When Attention is not Scarce - Detecting Boredom from Mobile Phone Usage

Conference Paper · September 2015

DOI: 10.1145/2750858.2804252

4 authors:

Martin Pielot
Google Inc.
85 PUBLICATIONS **2,895** CITATIONS
SEE PROFILE

Tilman Dingler
University of Melbourne
107 PUBLICATIONS **1,270** CITATIONS
SEE PROFILE

Jose San Pedro
Schibsted
33 PUBLICATIONS **903** CITATIONS
SEE PROFILE

Nuria Oliver
Data-Pop Alliance
225 PUBLICATIONS **13,374** CITATIONS
SEE PROFILE

When Attention is not Scarce - Detecting Boredom from Mobile Phone Usage

Martin Pielot[1], **Tilman Dingler**[2], **Jose San Pedro**[1], **Nuria Oliver**[1]

[1]Telefonica Research, Barcelona, Spain– {firstname.lastname}@telefonica.com
[2]University of Stuttgart, Germany – tilman.dingler@vis.uni-stuttgart.de

ABSTRACT

Boredom is a common human emotion which may lead to an active search for stimulation. People often turn to their mobile phones to seek that stimulation. In this paper, we tackle the challenge of automatically inferring boredom from mobile phone usage. In a two-week in-the-wild study, we collected over 40,000,000 usage logs and 4398 boredom self-reports of 54 mobile phone users. We show that a user-independent machine-learning model of boredom –leveraging features related to recency of communication, usage intensity, time of day, and demographics– can infer boredom with an accuracy (AUCROC) of up to 82.9%. Results from a second field study with 16 participants suggest that people are more likely to engage with recommended content when they are bored, as inferred by our boredom-detection model. These findings enable boredom-triggered proactive recommender systems that attune their users' level of attention and need for stimulation.

Author Keywords

Attention; Boredom; Mobile Devices; Killing Time; Attention Economy

ACM Classification Keywords

H.5.m Information interfaces and presentation: misc.

INTRODUCTION

In today's connected world, people are constantly exposed to external stimulation through technology – be it through connected TVs and desktop PCs at home or through tablets and mobile phones on the go. A large portion (43% according to Nielsen[1]) of this time is devoted to self-stimulation and entertainment activities, such as watching media, Web-browsing, playing games and social media. Further, an increasing number of services is requesting our attention. Many Internet companies primarily live off monetizing their users' attention by exposing them to advertisement. Consequently, attention has become a scarce resource [11]: knowing when a user is likely to pay attention to a specific piece of content is becoming increasingly valuable.

However, attention is not always scarce. One frequently occurring affective state [17] goes along with an abundance of attentional resources: *boredom*. Boredom is characterized by a "lack of stimulation" [14] and being "actively looking for stimulation" [12]. And, technology might even have changed our tolerance to boredom: over time people habituate to a constant exposure to stimuli [12, 25] such that, when the level of stimulation drops, they become bored. People who were asked to spend 24 hours without any media as part of a study[2] reported negative emotions, ranging from boredom to anxiety and even withdrawal symptoms.

Mobile phones are a commonly used tool to fill or kill time when bored [7, 25], especially while being on-the-go[3]. These devices are most likely to be present in all kinds of boredom-prone situations, such as subway rides, in class, or while waiting. In such situations, we turn to our phones to kill time, *i.e.*, for self-stimulation without having a particular task in mind.

For us, this reality represents an opportunity: if mobile phones are able to detect when their users are killing time, *i.e. when attention is not scarce*, then they could suggest a better use of those idle moments by,

- recommending content, services, or activities that may help to overcome the boredom;
- suggesting to turn their attention to more useful activities, such as revisiting *read later* lists, going over to-do lists, or participating in a research survey; or
- helping the user to make positive use of the boredom, such as using it for introspection, since mental downtime is essential to reflection, learning, and fostering creativity [34].

