How Do You Feel Online? Exploiting Smartphone Sensors to Detect Transitory Emotions during Social Media Use

MINTRA RUENSUK, UNIST, Republic of Korea
EUNYONG CHEON, UNIST, Republic of Korea
HWAJUNG HONG*, Seoul National University, Republic of Korea
IAN OAKLEY*, UNIST, Republic of Korea

Emotions are an intrinsic part of the social media user experience that can evoke negative behaviors such as cyberbullying and trolling. Detecting the emotions of social media users may enable responding to and mitigating these problems. Prior work suggests this may be achievable on smartphones: emotions can be detected via built-in sensors during prolonged input tasks. We extend these ideas to a social media context featuring sparse input interleaved with more passive browsing and media consumption activities. To achieve this, we present two studies. In the first, we elicit participant's emotions using images and videos and capture sensor data from a mobile device, including data from a novel passive sensor: its built-in eye-tracker. Using this data, we construct machine learning models that predict self-reported binary affect, achieving 93.20% peak accuracy. A follow-up study extends these results to a more ecologically valid scenario in which participants browse their social media feeds. The study yields high accuracies for both self-reported binary valence (94.16%) and arousal (92.28%). We present a discussion of the sensors, features and study design choices that contribute to this high performance and that future designers and researchers can use to create effective and accurate smartphone-based affect detection systems.

CCS Concepts: • **Human-centered computing** → **Smartphones**; *Ubiquitous and mobile computing systems and tools.*

Additional Key Words and Phrases: affective computing, emotion detection, classification, smartphones, social media

ACM Reference Format:
Mintra Ruensuk, Eunyong Cheon, Hwajung Hong, and Ian Oakley. 2020. How Do You Feel Online? Exploiting Smartphone Sensors to Detect Transitory Emotions during Social Media Use. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 150 (December 2020), 32 pages. https://doi.org/10.1145/3432223

1 INTRODUCTION

Social media use is widespread. In 2019, 72% of US adults reported use of at least one Social Networking Service (SNS) and most (74%) indicated they log in daily [21]. Facebook, one of the most popular sites, has more than 220 million US users and enables posting, browsing and commenting in and on diverse media formats including text status updates [7], images [124] and videos [95] that are either created personally or drawn from online sources such as other SNS [10] or news [117] and entertainment channels [102]. This material often evokes affective responses [6]. Indeed these reactions are arguably a key driver for use of the service and can have many positive

*Corresponding authors

Authors' addresses: Mintra Ruensuk, UNIST, Ulsan, Republic of Korea, mintra@unist.ac.kr; Eunyong Cheon, UNIST, Ulsan, Republic of Korea, beer@unist.ac.kr; Hwajung Hong, Seoul National University, Seoul, Republic of Korea, hwajunghong@snu.ac.kr; Ian Oakley, UNIST, Ulsan, Republic of Korea, ian.r.oakley@gmail.com.

2474-9567/2020/12-ART150 $15.00
https://doi.org/10.1145/3432223

benefits [85]. For example, it has been reported that Facebook users can feel closer to friends [17], achieve or experience increased levels of social capital [37] or better cope with stress [15] due to the consumption of contents on the service. However, the affective qualities of SNS have also been linked to a broad range of more negative outcomes and behaviors such as decreased affective well-being (potentially due to envy [113], or unflattering social comparison [97]), cyber-bullying [101], flaming [2], trolling [25], and "indignant disagreement"[73]. These phenomena have contributed to the emergence of terms such as "Facebook Depression" [87] defined as depression that develops when spending prolonged periods on social media sites.

Researchers have responded to the ubiquity of emotions on SNS by developing systems that detect emotions in order to achieve various goals such as protecting users from inappropriate content [24], identifying states such as excessive sadness [126], loneliness [92] or even conditions such as depression [32], postpartum depression [30] and the cumulative stress that signifies greater suicidal tendencies [74]. This work typically relies on analyzing users' messages and posts for language [59, 91], sentiment [79], or style [30, 32]. While valuable, we note these linguistic approaches suffer from several critical problems. First, they are reliant on access to highly private user messages and posts. Second, they require large data sets captured over the medium-to-long term—typically thousands to hundreds of thousands of posts [30, 50, 120]. Finally, they are based on explicit user-generated content, whereas research suggests that the vast majority of user activity on SNS is the consumption of content generated or posted by others [16], in the form of viewing and browsing feeds and media. These factors mean that while current emotion detection on SNS works well given access to long-term records of user posts, it is not applicable to live analysis of the typical day-to-day SNS experience of emotions that are evoked in the course of fragmented browsing and commenting on content posted by one's connections.

This is problematic because the transient emotional states that occur in everyday settings have wide ranging impacts. Although they are typically overlooked by the individuals who experience them [5], research indicates transient emotions influence a very broad set of general human qualities and activities, including: aspects of cognitive performance such as attention [38] and decision making [5]; social assessments in the form of trust [36] and; in the prevalence of behaviors such as procrastination [110]. Transient emotions can even reduce accuracy in the trivial repetitive work task of number entry [18]. Reflecting the importance of these passing affective states, and the mundane settings in which they typically occur, researchers have identified an opportunity to detect an individual's affective states using passive sensor data collected from mobile smart devices. This body of work leverages the advanced touch and motion sensors built into these platforms and has tended to rely on specific tasks that generate a substantial amount of user input, such as gameplay on a tablet computer [44] or Fitts' style tapping tasks and blocks of repeated swipes on a phone [28, 82], to achieve detection rates as high as 89.10% for binary affect. These studies highlight the potential for smartphone sensing systems that can detect the affective states of users and enable applications to adapt or respond appropriately.

We seek to build on this promising work by applying the beneficial properties of sensor-based emotion detection during smartphone use—the immediacy and reliance on data generated during generic user interface events such as taps or swipes—to a real-world SNS scenario. To achieve this, we focus on two major research gaps between prior work and our target SNS usage scenario. The first gap is between the highly structured, atomic and controlled tasks (e.g., Fitts' tapping [82], number entry [18], swipe-based gameplay [44]) that have been previously studied and the diverse, sporadic media consumption-orientated activities that characterize SNS use [113]. To address this gap, and inspired by related work in areas such as gaze-based personality detection [8], we study a novel sensor channel—mobile device-based eye-tracking—that directly and continuously captures user data during browsing and media consumption activities. We initially study this approach's feasibility in a controlled setting: we use pre-validated materials (both videos [98] and images [64]) to elicit emotions in artificial tasks designed to replicate common types of viewing, browsing, and posting interactions that occur during social media use. We measure the transitory emotional states evoked via a standard self-assessment questionnaire [115] and show that binary affect can be predicted with up to 93.20% accuracy, or with up to 88.49% using eye-tracking data alone. Encouraged by these results, we identify a second research gap in terms of the ecological validity of affect elicitation procedures used in both our study and prior work on smartphone-based affect detection. Specifically, we highlight a reliance on controlled exposure [8, 28, 82] to validated and vetted affect elicitation materials [64, 98] and suggest that reactions to such contents may differ from those that occur in response to naturally occurring affect. To address this gap, we conducted a follow-up study using an instrumented version of the popular SNS application Facebook and focused on predicting participants' spontaneous affect as it emerged from normal interactions (e.g., browsing the news feed, reacting and writing comments) with their private social media account. In this study, we captured self-reports of both valence and arousal [12] to support a more detailed analysis of transitory emotion. The results indicate that binary valence can be predicted with up to 94.16% accuracy and binary arousal with up to 92.28% (or 92.68% and 89.42%, respectively, for eye-tracking features alone). We argue that these figures are sufficient to support emotion-aware application and service design in the context of SNS use.

The core contribution of this paper is these demonstrations – we show that transitory emotions can be predicted using off-the-shelf smartphones during SNS use with accuracies of up to 94% in a binary classification task. This is meaningful in that it moves beyond prior work's focus on artificial and repetitive input tasks and vetted affect elicitation materials to a real-world application use case: the emotions we experience on social media matter [24, 32]. The techniques used to achieve these outcomes are also non-trivial, involving novel sensor channels, procedures and methods. To operationalize our work for future researchers, designers and developers who seek to include affective detection capabilities in their work, we provide a detailed explanation of how the different aspects of our studies led to the strong performance we report. Specifically, we discuss the impact of normalization procedures, assessment periods, different sensor channels (motion, touch and eye-tracking) and the role of individual features. This discussion unpacks the results of our studies and enables us to close each with recommendations for how future researchers and designers can successfully integrate affect detection functionality into their studies and applications in terms of what content they should monitor, when they should do it and how they should analyze it. In this way, this paper contributes practical knowledge that will facilitate the development of future affect detection systems on mobile devices.

2 RELATED WORK

2.1 Models of Emotions and Measurement

The research literature describes a wide range of models of emotion. A common and longstanding approach is to map emotions to different points on a multi-dimensional space [119]. Of the various two-dimensional emotion schemes that have been proposed, the circumplex model [96, 116] is based on perpendicular axes of valence and arousal and is arguably the most widely used [96]. It has been frequently applied in HCI research, including in examples such as De Choudhury et al. [29]'s work to quantify moods on Twitter through crowdsourced lexical analysis and Lee et al. [66]'s work on a Twitter client that can detect emotions during posting by analyzing data from smartphone sensors. Due to its longstanding and widespread use, including in closely related research [82], we opt to use the circumplex model throughout this paper.

We operationalize the emotions in this model as short-lived physiological and behavioral changes triggered by visual (pictures) and multi-modal (video) stimuli [111]. In this context, the emotional responses an individual experiences can be measured using traditional psychology-based approaches such as self-assessment, an established method that is widely viewed as simple, quick, and effective [100]. There are many well-known self-assessment instruments [72] such as the Positive Affect Negative Affect Schedule (PANAS) [115] and the Self Assessment Manikin (SAM) [12]. PANAS has been widely deployed for various purposes such as assessing life satisfaction [61] and monitoring mental health [11]. It consists of 20 words representing different emotions and respondents indicate their feelings with respect to these terms over various time periods including at the present moment [115]. It is primarily used to assess valence. SAM is a collection of pictorial manikins that reflect varying degrees of emotional valence, arousal and dominance, typically accompanied by a 9-point numeric scale [12]. Due to the simplicity of the SAM instrument [55], it has been widely employed in many research areas including HCI [52, 82]. To achieve different experimental objectives, in terms of the scope, time cost and reliability of the affect self-assessments captured, we deploy either PANAS or SAM to capture transient emotional states in our studies.

2.2 Emotion in Social Media

Considerable prior work has sought to understand how people experience emotions on social media [13, 67]. This work is predominantly reliant on corpora of user-generated data, most commonly textual content such as messages and posts. For example, De Choudhury and De [31] study how posted text can reveal emotional and mental distress [31], while Hasan et al. [50] examine similar data to infer people's basic emotional states, Mohammad et al. [80] propose a system that uses sentiment analysis of 160 character messages (tweets) to predict both emotion and purpose [80] and Gao et al. [43] present a tool that classifies Facebook chats and posts as exhibiting either positive or negative emotions. Furthermore, in addition to text-based analysis, other types of user data that have been deployed to detect or predict emotions includes pictures [39, 120, 124] and social interaction information [39, 70]. The diversity of this work highlights the importance and ubiquity of the emotions expressed, and the emotional content posted, on SNS.

However, from the point of view of detecting transient emotional states, there are several limitations with this prior work and, specifically, the data sources it relies on. Perhaps most fundamentally, posting content is not the only, nor even the dominant, activity on many SNS [16]. Rather, much activity on SNSs is browsing, viewing, reading or watching content posted by others [118], so called *passive* activities that researchers have suggested are disproportionately responsible for the reductions in affective well-being [113] that some SNS users experience. Techniques based on lexical or content analysis lack viable data sources to cover these important and commonplace use scenarios. Lexical analysis typically also operates over sustained time periods (e.g., over multiple posts [30, 31]) with the goal of detecting relatively persistent affective states [20, 80]. It is likely to be less effective at detecting the transient states that are the focus of the current article, as they may prevalently occur in response to passive consumption activities and thus express relatively weakly in posted text. Finally, we note there are privacy concerns inherent in systems that analyze a user's posted text and other content [43, 57]. These can exert a wide range of impacts including reducing the number of posts users produce [104].

Reflecting these concerns, in this paper we avoid reliance on any corpora of private user-generated data. Instead, our technique to detect emotions uses raw sensor information gathered from smartphones that represents essentially physiological data about the motions, touches and glances that users make with, on and at their devices. This approach allows us to capture moment-to-moment data during all social media activities, including viewing content as well as posting it, and has previously shown promise in capturing transitory affective states [82]. In addition, privacy concerns related to these types of data are reduced compared to content forms such as text messages and photographs [57].

2.3 Smartphone Sensors to Detect Emotion

A range of physiological responses accompany emotional changes. Previous studies indicated emotion can be detected reliably from physiological reactions such as the production of gestures and postures [48] and also from diverse physiological signals [108]. However, these traditional techniques typically require dedicated hardware such as, for example, high-end heart rate variability monitors [22, 106]. The cost of acquisition and lack of availability of these devices in the general population, together with the requirement that they be continuously worn, represent major barriers for widespread adoption of this type of emotion detection. To address these problems, several prior studies have built on the observation that smartphone use is associated with affective experiences [97, 122] to use these widely deployed platforms to detect emotional states [19, 28, 68]. While evidence for the effectiveness of this general approach is gathering weight, we note that existing studies focus on empirically elicited affect and artificial, sustained data input tasks. Work examining more applied scenarios, such as SNS use, is in its infancy— Zhang et al. [121], for example, simply propose their data collection infrastructure and procedure. The following subsections review the literature on mobile phone based affect detection, categorized by the sensing channels it relies on and highlighting how the studies reported in this article advance the field. We note that many of the articles in this review include data from multiple sensor sources: our review structure in a simplification for convenience and multi-modal approaches are, in fact, the default in this research area.

