Investigating Contextual Cues as Indicators for EMA Delivery

Varun Mishra
Dartmouth College
Hanover, NH
varun@cs.dartmouth.edu

Kelly Caine
Clemson University
Clemson, SC
caine@clemson.edu

Byron Lowens
Clemson University
Clemson, SC
blowens@g.clemson.edu

David Kotz
Dartmouth College
Hanover, NH
david.f.kotz@dartmouth.edu

Sarah Lord
Dartmouth College
Hanover, NH
sarah.e.lord@dartmouth.edu

UbiComp/ISWC'17 Adjunct, September 11–15, 2017, Maui, HI, USA.

ACM ISBN 978-1-4503-5190-4/17/09.
DOI: https://doi.org/10.1145/3123024.3124571

Abstract
In this work, we attempt to determine whether the contextual information of a participant can be used to predict whether the participant will respond to a particular Ecological Momentary Assessment (EMA) trigger. We use a publicly available dataset for our work, and find that by using basic contextual features about the participant's activity, conversation status, audio, and location, we can predict if an EMA triggered at a particular time will be answered with a precision of 0.647, which is significantly higher than a baseline precision of 0.41. Using this knowledge, the researchers conducting field studies can efficiently schedule EMAs and achieve higher response rates.

Author Keywords
Ecological Momentary Assessment; Notification; Interruptibility; Mobile sensing; Context-aware computing

ACM Classification Keywords
H.5.2 [Information interfaces and presentation (e.g., HCI)]: User Interfaces

Introduction
Ecological Momentary Assessment (EMA) [9], also known as the Experience Sampling Method (ESM) [4], is a commonly used technique designed to collect information about a participant's current behavior and experiences while they

are in their natural environment. EMA methods ask the participant to answer questions, in the moment, to assist researchers in collecting ecologically valid self-reported data [8]. By collecting responses "in the moment", EMA reduces recall bias relative to methods that query the participant at the end of the day or end of the study period. By using technology to collect responses, EMA reduces participant interaction with researchers relative to observational studies in which the researcher shadows the participant [9].

The ubiquitous presence of smartphones and wearable devices has enabled the common use of EMA in a broad range of studies. Researchers have successfully used EMA to collect ground truth for annotating the sensory measurements and the construction of training data for machine-learning models of human emotion, mood, stress, and personality [12, 2, 13].

The problem, however, is that the researchers are dependent on participants to correctly and diligently answer the EMA prompt. For the participants, responding to frequent or lengthy questionnaires can be burdensome. This burden may decrease participant responsiveness over the course of a study, an effect noted by several researchers [12, 10].

We anticipate that EMA participants would be more responsive if the prompts occur in a context where they are more likely to respond. We propose to time EMA prompts to suit the participant's context – reducing participant burden and increasing participant compliance. To do so, we must first understand how context affects participant compliance (responsiveness) to EMA prompts.

In this work, we evaluate the context of the participant to determine whether s/he is likely to answer an EMA. We investigate some basic features about participant context (which includes activity, audio, conversation, and location) to determine whether contextual information enables us to predict whether a given EMA prompt is likely to be answered quickly. Our work is the first to use activity, audio, conversation, and location data to predict whether an EMA prompt will be answered. Such a predictive model can help researchers develop effective strategies for delivery of EMA prompts without over-burdening the participant.

We use the publicly available StudentLife Dataset [12]. The dataset consists of longitudinal data from 48 participants over a period of 10 weeks. While the dataset itself contains a wide variety of data (including phone sensor and usage data, EMAs, surveys, dining data, and more), our work focuses particularly on the activity, audio, conversation, and location data along with the self-reported EMA data.

Background

Prior work has introduced an assortment of time-based sampling schemes for the delivery of EMA prompts to participants, selected by the researcher based on the goals of the study [1, 3, 9]. Prompts may be 'triggered' (1) at predetermined times, (2) at random times according to some parameters, or (3) at dynamic times according to some contextual policy (such as location, activity, physiological state, or combination thereof).

None of the above methods account for the participant's availability to respond to the prompt. A participant might not be available to answer a particular prompt, for many reasons: the EMA device may not be present, the social context may be require the participant's attention, or the participant's activity prevents him or her from seeing the prompt or from responding. In such cases, the participant may respond late (if permitted by the EMA protocol) or never. In some studies, even a delayed response may not meet the research goals.

Context	Values
Activity	0 : Stationary 1 : Walking 2 : Running 3 : Unknown
Audio	0 : Silence 1 : Voice 2 : Noise 3 : Unknown
Conversation	Start time, End time
GPS Location	Latitude, Longitude
WiFi Location	On-campus Location from WiFi scan

Table 1: Contextual information available in the StudentLife dataset.

