Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users

MARTIN PIELOT, Telefónica Research
BRUNO CARDOSO, Universidade Nova de Lisboa
KLEOMENIS KATEVAS, Queen Mary University of London
JOAN SERRÀ, Telefonica Research
ALEKSANDAR MATIC, Telefonica Alpha
NURIA OLIVER, Telefonica Research, now at Data-Pop Alliance and Vodafone Research

Many of today's mobile products and services engage their users proactively via push notifications. However, such notifications are not always delivered at the right moment, therefore not meeting products' and users' expectations. To address this challenge, we aim at developing an intelligent mobile system that automatically infers moments in which users are open to engage with suggested content. To inform the development of such a system, we carried out a field study with 337 mobile phone users. For 4 weeks, participants ran a study application on their primary phones. They were tasked to frequently report their current mood via a notification-administered experience-sampling questionnaire. In this study, however, we analyze whether they voluntarily engaged with content that we offered at the bottom of that questionnaire. In addition, the study app logged a wide range of data related to their phone use. Based on 120 Million phone-use events and 78,930 questionnaire notifications, we build a machine-learning model that before delivering a notification predicts whether a participant will click on the notification and subsequently engage with the offered content. When compared to a naïve baseline, which emulates current non-intelligent engagement strategies, our model achieves 66.6% higher success rate in its predictions. If the model also considers the user's past behavior, predictions improve 5-fold over the baseline. Based on these findings, we discuss the implications for building an intelligent service that identifies opportune moments for proactive user engagement, while, at the same time, reduces the number of undesirable interruptions.

CCS Concepts: • **Human-centered computing → Mobile computing**;

Additional Key Words and Phrases: Push Notifications; Conversion; Proactive Recommendations; Mobile Devices

ACM Reference format:
Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 91 (September 2017), 25 pages.
https://doi.org/10.1145/3130956

*While planning and conducting the study, all co-authors worked at Telefónica.

2474-9567/2017/9-ART91
https://doi.org/10.1145/3130956

1 INTRODUCTION

Many of today's mobile software products and services, such as games, brands, social networks, or news feeds, need to engage their users in order to be successful, where *engagement* refers to the involvement into something that attracts and holds our attention [8, 37]. Failing to engage users can endanger the sustainability of products and services, particularly if they are free to use and cover their costs through secondary streams of income, such as advertisements or upsells, which require repeated use of the service. However, engaging mobile users is increasingly challenging as we are exposed to an ever-growing number of online products and services which are all competing for our attention.

Given that user attention is a limited resource, we can observe a shift from pull-driven towards push-driven engagement[1], where mobile apps try to engage users proactively via, *e.g.*, push notifications. As a result of this paradigm shift, users are exposed to an increasing number of ill-timed notifications that are delivered in an as-soon-as-possible manner. Unfortunately, this naïve strategy leads to a number of negative outcomes, such as decreased satisfaction, uninstalls, negative emotions, and even hyperactivity or inattention [2, 25, 28, 58]. For that reason, a growing body of work from the UbiComp community has explored how to automatically detect moments when notifications are not interruptive [34, 39, 42, 43, 47, 54]. However, predicting if users are interruptible, *i.e.*, whether their attention may be attracted, does not imply that their attention will be held as well, *i.e.*, result into engagement.

In this paper, we describe our work towards the design of an intelligent system to detect *opportune moments* for products and services to *engage* users, *i.e.*, identifying moments where content is likely to attract and hold the user's attention. If done properly, such a system would contribute to reduce the cost of inopportune interruptions as well as to improve engagement.

To explore the feasibility of detecting opportune moments for engaging with users, we conducted a field study with 337 participants. For an average duration of 4 weeks, the participants used a mobile study app that (1) passively collected rich sensor data about their context and phone usage; (2) frequently prompted participants to fill out a very short mood questionnaire that served as a deception regarding the real goal of the study; (3) below the questionnaire responses, recommended diverse content from different categories, such as games, reader/news, music, and markets [51]; and (4) recorded when the participants voluntarily choose to engage with and consume that recommended content. The data collection yielded over 120 Million mobile phone use events and 78,930 instances of questionnaire notifications. In 30,689 cases, participants responded to the questionnaire, and in 3,367 cases, they voluntarily engaged with the suggest content. We used machine-learning methods to create a classifier that predicts, on the basis of the mobile phone use events, whether a notification will lead to a click and subsequently to voluntary engagement. The main contributions of this paper are:

- a machine-learning-based approach for predicting if a user will engage with proactively-recommended content, only relying on data collected via the mobile phone, achieving a 66.6% better precision than a baseline model;
- an analysis of the relative contribution of a range of variable categories to the developed model, namely demographics, phone status, phone use patterns, communication activity, and context, showing that context, communication activity, and phone use patterns contribute most to the top-ranked predictors; and
- an investigation into the predictive power of past behavior with respect to the engagement with the recommended content, which shows that the classification accuracy can be increased over 500% over the baseline model while significantly reducing engagement attempts with users who did not show much interest in the past recommended content.

[1] https://techcrunch.com/2015/04/21/notifications-are-the-next-platform/

With its focus on engagement prediction, this study fills a gap in a so far understudied aspect of interruptibility. Our findings and implications can inform the design and realization of an intelligent, engagement-timing service, that minimizes the disruption of their users while providing value to both users and service providers.

2 RELATED WORK

Attempting to proactively engage users with a product or service requires to draw their attention to it. The common form of attempting to attract user attention to content on their mobile phone is done via notification alerts. Traditionally being used for phone calls, SMS, and alarms, notifications are becoming increasingly pervasive. Previous work largely focused on ensuring that these alerts will not interrupt users. In this section, we review previous findings on what factors are known to affect the success of notifications and other types of alerts.

Source & Content. Alerts can originate from different sources and deliver different contents. For mobile phone users, the most common type of alerts are those from computer-mediated communication applications [51]; in particular if the user (a) communicates with the sender [44, 50] or (b) is close to the sender [33]. Alerts from non-communication apps are received less favorably [27, 33, 34, 48], and their content appears to have little impact on the perceived timeliness of the interruption [14]. We see this as an opportunity to improve notification delivery policies of this type of notifications, as we explore in this paper.

Context. Since mobile phones are typically close to their users all day long [11], alerts can take place in all kinds of situations, even the most inappropriate ones [45]. One popular strategy to avoid the negative impact of inopportune interruptions is to schedule alerts for specific times of the day [4, 15, 20, 31, 52, 57]. Using timing information, such as the hour of the day, has been found to be a useful factor in some use cases, such as predicting attentiveness to messages [46] or the suitability of a moment for health interventions [52]. Recent work on boredom [47], engagement [32], and ritualistic phone use [17] found that people exhibit more stimulation-seeking behaviors in the evenings. However, other works did not find time to be a good predictor for engagement push notifications [31, 57]. In a study with 126,000 users of a shopping brochure app, Westermann *et al.* [57] found that solely relying on the time of the day for sending notifications with recommendations had virtually no impact on how fast people engaged with the recommended content. All in all, there is no conclusive evidence that engaging users during certain hours of the day is a promising strategy. Further research has studied the impact of the user's location to determine her or his receptiveness to alerts [33, 49, 50, 52]. For example, Sarker *et al.* [52] found that, for health intervention alerts, participants were less available at work and more available outside work. Mehrotra *et al.* [33], however, found that response times to mobile phone notifications in general do not significantly vary depending on the location (home, work, other). These results indicate that for most types of engagement, the user location is not an important factor. The use of sensors to estimate the state of the user's surroundings, such as noise or light sensors in mobile phones, has been another approach to estimate interruptibility of mobile phone users [33, 49]. Ambient noise was not found to be a significant factor to determine a user's responsiveness to notifications in general [33]. We did not find any conclusive results related to light sensor data reported in the literature.

Current Activity. The user's current activity may indicate openness to interruptions as well [18, 33, 50, 52]. Ho *et al.* [18] found evidence that messages delivered between physical activities can be received more positively. Interruptions can be less opportune during certain modes of transportation, such as biking [33], which can be inferred from the phone's motion and orientation sensors [49, 52]. Furthermore, interruptibility is negatively affected by concurrent tasks that are challenging, require concentration, or in which the user is not skilled [43]. Rote work, *i.e.*, phases of work with high engagement on tasks which are not challenging, are correlated with openness to interruptions [29]. The use of entertainment apps, a possible proxy for openness to interruptions and engagement, is negatively correlated with alertness and the use of productivity apps [36]. Related work appears

to reveal the following pattern: if mobile phone users are already engaged with demanding tasks, moments are inopportune. Concurrent tasks that are less engaging, such as riding the metro or doing rote work, may indicate opportune moments for engaging users.

Phone Status. The phone state itself can give important insights to interruptibility. For example, past work has used the light sensor to detect if the screen of the phone is covered [27, 46, 49], which happens, *e.g.*, when the phone is stowed away. This state usually correlates with lower probability that the mobile phone owner will promptly react to an alert. Regarding the phone's ringer mode, Pielot *et al.* found that notifications of all types are attended fastest when the phone is set to vibration mode [45], and that the ringer mode can be an important predictor of how fast people attend to messages [46]. Without any perceivable alert, users are less likely to immediately attend to notifications [7]. Conversely, Mashhadi *et al.* [31] found that the modality of the alert did not affect attending times of the participants of their study. Furthermore, only a small fraction of mobile phone users consciously manage openness to interruptions through, *e.g.*, notification settings [56]. Thus, evidence is not conclusive whether if people do not perceive the alert, they will be less likely to engage.