2.3.1 Motion Sensors. Smart device motion sensors capture the movements of a mobile device in terms of accelerations, rotations and, via filtering, absolute changes in orientation. Their ubiquity, low cost and the richness of the data they capture has meant they have been widely used in work seeking to detect emotions during smartphone use [68, 89]. Early work used accelerometer data to distinguish people's activity (walking, sitting, running, standing) and showed this, in conjunction with other smartphone data such as location traces and communication patterns, could predict daily mood with an accuracy of around 50% on three different five-level dimensional scales [75]. More recent work has suggested that phone accelerometer data during walking alone (basically a proxy for gait) can yield detection accuracies of between 50% and 75% for valence and arousal, each segmented into the three levels of low, neutral and high [88]. Motion data has also proven useful in classification of persistent clinical conditions: Cao et al. [19] report that motion data during typing contributes to a 90.31% accuracy in predicting the score participants attain in clinical assessments of bipolar disorder and also provides insights into the physical behaviors underlying this, such as a tendency for participants with depressive or manic symptoms to hold their phone at more horizontal angles. These diverse results highlight that motion and pose data captured from smartphones represent a powerful channel that can be used to detect and predict a wide range of affective states. Based on these prior results, we use motion sensor data in our studies. We move beyond existing studies of this modality by exploring its veracity in detection of the transient affective states that occur in response to relatively short passive content consumption activities in an SNS use scenario.

2.3.2 Touch Sensors. Modern touchscreens also provide a rich source of information that can be used to detect affective states. Work has tended to focus on situations in which a large amount of data is generated rapidly such as during intense game play [54], text entry [45, 114] , or in repetitive scripted experimental tasks such as Fitts' law tapping or performance of sequences of swipes [82]. Application scenarios have tended to focus on these data collection scenarios with, for example, an emphasis on real-time monitoring [54] or personalization [44] of game experience. Researchers have also indicated that advanced touch features can be useful: touch pressure, for example, helped Gao et al. [44] achieve accuracy of between 69% and 77% in discriminating four different emotions: excited, relaxed, frustrated and bored. This study also suggests that specific emotions may be associated with specific touch patterns, such as exerting higher pressure during frustrating experiences. These findings indicate that a wide range of emotions are expressed in the currently observable qualities of touch interaction on smartphones. Based on these promising results, we include touch data in our work. We extend prior work on this modality by contributing data from a more ecologically valid setting involving diverse and sporadic input activities rather than the sustained and consistent tasks that have been previously studied.

2.3.3 Camera Sensors. Sensing data from mobile device camera feeds has been widely adopted to classify people's emotion. An obvious approach is to extract facial expressions [103]. Suk and Prabhakaran [105], for example, report on a real-time mobile application that can recognize facial expressions for six basic emotions (plus neutral) with an accuracy of 72%. A major disadvantage of this approach relates to concerns about the privacy issues involved in the collection and continuous analysis of face image data [1, 53]. We also note there may also be substantial changes in how the face expresses emotions during private mobile device use compared to, for example, social interaction with another person. Data about eye-gaze is another promising source of information about emotions. Prior examples use stand-alone eye-trackers to assess inner states elicited by images [63] or videos [4, 125]. Images, in general, have led to stronger performance, with a reported accuracy of up to 86.74% in a binary arousal classification task compared to accuracies of up to 66% using video media. Due to the nature of the information provided by eye-tracking systems, such as eye positions, pupil sizes, or event occurrence rates (e.g., of blinks), privacy concerns with this type of data are greatly reduced. We note that due to the relatively recent arrival of gaze tracking functionality on consumer smart devices, the potential of this information channel to support affect detection on mobile devices remains scantly explored compared to the more established channels of motion and touch. We are not aware of any prior work using mobile phone based eye-tracking for affect detection and believe we contribute the first data on the suitability of this modality for mobile phone based affect detection.

3 FEASIBILITY STUDY

This study seeks to adapt prior demonstrations of emotion detection using smartphones [82, 122] to a social media scenario. In order to do this, we both *extend* and *restrict* the data collected. Specifically, we extend the data collected through use of an additional sensor channel—we add eye-tracking to the typical channels of motion [19] and touch [44, 47] input. We also extend capture of motion data by recording filtered pose data in addition to raw sensor signals. In contrast, we restrict the data collected through carefully designed input tasks that mimic social media use patterns: images [64] and videos [99] are presented to participants to elicit emotions. Participants were also asked to perform sporadic taps and swipes, and enter short text responses, as they browse, navigate and respond to this content. These diverse and fragmented tasks stand in contrast to the repetitive and highly controlled tasks (e.g., number entry [18] or Fitts' tapping [82]) that have been previously studied in this area. We aimed to assess whether our extensions are sufficient to compensate for our restrictions: of whether emotions can still be accurately recognized by smartphones in this more data-rich but input-sparse setting. The study was approved by the host university's Institutional Review Board (IRB).

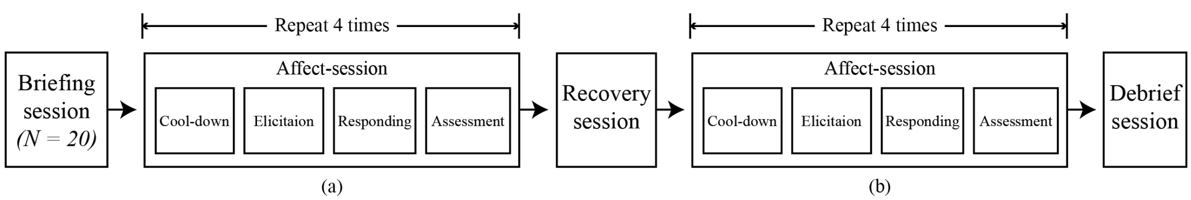

Fig. 1. Experimental Procedure of Feasibility Study. Each *affect-session* consists of four phases: a *cool-down* phase that captured baseline data; an *elicitation* phase in which media were presented; a *responding* phase that collected comments; and an *assessment* phase that recorded people's emotional state through questionnaires. In the first two *affect-sessions*, videos [99] were used in *elicitation* phase, while in the last two sessions, images [64] were shown. The study elicited two basic emotional states–positive and negative affect (marked as (a) and (b)). The order of emotion conditions was fully balanced.

3.1 Study Design

The study sought to assess the predictive power of sensor data from a smartphone in detecting emotions. Following prior work in this area [26, 82], it was structured as a series of *affect-sessions* each of which involved the following four phases: a *cool-down* phase in which baseline data was captured; an *elicitation* phase in which participants browsed evocative media; a *responding* phase in which they typed comments based on their reactions to the media; and an *assessment* phase in which they completed questionnaires to record ground truth about their emotional state in the session. Using this structure, the study elicited two basic emotional states—positive and negative affect—in a repeated measures design. The order in which these emotion conditions was presented was fully balanced among participants to control for order effects. In each emotion condition, participants completed four affect-sessions. In the first two sessions, videos were used to elicit affect, while in the last two sessions, images were used. Participants were also required to complete a recovery session in between the two emotion conditions. This was composed of a mandatory 90-second-video showing natural scenery accompanied by the suggestion they relax. In line with prior work, it was intended to minimize any carryover of emotions between the two conditions [26]. A summary of the study design and procedure is shown in Figure 1. We describe the materials and procedures used in the affect-sessions below.

3.2 Materials

Each affect-session was composed of four sequential phases: cool-down; elicitation; responding and; assessment. They were all displayed within a single mobile study app built for the iPhone X (screen size: 375 x 812 points, or 62.4 x 135.1 mm) using Swift 5. Each phase is described below.

3.2.1 Cool-down. In this phase, we followed typical procedures [8] to ensure participants have the opportunity to recover from prior affective stimuli and also to provide baseline data for normalizing responses to subsequent affective stimuli. Specifically, participants completed a set of four simple and affectively neutral tasks and the data captured during this phase was used to normalize data from the immediately subsequent elicitation, responding and assessment phases. This data was not used as training or testing data during model development. The four tasks generated data for a range of common interface events (tapping, scrolling, typing/chatting) and also uncued eye movement patterns. The tasks are illustrated in Figure 2. They were always presented in the following order and entailed:

- *Target Selection.* Participants tapped a sequence of five circular targets (diameter 80 points, or 13.3mm) that appeared sequentially on the phone in random locations.
- *Scrolling.* Participants were requested to scroll down until a button was reached, then select it. The scroll view was 3400 points high.
- *Text Entry.* Participants were presented with a single randomly selected phrase from the set introduced by MacKenzie and Soukoreff [76]. They were required to enter this phrase into a text box using the standard iPhone on-screen keyboard.
- *Eye movement.* Following prior work [40], participants were presented with a fixation stimuli—a black cross on a white background—for 15 seconds and instructed to rest.

3.2.2 Elicitation. Videos and images were presented to elicit emotions in task structures designed to mandate browsing the content so that touch interactions such as taps and scroll events could be recorded. The procedures and content for each media type differed and are described below.

- *Videos.* Libraries of emotionally labeled video clips are commonly used to reliably elicit emotions in empirical settings [8, 82]. In this study, each participant was exposed to a total of four film clips from the widely used FilmStim library [98]. We selected specific videos to achieve strong emotional peaks. The videos were FilmStim clips from the movies "Seven", "American History X", "Benny & Joone" and "There's Something about Mary" (subsequently referred to as "Mary" in this manuscript) that are rated as eliciting, respectively, fear, anger, tenderness, and amusement. We treated the first two clips as eliciting negative affect and the last two as eliciting positive affect. To reduce participant fatigue and provide more evenly long clips [8], we shortened them by conducting a brief study with three participants. Each participant watched the clips and verbally indicated periods in which they experienced emotional peaks. The clips were then edited, to

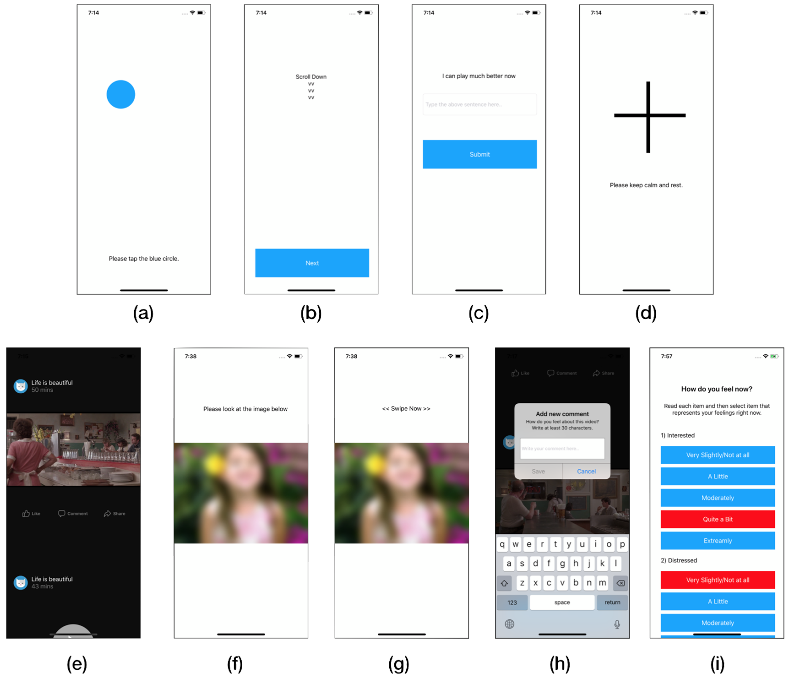

Fig. 2. User Interface of the study app. The *Cool-down* phase consists of four tasks. They are: (a) Target Selection, (b) Scrolling, (c) Text Entry, and (d) Eye movement. The *Elicitation* phase shows either a set of short video clips (e), or an image gallery (f). After eight seconds displaying each image, a right swipe control is enabled that moves to the next image (g). The *Responding* phase (h) displaying a popup window for entering a short comment. The *Assessment* phase (i) displayed the PANAS self-assessment questionnaire. Only the English language version of the app is shown.

between 57 and 80 seconds, to include only these affectively prominent periods (see Appendix A.1 for full details) To make the viewing experience more akin to browsing social media contents, we sliced each clip into five sequential segments arranged in a vertically tiled "news-feed" layout—see Figure 2e. Participants tapped to watch each segment, then needed to scroll down to view the following segment. Participants were not able to scroll past segments until they had been watched in full, ensuring the complete clip was watched in chronological order.

- *Images*. Sets of tagged images are also widely used to elicit emotions in empirical studies [8, 9, 27, 28, 41]. In this study, we used images from the International Affective Picture System (IAPS) dataset [64], one of the most common resources in this area. It provides images rated in terms of valence, arousal and dominance. To protect participants from potential emotional distress, we excluded images of extreme valence that could be disturbing as suggested by Goncalves et al. [49]. We then selected ten images with high valence (mean: 7.65, SD: 0.29) to elicit positive affect and ten with low valence (mean: 2.3, SD: 0.57) to elicit negative affect (see Appendix A.2). We grouped these images into four sets of five, one for each affect-session. To integrate image presentation with user input, we used a simple gallery design that showed each image in a set of five for a minimum of eight seconds, then required participants to swipe right to move to the next image—see Figure 2f, g.

3.2.3 Responding. After viewing and browsing the content to elicit emotions, participants were presented with a popup window that prompted them to enter a short (minimum of 30 characters) comment describing their feelings and reactions to the content—see Figure 2h. The goal of this phase was to capture touch input data during the typical social media task of posting a brief textual comment in response to viewed content. We did not store the entered text to ensure privacy.

3.2.4 Assessment. In this phase, participants completed the 20-item PANAS emotion self-assessment questionnaire. We instructed participants to report on their current feelings and emotions, an established use of this instrument [3]. To ensure comprehension we presented PANAS in both English and the local language [69]. We used a non-conventional presentation format for PANAS that stacked the questions vertically (see Figure 2i) while retaining the standard five response options. We opted for this design to capture additional touch data events (e.g., taps, scrolling, swiping) during questionnaire completion.

