Inferring a Mobile User's Valence and Arousal through On-Screen Text Analysis

Edita Džubur[1], Veljko Pejović[1]

[1]University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

Abstract

Understanding a user's emotional state is critical for building adaptive and intelligent mobile applications. In this paper we investigate the feasibility of inferring valence and arousal from the text displayed on smartphone screens. We developed AV-Sense, a mobile application that combines the Experience Sampling Method, a technique that prompts users to report their feelings in the moment, with passive screentext logging. In a two-week study with 12 participants, we collected 787 ESM responses and over 650,000 screentext entries. Data analysis revealed meaningful temporal and individual patterns in reported affect. We then explored the use of large language models to predict valence and arousal from screentext, but results indicated limited predictive power in this setting. Our findings highlight both the potential and current challenges of screentext-based affect inference, laying the groundwork for future research on emotion-aware applications and naturalistic psychological studies.

Keywords

Text analysis, experience sampling method, screentext sensing, valence, arousal, large language models

1. Introduction

Smartphones have become the primary medium for communication, information access, and digital services. Despite their ubiquity, these devices have little understanding of the emotional state of their users. Applications typically adapt based on context such as location or activity, but they rarely consider affective states such as valence and arousal.

Valence and arousal are the two fundamental dimensions of affect used in psychological models of emotion. Valence describes how pleasant or unpleasant an experience feels, whereas arousal reflects its intensity or level of activation, ranging from very calming to highly exciting [1]. This two-dimensional representation is widely used in affective science and has been established through decades of research.

Recognizing these states could enable more adaptive applications such as adjusting notifications, recommending suitable activities, or supporting well-being interventions. It could also provide valuable insights for psychological research in naturalistic settings [2].

Unlike physical activity or location, emotional state cannot be directly sensed with smartphone hardware. However, there is evidence that the content users engage with, particularly on-screen text, reflects their affective state [3, 4]. Simultaneously, recent advances in natural language processing, especially through large language models (LLMs), provide an attractive avenue for seamless analysis of texts in search for potential link between the content and the affect of a user consuming the content.

In this study, we developed *AV-Sense*, a mobile application based on the AWARE-Light [5] framework that integrates the Experience Sampling Method (ESM) with passive screentext logging. The application periodically prompted users to self-report their affective state on a two-dimensional valence–arousal grid, while continuously recording the textual content displayed on the smartphone screen. We conducted a two-week study with 12 participants, during which the system collected 787 ESM responses paired with approximately 650,000 screentext entries. This dataset enabled us to analyze temporal and individual patterns in reported affect, assess the availability of screentext data as contextual information, and explore the feasibility of applying large language models to predict valence and arousal from naturalistic

Human-Computer Interaction Slovenia 2025, October 13, 2025, Koper, Slovenia
veljko.pejovic@fri.uni-lj.si (V. Pejović)
0000-0002-9009-0024 (V. Pejović)

https://doi.org/10.26493/978-961-293-559-7.16

screen text. While the predictive performance of LLMs was limited, the study provides a first step towards understanding the potential and boundaries of screentext-based affect inference.

2. Related Work

2.1. Mobile sensing and the Experience Sampling Method

Mobile sensing has become an important way to study human behavior, since smartphones are always with users and carry a variety of built-in sensors [2]. When combined with ESM, this makes it possible to collect data about people's daily lives in a more ecologically valid way [6, 3]. ESM asks participants to provide in-the-moment self-reports, which helps reduce recall bias and capture affective states in their natural environment. Previous studies have shown that smartphones can measure mood and mental states by combining signals such as location, activity, and communication patterns [7, 8]. These works confirm that frequent affect measurement is feasible, but they also point out challenges around privacy, technical stability, and participant burden.

2.2. Inferring affect from text and screentext sensing

Language is one of the clearest ways people express emotions. Research has shown that linguistic markers are often linked to psychological and affective states [3]. This insight led to the development of sentiment analysis methods and affective lexicons, which are now widely used in psychology and computer science [9]. With the recent progress in machine learning, large language models have also been tested for emotion detection from text, offering better contextual understanding than earlier lexicon-based tools [10].