To address these issues, the EMA policy should pick a 'good' time to trigger prompts – times when the user is more likely to answer the prompt. Several studies have sought to find such *opportune moments*, when the user is likely to respond to a notification (any notification, not necessarily EMA prompts) [7, 11, 5, 6].

The most prominent example is *InterruptMe*, an interruption management library for Android smartphones, designed to allow researchers to look at opportune moments to interrupt the user. They consider contextual information like activity and location [7]. Their analysis also uses features computed from data reported by users, and the researchers achieve a precision of 0.64 in estimating whether a participant will respond to a notification prompt.

Turner et al. investigated whether to push or delay a notification based on contextual information about the phone, including motion, charging state, volume state, ambient light, and phone orientation [11]. They report preliminary results with accuracy of up to 60%.

Other researchers have looked at delivering notifications at activity 'breakpoints' [5] and discovered that delivering a notification at a breakpoint resulted in lower participant cognitive load as compared to those sent out "immediately" [6].

In contrast to the above research, we look at a broader set of contextual features, and use passively collected sensing data to predict whether a given prompt will be answered.

StudentLife Dataset

We use the publicly available StudentLife Dataset, which consists of a wide range of data collected from 48 participants over 10 weeks [12]. Table 1 lists some of the interesting sensor data. The study also used EMA to collect several types of self-report data: stress, affect, behavior, mood, sleep, and activity.

The StudentLife app triggered several EMA prompts each day, based on a predetermined schedule determined by the research team. The schedule was the same for all participants but changed every week. For each EMA response, the dataset recorded the time and content of the response – but does not indicate which prompt corresponds to which response, or when the prompt was triggered.

We seek to determine and then develop a model to predict, how quickly participants will respond to a prompt after it is triggered. Because StudentLife dataset does not include the trigger time, we must first estimate the time each EMA prompt was triggered, based on the responses available in the dataset.

In the following sections, we discuss our method to reconstruct the likely trigger times, followed by a discussion of our prediction model and the results obtained.

Trigger Time Estimation

In StudentLife, the trigger schedule was identical for all participants, but not recorded in the dataset. The trigger schedule varied from week to week, but the number of prompts per day was small (typically one or two). Our key insight is that the total number of responses to a EMA in a short time immediately following the EMA trigger time should be significantly higher than the number of responses at a later time. In other words, we expect that most participants respond quickly, leading to peaks in the number of responses over time, allowing us to infer that a prompt was triggered shortly before each such peak.

To pursue this approach, we had to verify that EMA prompts were triggered sparingly during the day; if prompts were frequent, responses may be frequent and distributed through-

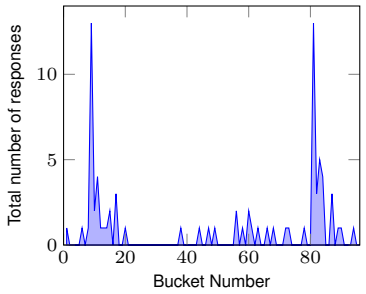

Figure 1: Histogram of responses across all participants in one day, in 15-minute buckets.

Context	Features
Activity	*before_activity* and *after_activity*, where each can take a value from 0-3, depending on the labels in Table 1
Audio	*before_audio* and *after_audio*, where each can take a value from 0-3, depending on the labels in Table 1
Conversation	*before_convo* and *after_convo*, where each can either be 'true' or 'false', depending on whether there was a conversation detected in that window
Location	*before_loc* and *after_loc*, where each can take be one label depending on the building type: study, dorm, food, gym, etc.
Time	*time* of the day, *day* of the week

Table 2: The features computed for the different contexts.

out the day, making it difficult to discern peaks. Although the StudentLife study triggered multiple prompts each day (about 8 per day) we focus on one category – the *stress* EMA – which was triggered only 1-2 times a day.

We group the responses by response time into 15-minute buckets, and count the number of responses in each bucket. Figure 1 plots the resulting histogram of responses to the stress EMA in one day of the StudentLife dataset. We can see that the number of responses increases drastically in the 9^{th} and the 81^{st} buckets. Based on our hypothesis, we conclude that the EMAs were triggered twice in that day – during the 9^{th} and the 81^{st} blocks.

To find the trigger times, we use a custom peak-detection algorithm to find the blocks in which the EMA was triggered. For every such block detected by our algorithm, we assume the corresponding prompt was triggered at the start time of that block. Using this approach, we determined 54 trigger times for the stress EMA over the length of the study. Although the EMA may have been triggered more times – we may have overlooked some peaks – we are confident of these 54 occasions.