Patterns of Phone Use. Iqbal and Bailey [21] proposed to defer notifications when the user is busy by applying the concept of *bounded deferral* [19]: for a limited period of time, alerts are attempted to be delivered during automatically-detected breakpoints between (work) tasks. When this period of time has been exceeded, notifications are delivered regardless of the state of the user. In the context of office work and email, this concept has been implemented by monitoring the user's interaction with the computer, such as app switches or mouse movements [22, 23, 30]. In the context of mobile phones, waiting for the end of episodes of mobile interaction can be a simple, yet powerful, approach to identify breakpoints, as demonstrated by Fischer *et al.* [13]. Monitoring interaction events through Android's accessibility service allows to detect breakpoints during episodes of mobile phone interaction [38–40]. Beyond breakpoints, certain types of phone usage, such as *killing time*, have been found to indicate openness to interruptions, even if the user is not at a task breakpoint. For example, recently observed battery drain, number of unlocks, or number of apps launched correlates with increased feelings of boredom, which in turn is correlated with higher openness to consume entertaining news articles [47]. Thus, related work shows the monitoring patterns of phone use is a promising strategy to identify opportune moments for user engagement.

Communication. During meetings or in the presence of co-workers, people sometimes rated themselves less open to interruptions – depending on whether they are speaking or listening [15, 20, 27]. Similarly, Pejovic *et al.* [42] found that changes in the devices seen through Bluetooth –a proxy for the number of nearby people– correlate with the users' perceptions of when a moment is opportune for interruption. In contrast, Schulze and Groh [53] found that during some types of conversations, such as small talk, people are even more open to interruptions by notifications. The use of computer-mediated communication is another indicator regarding the openness to interruptions [32, 44, 47]: recent incoming calls have been correlated with un-opportune moments while recent outgoing calls are correlated with opportune moments to deliver notifications. In sum, previous findings indicate that light-weight communication, such as small talk, messaging, or (terminated) phone calls indicate opportune moments, while more engaged types of communication indicate inopportune moments.

Affect and Personality Traits. Sarker *et al.* [52] found that participants were more available to health interventions when they were happy or energetic versus when they were stressed. With respect to proactively-recommended content, Kushlev *et al.* [24] found that when feeling good, mobile phone users are less likely to engage with mentally demanding tasks. When feeling calm, users are more likely to engage with diverting tasks. When feeling energetic, user are less likely to respond to engagement attempts altogether. Participants who scored high in the Boredom Proneness personality trait (measured by the Boredom Proneness Scale [12]) were more likely to click on notifications that suggested to read articles [47] than participants with low scores in boredom proneness. A

small (n = 11) data set by Mehrotra *el al.* [34] suggests that extroversion and neuroticism might play a role in how fast people attend to mobile phone notifications in general. Thus, previous work indicates that emotions and personality affect the user's openness to engagement as well.

Engagement with Content. As mentioned, most previous works focus on whether it is a good moment to attract attention via a notification [13, 34, 43, 45, 55]. As argued by Turner *et al.* [55], there is a difference between being *reachable*, *i.e.*, attracting the user's attention, and being *receptive*, *i.e.*, consuming the notifications content. In this work, we focus on predicting whether beyond attracting the user's attention, the moment permits that the attention will be held, *i.e.*, that the user will engage. Fischer *el al.* [14] studied how *receptive* users are to SMS depending on the content. They report that "the factors interest, entertainment, relevance and actionability influence people's receptivity significantly", indicating that ultimately the content plays an important role in user engagement. In a recent study, Mathur *el al.* [32] showed that patterns of mobile phone use can estimate the level of engagement while the user is already using a mobile application, having electroencephalogram (ECG) metrics as ground truth. This can be very useful to learn how successful an attempt to engage users was, but only if users decide to engage with the suggested content. Hiniker *el al.* [17] implemented a classifier to distinguish whether the phone is used in a goal-oriented fashion or in a ritualistic fashion without a clear goal in mind – hypothesizing that the latter state would be useful for recommender systems. Future work needs to prove whether ritualistic phone use equates opportune moments to engage users. Okoshi *el al.* [40] tested a breakpoint-detection system to time notifications of Yahoo! JAPAN. On the basis of data from over 680,000 users, they show that response times to notifications can be significantly reduced (27.32 instead of 54.30 minutes mean response time). The breakpoint detection increased click-through rates and engagement scores, but the effect was not statistically significant. Pielot *el al.* [47] demonstrated that it is possible to predict from mobile phone data whether users are bored, and that when predicted bored, users are more likely to engage with a specific type of content, namely entertaining news articles, on their smartphone. In this work, we advance the state of the art by focussing on predicting the likelihood that users will engage with a wide range of content a-priori to the engagement attempt, independent of how the phone is used, and independent of the current mood of the user.

3 GOAL AND HYPOTHESIS

The primary goal of this work is to enable the development of a mobile phone-based, intelligent service that uses a machine-learning component to estimate in real time when it is a good moment to engage with the user. *Engagement* can mean different things, depending on the product or service. Many content-based services, such as blogs, social networks, or news portals, offer their content for free and create revenue by exposing its users to advertisement: the more time users spent "on site", the higher the monetization. Other services, such as games, monetize by selling special upgrades or perks. These upsells are more likely to occur in engaged, frequently-visiting players. In other cases, such as apps from mobile phone providers, engagement can mean to remind customers to top-up the account before the balance reaches zero – which can equate to a loss in revenue.

The common factor in these examples is that they require the time and the openness of users to engage with the service. Therefore, simply finding a good moment where a user is *reachable* [55] by a notification or another type of alert, as covered extensively by previous work, is insufficient. Consequently, the intelligent service we envision needs to estimate whether users are in a situation where they are likely to actually engage with the suggested content.

The aim of this study is to explore, in a systematic way, to which extent it is possible to predict openness for engagement from data that can be collected by a mobile phone application, such as phone status (*e.g.* whether the screen is on or off), sensor data (*e.g.* the current location), and other types of information (*e.g.* the current time).

Our key hypothesis is that the inference of the probability to engage with content can be done *independently* of the actual content. This is important to enable the design of a smart notification system that would be useful

to a variety of products and services and a wide range of recommended contents. We hypothesize that we can accurately predict the probability that a mobile user will engage with recommended content from the data collected from a rich set of phone sensors. Note that we do not make any hypothesis about the types of features and patterns that would perform well. Hence, we compute a rich set of features and carry out state-of-the-art feature selection.

In the next section, we describe our methodology, driven by our goal and hypothesis.

4 METHODOLOGY

To systematically explore which type of information available in mobile phone applications indicates good moments to engage with users, we conducted a field study, which can be summarized as follows:

- For an average duration of 4 weeks, 337 participants installed a study application onto their primary mobile phones (Android) and kept it running in the background.
- The application created about 10 to 15 notifications per day. Such notifications led to a mood questionnaire with 4 Likert-scale items that could be answered in less than 10 seconds. We collected 30,689 responses to the questionnaire.
- At the bottom of the questionnaire, two types of content were offered to the users, randomly chosen from a pool of eight different types of content, corresponding to different services (*e.g.*, games, news, videos, etc.). In 3,367 cases, participants engaged with the offered content.
- While participants believed that the mood questionnaire was the main purpose of the study, what we were interested in was their engagement with the recommended content. Hence, we logged whether instead of simply closing the questionnaire, they opened one of the two suggested contents, which we refer to as *engaging* with the recommended content.
- At the same time, the application logged phone usage patterns from a wide range of sensors and other information sources.

In the following section, we explain the details of the methodology as well as the rationale behind our design choices.

4.1 Study Design

Different products and services may have different definitions for what it means to engage with their users. Thus, the best-performing, data-driven, intelligent algorithm would likely be one which is trained on data and observations from users of their target product or service. However, this requires to adapt the algorithm to each new product individually. This implies a cold-start period during which proactive recommendations are made using non-optimal decision-making criteria (*e.g.*, at random or at certain specific times), potentially harming the user experience. Depending on how frequently the product owner is willing to collect sample data (*e.g.*, by sending notifications), it may take weeks or months until the system has collected a sufficient amount of data to build a robust model. Hence, we decided to make the study as independent from the content as possible, so that its findings can be generalized more easily. To achieve this goal, we adopted the following three design decisions.

First, drawing on the services that we envision to serve, we exposed participants to content from a wide range of different products and services, namely games, articles, multimedia, entertainment, questionnaires, and stores. In total, there were 8 different options of content to engage with (Fig. 1):

- An action video game, randomly selected out of a curated set of three action games (from Silvergames.com).
- A puzzle video game, randomly selected out of a curated set of three puzzle games (from Silvergames.com).
- A curiosity/cultural article (the Wikipedia's "Today's Featured Article" web page).
- An assortment of daily news (Yahoo's "Latest News and Headlines" web page).
- A randomly selected funny fact (from Unkno.com).