More recently, researchers have looked at the text people actually see on their smartphone screens, known as *screentext*. Teng et al. introduced a screentext sensor for Android as part of the AWARE-Light framework, which enables continuous and privacy-conscious logging of on-screen text [11]. In later work, they showed that screentext data can be used to predict affective states, with LLM-based prompting approaches performing better than simple classifiers [4]. These studies suggest that screentext could serve as a useful, non-intrusive signal of user affect, while also highlighting the difficulty of modeling subtle and context-dependent emotions.

Our work continues in this direction by combining screentext sensing with real-time affect reporting on a two-dimensional valence–arousal grid. Where earlier studies often used weekly or retrospective questionnaires, our approach links screentext with immediate self-reports, giving us a closer look at how affect can be inferred in everyday settings.

3. Data collection

3.1. AV-Sense application

AV-Sense was built on top of the AWARE-Light framework, which is an open-source platform for mobile sensing. The main goal of the app is to combine active self-reporting with passive data collection. For self-reports, we introduce a new type of an ESM question in the form of a two-dimensional grid as shown in Figure 1. On this grid, participants rate their current valence, ranging from negative to positive on a scale from -3 to 3, and their arousal, ranging from calm to excited on the same scale. The grid also displays examples of emotions placed on the coordinate system to illustrate the combinations of the two dimensions. When a prompt appeared, participants saw the instruction: *"Please rate how you feel right now based on your current mood."* They were encouraged to report their general emotional state at that moment, not necessarily emotions related to the phone content. This clarification ensured consistency across self-reports and reduced ambiguity about whether responses should reflect momentary feelings or reactions to specific on-screen material.

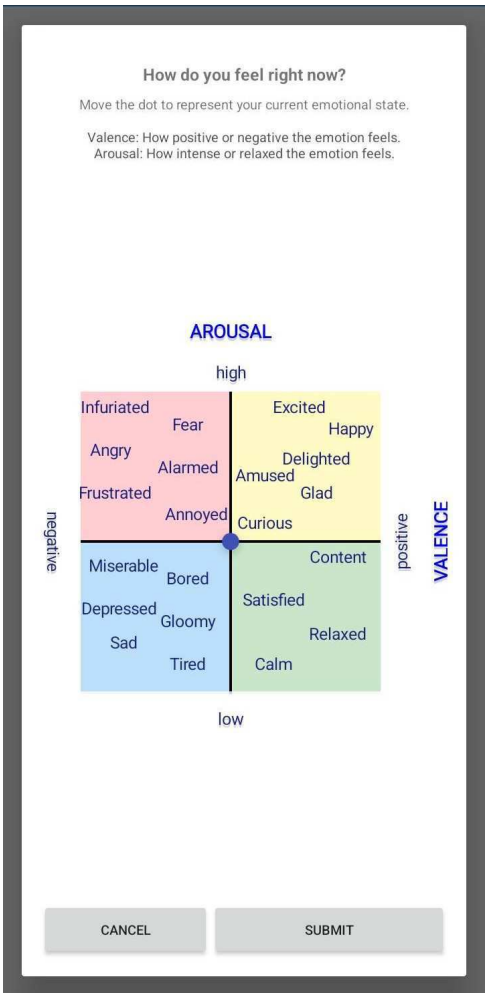

Figure 1: Valence-Arousal Questionnaire

The grid prompt is triggered after five minutes of continuous phone usage, based on screen-on and screen-off events. If the notification was not answered within five minutes, it disappeared. After each answered prompt, a timeout of two hours is applied to avoid overloading participants.

Alongside the ESM, the app also includes a screentext sensor. This sensor uses Android's accessibility service to access the view hierarchy of the operating system and extract the textual content displayed on the screen, without actually taking screenshots. To avoid collecting potentially highly-sensitive data, communication and banking apps are excluded from screentext sensing by default, with participants being able to add any other apps to the blacklist. Only the five minute window of screentext before an answered ESM prompt is saved into the database, while all other screentext is discarded.