In this paper, we report from two in-the-wild user studies that provide evidence to what extent killing time with the phone due to boredom –characterized as a stimulus-seeking state– manifests itself in detectable mobile usage patterns, and that being bored makes mobile phone users more open to consume suggested content. The main contributions of this paper are:

[1]http://www.nielsen.com/us/en/insights/news/2014/how-smartphones-are-changing-consumers-daily-routines-around-the-globe.html

[2]http://theworldunplugged.wordpress.com/

[3]42% of cell owners reported to have used their phone for entertainment when they were bored: http://www.pewinternet.org/2011/08/15/americans-and-their-cell-phones/

UbiComp '15, September 7–11, 2015, Osaka, Japan.

http://dx.doi.org/10.1145/2750858.2804252.

1. A machine-learning model to automatically infer boredom from demographics and mobile phone usage;
2. an analysis of the mobile phone usage patterns that are most related to boredom; and
3. evidence that people are more likely to engage with suggested content when the model infers that they are bored than when they are not.

BACKGROUND AND RELATED WORK

Boredom is defined as displeasure caused by "lack of stimulation or inability to be stimulated thereto." [14]. It goes along with a "pervasive lack of interest and difficulty concentrating on the current activity" [15]. Eastwood [12] highlights that "a bored person is not just someone who does not have anything to do; it's someone who is actively looking for stimulation but it is unable to do so".

Consequently, feeling bored often goes along with an urge to escape such a state [17]. This urge can be so severe that in one study reported by Wilson *et al.* [37], people preferred to self-administer electric shocks rather than being left alone with their thoughts for a few minutes.

"The cure to boredom is curiosity" is a famous quote by Dorothy Parker, an American writer and poet, which highlights that boredom has an actual purpose: it is an emotional state that signals that current activities or goals are not sufficiently satisfying and motivating. Potential benefits of boredom include the initiation of creative processes and self-reflection [34]. Given that bored people long for stimuli and that human attention has become scarce and increasingly valuable [11], there also is commercial value in knowing when a person is bored.

Detecting Boredom

According to Bixler and D'Mello [3], the most popular methods for detecting boredom have been facial expressions, speech and its paralinguistic features, text, and physiological signals.

In one of the early landmark studies in the field, Picard *et al.* [26] showed how emotional states –not including boredom– can be recognized by physiological sensors. For 30 days, one female subject recorded physiological sensors for 25 minutes each day. As sensors, they used an electromyogram for recognizing facial muscles tension, a photophletysmograph to measure blood volume pressure and HR, a skin conductance sensor, and a hall effect respiration sensor. Features derived from these sensors allowed to discriminate between 8 systematically elicited emotions with 81% accuracy.

However, to date, these sensors require extensive setup and may therefore not be available in most situation where boredom typically occurs. Thus, other researchers have explored other ways of detecting boredom, which do not require any explicit setup by the user.

For example, Bixler and D'Mello [3] explored how to detect boredom during writing tasks through logging the writers keystrokes. They found that "boredom" was named as an affective state in 26.4% out of the 5551 affect judgments – second most often after engagement (35.4%). Keystrokes alone had comparably low predictive power – roughly 11% above chance – for discriminating engagement-neutral and boredom-neutral states. However, when adding *stable traits* of the participants to the model, prediction performance could be notably improved.

Guo *et al.* [18] found that when users are engaged in a Web search task, mouse movements, clicks, page scrolls, and other fine-grained interaction events allow to predict the searchers openness to let themselves be distracted from their main task, which may be an indication of boredom.

Recently, Mark *et al.* [23] studied the rhythm of attention and affect, including boredom, in the workplace. In a 5-day in-situ study, they logged computer activity of 32 information workers and frequently (about 17 times per day) probed their affect. They found that boredom is related to, amongst other variables, the time of the day and computer interaction patterns, such as the frequency of window switches.