Table 1. Feature groups and specific features for each group.

Group	Features	Description
Motion [34]	Acceleration (x, y, z)	Acceleration in G's (gravitational force)
	Rotation (x, y, z)	The rotation rate as measured by the device's gyroscope.
	Core Motion (pitch, roll, yaw)	Processed device-motion data that remove environmental bias.
Touch [35]	Touch count	The number of times the screen was touched in 15s window.
	Hold duration	The duration of each touch in a 15s window
	Inter-tap interval	The time between each touch in a 15s window
	Speed	The overall movement speed of each touch in a 15s window
	Distance	The overall distance of each touch in a 15s window
	Touch pressure	The force applied during each touch (sampled at 60Hz)
	Touch area	The radius of each touch (sampled at 60Hz)
Eye-tracking [33]	Blink	The coefficient describing closure of the eyelids over the [left|right] eye.
	Look down	The coefficient describing movement of the [left|right] eyelids consistent with a downward gaze.
	Look in	The coefficient describing movement of the [left|right] eyelids consistent with a [right|left]ward gaze.
	Look out	The coefficient describing movement of the [left|right] eyelids consistent with a [left|right]ward gaze.
	Look up	The coefficient describing movement of the [left|right] eyelids consistent with an upward gaze.
	Eye squint	The coefficient describing contraction of the face around the [left|right] eye.
	Eye wide	The coefficient describing a widening of the eyelids around the [left|right] eye.

3.3 Procedure

The experiment was conducted in a quiet office environment with seated participants, and controlled light and temperature. Participants were first introduced to the experimental procedure in a briefing session. Participants were informed that the content seen in the study might evoke strong emotional responses and reminded of their right to terminate participation in the study at any time. If they opted to continue, participants then signed consent forms, completed demographics and were presented with detailed instructions and the mobile phone running the study. An experimenter was available to answer questions about the instructions and demonstrate use of the app. We provided no indication about the detailed purpose of the study, instead simply indicating it was intended to explore the design of new social media services. The study then began and each participant completed four affect-sessions, as defined in the materials section. After the study was complete, we debriefed participants as to the purposes of the study and solicited comments.

3.4 Participants

Twenty participants were recruited through word of mouth, social media, and a local university's online forum. To create a less variable environment for eye-tracking, participants were screened to ensure they did not wear corrective eye-glasses or contact lenses and had not previously undergone laser eye surgery (e.g., LASIK).

Additionally, we screened participants to ensure they were an account holder on at least one locally common social media platform (e.g., Facebook, YouTube, Twitter, Instagram, Facebook Messenger, KakaoTalk). They were requested not to wear eye-makeup during the study session. In total, 11 participants were male and 9 female, with a mean age of 24.25 (SD = 5.4). They came from seven different countries and 14 were undergraduate students, 3 were graduate students, and 3 were researchers. Using 5-point Likert scales (1 = Very poor, 5 = Very good), they self-reported a high familiarity with both computers (M=4.65, SD=0.62) and smartphones (M=4.8, SD=0.87) and were confident in their ability to understand English conversation (M=4.45, SD=0.67). Participants received approximately $10 as compensation for their participation in the experiment.

3.5 Data Collection and Preprocessing

3.5.1 Features. We retrieved three types of features from the iPhone during the study: 1) motion data from Core Motion framework [34], 2) touch data from UITouch [35], and 3) eye-tracking data from ARKit [33]. The frequency of motion data was set at 60Hz, eye-tracking data was 30Hz and touch data was recorded for each touch event, with continuous data (e.g., pressure, size) sampled at 60Hz. Features were chosen based on previous work in classification of human inner states using mobile devices (e.g., [45, 82, 94]), such as speed, rotation, and acceleration. We describe each of the 30 features we captured in Table 1. Data was collected in all four phases of each affect-session.

3.5.2 Data Preprocessing. We analyzed the data using Scikit-learn [93]. We first imputed data to replace missing or undefined values (0.001% of data) using a k-Nearest Neighbors approach based on the full set of other column values. We then divided the data from elicitation, responding and assessment into non-overlapping windows of 15 seconds, a typical period used in prior studies of smartphone-based emotion detection [123]. For motion and eye-tracking data in each window, we then calculated basic summary statistics for each feature: minimum, maximum, mean, median, standard deviation, and variance. For continuous touch features such as pressure and size, we calculated summary statistics for each touch, then averaged these over all touches within each window. Summary statistics were also calculated over all touches in a window for the measures of hold-duration, inter-tap interval and speed, while touch count was simply represented as a total.

In order to reduce the impact of features with difference ranges of magnitude and natural performance variations over time, we scaled and normalized all data. Specifically, for each affect-session, we expressed each metric captured in the elicitation, responding and assessment phases in terms of its magnitude relative to extremes of the data observed in the preceding cool-down phase [8]. We then normalized data for each feature. These processes preserved the shape of the original distributions, did not reduce the importance of outliers, and also ensured all features were on the same relative scale.

To define target labels, we analyzed the PANAS self-assessment score. We defined PANAS binary emotion via an affect balance score [61] computed by subtracting the negative affect score from positive affect score (PA - NA). Following prior work [61], we partitioned the affect scores into positive and negative zones using entropy minimization to derive a simple threshold value: -1.5 for our data set. This results in one target label per affect-session. Thus, each window in an affect-session is labelled as exhibiting positive or negative affect based on the self-reported data.

3.5.3 Classification. We selected eight classifiers for study based on closely related prior work (e.g. in [44, 45, 82]): ZeroR, AdaBoost (AB), Decision Tree (DT), k-Nearest Neighbour (kNN), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and RBF-kernel Support Vector Machine (SVM). We used default settings for these algorithms. We included the ZeroR classifier as a baseline; it simply predicts the most frequent label it encounters in the training set. To minimize the chance of over-fitting and reduce subject bias, we assessed the performance of each classifier using leave-one-subject-out cross-validation. This approach entails constructing an independent model for each participant in a study. Each participant's model uses the full set of their data for testing and all data from other participants for training [8, 60]. As our study involved 20 participants, we thus constructed 20 independent models and all results reported depict the mean performance over this set of models. The advantage of using this method is that it provides a robust and unbiased evaluation of how well models derived from training data fit independent testing data.

Furthermore, for each model, we executed both feature selection and data processing steps. Feature selection used both filter and wrapper methods [112] to screen and refine the final feature set and was conducted using only training data. The process began by first removing constant features with zero variance and quasi-constant features with variance less than 1%. We then calculated the correlation matrix of the remaining features. If two or more features were highly correlated (Pearson's r greater than 0.8), we retained only the one most highly correlated with the emotion labels. Second, we applied recursive feature elimination using a Linear-kernel Support Vector Machine (SVM) and five-fold cross-validation to the remaining features. Data processing procedures tackled the class imbalances in our data set (see Section 3.6.1). Imbalanced classes present a challenge for predictive modeling machine learning algorithms as these generally expect an equivalent number of examples for each class. This can lead to low performance for minority classes. To handle our imbalanced data sets, we over-sampled the minority class using Synthetic Minority Oversampling Technique (SMOTE) [23]. This method generates synthetic samples based on the nearest neighbors of feature values in the minority class and reduces the chance of over-fitting by providing more related minority class samples for classifiers to build larger decision regions that contain nearby minority class points [23]. We applied SMOTE only to the training data. We kept the independent testing data unbalanced to reflect the genuine distribution of classes we recorded in the study. Finally, we assessed the classification performance for each model by calculating mean accuracy and class-wise F_1-score.

Although leave-one-subject-out cross-validation approach confers strong benefits in terms of greatly reducing the impact of model over-fitting, the per-subject feature selection process it entails prevents ready production of a universal list of recognizer features. In order to be able to report on and discuss the features used in our recognizer, we generated a representative set of features by conducting an additional two-phase feature selection process on the full set of normalized data in the study. In this process, we filtered out 125 features (56%) from the original set of 222 features, then applied cross-validated recursive feature elimination to the remaining 97 features to yield a final set of 25 *representative* features, each of which is presented in Table 2. These representative features overlap substantially with those in the leave-one-out models—a mean of 75.2% appear in each leave-one-out model.

3.6 Results

3.6.1 Descriptive Statistics. For the media used in each affect-session, Table 3 describes the binary affect it was intended to elicit, the number of study participants self-reporting either positive or negative affect, and a summary of the amount of data that was recorded. Due to variations in sample rates, we logged a large set of motion data—approximately 1.802 million samples—and a somewhat reduced set of eye data—0.537 million samples. Eye data was recorded at a lower rate than expected due to tracking failures resulting from variations in how a user's face was positioned and oriented with respect to the phone [58]. Touch data, recorded only during actual touch input, occurred much less often. A total of 31,588 touches were recorded over the whole study. The table also notes the number of 15-second sessions we were able to extract from the data—1,927 in total, corresponding to 8 hours, 1 minute and 45 seconds of logging. It is also interesting to note that while elicitation of positive affect was largely successful, with 93.75% of participants providing a well aligned report of positive affect, elicitation of negative affect was achieved less reliably. Participants self-assessed their emotions as aligned with our intent to elicit negative affect in only 65% of affect-sessions.

3.6.2 Overall Classifier Performance. Table 4 displays mean accuracy and class-wise F_1-scores for predicting binary affect using all eight classifiers. RBF-kernel SVM performs best with a mean accuracy of 93.20%. This result

Table 2. The 25 representative features selected using filter and wrapper methods with linear-kernal SVM with five-fold cross-validation. Features listed in alphabetical order. Appendix A.3 lists ranked feature importance, calculated via feature ablation. $_L$=left eye, $_R$=right eye.

Feature	Sensor	Sub-feature
Acceleration, x-axis	Motion	maximum, mean, minimum
Acceleration, y-axis	Motion	maximum, mean, minimum, standard deviation
Acceleration, z-axis	Motion	mean
Eye squint	Eye-tracking	minimum_L, variance_L
Look in	Eye-tracking	$\text{maximum}_{L,R}$, mean_L, minimum_L
Roll	Motion	mean
Rotation around x-axis	Motion	mean, minimum
Rotation around y-axis	Motion	mean
Rotation around z-axis	Motion	mean, minimum
Touch area	Touch	maximum, mean
Touch pressure	Touch	minimum
Yaw	Motion	maximum, minimum

Table 3. Descriptive statistics of self-reported affect scores, number of records of data for each sensor type and number of sessions. *Elicit.*=Elicitation & Responding phases, *Assess.*=Assessment phase.

Material	Format	Elicited Emotion	Self-reported emotion		Number of records			# of sessions	
			Positive	Negative	Motion	Touch	Eye	Elicit.	Assess.
Benny & Joone	Video	Positive	20	0	283,821	4,669	107,443	257	45
Mary	Video	Positive	16	4	237,062	4,245	59,539	216	36
IAPS Set 1	Image	Positive	19	1	184,477	3,953	46,759	116	82
IAPS set 2	Image	Positive	20	0	170,011	3,769	42,532	105	80
Seven	Video	Negative	12	8	330,380	4,390	137,706	212	136
American History X	Video	Negative	3	17	261,478	3,671	59,924	179	100
IAPS Set 3	Image	Negative	8	12	178,408	3,503	43,510	135	59
IAPS Set 4	Image	Negative	5	15	156,316	3,388	39,735	122	47

Table 4. Accuracy and class-wise F_1-score (in %) for affect detection in the feasibility study for all eight classifiers and using data that was normalized relative to cool-down phase. Results from the best performing recognizer, RBF-kernel SVM, are highlighted in bold.

Metric	ZeroR	AB	DT	kNN	LR	NB	RF	SVM
Accuracy (%)	66.32	81.37	82.25	91.02	70.89	59.15	90.30	**93.20**
Positive Affect (F_1)	79.71	88.54	85.16	93.38	76.88	62.86	92.53	**94.83**
Negative Affect (F_1)	00.00	73.51	74.01	88.09	61.57	55.91	85.43	**89.93**

is 25.98% higher than the ZeroR baseline classifier [65] that simply selects the dominant class in the training set. Based on this strong performance, we limit reporting of follow-up tests to results from only the SVM classifier. We also note that the F_1-scores show a consistent trend in recognition accuracy—the dominant class of positive affect yields better performance than the minority negative affect class. While reduced performance with a minority class is not unexpected, we argue that the result from our selected RBF-kernel SVM recognizer (F_1-score of 89.93%) remains strong and is sufficient to warrant further study.

Table 5. Accuracy (in %) using RBF-kernel SVM models for different feature sets (by sensing channel), study phases, normalization procedures and media types. Class-wise F_1-scores for all models are included in Appendix A.4.

