of the heat maps of 80×64 pixels. To find the network parameters, we have employed the approach described by Bengio [6]. Specifically, we have increased the number of layers, and the number of features maps per layer until a good fit of the data was achieved (i.e., the loss was minimal). For the pressure heat maps, this resulted in a variational auto-encoder consisting of 2 layers (32 and 64 feature maps) for the encoder and decoder, a kernel size of 4×4, and a latent space with 10 dimensions. For the down-down speed and up-down speed heat maps, this resulted in a variational auto-encoder with 4 layers (32, 64, 128, and 256 feature maps) for the encoder and decoder, a kernel size of 3×3 and a latent space with twenty dimensions. In comparison to the network for the pressure heat maps, the down-down speed and up-down speed network was deeper and with a dimensionality of the latent space twice as high due to the higher complexity of the heat maps. For both networks, we have used a stride of 2×2 for each convolution. We chose a relatively small $\beta = 0.00001$ (compared to [26]) because of the difference in magnitude between reconstruction loss and the Kullback Leibler divergence. We trained the variational auto-encoders for 200 epochs with a batch size of 64 on 40162 heat maps and used 4463 heat maps as the validation set.

Fully-connected network. The network parameters for the fully-connected network used for classification were defined using a randomized search with 50 iterations. We trained the network using nested leave-one-user-out cross-validation for 100 epochs with a batch size of 8. All networks were implemented using the Keras framework with TensorFlow™ back-end and optimized using Adam optimization with standard parameters [36].

Experimental Validation

We conducted three Kruskal-Wallis tests to investigate whether the four text conversations elicited different levels of valence, arousal, and dominance. Results revealed significant differences in terms of valence ($H = 144.431$, 3 d.f., $p < 0.001$), arousal ($H = 19.461$, 3 d.f., $p < 0.001$) and dominance ($H = 39.982$, 3 d.f., $p < 0.001$). We performed five additional ANOVAs to investigate whether there were significant differences in terms of the basic emotions and stress reported by participants during the four conversations. For the ANOVAs, we added the times that participants reported a specific basic emotion or stress during each of the conversations. Here again, we found significant differences in terms of anger ($F(3,233) = 21.768$, $p < 0.001$), happiness ($F(3,233) = 238.068$, $p < 0.001$), sadness ($F(3,233) = 79.389$, $p < 0.001$), surprise ($F(3,233) = 6.158$, $p < 0.001$) and stress ($F(3,233) = 5.525$, $p = 0.001$). All tests are significant after Bonferroni correction. Table 1 presents the means and standard deviation for each of these variables (see supplemental material for additional statistics).

Table 1. Means and standard deviations (in brackets) for the self-reported SAM, four basic emotions, and stress during the four conversations. Percentages for the four basic emotions and stress do not add to 100% since participants could either simultaneously pick more than one emotion or not pick an emotion at all.

	Exciting	Shocking	Rude	Confusing	Total
Valence	7.3 (1.5)	3.3 (1.6)	4.8 (2.1)	5.2 (1.6)	5.2 (2.3)
Arousal	4.3 (2.1)	5.0 (2.2)	4.4 (2.2)	3.4 (1.9)	4.4 (2.2)
Dominance	6.3 (1.7)	4.8 (2.1)	5.3 (2.2)	5.1 (2.2)	5.4 (2.1)
Anger	0.7%	28.6%	25.4%	10.3%	15.7%
Happiness	77.5%	2.6%	16.7%	21.8%	33.1%
Sadness	2.9%	52.7%	10.5%	2.5%	21.0%
Surprise	7.6%	12.5%	18.7%	37.4%	16.0%
Stress	2.1%	8.0%	20.5%	15.6%	9.3%

We also performed a series of correlations to investigate the relationship between the SAM ratings for valence, arousal, and dominance and the four basic emotions and stress. Table 2 presents the results for each of these correlations. Notably, these results suggest a close match between the SAM ratings and the four basic emotions and stress.

Table 2. Effect sizes of the Pearson correlations between valence, arousal, and dominance (from the SAM) and the four basic emotions and stress. Asterisks denote correlations that survived Bonferroni correction (p = 0.003).