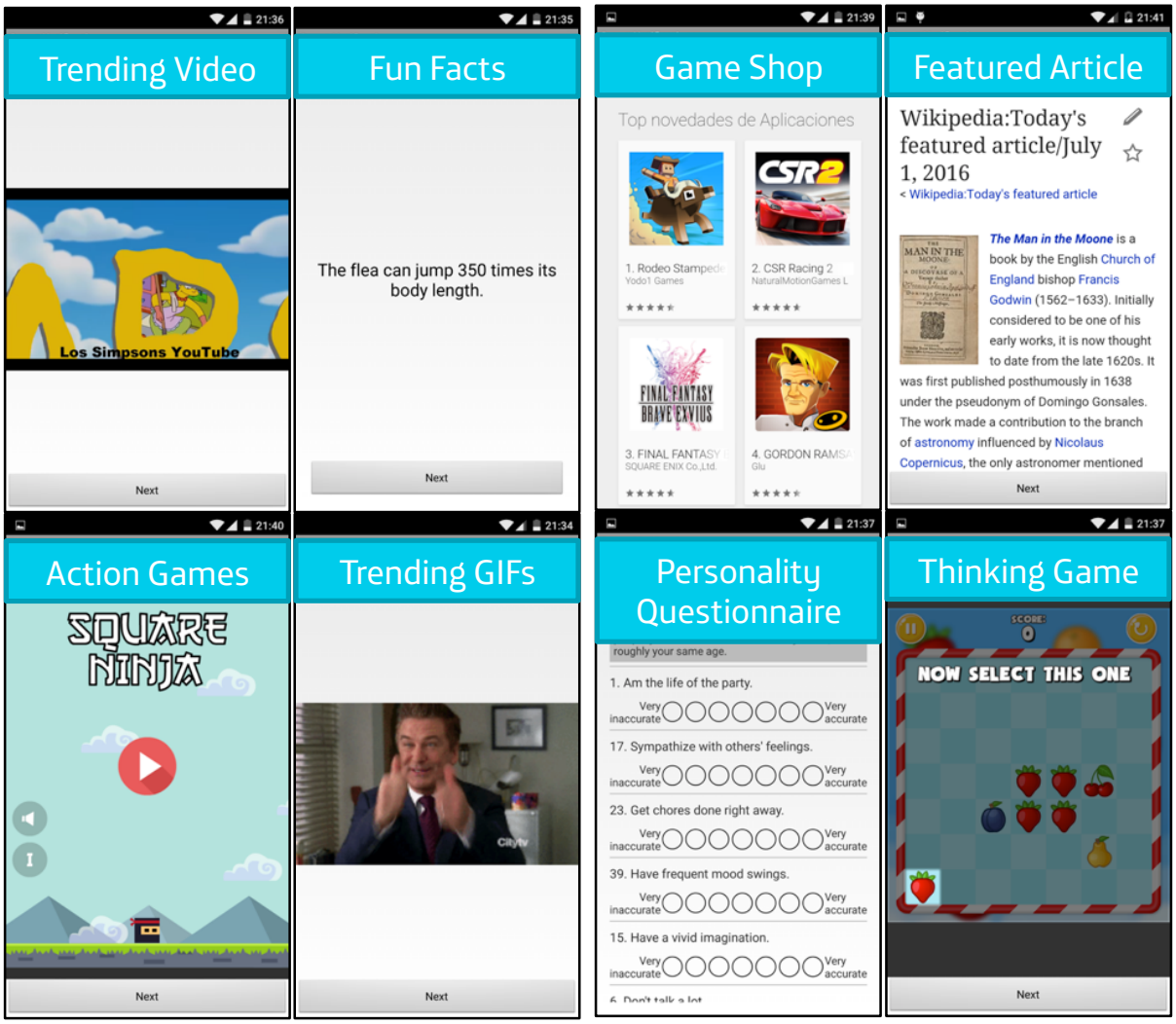

Fig. 1. Types of content offered to participants. Participants had two randomly chosen options to choose from.

- A trending animation (randomly selected from Giphy's database of trending GIFs)
- A trending video (randomly selected from YouTube's database of trending videos)
- A psychometric questionnaire (non-random, sequentially selected from the following list: the Big Five Personality Test, the Personal Health Questionnaire Depression Scale (PHQ-8), the Boredom Susceptibility Scale (SSS-BS), the Multidimensional State Boredom Scale (MSBS) and the Self-Assessment Manikin (SAM). Once all of the questionnaires were answered, the application started to cycle between the last two questionnaires, MSBS and SAM, which perform state assessments.

Second, each time the participants were exposed to the questionnaire, we only offered two options of content to engage with at a time. These options were chosen randomly. The rationale here was to dissociate the participant's

reaction from the content itself. If the same content was to be made available every time, this would allow participants to develop a stable attitude towards some type of content, such as becoming a devotee of a particular game. In this scenario, engagement would not necessarily have indicated an appropriate moment but would rather have been a reflection of the participants' attitude towards such content.

Third, to further dissociate the participants' reactions from the content, we attempted to ensure that the actual content was different every time the user engaged with a specific category. For example, participants would never encounter the same game or the same fun fact in a row. Instead, where possible, content was cycled through every time that it appeared as option in the questionnaire. Thus, prior to clicking on one of the content buttons, participants did not know exactly what content they would receive.

4.2 Proactive Content Suggestion

We employed the experience-sampling method [26] to explore openness to engagement throughout various moments of the day. As trigger, we used notifications generated by our study application (Fig. 2). These notifications were posted semi-randomly throughout the day (~10–15 notifications per day). Prior to triggering the notification, the application made sure that all information sources were updated. Most notably, we started collecting data from all sensors 30 seconds before sending the notification in order to capture the participants' context.

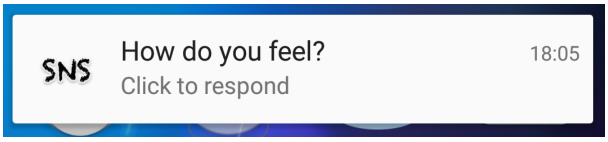

Fig. 2. Experience-Sampling Notification

If the participants did not respond to a notification within 10 minutes, it was removed from the notifications tray. In this case, we consider the moment of posting the notification as an inopportune moment for engagement. This time threshold was chosen since the majority of notifications are attended to within this time frame [45], and since previous work by Mehrotra *et al.* considered notifications as unattended if they were not clicked within this timespan [34].

4.3 Decoupling Content via an Intermediate Questionnaire

Upon clicking on the notification, participants were taken to the short questionnaire shown in Figure 3. The rationale for presenting a questionnaire instead of the actual content was twofold.

First, the questionnaire served as deception regarding the true purpose of the study. The experience sampling notification asked the participant to report "How do you feel?". The questionnaire contained four items about the participants' emotional state and took less than 10 seconds to complete. A fifth item asked the participants to select one of three buttons, and served as a means to keep our data clean from random responses (Fig. 3). The informed consent showed participants an illustration, explaining that we only expected responses to these questions. The suggested content was visually marked as optional, explaining that it only served as a gimmick and could be ignored. Past responses would be visualized on the main screen of the study application so that participants could see an overview of their history of emotional states. The recommended content was suggested through two buttons towards the bottom of the questionnaire (Fig. 3). The participants had been clearly and repeatedly informed that engaging with the suggested content was completely voluntary. Therefore, if participants engaged with the suggested content, we can assume that they were doing it out of free will and because their current context represented a good moment for doing so.

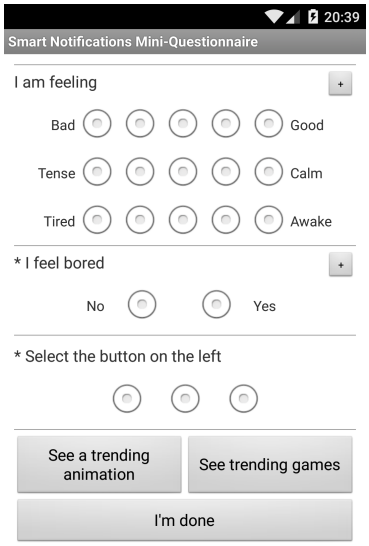

Fig. 3. Experience-Sampling Notification

The second rationale for introducing an intermediate questionnaire was to decouple the incentive mechanism from the observed target behavior, as there is an inherent challenge of incentivizing *voluntary* engagement with suggested content. If we had paid the incentive simply for joining the study by installing the study application, we could not have guaranteed exposure to the content suggestions. Participants could have simply disabled notifications altogether. If the incentive had been tied to a minimum number of engagements with suggested content, we would not have been able to distinguish between an opportune moment for engagement and engagement for the sake of collecting the incentive. Instead, participants were paid in full when they had accumulated 21 days during which they responded at least twice to the questionnaire, independently of whether they had opened any of the content suggestions.

4.4 Logging Patterns of Mobile Phone Use

One of the key hypotheses of our work is that certain patterns of phone use co-occur with opportune moments for engagement. Therefore, monitoring patterns of phone use can infer whether a given moment is opportune or not. In our intelligent system, we characterize mobile phone usage patterns by the information available to the app through mobile sensors or other information sources.

A priori, it is difficult to know which sensors and other sources of mobile phone usage are predictive of our target variable. As seen in the Related Work Section, the predictiveness of features can differ depending on the use case. An ideal approach for a product or service would thus be to conduct a learning phase, where all available sensors and other sources of information are collected, and then use state-of-the-art machine learning algorithms to identify which of these information sources are most predictive. However, this is impractical for two reasons: first, some sensors have significant resource requirements and cause the battery to deplete significantly faster. Second, some sensors and other information sources require special permissions, because they access

personally-identifiable information. Users have become quite sensitive to both battery consumption and required permissions. Product owners may not be willing to risk alienating the existing user base in order to follow this ideal approach. With this study we propose a more practical approach: establish the prediction power of available information sources in a general context. Once the prediction power is established, the value of each information source can be weighted against the added battery drain and the permission it requires when implementing a product-specific intelligent system.

To enable such informed decisions, the application built for this study collects a wide range of information obtainable through Android OS. Some examples are the user location, the foreground app, notifications, and screen events. Android OS was chosen since it provides access to a significantly larger number of information sources than iOS. To keep the overall energy consumption reasonable, we used three strategies to collect data:

(1) System-wide broadcasted events, such as battery level, were collected always by registering to the respective broadcast receivers and callbacks.
(2) Data from energy-intensive sensors, such as location, was collected every 10 minutes for 30 seconds.
(3) Data related to the user's interaction with the phone, such as the app in foreground, was only collected while the screen was unlocked.

With these strategies, we limited the battery consumption to a level that the study application was shown in the OS internal battery usage view as consuming about the same amount of battery as other popular apps, such as Facebook or WhatsApp.

Since the application was part of a dedicated study, we were also in a position to ask participants to grant two special permissions: access to accessibility events and notification events. To get access to these events, users currently have to visit a dedicated view in the settings and manually enable them for the requesting application. Since accessibility and notification events can contain potentially sensitive information, the informed consent contained a section specifically dedicated to those events. During the setup process, the application itself explained the participants how to grant these two special permissions and automatically sent them to the correct settings view.

4.5 Recruitment

The goal of the study was to obtain data from a representative sample. Hence, we recruited participants through a specialized agency. We requested a sample that matches the gender and age distribution of the country of study in Western Europe. The only restriction was that people were required to own an Android phone. Android phones account for the large majority (~ 90%) of the smartphone share in the country of study.