Feature Set	*Phase*	**Cool-down**			**Within session**		
		Both	**Video**	**Image**	**Both**	**Video**	**Image**
All features	All	93.20	92.04	93.56	83.83	82.17	90.64
	Elicitation & Responding	91.20	89.35	92.67	78.42	81.51	80.32
	Assessment	84.03	83.94	87.22	77.82	75.34	71.52
Motion Features	All	90.35	84.51	89.81	82.91	84.89	72.68
	Elicitation & Responding	89.79	83.58	87.90	83.11	78.14	84.14
	Assessment	82.62	82.87	83.15	70.82	68.51	70.54
Touch Features	All	77.32	78.40	82.39	67.22	68.08	76.94
	Elicitation & Responding	76.68	78.45	80.44	59.84	57.52	68.89
	Assessment	77.52	80.91	79.68	67.62	67.37	71.64
Eye-tracking Features	All	83.40	85.92	88.49	67.22	67.45	69.50
	Elicitation & Responding	83.61	84.30	85.42	65.78	66.49	67.38
	Assessment	59.49	69.54	64.21	54.56	53.27	54.05

3.6.3 Performance of Classifier Variants. We then constructed models from subsets of the data captured in the study to explore how various aspects of the experimental design impacted recognition performance. Specifically, for each leave-one-out model we looked at all combinations of the following four variables: the *sensor(s)* used to contribute features to the recognizer; the *phases* data is drawn from; the *normalization* procedures applied and; the *media* used to elicit affect. For sensors, we examined data from *all sensors* and from the individual channels of *motion*, *touch* and *eye-tracking*. This analysis sought to assess how the different sensor channels contributed to the overall recognition performance. In terms of phases, we considered three sets of data: *all* phases; *eliciting and responding* phases and; the *assessment* phase. The goal of considering these different phases was to assess classification accuracy of transitory emotions when participants are directly engaged with the viewed contents (*eliciting and responding* phases) against that when they were engaged in follow-up tasks, such as self-report of their emotions (*assessment* phase). For normalization procedures, we considered use of data from the *cool-down* phase to standardize data and also more typical normalization procedures based solely on data from *within session*. The goal here was to determine the extent to which the cool-down phase facilitated accurate affect detection. Finally, in terms of media, we examined data from *both* media types and also from *video* media and *image* media alone. Here we sought to determine whether the different forms of media led to better or worse classification accuracy. The goal of these manipulations, in general, was to inform the design of affect recognition systems in practice by shedding light on what data to capture, when to capture it, how to normalize it and what media format provides the most reliable basis for accurate classification. The results of these analyses, all using the optimal RBF-kernel SVM recognizer identified in the initial recognizer comparison, are presented in Table 5 while class-wise F_1-scores are in Appendix A.4. We note trends in F_1-scores largely follow those in accuracy data. Accordingly, in the interests of brevity, we base our discussion on this latter metric alone.

These data show the contribution of study methods to the high overall accuracy. In terms of the different sensor sources, motion shows performance that is most broadly similar to use of the full feature set—accuracy drops by between 0.92% and 17.96%. This reflects the large number of motion features in the representative feature set (16 from 25), the established links between motion and affect [75, 88] and also the richness, sensitivity and maturity of these sensor systems. It clearly indicates that use of motion sensors is imperative for any future mobile affect detection system. Touch performance was under-represented in the representative feature set (3 from 25) and shows mixed recognition performance of between 57.52% and 82.39%. This may relate to the relative scarcity of data captured in this modality compared to the continuous streams of sensor data from motion and eye-tracking. Regardless, its low performance, in comparison to the two other modalities in the study, casts doubt on its value as a mechanism for detecting affect in social media scenarios—in these passive, consumption orientated settings, touch features may be sparse and have limited salience. Finally, eye-tracking also led to mixed performance of between 53.27% and 88.49%. It performed better with data from elicitation and responding phases compared to data from assessment phases, with peak accuracies exceeding those achieved with motion sensors. We argue this indicates that eye-tracking has good potential to serve as an effective tool for affect detection during social media use on mobile devices and that it appears most valuable in the moments when affect is being directly elicited by viewed contents.

In terms of the phase used, data from the elicitation and responding phases is relatively similar to data from all phases (accuracy is reduced by between 0.56% and 10.56%) while data from the assessment phase shows more strongly reduced performance (of between 0.71% and 24.28%). This again highlights the importance of capturing behavioral measures during tasks that are affectively salient [44] rather than in dedicated post-session tasks [82] in order to support high detection accuracy. With regard to the normalization procedures used, the baseline provided by the cool-down phase provided clear benefits—it led to recognition accuracies of between 2.92% and 20.93% higher than those achieved with more standard in-session normalization procedures. Finally, the media format used to elicit affect also impacted recognizer accuracy, with images generally showing limited improvements in accuracy (of up to 11.37%) over video and both media formats. While our study is not designed to tease apart an explanation for these differences, candidate possibilities include the fact images more reliably elicited affect, led to more stable user data (e.g., less variable gaze patterns) or that the input tasks we mandated during image viewing (horizontal swipes) were particularly salient for affect detection. Alternatively, the fact that video and image media were not balanced (with images always following videos in each affect condition) could have influenced these results; while our results show a clear trend, this issue deserves further study in future work.

3.6.4 Feature Variability with Positive and Negative Affect. To shed light on the specific behaviors that characterize elicited positive and negative affect, we followed prior work [82] and ran a series of t-tests on the means of each of the representative features (see Table 2). As this totals 25 independent tests, we applied an alpha threshold of 0.002 (equivalent to applying Bonferroni correction). This analysis yielded a total of eight significant results, covering features derived from motion, touch and eye-tracking. In terms of motion, significantly higher mean acceleration in the x-axes and higher max acceleration in both x- and y-axis (all $p < 0.001$) were associated with positive affect. Similarly, in terms of filtered rotation data, higher maximum yaw data ($p < 0.001$) was also linked to positive affect. These results suggest that positive affect led to more extreme and substantial motions of the phone. The touch data corroborate this—significant results indicated that participants produced both larger (in terms of touch contact area) and more forceful (in terms of applied pressure) taps and swipes when positive affect was elicited (both $p < 0.001$). These more substantial touches correspond to faster and stronger physical motions and, indeed, likely contribute directly to the increased movements of the phone as the device shifts in response to the impacts they represent. Interpreting the remaining two significant results was more challenging. Eye tracking data showed a single significant result: lower minimum values of the "look in" feature for the left-eye, associated with gaze to the left of the device, were associated with positive affect ($p < 0.001$). Finally, positive affect was also associated with lower minimum rotation (sensed via the gyroscope) around the device x-axis ($p = 0.002$). Finally, we note these significant differences are well aligned with ranked feature importance data, as calculated via feature ablation: see Appendix A.3 for details. The eight features that showed significant differences appear in the top nine ranked features.

3.7 Discussion

This study builds on prior work that has demonstrated the feasibility of detecting the transitory emotions of users with smartphone sensors in lab settings involving controlled and protracted input tasks such as repeated data entry [18], prolonged game play [44] or Fitts' law targeting tasks [82]. It sought to assess whether or not these promising prior demonstrations can be applied in more realistic settings where the input tasks undertaken by users are relatively sparse and interleaved with passive activities such as viewing or reading media contents. The results suggest smartphones have very strong potential to detect emotions in these settings: we report accuracy levels as high as 93.20% in the task of distinguishing between explicitly elicited positive and negative affect. This compares well to prior work in this area which has reported between 76% accuracy in the classification of a set of multiple dimensional basic emotional states [122] to 90.31% accuracy in the arguably simpler task of predicting the presence of mood disorders [19].

It is worth discussing a range of factors that likely contribute to this high and improved performance. First and foremost, compared to prior work on smartphone affect detection, we collected novel forms of data, including both the previously unexplored channel of eye-tracking and filtered (rather than raw) phone pose data. Together these channels account for 36% (9 from 25) of the representative feature set shown in Table 2. Eye-tracking features alone achieved peak affect detection accuracies of 88.49%. The prevalence of these features, together with their standalone predictive power, strongly suggests that affect detection on smartphones can benefit from emerging sensor modalities to achieve more accurate performance. Based on the results of this study, we recommend that filtered pose and eye-tracking be deployed in future smartphone based affect detection systems in order to boost performance.

Aspects of our study design also likely contribute to the high recognition accuracy we recorded. Motion and eye-tracking sensor channels tended to yield highest accuracies using data from the Elicitation and Responding phases of our affect sessions and lower ones with data from the Assessment phase. This highlights the importance of collecting data during tasks that elicit affect, not in dedicated sessions immediately afterwards [28, 82]. The transient states we are concerned with in this paper are short lived and the results of this study indicate that the behavioral markers that can reveal them may be similarly fleeting. We recommend that future affect detection systems (or study designs) capture data during the critical moments which a user may be experiencing content that triggers an emotion. Deferring capture of sensor data to separate sessions that occur after viewing affect eliciting media (e.g., in dedicated post-exposure input tasks such as Fitts tapping [82]) will likely reduce detection accuracy.

The normalization procedures we applied also impacted recognition accuracy. In line with related prior work dealing with personality trait prediction from gaze [8], the use of a cool-down period to establish baseline data immediately prior to affect elicitation sessions was associated with more accurate recognition performance. While it is not clear how cool-down sessions can be integrated into regular application use on smartphones, the improved performance we observed (of up to 20.93%) indicates that it would be worthwhile to explore different approaches to achieve this. One possibility would be to use a rolling window for baseline data, simply normalizing each affect-session with data from the previous one. We also observed relatively minor variations in classification accuracy with changes in the media format. Specifically, affect elicitation via images led to generally increased accuracy (of up to 11.37%) over that achieved with videos. This may be due to differences in how effectively the images and video content elicited affect—images are reported to be more effective at evoking negative emotions than videos [8], an issue that may have been particularly prominent in the current study as many (35%, see Table 3) video based negative affect sessions did not successfully elicit their target affective response. In post-study comments, participants indicated many of the video clips intended to elicit negative affect were instead narratively engaging and resulted in mixed or positive emotions such as excitement or curiosity. In future studies in this area, it may be better to focus solely on image based media to more reliably elicit a full range of emotions.

It is also worth discussing the statistical differences between the distributions of feature data observed in the positive and negative affect sessions. In contrast to prior work suggesting that participants exposed to positive stimuli will exhibit steadier control of their device [82], we observed the opposite and saw more substantial and variable phone motion in response to eliciting positive affect. A possible explanation for this difference may be due to the less structured nature of our study tasks: periodic virtual or horizontal swipes, conducted freely with either the (opposite hand) index finger or (same hand) thumb versus structured repeated tasks such as Fitts' law tapping. Salient features in the artificial, performance focused input tasks used in prior work may differ from those in the more naturalistic tasks employed in the current study. Larger phone motions may also result from the significantly larger and heavier taps that participants produced in response to elicitation of positive affect. While these touch features are commonly associated with affect detection studies [54, 82, 83], no prior work has reported clear trends in this data between emotion classes. While our study can highlight a tentative link, we note this topic deserves research attention in the future, as identifying clear behavior patterns associated with particular affective states will do much to simplify, elucidate and expedite the development of this technology.

In summary, the high recognition performance that form the key results of this study are promising. Our study results suggest that affect detection in tasks reminiscent of social media use is achievable, at meaningful levels of performance, on commercial mobile devices. However, the ecological validity of these conclusions is limited by, firstly, the media we used and, secondly, by the tasks we deployed. Firstly, in terms of the media, while the study was designed to mimic the ways that image and video contents are displayed on social media platforms, the actual content differed substantially, particularly in terms of content that might yield negative affect. To put it another way: generic humorous or uplifting images, memes and videos proliferate on social media [84, 107], showing strong similarities to the content in the study. On the other hand, the material used to elicit negative affect arguably differed substantially from the types of personally meaningful content that might trigger such responses in real SNS use via processes such as envy [113] or social comparison [97]. This problem was particularly clear with the video content, in which a number of participants reported that clips intended to elicit negative affect were simply exciting or narratively engaging. Secondly, while the input tasks performed moved away from purely artificial procedures, such as Fitts tapping, they remained quite structured: each participant navigated through homogeneous content using the same set of basic tap and swipe operations. SNS use in real world scenarios will, inevitably, involve a much more diverse set of inputs, interactions and behavioral patterns. It is therefore unclear whether the results of this study will generalize to a more complex and naturalistic set of input tasks and actions. As the primary goal of this paper is to assess the viability of affect detection during SNS use, we designed a follow up study to tackle these issues of ecological validity.

4 APPLICATION STUDY

This study was designed to address limitations in the ecological validity of the feasibility study. This was achieved with a design that logged sensor data continuously but allowed participants to browse their own social media feeds and receive periodic prompts, which they could ignore or defer, to provide a self-assessment of their current emotional state. We argue that the use of a user's genuine social media contents to spontaneously elicit emotions and the lack of any scripted tasks to provide structured data to facilitate recognition represent key steps towards determining the viability of affect detection in real world SNS use scenarios. In this way this study moves beyond both the feasibility study presented in this paper and prior work in this area in general [28, 78, 82]. This study was approved by the host university's IRB.

4.1 Design and Materials

This study targeted a relatively naturalistic and realistic social media use scenario. Participants browsed their own social media feeds in an bespoke app on the same mobile device used in the feasibility study (an Apple iPhone X)

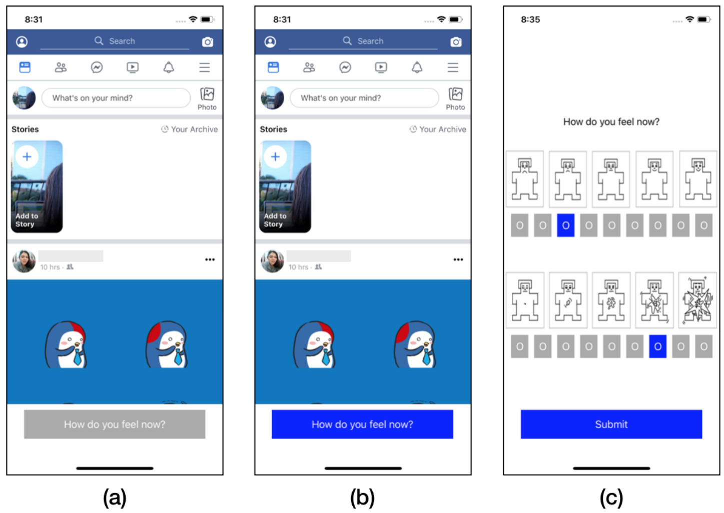

Fig. 3. User interface of the application: (a) Facebook's News Feed with a simple button for self-assessment questionnaire, (b) the button turned blue to signify the ability for submitting self-assessment questionnaire, and (c) self-assessment questionnaire [12]

and were periodically prompted to provide a self-assessment of their emotional state. The app, shown in Figure 3, achieved this by presenting two interface panels. The top panel, occupying the majority of the phone screen, showed Facebook News Feed's web view. This showed a participant's standard news feed including content such as videos, images, posts and advertisements and supported interactions such as entering reactions and comments in response to content or posting new status updates. Underneath this panel was a simple button inviting users to provide a self-assessment of their current emotional state. It was labelled "How do you feel now?" and was initially inactive (greyed out). Three minutes into the study it became active, signified by a change in color to blue, and participants were then able to select it at their convenience. After selecting it, they were presented with an emotion self-assessment questionnaire. Upon completion of the questionnaire, the button was again deactivated for a three minute period. The goal of this structure was to facilitate capture of several self-assessments from each participant, at times they self-identified as salient, but also to ensure that each self-assessment came after they had the opportunity to view new feed contents.