	Anger	Happiness	Sadness	Surprise	Stress
Valence	-0.55^*	$+0.79^*$	-0.62^*	-0.14	-0.30^*
Arousal	$+0.41^*$	$+0.07$	$+0.37^*$	$+0.03$	$+0.18$
Dominance	-0.19^*	$+0.43^*$	-0.24^*	-0.10	-0.33^*

Affective State Prediction

The performance of our model was evaluated with regards to the prediction of three classes (low $\in [1,3]$, medium $\in [4,6]$, high $\in [7,9]$) of valence (523, 712 and 660 data points), arousal (786, 758, 349) and dominance (375, 886, 632). We chose these three classes to cover the entire space considering all available ratings. Figure 6 and Table 3 present the performance of our model (ROC curves were calculated using the micro-averaging approach). See supplemental material for additional metrics.

Classification performance. Using all heat maps in combination, our model achieves an accuracy of 67% for valence, 63% for arousal, and 65% for dominance (chance level = 33%). Here, the slightly lower values for the macro-averaged AUC (0.83, 0.80, 0.80) compared to the micro-averaged AUC (0.84, 0.82, 0.82) may be attributed to class imbalance. If we

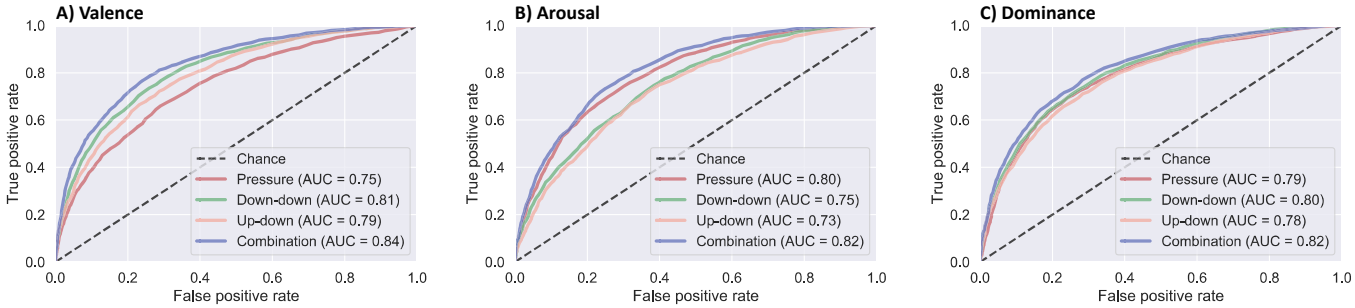

Figure 6. ROC curves and micro-averaged AUC scores for classification of three levels (low, medium, high) of A) valence, B) arousal, and C) dominance.

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

Dimension	Heat Map	AUC_{micro}	AUC_{macro}	Accuracy
Valence	Pressure	0.75	0.74	56%
	Down-down	0.81	0.81	64%
	Up-down	0.79	0.79	61%
	Combination	0.84	0.83	67%
Arousal	Pressure	0.80	0.78	62%
	Down-down	0.75	0.73	55%
	Up-down	0.73	0.70	53%
	Combination	0.82	0.80	63%
Dominance	Pressure	0.79	0.77	63%
	Down-down	0.80	0.78	63%
	Up-down	0.78	0.76	61%
	Combination	0.82	0.80	65%

consider the percentage of the most frequent class as baseline (valence = 38%, arousal = 42%, dominance = 47%), the predictions of our model are also above this baseline for all three dimensions. Figure 7 presents the confusion matrices for valence, arousal, and dominance based on the combination of all heat maps. The confusion matrices are calculated by predicting self-reports across all chat conversations. The matrices show that for valence, arousal, and dominance the low and high classes were often wrongly predicted as the medium class. As expected, the larger the distance between the classes, the easier it is to differentiate them for our model (i.e., the low class was only rarely confused with the high class and vice versa). Interestingly, for arousal, the medium class was most often wrongly predicted as the low class (Figure 7B), but medium dominance was more often confused with high dominance (Figure 7C).

Heat map comparison. Pressure is the best predictor of arousal (+0.05 AUC), while down-down speed and up-down speed are the best predictors for valence (+0.06 AUC). In terms of dominance, all three heat maps perform similarly (up to 0.80 AUC). Overall, the combination of all heat maps provides only marginal improvements compared to the individual heat maps (up to 0.03 AUC).