Over 500 people joined the study. However, we only consider data from 337 participants who participated for at least 10 days and gave at least 20 valid responses to the questionnaire. The participants' ages ranged from 18 to 66 years (M = 37.85, SD = 11.01), and the gender split we obtained was balanced (52.8% female, 47.2% male). The mean number of active participation days was 27.43 (Mdn = 27, SD = 11.49).

4.6 Procedure

People with interest in joining the study were first directed to the informed consent, which had been approved by the legal department of our institution. The consent form listed all data to be collected in the study and gave extra details about potential personally-identifiable information. The participants then were taken to an installation guide that explained how to install the mobile application. We ran informal usability tests to ensure that the installation process was fast and easy to understand. The data collection commenced once the app was installed, set-up properly, and once the participants confirmed their agreement with the informed consent from within the app.

To receive their compensation, participants were required to accumulate 21 active participation days before a fixed end date. Each day with at least 2 complete responses to the notification-administered questionnaire was considered to be a day of active participation. Through a visual annotation of a screenshot of the questionnaire, we emphasized that the engagement with the suggested content was completely voluntary and not required to receive the compensation. Hence, our incentive only ensured that participants got exposed to the recommended content, but it did not require them to engage with it. We informed participants that the app would typically notify them about 10–15 times a day. The rationale was to minimize the pressure to interact with each and every notification, and allow participants to ignore notifications during moments that were inopportune to fill out the questionnaire.

5 DATA ANALYSIS

For the analysis, we only consider data that was collected between the second and the last response to the questionnaires. This way, we exclude all data from situations where participants had already terminated or abandoned the study without uninstalling the app. We further exclude the first response to reduce bias from the phase during which participants were still familiarizing themselves with the application. The resulting data set contains over 120 million phone usage events and 78,930 notifications.

5.1 Target Variable

Our target or regressor is a binary variable with two values: {1} when the user opened the questionnaire and engaged with one of the two suggested contents by clicking on it, and {0} when the user either did not open the questionnaire or opened it but ignored the suggested content. Out of the 78,930 notifications, 30,689 (38.9%) were clicked, *i.e.*, the participants opened the questionnaire. In 3,367 of those 78,930 otifications (4.3%), participants further opened one of the recommended pieces of content, *i.e.*, they *engaged* with the content as explained in Section 3. In this case, we assume that the moment when the notification was posted represents an opportune moment for engagement. While the fraction of positive instances is comparably low, we observed a sufficient number of positive instances for training a machine learning algorithm and performing an in-depth analysis thanks to the large dataset.

5.2 Feature Extraction

Because we had no *a priori* hypothesis about which types of features would perform well, we computed a rich set of features from the available mobile phone use data and later employed a model with implicit feature selection. Our goal was to characterize the moment (and indirectly the context) before the experience-sampling notification was posted. Inspired by Choy *et al.* [10], we computed features corresponding to three different time windows: the *current moment* (*e.g.*, current screen status or number of unlocks in the last 5 minutes), *recent* (*e.g.*, fraction of time screen was on in the last hour), and *current day* (*e.g.*, fraction of time screen was on since 5 am today). For each of these time windows, and in line with related work, we compute 197 features that belong to 5 different groups of variables:

(1) *Communication Activity* contains 37 features related to computer-mediated communication. This group includes features that show how often a user is using the phone to communicate with others by, *e.g.*, sending or receiving messages, or making or replying to phone calls. For instance, a user that just got distracted by an incoming phone call might not be open to further interruptions. Examples of Communication Activity features are: number of SMS messages received in the past hour, time since the last incoming phone call, or category of the app that created the last notification.

(2) *Context* comprises 73 features related to the situation of the mobile phone user, *i.e.*, his or her environmental context. The context of use often determines whether it is appropriate or safe to interact with the mobile

phone. For instance, being at home during the weekend may indicate opportune moments for interruption, whereas being at work during the morning may indicate the opposite. Examples of Context features are: time of day, estimated current distance from home, recent levels of motion activity, or average ambient noise level during the last five minutes.

(3) *Demographics* refers to the age and gender of the user. These features can be important to determine openness to interruptions, as one can assume, for example, that adults tend to have significantly less personal time available to them when compared to other age ranges due to family or employment responsibilities. Demographics are an exception to the other groups, since they cannot directly be observed from the phone. However, they are comparably easy to obtain or infer [3].

(4) *Phone Status* includes 13 features related to the status of the mobile phone. For instance, a device with screen status 'unlocked' indicates that the user is currently using the phone, thus a notification might be interrupting a concurrent task. Examples of Phone Status features are: the current ringer mode, the charging state of the battery, or current screen status (off, on, unlocked).

(5) *Usage Patterns* spans 72 features that relate to the type and intensity of usage of the phone. For instance, a user engaged in playing a game or watching a video may be less open to an interruption, whereas surfing on the Internet might be a better moment. Examples of Phone Usage features are: number of apps launched in the 10 minutes prior to the notification, average data usage of the current day, battery drain levels in the last hour, number of device unlocks, screen orientation changes, or number of photos taken during the day.

Since our focus is not on predicting engagement related to computed-mediated communication, we did not compute any features regarding the sender-receiver relationship [33, 34]. In addition, we did not compute features related to the content of the notification [13, 33], since our aim was to create a model that would be content-independent. The feature extraction resulted into a table with 78,930 instances (one instance per notification). Each instance contained a user ID, the features, and the ground truth. This table served as input to the subsequent analyses steps.

5.3 Model Choice

To model the likelihood that a participant will open and engage with the recommended content, we use a machine learning approach [16, 35]. Machine learning models are able to leverage complex interactions (both linear and non-linear) between the available features and the ground truth target variable. Through the appropriate choice of our learning model, we are able to take such complex interactions into account when assessing feature importance (see below).

As learning model we used XGBoost [9]. XGBoost is a state-of-the-art gradient boosting regression tree algorithm that has successfully been used in several application domains. It is fast, scales beyond billions of examples, and yields state-of-the-art accuracies in standard classification benchmarks and challenges [9]. It belongs to the family of ensemble trees, sharing desirable properties with them, such as improved generalizability [16], robustness to different feature scales and distributions [6], and a principled methodology to deal with large numbers of features and assess their importance [5]. In pre-analysis, we saw that the default parameters yielded good performance, as compared to other state-of-the-art classifiers such as logistic regression [16] ($Precision$: 0.060, $F1Score$: 0.109) and Random Forests [5] ($Precision$: 0.061, $F1Score$: 0.108). We used sklearn [41] version 0.17.1 and its wrapper for the XGBoost Python package[2], version 0.4a30.

[2]http://xgboost.readthedocs.io

5.4 Evaluation Procedure

For each evaluation, we use a 10-fold cross-validation schema [16] to split the data set into training and test sets, ensuring that a participant does not get split across folds. Hence, the data from each participant is never used in both the training and testing sets and the classifier cannot exploit user-specific features or behaviors to improve its predictions.

As primary evaluation metric we use the average F1 score over the test folds [35]. In our case, precision measures the percentage of time that the model accurately predicts that the user will click on the recommended content from all the instances where participants clicked on the content. Recall measures the percentage of positive instances that were captured by our model. The F1 score is the harmonic mean of precision and recall [35].

Given the rationale to improve the success rate of engagement attempts, precision, at first glance, would be the most important metric: the higher the precision, the more accurately our model would predict when the user will engage with one of the suggested content items. For products, increased precision may directly translate into higher engagement and revenue. However, optimizing for precision typically lowers recall, which in our case can lead to an algorithm which cannot identify enough opportune moments anymore. Thus, it is important to also consider recall in the evaluation process. Depending on the application, the algorithm might not be given a lot of time to find opportune moments. In such a case, the algorithm might arrive to the end of the given time frame without having found an opportune moment if recall is too low. As we do not have an informed criterion on the relative importance of precision and recall without knowing the details of the target product or service, we used the F1 score as primary evaluation metric, balancing both precision and recall.

Another factor we had to consider was the different levels of participation and engagement across all participants: some participants just opened the minimum number of notifications ($n = 20$), while other participants provided data for more than 200 notifications each. If we learned a model without any further consideration on the number of sent notifications per participant, our model could be biased towards participants who received higher volumes of notifications. Hence, we normalized all observations per participant and class, so that their sum would be one. We apply these weights in training, by multiplying the classifier loss accordingly [16].

5.5 Classifier Tuning

XGBoost typically requires little effort to tune for a good performance. In pre-analysis, we saw that the only two parameters that had a non-negligible effect were `scale_pos_weight`, which is used to penalize misclassifications of one of the classes, and `max_depth`, which defines the maximum depth of the trees. To select the appropriate values for these two parameters, we did a grid search.

For `scale_pos_weight`, we tested the values [0.75, 0.9, 1.0, 1.1, 1.25] as multipliers for the expected presence of the weighted positive ground truth. For `max_depth`, we tested [3, 4, 5, 6] tree splits. The grid search iterated over all possible combinations of the two sets, computing their performance on the training set via 5-fold cross-validation.

Note that, by optimizing `scale_pos_weight`, we are learning and tuning the importance of the positive class in our problem in order to maximize performance. The optimal performance was achieved with a `scale_pos_weight` of 1.1 and a `max_depth` of 3.

5.6 Baseline Selection

Regarding the baseline, we compared the performance of different strategies of the `DummyClassifier` available in scikit-learn:

- *constant*, *i.e.*, setting all predictions to {1};
- *stratified*, *i.e.*, where predictions are generated randomly by respecting the probability of each class; and
- *uniform*, *i.e.*, where predictions are generated uniformly at random.

These strategies approximate the status quo, where product/service owners do not typically use real-time intelligence to determine the timing of sending the notifications. The strategies performed almost equally well, achieving a precision between 4.3% and 4.4%. Because of the higher recall value, the *constant* strategy was found to yield a higher F1 score than other strategies. Considering a precision of 4.3% and a recall of 100%, the baseline for the F1 score equates to 0.082. In our results, besides reporting precision, F1 scores, and confusion matrices, we compute the lift as the percentage of relative increase between the performance of the learned classifiers and this baseline.