The emotional self-assessment instrument used in this study captured a more detailed picture of emotional state than the tool used in the feasibility study (20-item PANAS self-assessment). This change was motivated by the fact that a much broader range of emotional states might occur in response to participants browsing their personal social media feeds than in response to the carefully selected and vetted affective media used in the previous study. In addition, we sought to minimize the time and cognitive load required to complete the self-assessment in order to reduce seams in the social media browsing experience. To achieve these objectives, we followed closely related prior work [82] and deployed two dimensions from the Self-Assessment Manikin (SAM) [12]. SAM shows sets of pictorial manikins that do not depend on textual guidance and emotion labels. The manikins depict emotions via facial expressions and bodily reactions and participants select the manikins most representative of their emotional state. This presentation is reported to be less reliant on abstract thinking and is not confined to a specific language and thus, broadly accessible for a wide variety of participant populations including non-English speaking individuals [12, 51]. To align with the two dimensions in the circumplex model, and following much recent related work [52, 81], we deployed the manikins sets corresponding to valence and arousal (see Figure 3c) but omitted those for dominance. The valence dimension is represented as a frowning to a smiling face, while arousal features manikins in states between sleeping and highly agitated or excited [12]. As in prior work [77, 82], we transformed both valence and arousal to binary measures by dividing scores at scale mid-points.

Table 6. The features showing the greatest predictive power for valence and arousal detection in the application study. Features shown in alphabetical order. $_L$=left eye, $_R$=right eye

Feature	Sensor	Valence	Arousal
Acceleration, y-axis	Motion	maximum	-
Acceleration, z-axis	Motion	-	standard deviation
Blink	Eye	mean$_L$, minimum$_L$, variance$_L$	mean$_L$, minimum$_L$, variance$_L$
Distance	Touch	variance	-
Eye squint	Eye	maximum$_L$, mean$_L$	minimum$_L$, mean$_L$
Look down	Eye	minimum$_L$, mean$_L$	mean$_L$, minimum$_L$, variance$_L$
Look in	Eye	maximum$_L$, mean$_L$, minimum$_{L,R}$	minimum$_R$, variance$_L$
Look out	Eye	maximum$_R$, mean$_R$, minimum$_R$	maximum$_L$, mean$_{L,R}$, median$_L$,variance$_L$
Look up	Eye	-	variance$_L$
Pitch	Motion	minimum	-
Roll	Motion	-	maximum, mean
Rotation around x-axis	Motion	maximum, mean	minimum
Rotation around y-axis	Motion	maximum	-
Rotation around z-axis	Motion	-	mean
Speed	Touch	minimum	-
Touch area	Touch	minimum, mean	minimum, median
Touch Pressure	Touch	maximum	minimum
Yaw	Motion	variance	maximum

4.2 Procedure

The study took place in a quiet office environment. Participants were first shown a five minute long briefing video on a laptop that introduced the study, tasks, and provided a short explanation of how to answer the SAM questionnaire. They were informed they could terminate the study at any time and, if they continued, then signed consent forms and completed demographics and then logged into their Facebook accounts using the study app. They then browsed their Facebook feeds for at least 25 minutes, providing self-assessments of their emotional state at their discretion. After 25 minutes, an experimenter verbally prompted participants to end the study after completing a final self-assessment at their convenience.

4.3 Participants

Participants were recruited via an online forum and from a different academic institution to the first study. They were screened to have a Facebook account that they self-reported using at least once per day. They were informed the study would involve viewing, browsing, or posting content on their private Facebook account for at least 25 minutes, but that no data from their profile or about the content seen or posted during the study would be logged. Participants were requested not to view their Facebook accounts for two hours prior to their study session. This was to ensure some previously unseen content would be available on their feeds.

Twenty new participants completed this study (7 male and 13 female) with a mean age of 25.5 (SD=3.53). Fourteen were undergraduate students and six were graduate students. They reported being experienced computer

(4.2/5) and smartphone (4.7/5) users. In addition, they actively used SNSs daily (mean hours/day = 2.25, SD=0.97). To make this study more ecologically valid, we relaxed screening for use of corrective lenses, eye surgery and eye makeup, targeting a more representative sample of the general population. Three participants reported being short-sighted. During the experiment, eight participants wore glasses, and two wore contact lenses. Participants each received $10 for approximately an hour of experimental participation.

4.4 Data Preprocessing

We collected the same features as in the feasibility study and followed similar processes for data imputation and windowing. We then defined *affect-sessions* as a period involving the following two phases: an *elicitation* phase in which participants browsed and reacted to content on Facebook; and an *assessment* phase in which they completed questionnaires. For each affect-session, we explored use of two normalization procedures: the standard *within session* approach used in the feasibility study and a new approach that used data from the *prior affect-session* to normalize data in the current affect-session. We used these options due to the infeasiblity of integrating a controlled and explicit cool-down phase into this study's more naturalistic design. Due to the brevity of the assessment phase (just 3.6% of the time spent on the study), we did not examine study phase as a variable and instead used all data from each affect-session for analysis. To define target labels, we computed binary valence (positive or negative) and arousal (low or high) from the self-report questionnaires completed in each affect-session by splitting the data at the scale mid-points. Each window in an affect-session was then assigned the calculated valence and arousal labels.

Procedures closely followed those in the feasibility study. We employed leave-one-subject-out cross-validation and the same initial set of eight classifiers. Furthermore, at each run, we followed the same classification pipeline for feature selection and handling imbalances in the data set. Finally, in order to produce a representative set of features for valence and arousal detection, we again employed the feature selection process on the full set of within session normalized data. However, unlike in the feasibility study, the recursive feature elimination procedure was manually configured to select the 25 top features. This ensured the representative feature sets were the same size as in the initial study. The sets of 25 representative features for valence and arousal detection are shown in Table 6. These representative features overlap substantially with those in each leave-one-out model—a mean of 74.3% for valence and 72.6% for arousal appear in each leave-one-out model. Appendix B.1 shows these features ranked in terms of importance, calculated via feature ablation.

4.5 Results

Table 7. Descriptive statistics of Application Study displaying number of records of data for each sensor type, number of sessions and self-reported emotion scores

Number of records			**# of sessions**	**Self-reported emotion**			
				Valence		**Arousal**	
Motion	Touch	Eye-tracking		Positive	Negative	Low	High
1,901,197	24,292	647,061	2022	94	26	85	35

4.5.1 *Descriptive Statistics.*

Participants completed the study in 25.58 minutes to 30.17 minutes (M = 27.98, SD = 1.61). In addition, Table 7 describes the amount of recorded data and the distribution of self-reported affect levels. Once again, we logged a large set of motion and eye-tracking data—respectively 1.901 and 0.647 million samples—and 24293 touches, which we divided into 2022 non-overlapping 15 second windows. This corresponds to 8 hours, 25 minutes and 30 seconds of data. Participants self-selected the moments and frequency with which they completed SAM questionnaires, ultimately submitting between one and eight (M = 6, SD = 1.98, total = 120) of these assessments. We removed data from one participant who reported no variations in their affective state and completed SAM only a single time. We also note the distribution of self-reported affective states was uneven. Participants tended to report positive valence (78.33%) and low-arousal (70.83%), trends that were consistent throughout the study—we calculated mean valence and arousal scores by submitted assessment and fit linear models to reveal slopes of just 0.081 and -0.099 respectively. To address these imbalances in the data set, we again over-sampled the minority class, on training data only, for each emotional scale using SMOTE [23].

4.5.2 Overall Classifier Performance. We calculated the mean accuracy and class-wise F_1-score in predicting both binary valence and binary arousal for the same set of eight classifiers used in the feasibility study. The results can be seen in Table 8; most closely follow the feasibility study and the RBF-kernel SVM again outperformed other classifiers; accordingly, all subsequent reporting of the study results include data only from this recognizer. There was also markedly improved recognition performance (of 5.58% and 8.56%) with the dominant class (either positive valence or low arousal) over the minority class (negative valence or high arousal) in both affective dimensions. Regardless accuracy with the minority classes remained relatively high at 87.61% and 90.16%, figures which we believe are sufficient to support a range of meaningful applications. We also note that, unlike the feasibility study, these results were achieved without screening participants for use of corrective lenses or makeup, suggesting they are robust to diversity in the use of such products.

Table 8. Mean accuracy and class-wise F_1-scores for valence detection and arousal detection for all eight classifiers, data normalized based on the prior session and including features from all sensors.

Affective Dim.	Metric	ZeroR	AB	DT	kNN	LR	NB	RF	SVM
Valence	Accuracy (%)	75.77	88.87	87.04	91.69	82.00	64.39	92.43	**94.16**
	Positive (F_1)	86.19	91.79	91.16	93.77	86.28	69.54	94.86	**96.17**
	Negative (F_1)	00.00	76.20	73.84	82.23	66.53	51.28	83.42	**87.61**
Arousal	Accuracy (%)	70.53	83.43	81.36	88.48	80.31	61.08	91.74	**92.28**
	Low (F_1)	82.68	89.10	88.86	92.44	87.87	68.62	94.83	**95.74**
	High (F_1)	00.00	76.48	73.07	82.48	72.00	53.37	86.10	**90.16**

4.5.3 Performance of Classifier Variants. We then explored how aspects of the experimental design impacted the classification performance. Specifically, we looked at combinations of two variables: *the sensor(s)* used to contribute features to the recognizer and the *normalization* procedures applied in this study. In terms of sensors, we again examined leave-one-subject-out models constructed from *all sensors* and from the individual modalities of *motion*, *touch* and *eye-tracking*. For normalization procedures, we examined the differences between scaling data solely *within session* or based on data from the immediately *prior session*. The goal here was to determine how normalization procedures impacted the accuracy of affect detection. Table 9 displays the results of these analyses, in terms of classification accuracy achieved by the RBF-kernel SVM recognizer. Corresponding class-wise F_1-scores appear in Appendix B.2.

In terms of the sensor modalities, peak performance relied on all features and reached 94.16% accuracy for binary valence and 92.28% for binary arousal. There were substantial variations in how strongly features from each modality appeared in the representative feature sets shown in Table 6. For predicting binary valence, 6 features were from motion, 5 features were from touch, and 14 features were from eye-tracking. In terms of predicting binary arousal, 6 features were from motion, 3 features were from touch, and 16 features were from eye-tracking. With motion features alone, performance remained high with modest reductions in accuracy of between 3.11% (valence) and 5.29% (arousal). Performance with touch features was more substantially reduced, by 9.00% (valence) and 15.28% (arousal). Finally, the prevalent eye-tracking features performed particularly well, leading to reductions of just 1.48% (valence) and 2.96% (arousal). This result is somewhat in contrast to performance in the feasibility study, where the dominantly selected motion features led to the best performance. This suggests that participants' eye movements while browsing their personal social media feeds were a rich source of information about their affective states, more so than when they were exposed to standardized elicitation videos and images. In terms of the normalization procedures used, the baseline provided by the prior session yielded clear benefits. It led to modestly higher accuracies of between 3.61% and 5.64% for valence and between 3.36% and 7.72% for arousal over those achieved with more standard within-session normalization procedures. This confirms the result from the first study that suggesting that dynamic normalization procedures are necessary in order to achieve high affect detection performance.

Table 9. Accuracy (in %) for RBF-kernel SVM in predicting binary valence and binary arousal using feature data from different *sensors* and subject to different *normalization* procedures.

	Valence		**Arousal**	
Normalization	**Within Session**	**Prior Session**	**Within Session**	**Prior Session**
All features	89.51	94.16	88.92	92.28
Motion features	86.40	92.04	83.63	88.77
Touch features	81.55	85.16	73.64	81.36
Eye-tracking features	88.62	92.68	85.96	89.42

4.5.4 Feature Variability with Binary Valence and Binary Arousal. In this section, we examine the specific behaviors that were associated with variations in binary valence and arousal by reporting on the results of a series of t-tests on the means feature values in the representative feature sets for both valance and arousal (see Table 6). We applied an alpha threshold of 0.002 for the 25 independent tests for valence and 25 independent tests for arousal. This examination revealed eight significant results for valence and four significant results for arousal, covering features derived from motion and eye-tracking. In terms of motion sensors, significantly lower mean rotation around the device x-axis ($p < 0.001$) and lower maximum rotation around the device y-axis ($p < 0.001$) were associated to positive valence. Results also revealed that significantly lower standard deviation acceleration in the z-axis ($p < 0.001$) and lower maximum yaw data ($p < 0.001$) were linked to low-arousal. In contrast to the feasibility study, we observed that participants with positive valence and low-arousal stimuli demonstrate steadier control of their device, a result also reported in prior work [82]. One possible explanation for this difference may be due to more natural settings in this study; participants in this study performed less structured tasks compared to the systematic and periodically repeated tasks presented in the feasibility study. With regard to eye-tracking data, we group significant results into two feature types: eye-blinks and eye movements. In terms of eye-blinks, we report significantly lower mean left-eye blink rates with self-assessments of positive valence ($p < 0.001$) and low arousal ($p < 0.001$). These findings are in line with prior research indicating that, for example, words associated with negative affect [86] and images associated with high arousal [42] lead to increased eye blink rates. Significant results from eye movement were more challenging to interpret. We observed that positive valence was associated with lower mean and maximum values of left-eye squint (both $p < 0.001$) as well as the lower look-in values: mean ($p < 0.001$) and minimum ($p = 0.001$) of the left-eye, and minimum ($p < 0.001$) of the right eye. Finally, low-arousal was associated with lower mean look-out values of the left-eye ($p = 0.001$). These results suggest that positive valence and low-arousal stimuli led to eye movements that were reduced in scale. Finally, we note that these significant results are well-aligned to the feature rankings, again calculated via feature ablation, presented in Appendix B.1.