Affective Sequence Analysis

Affective states can change over time, and this may be characterized either by smooth transitions or abrupt changes (e.g., from low to high states). We hypothesize that the performance of our classifier can be affected by the period over which affective states are constant. For example, if affective states are alternating in short time, it can be much harder to make an accurate prediction compared to when affective states are constant over a longer period. This potential fluctuation in affective states, cannot be taken into account if we consider all labeled data from the conversations. As such, we recalculated the accuracy measure by considering only the data points for which the affective state was constant over a certain period (i.e., a specific number of preceding data points with the same class). Figure 8 shows the result of this accuracy measure for valence, arousal, and dominance. Here, a sequence length of zero corresponds to considering all data while sequence lengths of one, two, and three imply that we only considered data points having at least one, two, and three preceding data points with the same label. By excluding only immediate jumps (sequence length of one), we observe a steep increase in accuracy, reaching 78%, 75%, and 77% for valence, arousal, and dominance. In contrast, increasing the sequence length to two or three preceding data points provides only marginal improvements.

Basic Emotion and Stress Prediction

With regard to the four basic emotions and stress, our classifier achieved a predictive performance of 87% (0.84 AUC) for anger, 81% (0.88 AUC) for happiness, 84% (0.87 AUC) for sadness, 84% (0.76 AUC) for surprise and 92% (0.80 AUC) for stress. The large differences between accuracy and AUC can be attributed to class imbalance (e.g., 164 vs. 1729 labels for stress). Altogether, these results reveal that our model is not only able to predict affective states measured in terms of valence, arousal, and dominance but is also predictive for a subset of the basic emotions and stress.

Runtime Analysis

For evaluating the applicability of our method for realtime predictions, we have conducted a runtime analysis of the different parts of our model. Our computing environment consisted of an Intel® Xeon® CPU E5-2698 v4 @ 2.20GHz and an NVIDIA GeForce® GTX 1080 Ti. Prediction of a new data point consisted of extracting heat maps (mean = 0.38 seconds, SD = 0.09 seconds), followed by extracting the low-dimensional embedding of the heat maps using the encoder (mean = 0.065 seconds, SD = 0.0089 seconds) and using the fully-connected network for prediction (mean = 0.002 seconds,

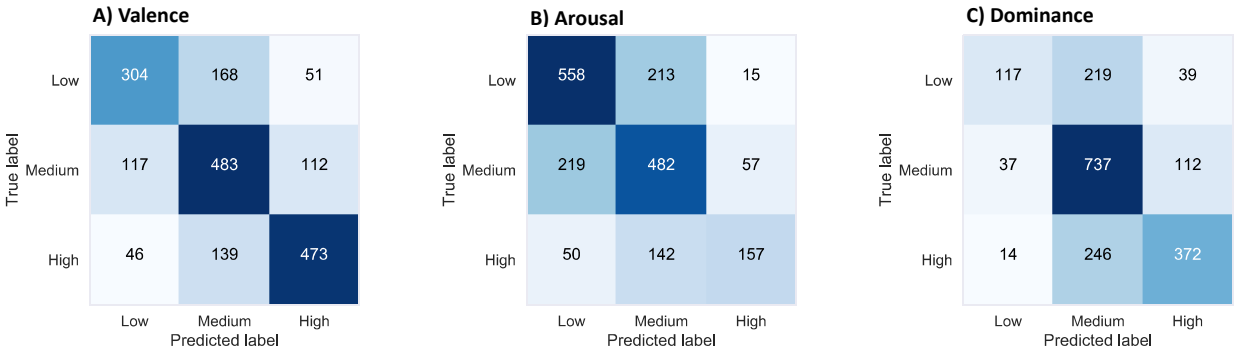

Figure 7. Confusion matrices for classification of three levels (low, medium, high) of A) valence, B) arousal, and C) dominance. The confusion matrices are calculated by predicting self-reports across all chat conversations.