3.2. User study

We conducted an AV-Sense field study with 12 participants, mostly university students between 21 and 30 years old. The study lasted for two weeks, during which participants had the app installed and running on their personal phones. The ESM prompts appeared several times per day, asking the users to report their current emotional state on the valence–arousal grid. Over the course of the study, we collected 787 ESM responses together with around 650,000 screentext entries. This provided us with a dataset where every self-reported emotional state could be linked to the text that was visible on the phone at the immediately preceeding time.

The collected data makes it possible to explore both how users' valence and arousal changes over time, as well as whether the screentext could serve as a signal for predicting the valence/arousal change.

4. Data overview

The ESM data are stored in form of individual records, each containing the notification time, the response time, and the selected valence and arousal values. The screentext data are stored separately and linked to ESM records by timestamps. For every answered prompt we only kept screentext entries from the five minutes before the response. This way each self-report can be directly connected to the text that was visible on the screen in the time leading up to it.

In our study, the number of answers varied among participants. Because of this, all distributions and averages in our analysis are calculated as weighted values, so that each participant contributed proportionally to their number of responses.

4.1. Temporal distribution

In Figure 2 we plot the temporal distribution of responses to ESM questionnaires. We observe a clear daily pattern already reported in similar previous studies [12]. Participants responded less during the night and early morning, with activity increasing throughout the day. We also compared workdays and weekends, with the distributions differing slightly between the two, but in both cases the majority of responses come during the active parts of the day.

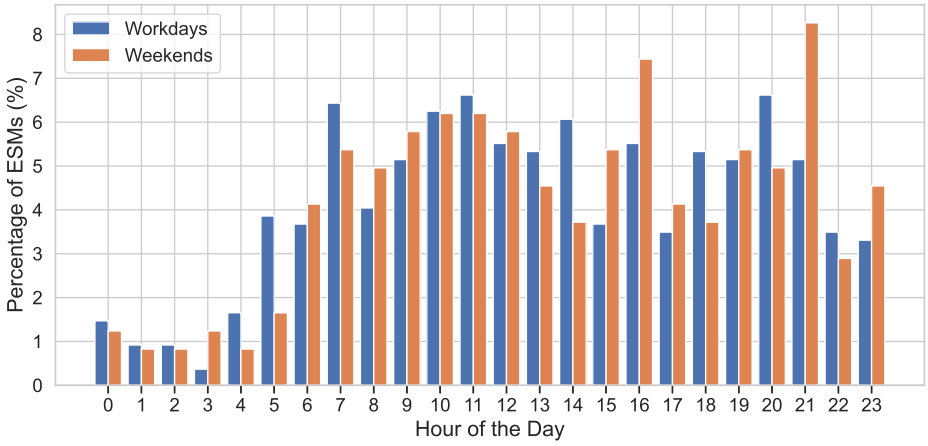

Figure 2: Average valence and arousal during workdays and weekends.

We also analyzed the response times to ESM prompts using a cumulative distribution function. The results showed that participants usually answered quickly. The median response time was about 13 seconds, and more than 90% of all responses were submitted within the first minute after the notification appeared. This means that most self-reports can be considered reliable in-the-moment reflections of the user's current state, since long delays that could introduce recall bias were rare [13]. The relatively fast response times also suggest that participants engaged with the study consistently and did not find the notifications overly disruptive.

4.2. Descriptive statistics of valence and arousal

Table 1 summarizes the mean and standard deviation of valence and arousal for each participant. Most valence averages are close to zero, with a slight shift toward positive values, while a few participants report mildly negative valence on the average. This suggests that neutral and slightly positive moods are most common, with extreme states appearing less often, which is typical for ESM data [6].

Arousal means were often slightly below zero, placing many responses in calmer or moderately active states. This is consistent with our notification scheme, which triggered prompts after five minutes of continuous phone use, usually during routine interaction rather than moments of high activation.

Standard deviations show clear individual differences, with some participants displaying more variability in arousal than valence. This aligns with the idea that arousal tends to fluctuate more with context, while valence often remains closer to neutral. Although group-level summaries are informative, they also hide strong individual differences, underlining the value of examining both perspectives.