5.7 Feature Preprocessing and Cleaning

We preprocessed the features under consideration before they were provided as input to the classifier. Feature cleaning is a standard procedure in machine learning to, for instance, deal with missing values or to convert categorical variables into numerical values. In our case, categorical variables such as ringer mode (one value from {"normal", "vibrate", "silent"}) were converted to numerical variables using the so-called one-hot encoding strategy [35]. We also found some missing values in our data due to temporally unresponsive resources or the fact that a feature represented the time since an action that had not yet been observed. In some cases, we were able to infer the value of the missing variable. For example, when launching the study app for the first time, the screen can be assumed to be on. For the rest of the cases, we performed distribution-based imputation by randomly sampling from the available variable values [16].

6 RESULTS

First, we report the performance of an XGBoost classifier built using all the 197 features from the previously described 5 groups. The classifier achieves an F1 score of 0.113. Table 1 shows the corresponding confusion matrix.

	Predict No	**Predict Yes**
No Engagement	63,304	12,259
Engagement	2,429	938

Table 1. Confusion matrix using all features as input.

If we had used the output of our algorithm to better time when to send notifications, 938 of 13,197 attempts would have led to conversions ($Precision$: 0.071), which constitutes of lift of 66.6% over the baseline precision of 0.043. If used to capture as many opportune moments as possible while reducing the number of engagement attempts, 938 of 3,367 total conversions would have been captured ($Recall$: 0.279) while only making 13,197 (16.7%) of the 78,930 engagement attempts. To ensure that the classifier would still be able to predict a sufficient number of opportune moments for each user, we investigated the fraction of positive predictions for each participants individually. None of the participants had fewer than 1% positive predictions and only 82 participants had fewer than 10% positive predictions. In summary, these results indicate that the algorithm can significantly increase the conversion rate of engagement attempts, but it does not cut off users from ever receiving notifications –which is important for this use case, where we try to find the sweet spot between maximizing engagement and minimizing disruption.

6.1 Adapting Positive-Prediction Frequencies to the User

We now turn our attention to the participants' conversion behavior during the study. Figure 4 visualizes the number of positive predictions (made by the classifier that used all features) for each user, alongside the number

of their actual conversions. We see that a notable fraction of the participants rarely or never clicked on any of the recommended content. In particular, 80 participants never engaged with any of the suggested contents and 68 participants engaged with suggested content only once. This phenomenon can be expected in real scenarios too, since some users may frequently ignore proactive recommendations from certain products and services. The classifier, however, does not appear to adapt to this. In the context of some products and services, it might be feasible and beneficial to reduce the number of notifications sent to participants who rarely or never convert, to avoid churn, such as uninstalls of the corresponding app, and/or to increase satisfaction. In this section, we analyze the value of taking the participants' past behavior regarding conversions into account when predicting future conversions.

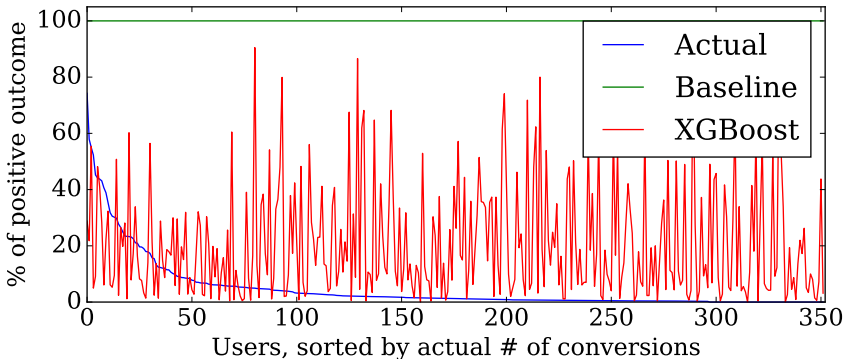

Fig. 4. Actual conversions (blue) overlaid by positive prediction (red). The figure visualizes that there is little adjustment by the model to the general amount of conversions of the participants.

To this end, we compute a new group of features called *PastActions* which models the participants' past interactions with the suggested content. The *PastActions* group consists of 4 features: (1) the *mean conversion rate* per participant, *i.e.*, the number of conversions divided by the number of notifications sent until this point, prior to each notification. This feature approximates the participant's general openness to the content that the app recommended; (2) the *response* to the last instance of recommended content; (3) a *slow rolling mean* –via exponential smoothing with exponent of 0.05– of the participant's conversions to capture the participant's behavior in the past few days; and (4) a *fast rolling mean* –with exponent of 0.2– to capture the participant's response to the last few notifications. The last three features in this group model the participant's recent reactions to the recommended content in order to take into account that his/her interest in the suggested content might vary over time.

We then built an XGBoost classifier adding the 4 *PastActions* features to the 197 features in the previously described groups. Again, we used grid search to find the optimal parameters. The best performance was achieved with a `scale_pos_weight` of 1.0 and a `max_depth` of 5. The performance of the resulting classifier was improved an order of magnitude, achieving an F1 score of 0.311. Table 2 shows the confusion matrix.

	Predict No	**Predict Yes**
No Engagement	69.051	6,512
Engagement	1,548	1,819

Table 2. Confusion matrix, including the *PastActions* feature group.

If we had used the output of this algorithm to better time attempts to engage users, 1,819 of 8,331 attempts would have led to conversions (*Precision* : 0.218), which constitutes over a 5-fold lift over the baseline precision of 0.043. If used to capture as many opportune moments as possible while reducing the number of engagement attempts, 1,819 of 3,367 total conversions would have been captured (*Recall* : 0.540), while only making 8,331 (10.6%) of the 78,930 engagement attempts.

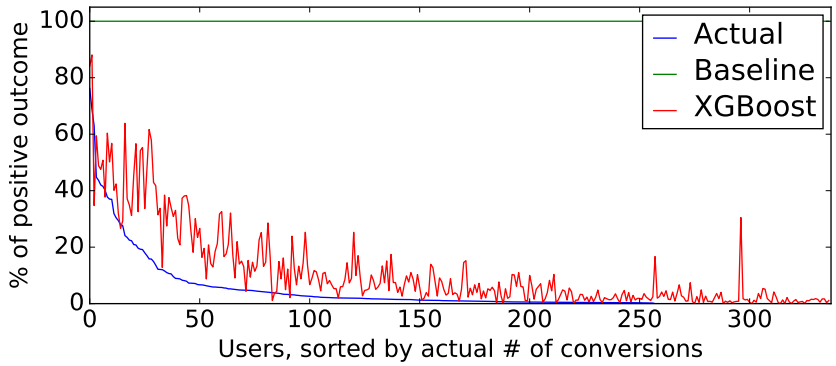

Fig. 5. Actual conversions (blue) overlaid by positive prediction (red). The number positive predictions of the optimized model is much closer to the number of conversions of the respective participants.

As shown in Figure 5, this new classifier adjusts the number of positive predictions much better to the number of actual conversions. To again ensure that the classifier would still be able to predict a sufficient number of opportune moments for each user, we investigated the fraction of positive predictions for each participant individually. 45 of the participants had fewer than 1% positive predictions and 225 participants had fewer than 10% positive predictions. This shows that if the algorithm would be deployed in this form, a fraction (in our case, about 15% of the participants) would almost never be predicted to be open for engagement. Thus, adding features reflecting past openness for suggested content may result in a classifier that for some fraction of the user base never predicts any opportune moments. Depending on the use case, this may or may not be a desirable property.

6.2 Feature Importance

In order to shed light on the role that different factors play to predict conversions to recommended content, we investigate the top-30 features ranked by their contribution to the model as reported by the XGBoost classifier (see Table 3). Below, we report significant correlations between the features and the ground truth. To compute correlations, we used Spearman's Rank Correlation, as most feature values were not normally distributed. Note that the importance of a feature does not necessarily imply a direct correlation with the ground truth. Given the non-linearity of the model, a variable may serve as mediator and gain prediction power only through its combination with other variables.

Communication. Seven features related to recent communication activity are amongst the top predictors. Higher conversion rates directly correlate with the more time had passed since the last outgoing phone call ($r = 0.017, p < 0.001$) or the last incoming phone call ($r = 0.008, p < 0.05$), the less time had passed since the last SMS was received ($r = -0.011, p < 0.01$), and the lower the volume of notifications received so far during that day ($r = -.013, p < 0.001$).

Importance	Feature
0.0365	Usage_Ringer_MinSinceChanged
0.0350	Context_HourOfDay
0.0274	Comm_PhoneCall_PickedUp_MinSinceLast
0.0259	Context_Noise_Last
0.0259	Comm_PhoneCall_Outgoing_MinSinceLast
0.0244	Comm_PhoneCall_MinSinceLast
0.0228	Comm_Notif_Posted_MinSinceLast
0.0228	Comm_Sms_Received_MinSinceLast
0.0198	Context_MidpointOfNight_HoursSinceLast
0.0198	Comm_PhoneCall_Incoming_MinSinceLast
0.0183	Usage_App_MinSinceLast
0.0183	Usage_NotifCenter_MinSinceLast
0.0183	Context_SemLoc_Work_Distance
0.0167	Demog_Age
0.0167	Comm_Notif_Posted_D_Count
0.0167	Usage_Screen_MinSinceChanged
0.0152	Usage_Data_Tx_D_Sum
0.0152	Context_Acc_Avg_60_Mad
0.0137	Usage_Screen_Unlocked_MinSinceLast
0.0137	Usage_BattDrainB_D_Mad
0.0122	Context_Acc_Max_60_Q50
0.0122	Usage_BattDrainB_D_Q50
0.0122	Context_Noise_D_Q50
0.0122	Context_MidpointOfNightHour
0.0122	Context_SemLoc_D_Percentage_Home
0.0122	Context_Loc_D_Sum_Distance
0.0122	Context_Light_D_Mad
0.0122	Phone_BattB_Last
0.0122	Context_UserAct_P_Last
0.0122	Context_Noise_60_Q50

Table 3. Top-30 most important features as reported by the XGBoost classifier. The features are named systematically: the first string (*e.g. Context* indicates the group that this feature belongs to; the second string (*e.g.*, *Noise*) indicates the sensor from which the feature was computed; the strings *Last*, *60*, and *D* indicate the time horizon that this feature models, namely the *current* moment, the last hour, or the current day (since 5am); the final suffixes indicate the type of feature, such as last-observed value (*Last*), minutes since the last occurrence (*MinSinceLast*), or lower quartile of the observed values in the five time span (*Q25*).