Inferring Emotions from Mobile Phone Usage

In the context of mobile phones, several works show that emotions are reflected in how we use our mobile phones. LiKamWa *et al.* [22] showed that daily mood (valence and arousal) can be inferred from monitoring social interactions via SMS, email, and phone calls, as well as routine activity, such as application usage. Similarly, Bogomolov *et al.* showed that daily happiness [5] and daily stress [4] can be inferred from mobile phone usage, personality traits, and weather data.

There has also been extensive research on using sensors to detect attentional states, such as a person's level of interruptibility. As boredom is defined in terms of attention [12] –*i.e.,* a state of looking for stimulation, insights from these studies may apply to boredom as well.

In particular, it has been shown that computing devices are able to detect a person's openness to receiving office visits [16], emails [21], messages from desktop instant messengers [1], SMS and mobile phone messages [30], mobile phone alerts [31], and mobile phone calls [20, 27].

This line of research shows that attention and openness to interruptions can be inferred from: time since recent usage of device [1, 16]; using of specific services –such as internet browsers, email inbox, calendar [23]– or usage of mobile phone messengers and notification center [27, 30]; level of activity –such as switching windows [21, 23]– or use of keyboard and mouse [16, 21]; ambient noise level (as proxies for the level of actives around the user) [16]; location (differences between home and work) [27, 31]; ringer mode (as an indicator of how we want to manage interruptions) [27, 31]; time –such as the hour of the day or the day of the week [1, 16, 20]; and proximity, *i.e.,* if a mobile phone's screen is free or covered (indicating if the phone is stowed away) [27, 30].

These studies show that level of attention, openness to interruption, and boredom measurably affect the way we interact with technology. However, three research questions remain unanswered to date, namely:

- If boredom, *i.e.* a state of actively looking for stimulation, measurably affects phone use (**RQ1**);
- what aspects of mobile phone usage are the most indicative of boredom (**RQ2**); and
- if people who are bored are more likely to consume suggested content on their mobile phones than when they are not bored (**RQ3**).

METHODOLOGY

In order to answer **RQ1** and **RQ2**, we conducted a field study with 54 participants who installed a dedicated data-collection application called *Borapp* on their personal mobile phones and actively contributed with their data for at least 14 days. The goal of this study was to collect: (1) mobile phone usage data; and (2) ground-truth about the participants' level of boredom through a refined Experience Sampling methodology [9].

Mobile Usage-Patterns Collection

Borapp ran on Android phones with OS 4.0 or newer. Usage patterns were inferred from the mobile phone's event listeners and sensor data. The data that *Borapp* collects is split into two groups: (1) data which is *always* collected and (2) data which is only collected when the phone is in use *i.e.*, the *screen is on and unlocked*. This approach enabled us to have a battery-efficient data-collection method.

Sensors that are constantly active are shown in Table 1 and sensors that were activated only when the phone was unlocked are shown in Table 2.

Sensor	Description
Battery Status	Battery level ranging from 0-100%
Notifications	Time and type (app) of notification
Screen Events	Screen turned on, off, and unlocked
Phone Events	Time of incoming and outgoing calls
Proximity	Screen covered or not
Ringer Mode	Silent, Vibration, Normal
SMS	Time of receiving, reading, and sending SMS

Table 1. List of sensors whose data was collected at all times.

Sensor	Description
Airplane Mode	Whether phone in airplane mode
Ambient Noise	Noise in dB as sensed by the microphone
Audio Jack	Phone connected to headphones or speakers
Cell Tower	The cell tower the phone is connected to
Data Activity	Number of bytes up/downloaded
Foreground app	Package name of the app in foreground
Light	Ambient light level in SI lux units
Screen Orient	Portrait or Landscape mode
Wifi Infos	The WiFi network the phone is connected to

Table 2. List of sensors whose data was collected only when the phone was unlocked.

Users were required to enable the *Android Accessibility Service*, as well as to grant *Borapp* access to notifications, in order to collect data about user activity that is otherwise not exposed via standard APIs. The accessibility service allowed us to *e.g.* monitor which app is in the foreground without having to run a busy-waiting polling service in the background. Notification access allowed us to learn when notifications from *e.g.* messengers or email applications were posted.