Table 1
Descriptive statistics of valence and arousal by participant

Participant	Valence μ	Valence σ	Arousal μ	Arousal σ
1	0.35	1.45	0.02	2.16
2	0.43	1.50	-0.97	1.73
3	1.35	1.14	-1.32	1.72
4	0.47	1.92	-0.13	1.32
5	0.61	1.06	0.11	1.08
6*	0.12	0.71	0.00	0.71
7	0.72	1.88	-0.80	1.85
8	-0.08	1.82	-0.12	1.66
9	0.02	1.13	-0.74	1.67
10	-0.20	1.34	0.16	1.34
11	-0.07	1.45	-0.14	1.12
12	0.35	1.36	-0.24	1.89

* Participant 6 reported near-constant values (zeros on the grid); retained here for completeness, but excluded from the heatmap in Fig. 3.

We also look at the overall distribution of valence and arousal answers across participants. One participant (User 6) is removed from the analysis because almost all of their answers are zeros, possibly indicating quick, meaningless answering. Figure 3 shows a heatmap of the normalized distribution for the remaining participants. The majority of answers are located around neutral to slightly positive valence and low to moderate arousal. Extremes on the edges of the grid are rare, which fits with the idea that everyday ESM data tends to capture mostly neutral and mild states [6]. There is also a weak positive relationship between valence and arousal, with higher valence sometimes paired with higher arousal, and lower valence with lower arousal, although the spread is wide.

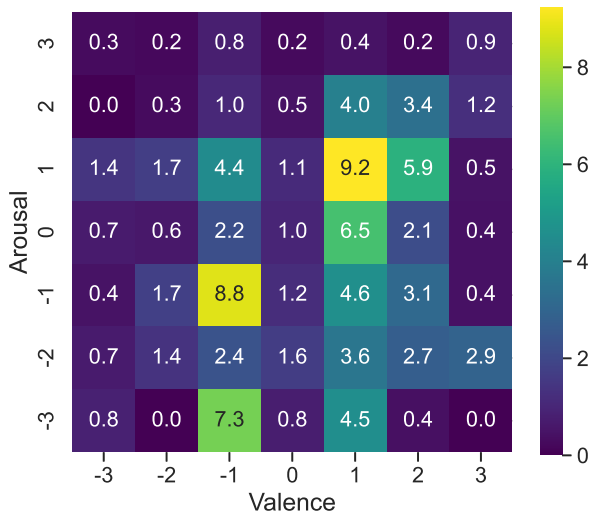

Figure 3: Heatmap of the normalized distribution of ESM answers (without User 6).

4.3. Daily pattern of valence and arousal

In our data we do not see a clear daily pattern for valence. As shown in Figure 4, the hourly averages move without a strong trend across the day, which differs from reports that often find higher positive valence in the morning and a decline toward evening [14]. Arousal shows a more typical daily rhythm. After waking up, subjective energy rises and reaches a peak in the first half of the day, followed by a noticeable afternoon dip and lower levels in the evening and at night [14, 15]. Hourly averages are computed with the weights described above so that participants contribute in proportion to their number of responses.

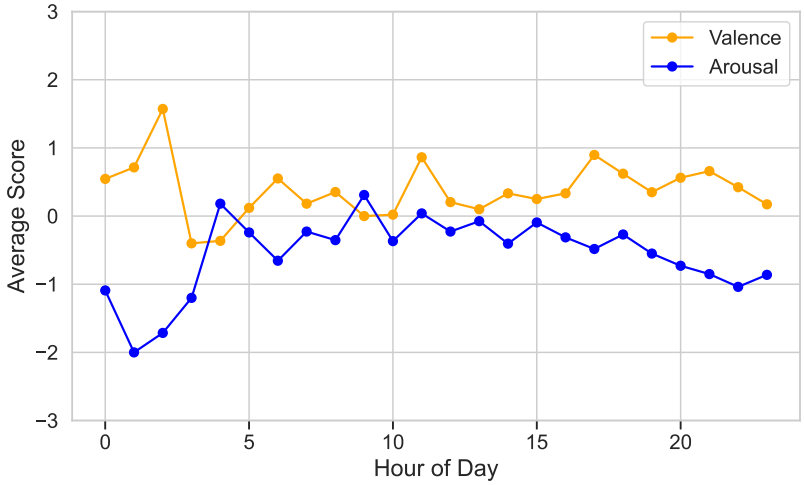

Figure 4: Hourly averages of valence and arousal across the study.