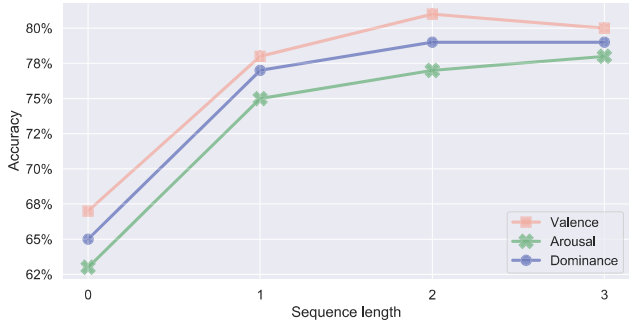

Figure 8. Accuracy only considering data points with a specific number of preceding data points with the same class label.

SD = 0.003 seconds). Summing up these values leads to a prediction time of 0.447 seconds. In other words, the system is capable of making two new predictions every second.

DISCUSSION

There has been a growing interest in the human-computer interaction community to provide interfaces that are sensitive to the emotions of users. In this paper, we presented a complete classification pipeline that is capable of accurately predicting three classes (low, medium, high) of valence (up to 0.84 AUC), arousal (up to 0.82 AUC) and dominance (up to 0.82 AUC). In addition, we also showed that we could accurately predict two levels (present vs. not present) of stress (0.80 AUC) and the basic emotions of anger (0.84 AUC), happiness (0.88 AUC), sadness (0.87 AUC) and surprise (0.76 AUC).

These predictions were based on heat maps generated from pressure and touch speed (i.e., down-down and up-down) collected during text conversations. We found that all three types of heat maps can predict valence, arousal, and dominance. Interestingly, while down-down speed showed the best performance for valence (0.81 AUC) and dominance (0.80 AUC), pressure was most predictive for arousal (0.80 AUC). These results may be related with the findings reported by Hernandez et al. [24], suggesting that people apply more pressure on keyboards under stressful conditions. Moreover, affective states characterized by higher valence (e.g., excitement) can lead to higher typing speed, increasing the down-down speed and up-down speed, which has also been reported in previous work (e.g., Lee et al. [46]).

The performance of our model cannot be directly compared with previous work due to differences in experimental setup. For example, Gao et al. [21] used a game-based setting and different measures of emotional states while Huang et al. [28] predicted mood on a regression scale. Our work did not focus on the comparison of performance but instead on automatic feature extraction in a different setting. Our use of heat maps also allowed us to investigate the distributions of keystrokes as a measure of affective states (e.g., use of more backspaces when experiencing negative emotions). Interestingly, running our model using only spatial heat maps, we achieved a performance of only up to 0.60 AUC. Thus, we conclude that the distribution of keystrokes alone has only little predictive power.

We have also shown that accuracy depends on the sequence of previous affective states and that accuracy tends to drop if affective states alternate. The reason for this is that when there is a preceding state belonging to a different class (e.g., low), noise is added to the window used for calculating the heat maps because this window contains touch data from both states whereby 1) the touch data is very different (e.g., low and high classes) or 2) the touch data is similar, but the class is different (e.g., low and medium classes).

Another noteworthy property of our model is its efficiency, which is particularly relevant for interactive applications. The computation of the heat maps, embedding, and prediction takes 0.447 seconds in total, meaning that the system can provide feedback on the user's emotional state in less than a second.

Applications

The ability to predict affective states based on touch patterns during text conversations has a broad range of applications. In the following, we present two possible applications that can benefit from affective predictions.

Woebot. Woebot [64] is one of many therapeutic chatbots available for Android and iOS devices. Using methods from cognitive behavioral therapy, Woebot aims to increase the overall mood of users and has shown to reduce symptoms of depression and anxiety [20]. Woebot uses predefined questions

to adequately adapt the conversation to the mood of the user, inferring the mood directly from the chat messages provided by the user. Our approach could provide predictions of users' affective states while they are engaged in chat conversations with Woebot and has the potential of improving human-bot interaction. Here, the bot would be able to adapt the responses to the users' changing affective state determined by typing pressure and speed. Similarly, other typing based chat applications could benefit from affective predictions from our model, such as customer service applications (e.g., Zendesk [66]).

Awareness. Knowledge about affective states can be leveraged to increase self-awareness and to convey awareness of affective states to others. Here, textual or graphical elements can be used to make users aware of their affective states. Such feedback can make users think about their affective state and encourage them to take regulatory actions (e.g., taking a break). If the user agrees, these affective states can be communicated to others using status messages that are common on social networks and chat applications. Figure 9A provides an example of our visualization for valence, arousal, and dominance. The circle is divided into three equal-sized segments, one for each dimension, and colored using color-blind friendly palettes. Each segment is further subdivided into nine parts representing the nine possible levels of the SAM. The parts of the segments are filled according to the level of the corresponding dimension. Figure 9B shows how our visualization could be used as part of the header in a chat application.