5 DISCUSSION

In this article, we investigated the possibility of detecting people's transitory emotional states during social media use with currently available smartphone sensors: device motion, touch and eye-tracking. To achieve this, we first conducted a feasibility study in a controlled setting to determine whether promising prior demonstrations of using such systems to detect binary affect (e.g., in Fitts' law targeting tasks [82]) transfer to tasks more akin to those that occur in unscripted social media use. The results indicate that data from smartphone sensors is sufficient to distinguish between positive and negative affect with an accuracy of up to 93.20% and highlights elements of our study design that contribute to this strong performance: the use of novel sensor channels such as eye-tracking; regular calibration sessions; and the capture of sensor data during (as opposed to after) affect elicitation sessions. We apply these lessons in a follow up study that tackles issues of ecological validity. It involves logging sensor data and ground truth valence and arousal ratings while participants freely browse and use their own social media accounts. The results of this study revealed that smartphone sensors can predict binary valence with an accuracy of 94.16% and binary arousal with 92.28%. These results suggest it is feasible to detect transitory emotions while consuming or responding to content on social media.

It is worth contrasting the results of our second application study with both our initial feasibility study and against related prior work. While affect detection accuracies in both our studies are high, one of the most notable differences is terms of the salient features. In the feasibility study, motion features are prevalent in the representative feature set and support the highest levels of accuracy in the leave-one-subject-out models. In the application study, eye-tracking features supplant them on both these measures. Furthermore, the accuracies we observe from eye-tracking data in the application study compare favourably to those in the literature: 87.59% to 92.36% in the current study versus 58.90% [125] to 86.74% [63] in prior work on eye-tracking based affect detection. One possible explanation for the high value and performance of eye-tracking features in our application study is its naturalistic design. Rather than rely on vetted affect elicitation materials [4, 125], we had participants freely browse their personal social media feeds. This may result in eye behaviors that are simply more salient than those that occur in response to standard elicitation materials. This observation is in line with prior work that has suggested that capturing user behaviors in more naturalistic settings may improve the accuracy of sensor based affect detection systems [66]. Based on these results, we identify eye-tracking as a key sensor modality for future work on smartphone based affect detection. However, this result is preliminary—we highlight a need for future studies that more formally compare the use of eye-tracking features for affect detection in a range of scenarios to develop a more complete understanding of the settings and situations in which it is most effective.

Other aspects of our study design also contributed to the high accuracies we observed. For example, in line with prior work, our feasibility study showed the benefits of an explicit cool-down phase [8] that can provide regularly updated baseline data in order to support high affect detection accuracy. However, it is challenging to integrate this concept into real-world scenarios such as social media use. Our application study took a simple approach to solving this problem and used data from preceding affect-sessions to normalize current ones, with resulting improvements in accuracy of up to 4.21% over standard within session normalization. Accordingly, we recommend that future real-world affect detection systems establish dynamic normalization procedures. Further work is required to identify the most effective techniques for this—our current approach leveraged aspects of our study design to identify critical moments (when self-reports of affect were submitted) at which to segment data into sessions. Such neat seams are unlikely to be observable outside of study settings. Possible alternative approaches for determining appropriate periods for baselines include the use of simple fixed temporal windows, or applying the ongoing outcomes of a live affect detection recognizer. The idea here would be to identify affectively salient moments and use these as breakpoints for "sessions" that can be used to normalize subsequent data.

It is also worth contextualizing our application study in the broader literature dealing with the emotions that occur during social media use. Our work is partly motivated by prior reports that social media use may, for example, lead to unfavorable social comparison [14], undermine affective well-being [113] or lead to feelings of envy [71]. These are, arguably, undesirable outcomes. However, the application study saw relatively scant examples of the types of affective state associated with these experiences. Specifically, 78.33% of self-reported emotions in our application study were categorized as positive valence and 70.83% were labelled as low-arousal. These results are more in line with prior work indicating that browsing social media makes people feel positive [62] and relaxed [90], arguably desirable outcomes. While the prevalence of these states does not diminish the importance of the potentially negative emotional impacts of social media, it does impact how we might consider deploying an affect detection system to mitigate them. The relative infrequency of negative affect and high valence emotions (and thus, the content that presumably triggers them) may mean that such systems might be best designed to support awareness of key affective states (by flagging content that negatively impacts users) or for specific groups of high risk users (such as those with a history of prior issues).

6 FUTURE WORK AND CONCLUSIONS

Several limitations impact our studies; we consider these as opportunities for future work. In technical terms, our system is inefficient. In particular, the package we used to retrieve eye-tracking data [33] substantially impacted CPU use and, therefore, phone battery life. An informal examination of this issue suggests CPU loads with our study applications approached 75% during eye-tracking and were approximately 30% without it. Until mobile device eye-tracking becomes less resource intensive, it may be impractical to use it in real world affect detection scenarios. In addition, the reliability of current mobile device eye-trackers remains questionable. The data described in this article, for example, features variations between the right and left eyes in terms of behaviors that are typically coupled, such as blinks. These variations likely reflect a range of sensor quality and environmental factors (e.g., lighting, handedness) that are beyond the scope of the current article to quantify. They deserve more detailed study in the future. We see clear benefits in research that seeks to improve the hardware and software systems used for smartphone eye-tracking and in high quality, independent verification of the objective performance of such systems in real world use scenarios. In terms of methods, while our application study sought to improve on the ecological validity of our feasibility study, issues remain. For example, we requested participants to use their Facebook accounts for at least 25 minutes. While this remains lower than the mean daily use time for our participants, a single prolonged session may be quite artificial [56] and may have resulted in atypical behaviors. Future work should examine affect detection in more natural (i.e., likely shorter and more sporadic) use sessions, for example, in field or diary style studies that take place outside of the lab [46, 122]. Furthermore, the duration of affect-sessions in our application study varies between 3 and 14 minutes. This means that a pair of binary valance and arousal labels is used to tag diverse periods of social media browsing. We suggest that further research is needed to define and/or verify the appropriate duration of affect-sessions in order to maintain validity and achieve high classification accuracy. In addition, future work should explore whether the findings we present generalize to other SNS services, such as Twitter, Instagram or TikTok, that exhibit substantially different presentation formats and usage patterns. Finally, while our treatment of affect, valence and arousal as binary states is common in the literature [26, 63, 82], it is also arguably somewhat coarse and abstract. Future work should examine detecting affective states with more granularity [28, 45, 46], focus on a greater variety of affective states (e.g., envy [71]) or tackle more concrete and applied topics. For example, detecting emotions in social media may be able to contribute to a wide range of critical issues, such as promoting mental health [109] or identifying depression [90]. By continuing research on these topics and issues, we believe we can transform the promising results reported in both our feasibility and application studies (respectively, peak binary affect, valence and arousal detection accuracies of 93.20%, 94.16% and 92.28%) into valuable and effective tools to support emotionally aware social media services and experiences on smartphones.

ACKNOWLEDGMENTS

We thank reviewers for their thoughtful feedback, which significantly improved this paper. We would also like to thank Taewan Kim for his support on experiments, members of Interactions lab and Hyunmi Oh for their constructive discussions and feedback, and all participants for their enthusiasm for this study. This work was generously supported by the Graduate School of Creative Design Engineering, UNIST and a grant of the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2020R1F1A1070699) and a grant of the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017M3C4A7083534).

REFERENCES

[1] Andy Adler. 2004. Reconstruction of source images from quantized biometric match score data. In *In Biometrics Conference, Washington, DC (September 2004)*, Vol. 18. Citeseer, 43–98.

[2] June Ahn and Jinyoung Kim. 2013. The Show Must Go on: The Presentation of Self During Interpersonal Conflict on Facebook. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries (ASIST '13)*. American Society for Information Science, Silver Springs, MD, USA, Article 94, 10 pages. http://dl.acm.org/citation.cfm?id=2655780.2655874

[3] Anthony H Ahrens and David AF Haaga. 1993. The specificity of attributional style and expectations to positive and negative affectivity, depression, and anxiety. *Cognitive therapy and research* 17, 1 (1993), 83–98.

[4] Sharifa Alghowinem, Majdah AlShehri, Roland Goecke, and Michael Wagner. 2014. *Exploring Eye Activity as an Indication of Emotional States Using an Eye-Tracking Sensor*. Springer International Publishing, Cham, 261–276. https://doi.org/10.1007/978-3-319-04702-7_15

[5] Eduardo B Andrade and Dan Ariely. 2009. The enduring impact of transient emotions on decision making. *Organizational Behavior and Human Decision Processes* 109, 1 (2009), 1–8.

[6] Natalya N. Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. 2015. Social Sharing of Emotions on Facebook: Channel Differences, Satisfaction, and Replies. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work (CSCW '15)*. ACM, New York, NY, USA, 154–164. https://doi.org/10.1145/2675133.2675297

[7] Natalya N. Bazarova, Jessie G. Taft, Yoon Hyung Choi, and Dan Cosley. 2013. Managing Impressions and Relationships on Facebook: Self-Presentational and Relational Concerns Revealed Through the Analysis of Language Style. *Journal of Language and Social Psychology* 32, 2 (2013), 121–141. https://doi.org/10.1177/0261927X12456384 arXiv:https://doi.org/10.1177/0261927X12456384

[8] Shlomo Berkovsky, Ronnie Taib, Irena Koprinska, Eileen Wang, Yucheng Zeng, Jingjie Li, and Sabina Kleitman. 2019. Detecting Personality Traits Using Eye-Tracking Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 221, 12 pages. https://doi.org/10.1145/3290605.3300451

[9] A. Bhardwaj, A. Gupta, P. Jain, A. Rani, and J. Yadav. 2015. Classification of human emotions from EEG signals using SVM and LDA Classifiers. In *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*. 180–185. https://doi.org/10.1109/SPIN.2015.7095376

[10] Bloglingo. 2019. 12 Simple Facebook Tactics That Will Seduce People And Win Clients Bloglingo. https://bloglingo.com/facebook-traffic/

[11] Mehdi Boukhechba, Alexander R. Daros, Karl Fua, Philip I. Chow, Bethany A. Teachman, and Laura E. Barnes. 2018. DemonicSalmon: Monitoring mental health and social interactions of college students using smartphones. *Smart Health* 9-10 (2018), 192 – 203. https://doi.org/10.1016/j.smhl.2018.07.005 CHASE 2018 Special Issue.

[12] 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. https://doi.org/10.1016/0005-7916(94)90063-9

[13] Jed Brubaker, Funda Kivran-Swaine, Lee Taber, and Gillian Hayes. 2012. Grief-Stricken in a Crowd: The Language of Bereavement and Distress in Social Media. (01 2012).

[14] Moira Burke. 2018. Reading, Writing, Relationships: The Impact of Social Network Sites on Relationships and Well-Being. https://doi.org/10.1184/R1/6721256.v1

[15] Moira Burke and Robert Kraut. 2013. Using Facebook after Losing a Job: Differential Benefits of Strong and Weak Ties. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 1419–1430. https://doi.org/10.1145/2441776.2441936

[16] Moira Burke, Robert Kraut, and Cameron Marlow. 2011. Social capital on Facebook: Differentiating uses and users. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 571–580.

[17] Moira Burke and Robert E. Kraut. 2014. Growing Closer on Facebook: Changes in Tie Strength through Social Network Site Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 4187–4196. https://doi.org/10.1145/2556288.2557094

[18] Paul Cairns, Pratyush Pandab, and Christopher Power. 2014. The Influence of Emotion on Number Entry Errors. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2293–2296. https://doi.org/10.1145/2556288.2557065

[19] Bokai Cao, Lei Zheng, Chenwei Zhang, Philip S. Yu, Andrea Piscitello, John Zulueta, Olu Ajilore, Kelly Ryan, and Alex D. Leow. 2017. DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 747–755. https://doi.org/10.1145/3097983.3098086

[20] R. C.Balabantaray, Mudasir Mohammad, and Nibha Sharma. 2012. Multi-Class Twitter Emotion Classification: A New Approach.

[21] Pew Research Center. 2019. Demographics of Social Media Users and Adoption in the United States. https://www.pewresearch.org/internet/fact-sheet/social-media/

[22] Hao Chao, Liang Dong, Yongli Liu, and Baoyun Lu. 2019. Emotion Recognition from Multiband EEG Signals Using CapsNet. *Sensors* 19, 9 (May 2019), 2212. https://doi.org/10.3390/s19092212

[23] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[24] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 71–80.

[25] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1217–1230. https://doi.org/10.1145/2998181.2998213

[26] Pietro Cipresso, Silvia Serino, Daniela Villani, Claudia Repetto, Luigi Sellitti, Giovanni Albani, Alessandro Mauro, Andrea Gaggioli, and Giuseppe Riva. 2012. Is your phone so smart to affect your state? An exploratory study based on psychophysiological measures. *Neurocomputing* 84 (2012), 23 – 30. https://doi.org/10.1016/j.neucom.2011.12.027 From Neuron to Behaviour: Evidence from Behavioral Measurements.

[27] Jeffrey F Cohn and Gary S Katz. 1998. Bimodal expression of emotion by face and voice. In *International Multimedia Conference: Proceedings of the sixth ACM international conference on Multimedia: Face/gesture recognition and their applications*, Vol. 1998. 41–44.

[28] D. Dai, Q. Liu, and H. 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

[29] Munmun De Choudhury, Scott Counts, and Michael Gamon. 2012. Not all moods are created equal! exploring human emotional states in social media. In *Sixth international AAAI conference on weblogs and social media*.