Average values alone can sometimes hide the variety of affective experiences during the day. To capture this diversity we grouped the answers into categories of valence (negative, neutral, positive) and arousal (low, neutral, high). Figures 5 and 6 show how the shares of these categories change across hours of the day. This view reveals aspects that averages miss. For example, a few very positive ratings can push the average upward even if there are many slightly negative ones, creating a misleading impression.

In our data, average valence stays slightly positive through most of the day, but the categorical view shows that negative moods became more common in the evening. For arousal we saw a similar pattern. Although the average drops in the afternoon, the distributions make it clear that this is driven by a higher share of low-energy states such as tiredness or calmness, together with a decrease in high-energy states. These categorical distributions therefore complement the averages and provide a more detailed look at daily affective dynamics.

4.4. Availability of screentext data

Figure 7 shows, for each participant, the share of ESM answers where screentext data are available in the five minutes before the response. For most participants, more than half of their ESM answers are linked with screentext, and in many cases the coverage is above 90 %. A few participants have very low availability, with one showing no screentext data at all and another only about 22 %. These cases were most likely caused by missing permissions or technical issues that prevented the sensor from working, or by frequent use of apps that were blocked from screentext capture. Such answers were excluded from further analyses, since models in the next section require at least some text context. For the remaining participants, the high availability rates mean that ESM answers could usually be interpreted together with surrounding screentext, giving us a rich basis for later analysis.

To better understand the context of the collected screentext, we also looked at the categories and

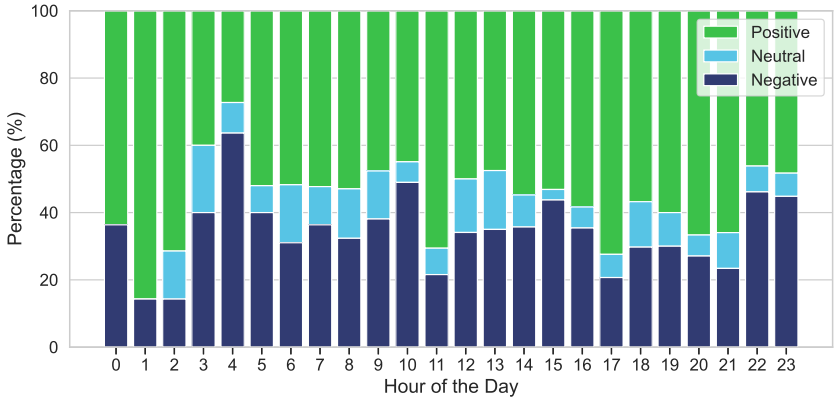

Figure 5: Share of answers by valence categories (negative, neutral, positive) across hours of the day.

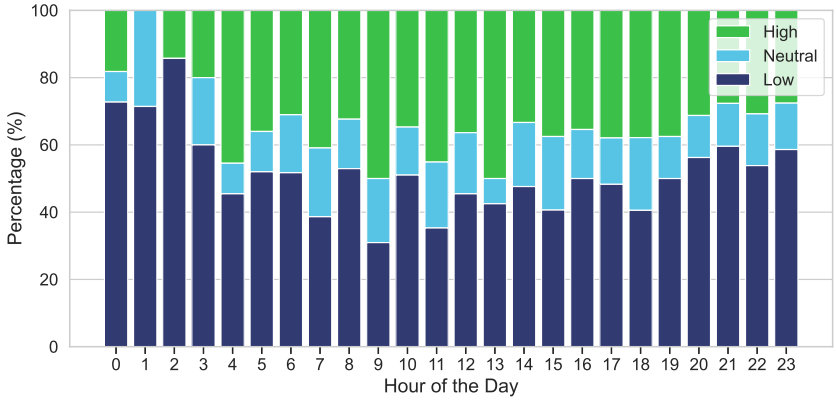

Figure 6: Share of answers by arousal categories (low, neutral, high) across hours of the day.