Context. Of the 30 most-predictive features, 13 belong to the *Context* group. Higher conversion rates correlate with higher the fraction of time spent at home during the day so far ($r = 0.034$, $p < 0.001$), increased distance to work ($r = 0.022$, $p < 0.001$), and larger distances traveled during the day so far ($r = 0.010$, $p < 0.01$), Regarding motion activity, higher conversion rates correlate with higher median of spikes in the level of physical activity during the last 60 minutes ($r = 0.017$, $p < 0.001$) and higher variance in the level of physical activity during the

last 60 minutes, as reported by the acceleration sensor ($r = 0.034$, $p < 0.001$). Regarding ambient noise and light levels, higher conversion rates correlate with higher noise levels ($r = 0.013$, $p < 0.001$), higher median levels of noise during the day so far ($r = 0.025$, $p < 0.001$), higher median levels of noise during the last hour ($r = 0.016$, $p < 0.001$), and less variance in the ambient light level recorded by the phone so far ($r = -0.043$, $p < 0.01$).

With respect to the time of the day, the later the notification was posted during the day, the higher the conversion rates ($r = 0.011$, $p < 0.01$). An inspection of the histograms revealed that conversion rates are largely the same, and only lower during the early hours of the day (7-8 am). In addition, inspired by [36], we computed the time since the midpoint of the night, that is, when a device was unused between 11pm and 7am, we consider 3am to be the midpoint of the night. We found a positive correlation between the time since the midpoint of the night and the openness for engagement ($r = 0.008$, $p < 0.05$). An inspection of the histograms revealed a peak in openness to engagement from 6 to 7 hours after the midpoint of night. Furthermore, the earlier the midpoint of the night, the higher the conversion probability ($r = 0.033$, $p < 0.001$).

Finally, higher conversion rates correlate with lower certainty values reported by Google's Recognition API with respect to the estimate activity prior to posting an experience-sampling notification ($r = -0.024$, $p < 0.01$). We hypothesize that activities, such as driving, cycling, walking, are in general more difficult to predict with high certainty than the phone being still. Hence, we assume that the presence of difficult-to-predict activities correlate with higher openness for engagement.

Demographics. Age turned out to be amongst the top predictors. We found a positive correlation between age and conversion rates ($r = 0.044$, $p < 0.001$), *i.e.*, the older the participant, the more likely to click on the recommended content.

Phone State. One feature describing the status of the phone was amongst the top predictors: participants were more likely to convert the higher the remaining battery charge was ($r = 0.008$, $p < 0.05$).

Usage Patterns. Eight features related to usage of the phone were amongst the top-30 features. Participants were more likely to convert the less time had passed: since the last launch of an app on their phone ($r = -0.065$, $p < 0.001$), since the last access the notification center ($r = -0.035$, $p < 0.001$), since the last change of the screen status (turning the screen on, off, or unlocking it), ($r = -0.063$, $p < 0.001$), and since the last unlock of the screen ($r = -0.069$, $p < 0.001$). Furthermore, participants were more likely to convert the more data the phone had transmitted during the day so far ($r = 0.009$, $p < 0.05$), the higher the variance of the battery drain had been during the day so far ($r = 0.008$, $p < 0.05$), and the higher the median battery drain had been during the day so far ($r = 0.015$, $p < 0.001$).

7 DISCUSSION

The study results show that features from all five groups of variables (communication activity, user context, demographics, the state of the phone, and user activity) were useful for predicting whether a participant would engage with suggested content. In the following, we discuss the most predictive features and the feasibility to compute them on the two major smartphone operating systems: Android and iOS.

Communication. Features related to communication activity were amongst the lower half of the most predictive features. The analysis of the individual features from this group reveals a direct correlation between engagement and incoming phone calls & incoming notifications. This indicates that our participants were more likely to engage when there was less incoming communication. This finding corroborates previous work by Pielot *et al.* [47], who found that boredom –a state with desires for stimuli– correlates with less incoming communication as well. Since most of these features are computed from events obtained through callback functions, the battery impact is negligible. Accessing them, however, is challenging. On Android, special permissions are required to access phone call logs, SMS logs, and notification events. On iOS, this is not possible at all. Thus, this feature

group makes most sense for communication products and services, where communication activity can be inferred through the product itself.

Context - Hour of the Day. Features related to the user's context were the most important to predict engagement. Almost half of the top-30 features belong to this group. With respect to the time, we found that the hour of the day was a good predictor of openness for engagement: caused by a significant drop of openness to engagement in the morning hours (7-8am). Otherwise, we found that the time of the day had relatively little influence, which supports the finding by Westermann *et al.* [57], where the timing of advertisement notifications had no statistically significant effect. Furthermore, we found that the users biological clock, represented by the time (in hours) since the midpoint of the night was a strong predictor: the earlier the midpoint of the night, and the more time had passed since said midpoint, the higher the openness for engagement. This is in line with findings by Murnane *et al.* [36] where the use of entertainment apps was higher compared to the use of productivity apps during the "evening" in relation to the personal biorhythm. The hour of the day is usually available through the system clock of any OS and does not require any special permission.

Context - Location. With respect to location, analyses of the direct correlations between location-derived features and the ground truth variable revealed that conversions were more likely to take place at home or when traveling compared to being close to work. This corroborates findings by Sarker *et al.* [52], where participants were less open to health intervention alerts at work. The use of semantic location as a feature is a borderline decision. If implemented via location sensors, it requires permission on iOS and Android. Users may react cautiously if an application asks for location without having a clear need for it. Alternative implementations can be found that, *e.g.*, can approximate important places (work, home) from the WiFi stations or cell towers the phones see nearby, which is possible to implement at least on Android with possibly less controversial permissions.

Context - Acceleration Levels. Higher levels of motion during the last hour, as measured by the phone's accelerometer, were positively correlated with openness for engagement. Previous work only reported findings related to the concurrent physical motion activity, showing that it correlates with less opportune moments [33, 49, 52]. Given these previous findings and our finding that the most predictive motion-related features do not indicate activity at the moment of posting the notification, but during the last hour, motion-related features might allow to capture opportune moments between physical activities [18] or simply indicate that the phone has been used recently. Motion sensors can be accessed on Android and iOS without special permissions. Battery drain can be kept at reasonable levels by adjusting the sample rate (*e.g.*, 15 seconds every 10 minutes).

Context - Ambient Noise. Higher noise levels at the time of the notification and during the day so far were indicating for openness for engagement. The role of ambient noise level hasn't been studied much in related work and was not found to be indicative of a user's responsiveness to notifications [33] or how immersed people are into their phones [32]. Our results indicate that quiet environments and quiet days indicate less opportune moments for engagement. Both, Android and iOS allow to obtain noise levels via the built-in microphone, which can be accesses if the user grants the permission to record audio. Again, this approach requires a conservative sampling strategy to limit battery drain. The nature of the permission, however, makes it unlikely that these features are feasible in practice.

Context - Ambient Light. More stable light conditions during the day so far correlated with higher likelihood to engage with suggested content. Light values reported by the mobile phone sensor are most likely to be stable if the ambient light is artificial, if it is dark, or if the phone is covered, *e.g.* from resting in a pocket. Previous work reports no conclusive insights related to ambient light. One general theme emerging from these correlations is an indication that being at home in the evening constitutes a context in which people are likely to convert to suggested content. This findings are in line with recently-reported findings related to boredom [47],

alertness [1, 36], engagement [32], and ritualistic phone use [17]. The light sensor can be accessed without special permissions on Android, and the light-related features can be realized with a battery-conserving sampling strategy.

Demographics. The participants' age was found to be correlated with openness to engagement: participants who were older were slightly more open to the suggested content. Previous work [47] found that participants in their 30 were less prone to boredom than participants in their 20s and 40s. The difference in the results may be explained by the different demographics: the participants of the study presented in this article were older on average (M = 37.85 compared to M = 31.0) and therefore had comparably more participants in age groups older than 30-39. Many products and services won't have access to demographics. However, there is still a good share of applications who collects or has access to basic demographic information, *e.g.*, when working with profiles or when using logins of social networks which may disclose this information.

Phone state. One feature describing the status of the phone made it into the list of the top predictors. Participants were more open for engagement the more the battery was charged. This finding has not been brought up in related work. One explanation might be that in a low-battery situation, people are less likely to engage into non-targeted behavior with their mobile phone, to not risk running out of battery. The battery level can be obtained without explicit permission on Android and iOS. Beyond the battery level, Android allows to capture further phone states, such as the ringer mode, which can be useful if products decide to consciously only attempt to engage users that do not have their phone in silent mode.

Usage. Eight features related to phone usage were amongst the top predictors. Features derived from screen use, app launches, and access to the notification center indicate that openness for engagement is higher if the user has recently interacted with the phone. Additionally, higher data transmission rates and higher battery drain were indicative of openness for engagement as well. These findings are inline with previous work, which found that general app use, with the exception of communication apps, predicts phases of boredom [47] and stimulation-seeking behavior [17]. Both, Android and iOS allow to obtain data related to battery drain and data transmission rates. iOS allows to learn whether the screen is unlocked, whereas Android allows to register to callbacks that are called when the screen is turned on, off, or unlocked. All of this information is available without special permissions. App launches can be tracked on Android OS with a special permission. When the study was taking place, the majority of the phones still ran a version of Android OS that allows to register which application exactly is in the foreground. In newer versions, access has been more restricted so that an app can only know whether the launcher or another app is in foreground – which is sufficient to implement the app-related feature from the top predictors. Android further allows to register applications as accessibility service, which allows to track, amongst many other things, interactions with the notification center, as well as the name of the application in foreground. However, giving accessibility service access requires explicit action by the user and is highly unlikely to work with a product. All of these features can be implemented in a battery-preserving fashion.