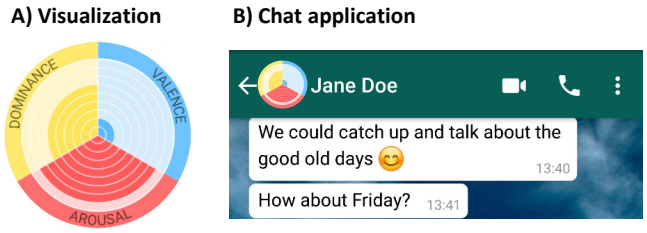

Figure 9. A possible visualization of affective states. A) One segment for each affective dimension. The parts of the segments are filled according to the level of the corresponding dimension. B) Example of how the visualization can be used in a chat application.

Limitations

The experiment was restricted to a controlled lab environment and a population consisting of bachelor and master students. As such, generalizations to real-world situations with a more diverse population requires further studies. The next step is to evaluate the system in real-life settings on different devices. Another limitation is the difficulty associated with collecting a direct measure of affect or emotions [58]. In our experiment, we have used self-reports, which can restrict our results to the specific conceptualization of affect that we have chosen. In addition, by querying emotions every 90 seconds, we might miss finer changes in emotions. A remedy would be to allow users to manually fill in self-reports when they face changes in emotions or to allow retrospective ratings. Finally, we acknowledge that using the pressure signal is limited to devices supporting pressure measurement. Pressure can also be measured using the contact area of the fingertips, which is a supported measure by many smartphones nowadays, but this might negatively affect the performance of the prediction because the contact area can only approximate real pressure.

Future Work

Future work could move data collection outside the laboratory to match real-world settings. This change would also allow for the collection of other types of interaction data, including acceleration and gyroscope, which could be used to complement touch data. Moreover, the presented prototype visualization for affective states could be connected to our model to provide visual feedback in real-time. Lastly, personality traits could be considered, which we have already measured in the current experiment. This would complement ongoing research that has shown feasibility of predicting personality based on keyboard input [34].

CONCLUSION

In this paper, we presented a semi-supervised pipeline for predicting affective states and emotions based on heat maps generated from smartphone touch data. We validated our pipeline on touch data collected from text conversations in a lab experiment with 70 participants. We conducted the evaluation using a leave-one-user-out cross-validation, which ensures that our results generalize among users, and similar results can be expected when applying our pipeline to data from new users. We demonstrated that our pipeline could accurately predict three classes (low, medium, high) of valence (up to 0.84 AUC), arousal (up to 0.82 AUC) and dominance (up to 0.82 AUC). We also presented results for the prediction of two levels (present vs. not present) of anger (0.84 AUC), happiness (0.88 AUC), sadness (0.87 AUC), surprise (0.76 AUC), and stress (0.80 AUC). Considering the real-time applicability of our method (predictions are possible in less than one second), our pipeline can be useful in combination with our proposed visualization of affective states. The novelty of our contribution consists of the semi-supervised deep learning pipeline and efficient feature embedding of 2D heat maps. Our model provides an elegant way to combine features (i.e., the features are learned automatically by the encoder as part of the low-dimensional embedding) without explicit feature engineering. By using heat maps in contrast to raw data, we are also taking into account the spatial distribution of the data. In contrast to other work using sentiment and video analysis, our approach is light-weight, less invasive, and can be used on different types of mobile devices. The findings of this work are important because they show a promising possibility of leveraging touch data to create emotion-aware chat conversations.

ACKNOWLEDGMENTS

We thank Christian Holz, Tobias Günther, Alexandra Ion, Katja Wolff and Anna Lisa Martin-Niedecken for insightful discussions. We would also like to thank Oliver Glauser for his assistance in creating the video figure 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] Emre Aksan, Fabrizio Pece, and Otmar Hilliges. 2018. DeepWriting: Making Digital Ink Editable via Deep Generative Modeling. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.

[2] Androidrank 2019. Website. (2019). Retrieved January 8, 2020 from `https://www.androidrank.org`.