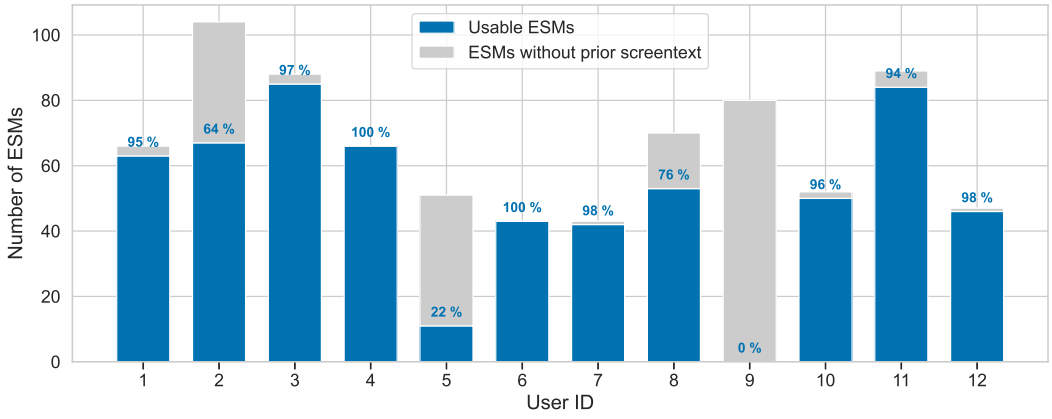

Figure 7: Share of ESM answers per participant with screentext available in the five minutes before the response.

specific apps that appeared most often before ESM answers. Social, messaging, and video apps were by far the most common, while other categories like browsing, reading, games, or productivity were much less frequent. At the app level, Instagram, YouTube, and Reddit were the top sources of screentext, followed by Discord, Chrome, and TikTok. This confirms that most of the captured screentext came from social and media-related contexts, which is important for interpreting the results of our predictive models.

5. LLM inference

After collecting ESM responses and screentext, we explored whether LLMs could predict valence and arousal from the text captured on participants' devices. We tested two prompting strategies and compared them to a simple baseline. For each participant, we divided their examples into a training group and a test set (80/20 split). The first approach used few-shot prompting, where a subset of training examples (we included $k = 3$–5 per participant) was inserted into the prompt together with their associated arousal–valence scores. The model was then asked to predict the score for a new example from the test set. The second approach used rubric-based prompting, where instead of concrete examples, the model was given a set of short written instructions. These instructions described how to assign arousal and valence values based on emotional tone and intensity (e.g., positive language as higher valence, negative language as lower valence, urgent text as higher arousal, calm text as lower arousal). The baseline model simply predicted the average score of the training group for each participant. We evaluated all approaches using Mean Absolute Error (MAE).

We tested these strategies with GPT-5-nano accessed via API.

Table 2
MAE of predicted valence and arousal per each participant with different prediction methods. Highlighted is the best performing approach for each user and separately for valence and arousal.

Participant	Valence (MAE)			Arousal (MAE)		
	Basic approach	Few-shot approach	Rubric approach	Basic approach	Few-shot approach	Rubric approach
1	1.385	1.692	1.154	1.154	1.615	1.385
2	0.692	1.538	1.538	1.615	1.846	1.846
3	1.000	1.235	1.176	1.235	1.529	1.529
4	1.154	1.231	1.308	1.692	1.846	1.923
6	0.375	0.500	0.500	0.250	0.375	0.375
7	1.500	1.625	1.250	1.000	1.250	1.125
8	2.000	1.818	1.455	1.455	1.545	1.455
10	1.100	1.100	1.200	1.400	1.500	1.000
11	1.188	1.438	1.375	1.813	1.875	1.938
12	0.556	1.333	1.222	0.889	0.778	0.889

Table 2 summarizes the results. Prediction performance varied substantially among participants and prompting strategies. For some participants, the rubric-based prompt gave the lowest MAE, while for others, the few-shot prompt worked slightly better. However, in most cases, the baseline model outperformed both prompting strategies. This demonstrates that predicting valence and arousal from screentext is highly challenging and that simple prompting setups with LLMs do not yield consistent improvements over a trivial average-based predictor.