The collected data was saved locally until the mobile phone was connected to a Wifi network. Only then, *Borapp* transmitted the logged data to our server so that the data transfer would not impact our participants' data plans.

Demographics

During the setup phase, participants were asked to enter their age, gender, and an email address for follow-up communication. Due to the open nature of the participation, the introduction of this information was voluntary and this was clearly explained in the application.

Experience-Sampling Probes

We collected ground truth about the participants' state of boredom via *experience-sampling* (ESM) [10]. Generally, this research method entails to probe users at certain times throughout the day to collect their subjective feedback about their current experience or situation. In our case, we gathered *in-situ* self-reports on the subjective level of boredom.

Borapp delivered self-report probes through mobile phone notifications (see Figure 1). These notifications were scheduled in semi-regular intervals whenever the phone was in use and a minimum amount of 60 minutes had passed since the last probe was answered. Because we were interested in understanding boredom while using the phone, a probe was more likely to be triggered when a participant was interacting with the mobile phone.

If the participant clicked on an probe notification, a view with a mini-questionnaire opened. The questionnaire asked participants to respond on a 5-point Likert scale to the question: "*To what extend do you agree to the following statement: 'Right now, I feel bored.'?*". The extremes were labelled with *disagree* and *agree*. Internally, the responses were stored with values from 0 (disagree) to 4 (agree).

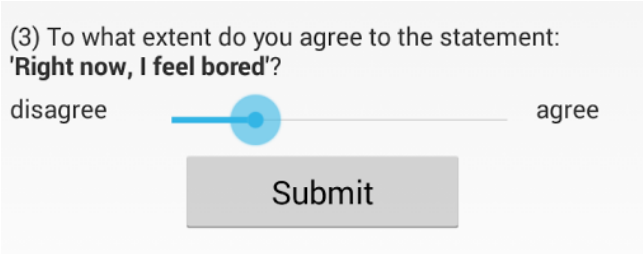

Figure 1. Screenshot of the ESM probe.

Note that the mini-questionnaire also probed participants' about their levels of valence and arousal. However, we do not report these results here, since they are out of the scope of this work.

Procedure

We launched the study in June 2014. For widespread distribution we made *Borapp* available to download for free on the Google Play store, which means that anyone could join the

study at any time by simply downloading the app. Since *Borapp* does not provide an immediate user benefit, we advertised the study through various email lists and social network channels.

Once participants had downloaded and installed *Borapp*, they were asked for their explicit consent to their participation in the study. Therefore the initial screen explained the background of the study, what kind of data was collected, how and where it was stored and how it was going to be used. The consent explicitly pointed out which personally identifiable information was stored, namely the device location. Also, the terms and conditions of participating in the study and to collect the study reward were disclosed here.

After consenting, *Borapp* walked participants through the setup, which includes giving access to the Android Accessibility Services and grant the app access to notifications. In the final step, participants could optionally specify their age, gender, and an email address. Once consent was given and *Borapp* was set up successfully, it started collecting data and triggering probes via experience sampling.

To successfully participate in the study, participants had to keep *Borapp* running for at least 2 weeks and answer an average of 6 probes per day. Participants could check their progress in a status screen. Those who successfully completed were rewarded with a 20 Euro gift card of a large online store.

Participants

Recruitment was primarily done via two mailing lists. One list contained email addresses of computer-science students at a German university. The other contained one thousand volunteers from Spain who had signed up to be informed about opportunities to participate in research of a large organization. In addition, we announced the study via social networks.

At the beginning of July 2014, we created a snapshot of the data of all participants who had completed the study so far. The raw data set contains 43,342,860 mobile phone sensor entries and 4,826 responses to the ESM probes from 61 unique mobile devices.