[3] Lívia C. F. Araújo, Luiz H. R. Sucupira, Miguel Gustavo Lizarraga, Lee Luan Ling, and Joao Baptista T. Yabu-Uti. 2005. User authentication through typing biometrics features. *IEEE transactions on signal processing* 53, 2 (2005), 851–855.

[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*. ACM, 693–702.

[5] Gerald Bauer and Paul Lukowicz. 2012. Can smartphones detect stress-related changes in the behaviour of individuals?. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, 423–426.

[6] Yoshua Bengio. 2012. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*. Springer, 437–478.

[7] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. 2014. Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. In *Proceedings of the 22Nd ACM International Conference on Multimedia (MM '14)*. ACM, New York, NY, USA, 477–486.

[8] Andrey Bogomolov, Bruno Lepri, and Fabio Pianesi. 2013. Happiness recognition from mobile phone data. In *2013 International Conference on Social Computing*. IEEE, 790–795.

[9] 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, 1 (1994), 49–59.

[10] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[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, 18 (2012), 13376–13389.

[12] Rollo Carpenter. 2011. Cleverbot. (2011).

[13] Alexander De Luca, Alina Hang, Frederik Brudy, Christian Lindner, and Heinrich Hussmann. 2012. Touch me once and i know it's you!: implicit authentication based on touch screen patterns. In *proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 987–996.

[14] Willem G. De Ru and Jan H. P. Eloff. 1997. Enhanced password authentication through fuzzy logic. *IEEE Expert* 12, 6 (1997), 38–45.

[15] Panteleimon Ekkekakis. 2012. Affect, mood, and emotion. *Measurement in sport and exercise psychology* (2012), 321–332.

[16] Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion* 98, 45-60 (1999), 16.

[17] 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*. ACM, 715–724.

[18] Marc Exposito, Javier Hernandez, and Rosalind W. Picard. 2018. Affective keys: towards unobtrusive stress sensing of smartphone users. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. ACM, 139–145.

[19] Raihana Ferdous, Venet Osmani, and Oscar Mayora. 2015. Smartphone app usage as a predictor of perceived stress levels at workplace. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE, 225–228.

[20] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e19.

[21] Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. 2012. What does touch tell us about emotions in touchscreen-based gameplay? *ACM Transactions on Computer-Human Interaction (TOCHI)* 19, 4 (2012).

[22] Enrique Garcia-Ceja, Venet Osmani, and Oscar Mayora. 2015. Automatic stress detection in working environments from smartphones' accelerometer data: a first step. *IEEE journal of biomedical and health informatics* 20, 4 (2015), 1053–1060.

[23] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2017. 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*. ACM.

[24] Javier Hernandez, Pablo Paredes, Asta Roseway, and Mary Czerwinski. 2014. Under pressure: sensing stress of computer users. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 51–60.

[25] Irina Higgins, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. 2016. Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579* (2016).

[26] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *In Proceedings of the 8th International Conference on Learning Representations* 2, 5 (2017).

[27] Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49 (2015), 245–250.

[28] 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)*. IEEE, 157–166.

[29] Stephen Hutt, Joseph F. Grafsgaard, and Sidney K. D'Mello. 2019. Time to Scale: Generalizable Affect Detection for Tens of Thousands of Students across An Entire School Year. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

[30] Oliver P. John, Eileen M. Donahue, and Robert L. Kentle. 1991. The big five inventory - versions 4a and 54. (1991).

[31] Oliver P. John, Laura P. Naumann, and Christopher J. Soto. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research* 3, 2 (2008), 114–158.

[32] Eiman Kanjo, Luluah Al-Husain, and Alan Chamberlain. 2015. Emotions in context: examining pervasive affective sensing systems, applications, and analyses. *Personal and Ubiquitous Computing* 19, 7 (2015), 1197–1212.

[33] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4401–4410.

[34] Iftikhar Ahmed Khan, Willem-Paul Brinkman, Nick Fine, and Robert M. Hierons. 2008. Measuring personality from keyboard and mouse use. In *Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction*. ACM.

[35] Kyung Hwan Kim, Seok Won Bang, and Sang Ryong Kim. 2004. Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing* 42, 3 (2004), 419–427.

[36] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference for Learning Representations* (2015).

[37] Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*. 3581–3589.