Participant-level variability also shaped the results. Participant 6 had uniformly low errors for all approaches because their ESM responses were highly repetitive and clustered around neutral values. This made predictions easier but at the same time less informative. In contrast, participants with more varied responses introduced more complexity for the models, and errors rose accordingly. A broader limitation is that the models had difficulty adapting to different personal styles of smartphone use, from routine messaging to more expressive content. Personalization may therefore be important for making these models useful in practice.

5.1. Correlation of sentiment with valence and arousal

The premise behind the analysis above was that the text a user sees on the screen reflects the user's valence and arousal. This link was further reinforced in the few-shot prompting setup, where texts were explicitly paired with the reported valence and arousal. Nevertheless, the LLM may sometimes

infer only the sentiment of the text itself rather than the user's emotional state. In such cases, it is useful to know whether the sentiment of the text, on average, correlates with the expressed sentiment of the user.

We therefore conducted an additional experiment in which we classified screentext into coarse sentiment categories of polarity (negative, neutral, positive) and intensity (none, low, medium, high). These categories were then mapped to numerical values to enable comparison with the self-reported ESM scores. Polarity was mapped to $-1, 0, 1$ and intensity to the interval $[0, 1]$. We then calculated Pearson and Spearman correlations with the valence and arousal dimensions.

The results showed a weak but statistically significant correlation between sentiment polarity and valence ($n = 591$, Pearson $r = 0.143$, $p = 0.00048$; Spearman $\rho = 0.132$, $p = 0.0013$). This indicates that more positive sentiment in the screentext slightly increased the likelihood of a higher reported valence, although the effect was small. For intensity and arousal, however, no meaningful relationship was found ($n = 591$, Pearson $r = 0.004$, $p = 0.919$; Spearman $\rho = 0.002$, $p = 0.959$).

These findings highlight a key problem. If the models rely heavily on sentiment cues in the text, they only recognize whether text is broadly positive or negative. This overlaps slightly with reported valence, but it does not capture differences in arousal. As a result, calm and excited states with the same polarity may be confused by the model. This gap reflects a well-known limitation of sentiment analysis approaches, which emphasize polarity while overlooking the activation dimension. In naturalistic screentext, this limitation becomes even more pronounced because text fragments are short, noisy, and lack broader context.

6. Discussion and Conclusions

Our results showed that combining ESM with screentext collection was feasible in everyday settings. Most ESM answers were linked with screentext, which provided useful context for analysis. The data was often noisy and sometimes missing due to blocked apps or technical limits, highlighting the trade-off between privacy and coverage in real-world deployments.

In terms of affect patterns, arousal showed a clearer daily rhythm than valence. Category distributions revealed changes that averages alone would have hidden, such as more negative moods in the evening and more low-energy states in the afternoon. These patterns are consistent with known affect dynamics but were also shaped by the way prompts were triggered.

LLM inference proved more challenging. Performance varied across participants and prompting strategies, with neither few-shot nor rubric-based prompting showing a consistent advantage over the baseline. This suggests that models were sensitive to individual differences and that simple prompting was not sufficient to capture subtle affective variation. The sentiment-based experiment confirmed this limitation. The weak correlation with valence and almost no correlation with arousal showed that models relied heavily on sentiment cues, which explains why predictions often missed the difference between calm and excited states of the same polarity. The gap between sentiment detection and full valence–arousal inference therefore remains large, especially in naturalistic data.