Past Behavior. Dramatic increases in the classification performance were achieved by including features that model recent and past behavior related to the suggested content. In general, the more frequently participants had opened recommended content in the recent past, the more likely they were to engage with it in the future. Our results demonstrate that these features are a powerful means to match the frequency of positive predictions (*i.e.*, predicting that a participant would be open for recommendations) with the frequency of actual conversions. Past behavior related features can be computed for every product or service that keeps track of the reactions of the users to past engagement attempts.

Limitations of the Study. One limitation of our study springs from the design choice of not having a direct link between notification and content. The study shows that phone use can predict whether people will interact

with unknown content after they have responded to a notification. Thus, the model was trained to predict opportune moments for engaging users, but it did not consider the notification itself and its content. Whether users finally engage with an app should depend on many more factors than the timing of the engagement notification. Furthermore, responding to a minimum number of notifications was incentivized, which in a real context would not be the case. The findings may therefore not generalize situations, where the notification itself contains the content, as the user can already factor in the content to decide whether to engage or not. The proposed model is more conservative in the sense that the intelligent system will not take into account the propensity of the user to the content itself.

8 IMPLICATIONS

As stated earlier, the goal of this research is to inform the development of an intelligent system for proactive content delivery that finds a balance between (1) reducing the number of unwanted interruptions while (2) increasing conversion rates of engagement attempts to be attractive to product owners. In this section we present two implications that can be derived from the results of our study.

8.1 Optimizing the Timing of Engagement Attempts

If we had used our classifier in the study as an intelligent engagement system –sending notifications only when our system predicted that the user would engage– conversion rates would have been 7.1%. While this number in isolation may not seem impressive, it constitutes an increase of 66.6% over the baseline conversion rate of 4.3%. For products and services, an increase of 66.6% in engagement from notification campaigns can result into a significant increase in revenue, allowing to easily justify the introduction of such an intelligent engagement system from a business perspective.

One limitation of the classifier is that it only would have recalled 27.9% of the opportune moments. A product would need to give each user-engagement campaign enough time so that its users encounter a sufficient number of opportune moments. For example, given a recall of 0.279, the user would have to experience 10 or more opportune moments for the cumulative probability to exceed 95% to recall at least one opportune moment. Since a number of opportune moments cannot be guaranteed, products and services should employ a timeout, *i.e.*, handling the notification in a different way if no opportune moment has been found within the permitted time frame. Depending on the focus, the system could either post the notification when it times or simply abandon the engagement attempt.

The former strategy results into campaigns where engagement attempts are less likely to occur in moments when users are not paying attention to their phone or otherwise occupied. The second strategy results into campaigns where the notification volume is reduced for users who have not shown any interested in the product. Both strategies have the potential to reduce churn (*e.g.*, through disabling notifications or uninstalling the app) and hence, increase revenue and customer satisfaction in the long run. In an ideal scenario, product owners could achieve a higher absolute number of conversions while reducing the amount of unwanted engagement attempts.

8.2 Reducing Unwanted Interruptions through Past-Behavior Modeling

While it may not be too surprising that past behavior predicts future behavior in this context, our finding of a hypothetical success rate of 21.8% shows how powerful observing past behavior is compared to modeling openness for engagement from other phone use patterns. This stresses the importance of respecting users who have repeatedly ignored suggested content in the past. Adapting the notification frequency to past behavior has huge positive impacts on the prediction performance, while maintaining a high recall (54.0%). The positive side of the predictive power of past behavior is that it does not require to capture mobile sensor data or any personal

data that would require explicit consent and could have potential privacy implications. Moreover, the use of rolling averages allows to adapt to changes in the recent conversion behavior of specific users.

One challenge when employing past behavior as a feature is the cold start problem, *e.g.*, when there are new users or new content offerings there is no data about past behavior. Depending on the frequency of engagement campaigns that a product or service is willing to undertake, building up a reliable set of features related to past behavior could take weeks or months. In these cases, an intelligent notification delivery system could fall back to a model without past behavior features until a sufficient number of data points have been collected for the respective user.

A second challenge is that the use of *PastAction* features requires product owners to accept the reality that not all users can be engaged. Roughly one third of the participants in our study were never or only once predicted to be open for engagement. Thus, the use of past behavior is particularly helpful if products and services are willing to emphasize the improvement of user experience and the decrease churn from unwanted interruptions over attempts to generate conversions at any cost.

9 CONCLUSIONS

We conducted a field study with 337 participants in which, for an average of 4 weeks, they installed a study application onto their primary mobile phones. The participants considered the primary purpose of the study to self-report their emotions to a notification-triggered mini questionnaire. In reality, we were interested whether they would voluntarily engage with content suggested at the bottom of those questionnaires.

We show that features derived from data collected from the mobile phone allow to train a classifier that predicts – at the time of posting a non-specific notification – whether participants will engage with suggested content that is offered by the notification-sending app. The classifier achieves a 66.6% higher precision than the baseline. Furthermore, we show how modeling past interest with the suggested content can be used significantly increase the precision 5 times over the baseline, while avoiding failed engagement attempts of about one third of the participants.

We describe how many of the most predictive features can be derived from popular smartphones operating systems with reasonable impact on battery life and without requiring explicit permissions. We also discuss how such a classifier could be used in products to increase conversion rates, improve user experience, and lower churn by reducing undesired interruptions.

Future work includes (1) testing to which degree the performance that we achieved in this study can be achieved in a product or service that uses notifications to engage its users; (2) exploring the role of the content on the prediction performance; and (3) investigating the performance on an individual level to better account for the fact that many individuals will not be open to certain types of content, no matter how well the engagement attempt is timed.

10 ACKNOWLEDGMENTS

We thank the team of the Smart Notifications Project for their support in conducting this study. In particular, we thank Michal Ficek and Pablo Lanaspa for the invaluable help with processing the data collected in this study.

REFERENCES

[1] Saeed Abdullah, Elizabeth L. Murnane, Mark Matthews, Matthew Kay, Julie A. Kientz, Geri Gay, and Tanzeem Choudhury. 2016. Cognitive Rhythms: Unobtrusive and Continuous Sensing of Alertness Using a Mobile Phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 178–189. DOI: http://dx.doi.org/10.1145/2971648.2971712

[2] Piotr D. Adamczyk and Brian P. Bailey. 2004. If Not Now, when?: The Effects of Interruption at Different Moments Within Task Execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 271–278. DOI: http://dx.doi.org/10.1145/985692.985727

[3] Ionut Andone, Konrad Blaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. 2016. How Age and Gender Affect Smartphone Usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 9–12. DOI: http://dx.doi.org/10.1145/2968219.2971451

[4] Daniel Avrahami and Scott E. Hudson. 2006. Responsiveness in Instant Messaging: Predictive Models Supporting Inter-personal Communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 731–740. DOI: http://dx.doi.org/10.1145/1124772.1124881

[5] L. Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

[6] L. Breiman and J. Friedman. 1984. *Classification and regression trees*. Chapman and Hall/CRC, Monterey, USA.

[7] Yung-Ju Chang and John C. Tang. 2015. Investigating Mobile Users' Ringer Mode Usage and Attentiveness and Responsiveness to Communication. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. ACM, New York, NY, USA, 6–15. DOI: http://dx.doi.org/10.1145/2785830.2785852

[8] Peter McFaul Chapman. 1997. *Models of engagement: Intrinsically motivated interaction with multimedia learning software*. Ph.D. Dissertation. University of Waterloo.

[9] T. Chen and C. Guestrin. 2016. XGBoost: a scalable tree boosting system. In *Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*. 785–794.

[10] Minsoo Choy, Daehoon Kim, Jae-Gil Lee, Heeyoung Kim, and Hiroshi Motoda. 2016. Looking Back on the Current Day: Interruptibility Prediction Using Daily Behavioral Features. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 1004–1015. DOI: http://dx.doi.org/10.1145/2971648.2971649

[11] Anind K Dey, Katarzyna Wac, Denzil Ferreira, Kevin Tassini, Jin-Hyuk Hong, and Julian Ramos. 2011. Getting closer: an empirical investigation of the proximity of user to their smart phones. In *Proc. UbiComp '11*. ACM. http://www.mediateam.oulu.fi/publications/pdf/1432.pdf

[12] Richard Farmer and Norman D. Sundberg. 1986. Boredom Proneness – The development and correlates of a new scale. *Journal of Personality Assessment* 50 (1986), 4–17. Issue 1.

[13] Joel E. Fischer, Chris Greenhalgh, and Steve Benford. 2011. Investigating Episodes of Mobile Phone Activity As Indicators of Opportune Moments to Deliver Notifications. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*. ACM, New York, NY, USA, 181–190. DOI: http://dx.doi.org/10.1145/2037373.2037402

[14] Joel E. Fischer, Nick Yee, Victoria Bellotti, Nathan Good, Steve Benford, and Chris Greenhalgh. 2010. Effects of Content and Time of Delivery on Receptivity to Mobile Interruptions. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '10)*. ACM, New York, NY, USA, 103–112. DOI: http://dx.doi.org/10.1145/1851600.1851620

[15] James Fogarty, Scott E. Hudson, Christopher G. Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C. Lee, and Jie Yang. 2005. Predicting Human Interruptibility with Sensors. *ACM Trans. Comput.-Hum. Interact.* 12, 1 (March 2005), 119–146. DOI: http://dx.doi.org/10.1145/1057237.1057243

[16] T. Hastie, R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning* (2nd ed.). Springer, Berlin, Germany.