[38] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[39] Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2012. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing* 4, 1 (2012), 15–33.

[40] Severin Klingler, Rafael Wampfler, Tanja Käser, Barbara Solenthaler, and Markus H. Gross. 2017. Efficient Feature Embeddings for Student Classification with Variational Auto-encoders. In *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2019)*. 72–79.

[41] 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)*. IEEE, 548–555.

[42] Sarah Martina Kolly, Roger Wattenhofer, and Samuel Welten. 2012. A personal touch: Recognizing users based on touch screen behavior. In *Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones*. ACM.

[43] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine* 16, 9 (2001), 606–613.

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

[45] 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)*. IEEE, 260–264.

[46] Poming Lee, Wei-Hsuan Tsui, and Tzu-Chien Hsiao. 2015. The Influence of Emotion on Keyboard Typing: An Experimental Study Using Auditory Stimuli. *PLoS ONE* 10 (2015).

[47] Alex Leow, Jonathan Stange, John Zulueta, Olusola Ajilore, Faraz Hussain, Andrea Piscitello, Kelly Ryan, Jennifer Duffecy, Scott Langenecker, Peter Nelson, and Melvin McInnis. 2019. BiAffect: Passive Monitoring of Psychomotor Activity in Mood Disorders Using Mobile Keystroke Kinematics. *Biological Psychiatry* 85, 10 (2019), S102–S103.

[48] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. ACM, 389–402.

[49] Katharina Lochner and Michael Eid. 2016. *Successful emotions: how emotions drive cognitive performance*. Springer.

[50] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 351–360.

[51] H. Lv and Wen-Yuan Wang. 2006. Biologic verification based on pressure sensor keyboards and classifier fusion techniques. *IEEE Transactions on Consumer Electronics* 52, 3 (2006), 1057–1063.

[52] Hai-Rong Lv, Zhong-Lin Lin, Wen-Jun Yin, and Jin Dong. 2008. Emotion recognition based on pressure sensor keyboards. In *2008 IEEE International Conference on Multimedia and Expo*. IEEE, 1089–1092.

[53] Alban Maxhuni, Pablo Hernandez-Leal, L. Enrique Sucar, Venet Osmani, Eduardo F. Morales, and Oscar Mayora. 2016. Stress modelling and prediction in presence of scarce data. *Journal of biomedical informatics* 63 (2016), 344–356.

[54] Albert Mehrabian and James A. Russell. 1974. *An approach to environmental psychology.* MIT Press.

[55] Fabian Monrose, Michael K. Reiter, and Susanne Wetzel. 2002. Password hardening based on keystroke dynamics. *International Journal of Information Security* 1, 2 (2002), 69–83.

[56] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When attention is not scarce-detecting boredom from mobile phone usage. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 825–836.

[57] Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J. Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 281–290.

[58] Robert Rosenthal and Ralph L. Rosnow. 1991. *Essentials of behavioral research: Methods and data analysis*. Vol. 2. McGraw-Hill New York.

[59] Mar Saneiro, Olga C. Santos, Sergio Salmeron-Majadas, and Jesus G. Boticario. 2014. Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches. *The Scientific World Journal* 2014 (2014).

[60] Zhanna Sarsenbayeva, Niels van Berkel, Danula Hettiachchi, Weiwei Jiang, Tilman Dingler, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2019. Measuring the effects of stress on mobile interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019).

[61] Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social science information* 44, 4 (2005), 695–729.

[62] Wenjun Sun, Siyu Shao, Rui Zhao, Ruqiang Yan, Xingwu Zhang, and Xuefeng Chen. 2016. A sparse auto-encoder-based deep neural network approach for induction motor faults classification. *Measurement* 89 (2016), 171–178.

[63] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R. Schinazi, and Markus Gross. 2019. Affective State Prediction in a Mobile Setting using Wearable Biometric Sensors and Stylus. In *Proceedings of the 12th International Conference on Educational Data Mining*. 198–207.

[64] Woebot 2019. Website. (2019). Retrieved January 8, 2020 from `https://woebot.io`.

[65] Alex J. Zautra. 2006. *Emotions, stress, and health*. Oxford University Press, USA.

[66] Zendesk 2019. Website. (2019). Retrieved January 8, 2020 from `https://www.zendesk.com/`.