[30] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting Postpartum Changes in Emotion and Behavior via Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 3267–3276. https://doi.org/10.1145/2470654.2466447

[31] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

[32] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 13 (2013), 1–10.

[33] Apple Developer Documentation. 2020. ARKit | Apple Developer Documentation. https://developer.apple.com/documentation/arkit

[34] Apple Developer Documentation. 2020. Core Motion | Apple Developer Documentation. https://developer.apple.com/documentation/coremotion

[35] Apple Developer Documentation. 2020. UITouch - UIKit | Apple Developer Documentation. https://developer.apple.com/documentation/uikit/uitouch

[36] Jennifer R Dunn and Maurice E Schweitzer. 2005. Feeling and believing: the influence of emotion on trust. *Journal of personality and social psychology* 88, 5 (2005), 736.

[37] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. 2007. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication* 12, 4 (07 2007), 1143–1168. https://doi.org/10.1111/j.1083-6101.2007.00367.x arXiv:https://academic.oup.com/jcmc/article-pdf/12/4/1143/22316419/jjcmcom1143.pdf

[38] Hamdi Eryilmaz, Dimitri Van De Ville, Sophie Schwartz, and Patrik Vuilleumier. 2011. Impact of transient emotions on functional connectivity during subsequent resting state: a wavelet correlation approach. *Neuroimage* 54, 3 (2011), 2481–2491.

[39] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Predicting personality traits with instagram pictures. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015*. ACM, 7–10.

[40] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Predicting Personality Traits with Instagram Pictures. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015 (EMPIRE '15)*. ACM, New York, NY, USA, 7–10. https://doi.org/10.1145/2809643.2809644

[41] C. A. Frantzidis, C. Bratsas, C. L. Papadelis, E. Konstantinidis, C. Pappas, and P. D. Bamidis. 2010. Toward Emotion Aware Computing: An Integrated Approach Using Multichannel Neurophysiological Recordings and Affective Visual Stimuli. *IEEE Transactions on Information Technology in Biomedicine* 14, 3 (May 2010), 589–597. https://doi.org/10.1109/TITB.2010.2041553

[42] Philip A Gable and Eddie Harmon-Jones. 2009. Postauricular reflex responses to pictures varying in valence and arousal. *Psychophysiology* 46, 3 (2009), 487–490.

[43] Bo Gao, Bettina Berendt, and Joaquin Vanschoren. 2015. Who is More Positive in Private? Analyzing Sentiment Differences across Privacy Levels and Demographic Factors in Facebook Chats and Posts. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15)*. Association for Computing Machinery, New York, NY, USA, 605–610. https://doi.org/10.1145/2808797.2809421

[44] Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. 2012. What Does Touch Tell Us About Emotions in Touchscreen-Based Gameplay? *ACM Trans. Comput.-Hum. Interact.* 19, 4, Article 31 (Dec. 2012), 30 pages. https://doi.org/10.1145/2395131.2395138

[45] 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 (MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article Article 2, 12 pages. https://doi.org/10.1145/3098279.3098564

[46] S. Ghosh, N. Ganguly, B. Mitra, and P. De. 2019. Designing An Experience Sampling Method for Smartphone based Emotion Detection. *IEEE Transactions on Affective Computing* (2019), 1–1. https://doi.org/10.1109/TAFFC.2019.2905561

[47] Surjya Ghosh, Kaustubh Hiware, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Emotion detection from touch interactions during text entry on smartphones. *Int. J. Hum.-Comput. Stud.* 130 (2019), 47–57.

[48] Donald Glowinski, Nele Dael, Antonio Camurri, Gualtiero Volpe, Marcello Mortillaro, and Klaus Scherer. 2011. Toward a minimal representation of affective gestures. *IEEE Transactions on Affective Computing* 2, 2 (2011), 106–118.

[49] Jorge Goncalves, Pratyush Pandab, Denzil Ferreira, Mohammad Ghahramani, Guoying Zhao, and Vassilis Kostakos. 2014. Projective Testing of Diurnal Collective Emotion. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 487–497. https://doi.org/10.1145/2632048.2636067

[50] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. 2014. Emotex: Detecting emotions in twitter messages. (2014).

[51] 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 (IHC '16)*. Association for Computing Machinery, New York, NY, USA, Article 27, 10 pages. https://doi.org/10.1145/3033701.3033728

[52] Martin Hibbeln, Jeffrey L. Jenkins, Christoph Schneider, Joseph S. Valacich, and Markus Weinmann. 2017. How is Your User Feeling? Inferring Emotion through Human-Computer Interaction Devices. *MIS Q.* 41, 1 (March 2017), 1–21. https://doi.org/10.25300/MISQ/2017/41.1.01

[53] P. Hu, H. Ning, T. Qiu, H. Song, Y. Wang, and X. Yao. 2017. Security and Privacy Preservation Scheme of Face Identification and Resolution Framework Using Fog Computing in Internet of Things. *IEEE Internet of Things Journal* 4, 5 (Oct 2017), 1143–1155. https://doi.org/10.1109/JIOT.2017.2659783

[54] Sinh Huynh, Seungmin Kim, JeongGil Ko, Rajesh Krishna Balan, and Youngki Lee. 2018. EngageMon: Multi-Modal Engagement Sensing for Mobile Games. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 13 (March 2018), 27 pages. https://doi.org/10.1145/3191745

[55] Gonzalo Iturregui-Gallardo and Jorge Luis Méndez-Ulrich. 2019. Towards the Creation of a Tactile Version of the Self-Assessment Manikin (T-SAM) for the Emotional Assessment of Visually Impaired People. *International Journal of Disability, Development and Education* 0, 0 (2019), 1–18. https://doi.org/10.1080/1034912X.2019.1626007 arXiv:https://doi.org/10.1080/1034912X.2019.1626007

[56] Branka Janković, Milan Nikolić, Jelena Vukonjanski, and Edit Terek. 2016. The impact of Facebook and smart phone usage on the leisure activities and college adjustment of students in Serbia. *Computers in Human Behavior* 55 (2016), 354 – 363. https://doi.org/10.1016/j.chb.2015.09.022

[57] Mohsen Jozani, Emmanuel Ayaburi, Myung Ko, and Kim-Kwang Raymond Choo. 2020. Privacy concerns and benefits of engagement with social media-enabled apps: A privacy calculus perspective. *Computers in Human Behavior* 107 (2020), 106260. https://doi.org/10.1016/j.chb.2020.106260

[58] Mohamed Khamis, Anita Baier, Niels Henze, Florian Alt, and Andreas Bulling. 2018. Understanding Face and Eye Visibility in Front-Facing Cameras of Smartphones Used in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article Paper 280, 12 pages. https://doi.org/10.1145/3173574.3173854

[59] Funda Kivran-Swaine, Sam Brody, Nicholas Diakopoulos, and Mor Naaman. 2012. Of Joy and Gender: Emotional Expression in Online Social Networks. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion (CSCW '12)*. ACM, New York, NY, USA, 139–142. https://doi.org/10.1145/2141512.2141562

[60] Atesh Koul, Cristina Becchio, and Andrea Cavallo. 2018. Cross-Validation Approaches for Replicability in Psychology. *Frontiers in Psychology* 9 (2018), 1117. https://doi.org/10.3389/fpsyg.2018.01117

[61] Selda Koydemir, Ömer Faruk Şimşek, Astrid Schütz, and Arun Tipandjan. 2013. Differences in how trait emotional intelligence predicts life satisfaction: The role of affect balance versus social support in India and Germany. *Journal of Happiness Studies* 14, 1 (2013), 51–66.

[62] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

[63] Antonio Lanata, Gaetano Valenza, and Enzo Pasquale Scilingo. 2013. Eye gaze patterns in emotional pictures. *Journal of Ambient Intelligence and Humanized Computing* 4, 6 (2013), 705–715.

[64] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention* 1 (1997), 39–58.

[65] Scikit learn developers Documentation. 2020. sklearn.dummy.DummyClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html

[66] H. 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

[67] X. Li, Q. Peng, Z. Sun, L. Chai, and Y. Wang. 2019. Predicting Social Emotions from Readers' Perspective. *IEEE Transactions on Affective Computing* 10, 2 (April 2019), 255–264. https://doi.org/10.1109/TAFFC.2017.2695607

[68] 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 (MobiSys '13)*. Association for Computing Machinery, New York, NY, USA, 389–402. https://doi.org/10.1145/2462456.2464449

[69] Young-Jin Lim, Bum-Hee Yu, Doh-Kwan Kim, and Ji-Hae Kim. 2010. The positive and negative affect schedule: Psychometric properties of the Korean version. *Psychiatry investigation* 7, 3 (2010), 163–169.

[70] H. Lin, J. Jia, J. Qiu, Y. Zhang, G. Shen, L. Xie, J. Tang, L. Feng, and T. Chua. 2017. Detecting Stress Based on Social Interactions in Social Networks. *IEEE Transactions on Knowledge and Data Engineering* 29, 9 (Sep. 2017), 1820–1833. https://doi.org/10.1109/TKDE.2017.2686382

[71] Ruoyun Lin and Sonja Utz. 2015. The emotional responses of browsing Facebook: Happiness, envy, and the role of tie strength. *Computers in Human Behavior* 52 (2015), 29 – 38. https://doi.org/10.1016/j.chb.2015.04.064

[72] Zhe Liu, Anbang Xu, Yufan Guo, Jalal U. Mahmud, Haibin Liu, and Rama Akkiraju. 2018. Seemo: A Computational Approach to See Emotions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article Paper 364, 12 pages. https://doi.org/10.1145/3173574.3173938

[73] Danielle Lottridge and Frank R. Bentley. 2018. Let's Hate Together: How People Share News in Messaging, Social, and Public Networks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 60, 13 pages. https://doi.org/10.1145/3173574.3173634

[74] Kim Luyckx, Frederik Vaassen, Claudia Peersman, and Walter Daelemans. 2012. Fine-grained emotion detection in suicide notes: a thresholding approach to multi-label classification. *Biomedical informatics insights* 5 (2012), BII–S8966.

[75] Y. Ma, B. Xu, Y. Bai, G. Sun, and R. Zhu. 2012. Daily Mood Assessment Based on Mobile Phone Sensing. In *2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks*. 142–147. https://doi.org/10.1109/BSN.2012.3

[76] I. Scott MacKenzie and R. William Soukoreff. 2003. Phrase Sets for Evaluating Text Entry Techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03)*. ACM, New York, NY, USA, 754–755. https://doi.org/10.1145/765891.765971

[77] Laura Alejandra Martínez-Tejada, Yasuhisa Maruyama, Natsue Yoshimura, and Yasuharu Koike. 2020. Analysis of Personality and EEG Features in Emotion Recognition Using Machine Learning Techniques to Classify Arousal and Valance Labels. *Machine Learning and Knowledge Extraction* 2, 2 (2020), 99–124.

[78] Abhinav Mehrotra, Fani Tsapeli, Robert Hendley, and Mirco Musolesi. 2017. MyTraces: Investigating Correlation and Causation between Users' Emotional States and Mobile Phone Interaction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article Article 83 (Sept. 2017), 21 pages. https://doi.org/10.1145/3130948

[79] Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*. Elsevier, 201–237.

[80] Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management* 51, 4 (2015), 480–499.

[81] Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the affective norms for English words (ANEW) for Italian. *Behavior research methods* 46, 3 (2014), 887–903.

[82] 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 (UbiComp '16)*. ACM, New York, NY, USA, 781–792. https://doi.org/10.1145/2971648.2971654

[83] Aske Mottelson, Jarrod Knibbe, and Kasper Hornbæk. 2018. Veritaps: Truth Estimation from Mobile Interaction. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 561, 12 pages.

https://doi.org/10.1145/3173574.3174135
[84] Jessica Gall Myrick. 2015. Emotion regulation, procrastination, and watching cat videos online: Who watches Internet cats, why, and to what effect? *Computers in Human Behavior* 52 (2015), 168 – 176. https://doi.org/10.1016/j.chb.2015.06.001
[85] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is It Really about Me? Message Content in Social Awareness Streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 189–192. https://doi.org/10.1145/1718918.1718953
[86] Hideki Ohira, Ward M Winton, and Makiko Oyama. 1998. Effects of stimulus valence on recognition memory and endogenous eyeblinks: Further evidence for positive-negative asymmetry. *Personality and Social Psychology Bulletin* 24, 9 (1998), 986–993.
[87] Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, et al. 2011. The impact of social media on children, adolescents, and families. *Pediatrics* 127, 4 (2011), 800–804.
[88] A. F. Olsen and J. Torresen. 2016. Smartphone accelerometer data used for detecting human emotions. In *2016 3rd International Conference on Systems and Informatics (ICSAI)*. 410–415. https://doi.org/10.1109/ICSAI.2016.7810990
[89] Martín G Salido Ortega, Luis-Felipe Rodríguez, and J Octavio Gutierrez-Garcia. 2019. Towards emotion recognition from contextual information using machine learning. *Journal of Ambient Intelligence and Humanized Computing* (2019), 1–21.
[90] Galen Panger. 2018. People Tend to Wind Down, Not Up, When They Browse Social Media. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article Article 133 (Nov. 2018), 29 pages. https://doi.org/10.1145/3274402
[91] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology* 108, 6 (2015), 934.
[92] Sungkyu Park, Inyeop Kim, Sang Won Lee, Jaehyun Yoo, Bumseok Jeong, and Meeyoung Cha. 2015. Manifestation of Depression and Loneliness on Social Networks: A Case Study of Young Adults on Facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 557–570. https://doi.org/10.1145/2675133.2675139
[93] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html
[94] 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 (UbiComp '15)*. ACM, New York, NY, USA, 825–836. https://doi.org/10.1145/2750858.2804252
[95] Katharina Rein and Tommaso Venturini. 2018. Ploughing digital landscapes: How Facebook influences the evolution of live video streaming. *New Media & Society* 20, 9 (2018), 3359–3380. https://doi.org/10.1177/1461444817748954 arXiv:https://doi.org/10.1177/1461444817748954
[96] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.
[97] Zhanna Sarsenbayeva, Gabriele Marini, Niels van Berkel, Chu Luo, Weiwei Jiang, Kangning Yang, Greg Wadley, Tilman Dingler, Vassilis Kostakos, and Jorge Goncalves. 2020. Does Smartphone Use Drive Our Emotions or Vice Versa? A Causal Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376163
[98] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion* 24, 7 (2010), 1153–1172. https://doi.org/10.1080/02699930903274322 arXiv:https://doi.org/10.1080/02699930903274322
[99] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion* 24, 7 (2010), 1153–1172.
[100] Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social science information* 44, 4 (2005), 695–729.
[101] Vivek K. Singh, Souvick Ghosh, and Christin Jose. 2017. Toward Multimodal Cyberbullying Detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 2090–2099. https://doi.org/10.1145/3027063.3053169
[102] Manya Sleeper, Rebecca Balebako, Sauvik Das, Amber Lynn McConahy, Jason Wiese, and Lorrie Faith Cranor. 2013. The Post That Wasn't: Exploring Self-Censorship on Facebook. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 793–802. https://doi.org/10.1145/2441776.2441865
[103] I. Song, H. Kim, and P. B. Jeon. 2014. Deep learning for real-time robust facial expression recognition on a smartphone. In *2014 IEEE International Conference on Consumer Electronics (ICCE)*. 564–567. https://doi.org/10.1109/ICCE.2014.6776135
[104] Fred Stutzman, Robert Capra, and Jamila Thompson. 2011. Factors mediating disclosure in social network sites. *Computers in Human Behavior* 27, 1 (2011), 590–598.