We observed that all 54 occasions discovered by our algorithm were either at the 0^{th} or the 30^{th} minute of an hour. We contacted the authors of StudentLife and they confirmed our findings, saying that their EMA trigger times were always at the 0^{th} or 30^{th} of the hour. This confirmation gave us confidence that our estimated trigger times were accurate.

Next, for every participant, we check for a response within 4 hours of the estimated EMA trigger time. If a response exists, we assume the participant answered that the prompt, whereas if there is no response within the 4 hours window, we assume the participant did not answer that EMA prompt. The reason being that the StudentLife system does not save a prompt 'id' with the responses, so if we look at a longer time period, then the response might be to a later prompt, instead of the current prompt. We report a total of 906 responses from 2,179 prompts across all participants.

With solid estimates for the trigger time of the EMA prompts, we explain our prediction model in the next section.

Prediction Model

In this work, we use contextual information – activity, audio, conversation and location – all of which are readily available in the StudentLife dataset. As shown in Table 1, the dataset consists of two different types of locations – (1) GPS based location, the latitude and longitude of the participants' current location, and (2) Wi-Fi based location, which provides the on-campus building name in or around which the student is present. Since the building name can give us more information about a participant's location, we use the Wi-Fi based location in our model. We then map each building name to a particular category (e.g., study, dorm, food, street) and use these labels in our predictive model.

In our model, we look not only at the contextual information leading up to the time when an EMA was triggered, but also if there was any *change* in context during that time. We consider a window of time before an EMA trigger time, and compute the "before" and "after" contextual features on that window, so that we can capture the context change in that window. Table 2 lists the features we compute. For example, if the time at which the EMA prompt was triggered was t, and the size of the time window we use to compute features is Δt, then the "before" features will be computed in the time range $[t - \Delta t,\ t - \frac{\Delta t}{2}]$, and the "after" features will be computed at $[t - \frac{\Delta t}{2},\ t]$. For our experiments, we set $\Delta t = 10$ minutes.

Classifier	Precision	Recall
SVM	0.647	0.526
Random Forest	0.633	0.551
Naive Bayes	0.635	0.546
Baseline	0.41	0.42

Table 3: Predicting if an EMA will be answered based on the context at the time of prompt.

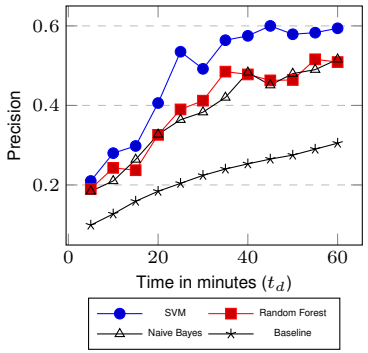

Figure 2: Predicting whether an EMA will be answered within a given time threshold (t_d), based on the context at the time of the prompt.

Since we aim at modeling interruptability, we predict the following outcomes: (1) whether a participant will respond to an EMA prompt, ever, and (2) whether a participant will respond to an EMA prompt within a given time interval, t_d. Prior works like InterruptMe have also measured similar outcomes for measuring interrupt ability [7]. For both the outcomes, we report the *precision*, i.e., the proportion of the instances our model predicted the prompt will be answered which actually were answered, and the *recall*, i.e., the proportion of all the instances the prompt was actually answered and was identified by our model.

To evaluate the first outcome: for each EMA prompt we calculate the notification context for every participant and label it *true* if that participant provided a response to that EMA prompt,and *false* otherwise. We then perform 10-Fold cross validation using three different classifiers – SVM, Random Forest and Naive Bayes – and report the results in Table 3. We also report the baseline classification results for comparison.This baseline is calculated by classifying the instances with a probability based on the proportion of EMA prompts that were actually answered in the training set.

We observe that all the context-based models perform significantly better than the baseline model, consistently achieving a precision above 0.63, with a highest precision precision of 0.647, which is similar to the precision achieved by InterruptMe. Furthermore, for the highest precision, we achieve recall greater than 0.52, which is significantly better than the recall reported in InterruptMe for a similar precision. This suggests that in comparison to InterruptMe, our model finds a greater proportion of *opportune* moments, with comparable precision.

To evaluate the second outcome: for each EMA prompt we calculate the notification context for every participant and label it *true* if that participant provided a response to that EMA prompt, within a *threshold time* (t_d), and *false* otherwise. Figure 2 shows the 10-fold cross-validation results across different classifiers. Observe that as we increase the time boundary (t_d), the precision also improves.