Checking the data for validity, we found that responses to the ESM probes from 7 devices barely varied, which might be an indication that their users did not seriously try to reflect their emotional states. Hence, we filtered the data from these devices, which led to 54 remaining participants with 4,398 valid self-reports.

All results reported subsequently will be based on the data from the 54 valid participants. Each participants contributed 84 and 173 ($M = 110.3, SD = 25.8$) self reports. As explained earlier, it was voluntary to specify demographics: 39 participants specified their age in a range from 21 to 57, with a mean age of 31.0 ($SD = 7.9$). In terms of gender, 11 participants reported to be female and 23 reported to be male. The remaining 19 participants either chose the 'other' option or did not specify their gender. According to the most frequent device locales (52% es_ES, 18% de_DE, 13% en_US) and timezones (79% UTC+1, 6% UTC+0 and 5% UTC+8), most participants were from Spain, Germany, and the US.

RESULTS

To explore the relation between boredom and mobile phone usage, we approached the data analysis as a machine-learning classification task. Our rationale for following such an approach was two-fold. First, machine-learning techniques would allow us to explore to which degree different usage patterns were related to boredom and killing time on the phone, and second, they would allow to quantify to which degree boredom can be inferred from mobile phone usage.

Features

We extracted 35 features related to phone-usage patterns in 7 categories: **context, demographics, time since last activity**, **intensity of usage, external triggers, "idling"** –our assumption being that short, frequent phone interactions relate to less goal-oriented activity– and **type of usage**. Table 3 and Table 4 list their description.

Context	
audio	Indicates whether the phone is connected to a headphone or a bluetooth speaker
charging	Whether the phone is connected to a charger or not
day_of_week	Day of the week (0-6)
hour_of_day	Hour of the day (0-23)
light	Light level in lux measured by the proximity sensor
proximity	Flag whether screen is covered or not
ringer_mode	Ringer mode (silent, vibrate, normal)
semantic_location	Home, work, other, or unknown
Demographics	
age	The participant's age in years
gender	The participant's gender
Last Communication Activity	
time_last_incoming_call	Time since last incoming phone call
time_last_notif	Time since last notification (excluding Borapp probe)
time_last_outgoing_call	Time since the user last made a phone call
time_last_SMS_read	Time since the last SMS was read
time_last_SMS_received	Time since the last SMS was received
time_last_SMS_sent	Time since the last SMS was sent

Table 3. List of features related to context, demographics, and time since last communication activity.

Some of the data collected from the mobile sensors –such as the time since the last phone call, or ringer mode status– could be used directly as a feature. We computed other features –such as recent battery drain or network usage– by applying a *time window* prior to submitting the subjective ratings, *e.g.* battery drain in the last n minutes before self-reporting the current level of boredom. We tested time windows of 1, 5, 10, 30, and 60 minute-length prior to the probe. We achieved the best classification results with 5-minute time windows. Hence all time window-dependent features reported hereafter are based on a 5-minute window.

Usage (related to usage intensity)	
battery_drain	Average battery drain in time window
battery_level	Battery change during the last session
bytes_received	Number of bytes received during time window
bytes_transmitted	Number of bytes transmitted during time window
time_in_comm_apps	Time spent in communication apps, categorized to none, micro session, and full session
Usage (related to whether it was triggered externally)	
num_notifs	Number of notifications received in time window
last_notif	Name of the app that created the last notification
last_notif_category	Category of the app that created the last notification
Usage (related to the user being idling)	
apps_per_min	Number of apps used in time-window divided by time the screen was on
num_apps	Number of apps launched in time window before probe
num_unlock	Number of phone unlocks in time window prior to probe
time_last_notif_access	Time since the user last opened the notification center
time_last_unlock	Time since the user last unlocked the phone
Usage (related to the type of usage)	
screen_orient_changes	Flag whether there have been screen orientation changes in the time window
app_category_in_focus	Category of the app in focus prior to the probe
app_in_focus	App that was in focus prior to the probe
comm_notifs_in_tw	received in the time window prior to the probe
most_used_app	Name of the app used most in the time window
most_used_app_category	Category of the app used most in the time window
prev_app_in_focus	App in focus prior to *app_in_focus*

Table 4. List of features related to usage intensity, external trigger, idling and type.