[17] Alexis Hiniker, Shwetak N. Patel, Tadayoshi Kohno, and Julie A. Kientz. 2016. Why Would You Do That? Predicting the Uses and Gratifications Behind Smartphone-usage Behaviors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 634–645. DOI: http://dx.doi.org/10.1145/2971648.2971762

[18] Joyce Ho and Stephen S. Intille. 2005. Using Context-aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, New York, NY, USA, 909–918. DOI: http://dx.doi.org/10.1145/1054972.1055100

[19] Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. 2003. Models of Attention in Computing and Communication: From Principles to Applications. *Commun. ACM* 46, 3 (March 2003), 52–59. DOI: http://dx.doi.org/10.1145/636772.636798

[20] Eric Horvitz, Paul Koch, Raman Sarin, Johnson Apacible, and Muru Subramani1. 2005. Bayesphone: Precomputation of Context-Sensitive Policies for Inquiry and Action in Mobile Devices. In *Proc UM '05*.

[21] Shamsi T. Iqbal and Brian P. Bailey. 2008. Effects of Intelligent Notification Management on Users and Their Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 93–102. DOI: http://dx.doi.org/10.1145/1357054.1357070

[22] Shamsi T. Iqbal and Brian P. Bailey. 2010. Oasis: A Framework for Linking Notification Delivery to the Perceptual Structure of Goal-directed Tasks. *ACM Trans. Comput.-Hum. Interact.* 17, 4, Article 15 (Dec. 2010), 28 pages. DOI: http://dx.doi.org/10.1145/1879831.1879833

[23] Yasumasa Kobayashi, Takahiro Tanaka, Kazuaki Aoki, and Kinya Fujita. 2015. Automatic Delivery Timing Control of Incoming Email Based on User Interruptibility. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 1779–1784. DOI: http://dx.doi.org/10.1145/2702613.2732825

[24] Kostadin Kushlev, Bruno Cardoso, and Martin Pielot. 2017. Too Tense for Candy Crush: Affect Influences User Engagement With Proactively Suggested Content. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*. ACM, New York, NY, USA.

[25] Kostadin Kushlev, Jason Proulx, and Elizabeth W. Dunn. 2016. "Silence Your Phones": Smartphone Notifications Increase Inattention and Hyperactivity Symptoms. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1011–1020. DOI: http://dx.doi.org/10.1145/2858036.2858359

[26] R. Larson and M. Csikszentmihalyi. 1983. The experience sampling method. *New Directions for Methodology of Social and Behavioral Science* 15 (1983), 41?–56.

[27] Hugo Lopez-Tovar, Andreas Charalambous, and John Dowell. 2015. Managing Smartphone Interruptions Through Adaptive Modes and Modulation of Notifications. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 296–299. DOI: http://dx.doi.org/10.1145/2678025.2701390

[28] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The Cost of Interrupted Work: More Speed and Stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 107–110. DOI: http://dx.doi.org/10.1145/1357054.1357072

[29] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2015. Focused, Aroused, but So Distractible: Temporal Perspectives on Multitasking and Communications. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 903–916. DOI: http://dx.doi.org/10.1145/2675133.2675221

[30] Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, and Paul Johns. 2014. Bored Mondays and Focused Afternoons: The Rhythm of Attention and Online Activity in the Workplace. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3025–3034. DOI: http://dx.doi.org/10.1145/2556288.2557204

[31] Afra Mashhadi, Akhil Mathur, and Fahim Kawsar. 2014. The Myth of Subtle Notifications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 111–114. DOI: http://dx.doi.org/10.1145/2638728.2638759

[32] Akhil Mathur, Nicholas D. Lane, and Fahim Kawsar. 2016. Engagement-aware Computing: Modelling User Engagement from Mobile Contexts. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 622–633. DOI: http://dx.doi.org/10.1145/2971648.2971760

[33] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. 2015. Designing Content-driven Intelligent Notification Mechanisms for Mobile Applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 813–824. DOI: http://dx.doi.org/10.1145/2750858.2807544

[34] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1021–1032. DOI: http://dx.doi.org/10.1145/2858036.2858566

[35] T. M. Mitchell. 1997. *Machine Learning*. McGraw-Hill, New York, USA.

[36] Elizabeth L. Murnane, Saeed Abdullah, Mark Matthews, Matthew Kay, Julie A. Kientz, Tanzeem Choudhury, Geri Gay, and Dan Cosley. 2016. Mobile Manifestations of Alertness: Connecting Biological Rhythms with Patterns of Smartphone App Use. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '16)*. ACM, New York, NY, USA, 465–477. DOI: http://dx.doi.org/10.1145/2935334.2935383

[37] Heather L O'Brien and Elaine G Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology* 59, 6 (2008), 938–955.

[38] Tadashi Okoshi, Hiroki Nozaki, Jin Nakazawa, Hideyuki Tokuda, Julian Ramos, and Anind K. Dey. 2016. Towards attention-aware adaptive notification on smart phones. *Pervasive and Mobile Computing* (2016).

[39] Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K. Dey, and Hideyuki Tokuda. 2015. Reducing Users' Perceived Mental Effort Due to Interruptive Notifications in Multi-device Mobile Environments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 475–486. DOI: http://dx.doi.org/10.1145/2750858.2807517

[40] Tadashi Okoshi, Kota Tsubouchi, Masaya Taji, Takanori Ichikawa, and Hideyuki Tokuda. 2017. Attention and Engagement-Awareness in the Wild: A Large-Scale Study with Adaptive Notifications. In *IEEE International Conference on Pervasive Computing and Communications*.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[42] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, USA, 897–908. DOI: http://dx.doi.org/10.1145/2632048.2632062

[43] Veljko Pejovic, Mirco Musolesi, and Abhinav Mehrotra. 2015. Investigating The Role of Task Engagement in Mobile Interruptibility. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '15)*. ACM, New York, NY, USA, 1100–1105. DOI: http://dx.doi.org/10.1145/2786567.2794336

[44] Martin Pielot. 2014. Large-scale Evaluation of Call-availability Prediction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, USA, 933–937. DOI: http://dx.doi.org/10.1145/2632048.2632060

[45] Martin Pielot, Karen Church, and Rodrigo de Oliveira. 2014. An In-situ Study of Mobile Phone Notifications. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '14)*. ACM, New York, NY, USA, 233–242. DOI: http://dx.doi.org/10.1145/2628363.2628364

[46] Martin Pielot, Rodrigo de Oliveira, Haewoon Kwak, and Nuria Oliver. 2014. Didn't You See My Message?: Predicting Attentiveness to Mobile Instant Messages. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3319–3328. DOI: http://dx.doi.org/10.1145/2556288.2556973

[47] 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. DOI: http://dx.doi.org/10.1145/2750858.2804252

[48] Martin Pielot and Luz Rello. 2017. Productive, Anxious, Lonely ? 24 Hours Without Push Notifications. In *MobileHCI '17: Proceedings of the 18th International Conference on Human-computer Interaction with Mobile Devices and Services*.

[49] Benjamin Poppinga, Wilko Heuten, and Susanne Boll. 2014. Sensor-Based Identification of Opportune Moments for Triggering Notifications. *IEEE Pervasive Computing* 13, 1 (Jan. 2014), 22–29. DOI: http://dx.doi.org/10.1109/MPRV.2014.15

[50] Stephanie Rosenthal, Anind K. Dey, and Manuela Veloso. 2011. Using decision-theoretic experience sampling to build personalized mobile phone interruption models. In *Proc. Pervasive '11*. Springer-Verlag, 18. http://dl.acm.org/citation.cfm?id=2021975.2021991

[51] Alireza Sahami Shirazi, Niels Henze, Tilman Dingler, Martin Pielot, Dominik Weber, and Albrecht Schmidt. 2014. Large-scale Assessment of Mobile Notifications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3055–3064. DOI: http://dx.doi.org/10.1145/2556288.2557189

[52] Hillol Sarker, Moushumi Sharmin, Amin Ahsan Ali, Md. Mahbubur Rahman, Rummana Bari, Syed Monowar Hossain, and Santosh Kumar. 2014. Assessing the Availability of Users to Engage in Just-in-time Intervention in the Natural Environment. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, USA, 909–920. DOI: http://dx.doi.org/10.1145/2632048.2636082

[53] Florian Schulze and Georg Groh. 2016. Conversational Context Helps Improve Mobile Notification Management. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '16)*. ACM, New York, NY, USA, 518–528. DOI: http://dx.doi.org/10.1145/2935334.2935347

[54] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. 2015. Interruptibility Prediction for Ubiquitous Systems: Conventions and New Directions from a Growing Field. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 801–812. DOI: http://dx.doi.org/10.1145/2750858.2807514

[55] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. 2017. Reachable but not Receptive: Enhancing Smartphone Interruptibility Prediction by Modelling the Extent of User Engagement with Notifications. *Pervasive and Mobile Computing* (2017). DOI: http://dx.doi.org/10.1016/j.pmcj.2017.01.011 accepted Jan 31, 2017.

[56] Tilo Westermann, Sebastian Möller, and Ina Wechsung. 2015. Assessing the Relationship between Technical Affinity, Stress and Notifications on Smartphones. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '15)*. ACM, New York, NY, USA, 652–659. DOI: http://dx.doi.org/10.1145/2786567.2793684

[57] Tilo Westermann, Ina Wechsung, and Sebastian Möller. 2016. Smartphone Notifications in Context: A Case Study on Receptivity by the Example of an Advertising Service. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 2355–2361. DOI: http://dx.doi.org/10.1145/2851581.2892383

[58] Fred Zijlstra, Robert Roe, Anna B. Leonora, and Irene Krediet. 1999. Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology* 72 (June 1999), 163–185. Issue 2. DOI: http://dx.doi.org/10.1348/096317999166581

Received February 2017; revised May 2017; accepted June 2017