We further sought to understand how context affected response to EMA prompts. For our purposes, we define *responsiveness* as the percentage of prompts a participant answered. We look at how a *change* in context in the time just before a prompt increased or decreased the responsiveness of a participant in that context, as compared to the baseline measure, i.e., overall responsiveness across all prompts. In Table 4 we observe that context changes had only a slight impact on participant responsiveness, when we consider the average across all the participants. We found, however, a substantial change in responsiveness when we look at individual participants (in this table we examine two randomly selected participants): note, for example, how a change in *location* decreased the responsiveness of Participant 1 (P1) by 13.6%, whereas it increased the responsiveness of Participant 2 (P2) by 16%. It is interesting to observe how a context can have opposite effects on the responsiveness of different participants.

Conclusion and Future Work

In this paper we evaluate the use of contextual information to predict whether a participant will respond to an EMA prompt. Specifically, we explored activity, conversation, audio and location context from the StudentLife dataset. While we understand that interruptability is based on a wide range of factors, our preliminary results give us the confidence to explore deeper. In future work, we hope to explore factors like telephone and SMS logs, phone-app usage, phone-charging events, and calendar events. We also aim to develop an application that triggers EMA prompts according

Contexts	All	P1	P2
Baseline	43.3%	80.0%	59.0%
Activity			
Change	3.9%	2.2%	41.0%
No Change	-0.4%	-0.2%	-1.9%
Audio			
Change	1.2%	10.8%	-34.0%
No Change	-0.3%	-4.0%	3.5%
Conversation			
Change	-2.1%	3.3%	21.0%
No Change	0.4%	-2.0%	-2.6%
Location			
Change	3.2%	-13.6%	16.0%
No Change	-0.7%	2.0%	-1.5%

Table 4: Affect of context on the change in *responsiveness* towards EMA prompts: across **all** participants, and two randomly chosen participants (**P1** and **P2**).

to context so we can evaluate the effect on participant's response time, quality of response, and number of responses.

Acknowledgement

This research results from a research program at the Institute for Security, Technology, and Society, supported by the National Science Foundation under award numbers CNS-1314281, CNS-1314342, CNS-1619970, and CNS-1619950, The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

References

[1] G. Affleck, H. Tennen, S. Urrows, P. Higgins, M. Abeles, C. Hall, P. Karoly, and C. Newton. Fibromyalgia and women's pursuit of personal goals: a daily process analysis. *Health Psychology*, 17(1):40, 1998.

[2] K. Hovsepian, M. al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar. cStress: Towards a gold standard for continuous stress assessment in the mobile environment. In *UbiComp'15*, 493–504. ACM, 2015.

[3] S. S. Intille, A. Stone, and S. Shiffman. Technological innovations enabling automatic, context-sensitive ecological momentary assessment. *The science of real-time data capture: Self-reports in health research*, 308–337, 2007.

[4] R. Larson and M. Csikszentmihalyi. The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, 1983.

[5] M. Obuchi, W. Sasaki, T. Okoshi, J. Nakazawa, and H. Tokuda. Investigating interruptibility at activity breakpoints using smartphone activity recognition api. In *UbiComp'16: Adjunct*, 1602–1607. ACM, 2016.

[6] T. Okoshi, J. Ramos, H. Nozaki, J. Nakazawa, A. K. Dey, and H. Tokuda. Attelia: Reducing user's cognitive load due to interruptive notifications on smart phones. In *PerCom'15*, 96–104, March 2015.

[7] V. Pejovic and M. Musolesi. Interruptme: Designing intelligent prompting mechanisms for pervasive applications. In *UbiComp'14*, 897–908. ACM, 2014.

[8] S. Shiffman. Ecological momentary assessment. In *The Oxford Handbook of Substance Use and Substance Use Disorders*. 1998.

[9] S. Shiffman, A. A. Stone, and M. R. Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[10] V. W. S. Tseng, M. Merrill, F. Wittleder, S. Abdullah, M. H. Aung, and T. Choudhury. Assessing mental health issues on college campuses: Preliminary findings from a pilot study. In *UbiComp'16: Adjunct*, 1200–1208. ACM, 2016.

[11] L. D. Turner, S. M. Allen, and R. M. Whitaker. *Push or Delay? Decomposing Smartphone Notification Response Behaviour*, 69–83. Springer International Publishing, 2015.

[12] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell. Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *UbiComp'14*, 3–14. ACM, 2014.

[13] R. Wang, G. Harari, P. Hao, X. Zhou, and A. T. Campbell. Smartgpa: How smartphones can assess and predict academic performance of college students. In *UbiComp'15*, 295–306. ACM, 2015.