The study has several limitations. The sample was small, with only twelve participants, and the duration was short, which limits generalizability. Participants were also relatively homogeneous in age and background. The data itself was restricted to screentext, without additional modalities such as audio, images, or sensor data that might capture arousal more effectively. Screentext was often noisy, containing UI elements and symbols rather than only meaningful content. Privacy rules also reduced coverage, since communication and banking apps were blocked by default and participants could blacklist additional apps. Finally, the prompting scheme, which triggered questions after five minutes of continuous screen use followed by a two-hour timeout, shaped the contexts in which responses were collected and likely emphasized routine phone use.

Even with these limits, the main takeaway is clear. ESM plus screentext is feasible on personal phones and yields a dataset that links self-reports with the text people actually see. Text alone is not enough for robust affect inference, especially for arousal. Looking forward, there is potential to combine screentext

with other lightweight signals, such as app usage context, time of day, or sensor data, to capture a fuller picture of affect. With larger and more diverse samples, and with careful attention to privacy, these approaches could support the development of more reliable affect-aware mobile applications. Possible applications include digital well-being tools that adapt notifications based on the user's state, chat or messaging apps that detect emotionally charged interactions, or learning and productivity apps that adjust difficulty and pacing depending on energy and mood. They could also support psychological research by providing ecologically valid affect data without requiring intrusive sensing.

References

[1] E. A. Kensinger, D. L. Schacter, Processing emotional pictures and words: Effects of valence and arousal, Cognitive, Affective, & Behavioral Neuroscience 6 (2006) 110–126.

[2] G. Miller, The smartphone psychology manifesto, Perspectives on Psychological Science 7 (2012) 221–237.

[3] Y. R. Tausczik, J. W. Pennebaker, The psychological meaning of words: Liwc and computerized text analysis methods, Journal of language and social psychology 29 (2010) 24–54.

[4] S. Teng, T. Zhang, S. D'Alfonso, V. Kostakos, Predicting affective states from screen text sentiment, in: Companion of the ACM International Joint Conference on Pervasive and Ubiquitous Computing, Melbourne, Australia, 2024.

[5] N. van Berkel, S. D'Alfonso, R. K. Susanto, J. Goncalves, V. Kostakos, Aware-light: a smartphone tool for experience sampling and digital phenotyping, Personal and Ubiquitous Computing 27 (2023) 435–445.

[6] S. Shiffman, A. A. Stone, M. R. Hufford, Ecological momentary assessment, Annual Review of Clinical Psychology 4 (2008) 1–32.

[7] R. LiKamWa, Y. Liu, N. D. Lane, L. Zhong, Moodscope: Building a mood sensor from smartphone usage patterns, in: Proceeding of the 11th annual international conference on Mobile systems, applications, and services, 2013.

[8] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, A. Aucinas, Emotionsense: a mobile phones based adaptive platform for experimental social psychology research, in: Proceedings of the 12th ACM international conference on Ubiquitous computing, 2010.

[9] S. M. Mohammad, Sentiment analysis: Detecting valence, emotions, and other affectual states from text, in: Emotion measurement, Elsevier, 2016.

[10] S. M. Mohammad, P. D. Turney, Crowdsourcing a word–emotion association lexicon, Computational intelligence 29 (2013) 436–465.

[11] S. Teng, S. D'Alfonso, V. Kostakos, A tool for capturing smartphone screen text, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 2024.

[12] V. Pejovic, M. Musolesi, Interruptme: Designing intelligent prompting mechanisms for pervasive applications, in: ACM UbiComp, Seattle, WA, USA, 2014.

[13] A. A. Stone, S. Schneider, J. M. Smyth, Evaluation of pressing issues in ecological momentary assessment, Annual Review of Clinical Psychology 19 (2023) 107–131.

[14] F. Bu, J. K. Bone, D. Fancourt, Will things feel better in the morning? a time-of-day analysis of mental health and wellbeing from nearly 1 million observations, BMJ Ment Health 28 (2025).

[15] A. A. Stone, J. M. Smyth, T. Pickering, J. Schwartz, Daily mood variability: Form of diurnal patterns and determinants of diurnal patterns, Journal of Applied Social Psychology 26 (1996) 1286–1305.