[105] M. Suk and B. Prabhakaran. 2014. Real-Time Mobile Facial Expression Recognition System – A Case Study. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 132–137. https://doi.org/10.1109/CVPRW.2014.25

[106] David Sun, Pablo Paredes, and John Canny. 2014. MouStress: Detecting Stress from Mouse Motion. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 61–70. https://doi.org/10.1145/2556288.2557243

[107] Viriya Taecharungroj and Pitchanut Nueangjamnong. 2015. Humour 2.0: styles and types of humour and virality of memes on Facebook. *Journal of Creative Communications* 10, 3 (2015), 288–302.

[108] Kazuhiko Takahashi et al. 2004. Remarks on emotion recognition from bio-potential signals. In *2nd International conference on Autonomous Robots and Agents*, Vol. 3. 1148–1153.

[109] Edson C Tandoc Jr, Patrick Ferrucci, and Margaret Duffy. 2015. Facebook use, envy, and depression among college students: Is facebooking depressing? *Computers in Human Behavior* 43 (2015), 139–146.

[110] Dianne M Tice, Ellen Bratslavsky, and Roy F Baumeister. 2001. Emotional distress regulation takes precedence over impulse control: If you feel bad, do it! *Journal of personality and social psychology* 80, 1 (2001), 53.

[111] Meike K. Uhrig, Nadine Trautmann, Ulf Baumgärtner, Rolf-Detlef Treede, Florian Henrich, Wolfgang Hiller, and Susanne Marschall. 2016. Emotion Elicitation: A Comparison of Pictures and Films. *Frontiers in Psychology* 7 (2016), 180. https://doi.org/10.3389/fpsyg.2016.00180

[112] Gert Van Dijck and Marc M. Van Hulle. 2006. Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis. In *Artificial Neural Networks – ICANN 2006*, Stefanos D. Kollias, Andreas Stafylopatis, Włodzisław Duch, and Erkki Oja (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 31–40.

[113] Philippe Verduyn, David Seungjae Lee, Jiyoung Park, Holly Shablack, Ariana Orvell, Joseph Bayer, Oscar Ybarra, John Jonides, and Ethan Kross. 2015. Passive Facebook usage undermines affective well-being: Experimental and longitudinal evidence. *Journal of Experimental Psychology: General* 144, 2 (2015), 480.

[114] 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 (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376504

[115] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[116] David Watson and Auke Tellegen. 1985. Toward a consensual structure of mood. *Psychological bulletin* 98, 2 (1985), 219.

[117] Stephan Winter, Caroline Brückner, and Nicole C Krämer. 2015. They came, they liked, they commented: Social influence on Facebook news channels. *Cyberpsychology, Behavior, and Social Networking* 18, 8 (2015), 431–436.

[118] Kevin Wise, Saleem Alhabash, and Hyojung Park. 2010. Emotional responses during social information seeking on Facebook. *Cyberpsychology, Behavior, and Social Networking* 13, 5 (2010), 555–562.

[119] Wilhelm Max Wundt. 1907. *Outlines of psychology*. W. Engelmann.

[120] Yang Yang, Jia Jia, Shumei Zhang, Boya Wu, Qicong Chen, Juanzi Li, Chunxiao Xing, and Jie Tang. 2014. How do your friends on social media disclose your emotions?. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

[121] Heng Zhang, Shkurta Gashi, Hanke Kimm, Elcin Hanci, and Oludamilare Matthews. 2018. Moodbook: An Application for Continuous Monitoring of Social Media Usage and Mood. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 1150–1155. https://doi.org/10.1145/3267305.3274760

[122] Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. 2018. MoodExplorer: Towards Compound Emotion Detection via Smartphone Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 176 (Jan. 2018), 30 pages. https://doi.org/10.1145/3161414

[123] Xiao Zhang, Fuzhen Zhuang, Wenzhong Li, Haochao Ying, Hui Xiong, and Sanglu Lu. 2019. Inferring Mood Instability via Smartphone Sensing: A Multi-View Learning Approach. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery, New York, NY, USA, 1401–1409. https://doi.org/10.1145/3343031.3350957

[124] S. Zhao, H. Yao, Y. Gao, G. Ding, and T. Chua. 2018. Predicting Personalized Image Emotion Perceptions in Social Networks. *IEEE Transactions on Affective Computing* 9, 4 (Oct 2018), 526–540. https://doi.org/10.1109/TAFFC.2016.2628787

[125] W. Zheng, B. Dong, and B. Lu. 2014. Multimodal emotion recognition using EEG and eye tracking data. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 5040–5043. https://doi.org/10.1109/EMBC.2014.6944757

[126] Hussain D. Zuorba, Celine Louise O. Olan, and Anthonette D. Cantara. 2017. A Framework for Identifying Excessive Sadness in Students Through Twitter and Facebook in the Philippines. In *Proceedings of the International Conference on Bioinformatics Research and Applications 2017 (ICBRA 2017)*. ACM, New York, NY, USA, 52–56. https://doi.org/10.1145/3175587.3175600

A FEASIBILITY STUDY

A.1 Videos

We list shortened clips from FilmStim library [98] used in this study. The values in a bracket indicate start time and end time, respectively: Benny Joone [00:00,01:08], There's Something about Mary (1) [01:13,02:10], Seven (2) [00:00,01:48], and American History X [00:00,01:20].

A.2 Images

We list IAPS image set [64] for eliciting positive affect: 2035, 2395, 4623, 5200, 5760, 8080, 8190, 8370, 8470, 8499 and another set for eliciting negative affect: 2301, 2455, 2490, 2900, 6520, 6563, 9413, 9902, 9921, 9940.

A.3 Feature Ranking

Table 10. Feature ranking score calculated using feature ablation technique. The *p-value* is listed for those features that show significant differences (at an alpha threshold of 0.002) between classes. $_L$=left eye, $_R$=right eye.

Ranking	Feature	p-value
1	Acceleration, x-axis (mean)	= 0.0002
2	Yaw (max)	= 0.0002
3	Acceleration, x-axis (max)	= 0.0002
4	Acceleration, y-axis (max)	= 0.0001
5	Look in$_L$ (min)	< 0.0000
6	Rotation around x-axis (min)	= 0.0018
7	Rotation around y-axis (mean)	-
8	Touch area (max)	< 0.0000
9	Touch pressure (min)	< 0.0000
10	Rotation around z-axis (mean)	-
11	Eye squint$_L$ (min)	-
12	Roll (mean)	-
13	Eye squint$_L$ (var)	-
14	Rotation around x-axis (mean)	
15	Acceleration, y-axis (std)	
16	Look in$_L$ (mean)	
17	Acceleration, x-axis (min)	
18	Yaw (min)	
19	Acceleration, z-axis (mean)	
20	Acceleration, y-axis (mean)	
21	Acceleration, y-axis (min)	
22	Look in$_L$ (max)	
23	Touch area (mean)	
24	Look in$_R$ (max)	
25	Rotation around z-axis (min)	

A.4 Class-wise Classification

Table 11. Class-wise classification F_1-score using RBF-kernel SVM models for recognizing positive and negative affect for different feature sets (by sensing channel), study phases, normalization procedures and media types. *Elic. Res.*=Elicitation & Responding, *Pos.*=Positive Affect, *Neg.*=Negative Affect

Feature	*Phase*	**Cool-down**						**Within session**					
		Both		**Video**		**Image**		**Both**		**Video**		**Image**	
		Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
All	All	94.83	89.93	93.91	88.34	95.25	89.78	87.75	76.02	86.96	70.37	92.92	85.99
	Elic. & Res.	93.50	86.16	92.03	83.50	94.85	86.70	83.84	66.94	86.65	68.30	84.91	70.62
	Asses.	88.29	74.30	88.30	73.63	90.43	79.19	83.83	64.13	88.30	73.63	90.43	79.19
Motion	All	92.67	85.69	87.50	79.30	92.46	83.56	87.17	74.22	88.50	77.46	79.05	59.45
	Elic. & Res.	92.44	83.87	87.11	76.58	91.13	79.31	87.19	74.72	68.48	82.91	88.51	71.58
	Asses.	87.16	72.43	86.67	74.72	86.95	75.72	78.25	55.01	75.90	51.29	76.62	56.58
Touch	All	82.65	66.75	83.89	65.97	86.46	74.46	75.01	51.99	74.76	55.57	82.91	63.16
	Elic. & Res.	82.66	64.20	84.60	62.12	85.12	70.99	68.11	45.16	64.80	45.87	77.27	49.80
	Asses.	83.35	64.56	85.39	70.57	85.12	65.91	75.60	51.29	76.38	43.44	79.44	53.43
Eye-tracking	All	87.13	76.36	89.36	78.44	90.97	83.44	71.78	60.61	72.08	60.41	73.48	63.83
	Elic. & Res.	87.63	75.50	88.86	72.57	88.65	78.83	70.98	57.69	71.47	57.91	72.00	59.51
	Asses.	64.54	51.49	73.25	63.11	69.33	56.12	59.17	48.05	56.94	46.53	57.78	48.24

B APPLICATION STUDY

B.1 Feature Ranking for Valence and Arousal Detection

Table 12. Feature ranking score for valence detection calculated using feature ablation technique. The *p-value* is listed for those features that show significant differences (at an alpha threshold of 0.002) between classes. $_L$=left eye, $_R$=right eye.

Ranking	**Feature**	**p-value**
1	Blink_L (mean)	< 0.0000
2	Rotation around x-axis (mean)	= 0.0004
3	Eye squint_L (mean)	< 0.0000
4	Look in_L (mean)	= 0.0001
5	Rotation around y-axis (max)	= 0.0002
6	Look in_L (min)	= 0.0012
7	Look in_R (min)	< 0.0000
8	Eye squint_L (max)	= 0.0005
9	Touch pressure (max)	-
10	Look in_L (max)	-
11	Distance (var)	-
12	Look down_L (mean)	-
13	Look down_L (min)	-

Ranking	**Feature**
14	Rotation around x-axis (max)
15	Touch area (mean)
16	Touch area (min)
17	Look out_R (min)
18	Pitch (min)
19	Blink_L (min)
20	Speed (min)
21	Blink_L (var)
22	Yaw (var)
23	Look out_R (max)
24	Look out_R (mean)
25	Acceleration, y-axis (max)

Table 13. Feature ranking score for arousal detection calculated using feature ablation technique. The *p-value* is listed for those features that show significant differences (at an alpha threshold of 0.002) between classes. $_L$=left eye, $_R$=right eye.

Ranking	Feature	p-value	Ranking	Feature
1	Blink_L (mean)	< 0.0000	14	Look out_L (median)
2	Yaw (max)	= 0.0005	15	Look down_L (var)
3	Look out_L (mean)	= 0.0011	16	Look up_L (var)
4	Acceleration, z-axis (std)	= 0.0002	17	Blink_L (min)
5	Rotation around z-axis (mean)	-	18	Touch area (min)
6	Look in_R (min)	-	19	Look in_L (var)
7	Look out_R (mean)	-	20	Look out_L (max)
8	Look down_L (min)	-	21	Roll (mean)
9	Roll (max)	-	22	Rotation around x-axis (min)
10	Look out_L (var)	-	23	Blink_L (var)
11	Eye squint_L (min)	-	24	Touch pressure (min)
12	Eye squint_L (mean)	-	25	Look down_L (mean)
13	Touch area (median)	-		

B.2 Class-wise Classification

Table 14. Class-wise classification F_1-score using RBF-kernel SVM models for valence detection and arousal detection for different feature sets (by sensing channel) and normalization procedures.

	Valence				Arousal			
Normalization	Within Session		Prior Session		Within Session		Prior Session	
Features	Positive	Negative	Positive	Negative	Low	High	Low	High
All	93.11	77.63	96.17	87.61	92.27	80.26	95.74	90.16
Motion	91.01	71.99	94.72	83.81	88.30	72.67	91.86	81.57
Touch	87.78	62.16	90.23	69.08	88.30	72.67	91.86	81.57
Eye-tracking	92.43	76.95	95.15	85.02	89.92	76.50	92.38	82.61