Please note that in case of application and notification-related features, we applied a blacklist, so that *Borapp* and system services[4], were excluded.

Feature Cleaning

Since linear models are heavily affected by outliers, we inspected all numeric features to determine if they require saturation, *i.e.* reducing outliers to a certain threshold. All of the numeric features were long-tail distributions, hence, there were only positive outliers. Depending on the skewness of each feature, we chose the appropriate percentile out of 90%, 95%, and 99%, and used it as upper limit.

Many entries in the app-related features (*e.g.* last app in foreground, most-used app, last notification) were sparse, that is, many of the recorded apps appeared only a few times or once. Since such sparsity makes it difficult to properly learn the meaning of sparse elements, we reduced the dimensionality of these features by mapping rarely used apps into an 'other' category. The distribution was again heavily skewed, thus we kept the 10 most frequent applications and mapped the rest into the 'other' category. In the features describing the application categories, we kept the three major categories –namely, Communication, Productivity, and Society– which account for two-third of the instances.

[4]For example, on some Android devices, a notification event is fired every time the keyboard is opened.

Ground Truth

We define the modeling of boredom as a binary classification problem: detecting whether the user is in a bored state or not.

Figure 2 shows a histogram of the boredom ratings collected in our study. The average boredom rating is $M = 1.17$ and $Mdn = 1$, *i.e.,* in general our participants tended to disagree with the statement that they felt bored in the moment in which the question was asked.

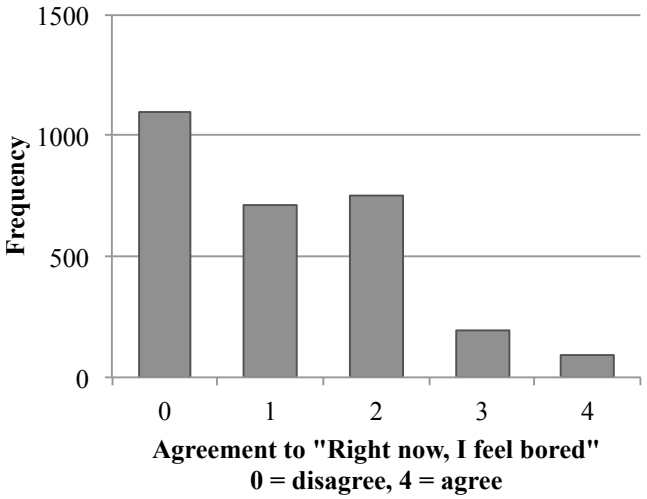

Figure 2. Histogram of self-report probe responses.

We computed two different ground truths: first, we took the straightforward approach and mapped participants into the *bored* class when they agreed to feeling bored (scores 3 and 4). We will refer to this as *absolute* ground truth.

Investigating the data, we found that some of the participants had different anchor points, such that they rated themselves as being much more bored on average than the rest. One explanation might be that people have different predisposition to boredom [13], hence they tend have different **normal** or baseline levels of boredom.

Therefore, we decided to consider an personalized ground truth definition that reflected when participants **felt more bored than usual**. Hence, we transformed the absolute responses into personalized z-Scores, where 0 indicates that the participant felt as (non-)bored as on average during the study. We labeled samples with a value over +.25 in this personalized scale as positive. We will refer to this as *normalized* ground truth.

Data Sets

While our main insights are based on a primary data set, we explored the effect of altering two different factors.

The first aspect was whether the ground truth was computed from *absolute* or *normalized* boredom scores. The data set with *normalized* ground truth contains 4398 instances, with 1518 (34.5%) instances classified as *bored*, and 2880 (65.5%) instances classified as *baseline*. This distribution is well aligned with the values reported from boredom assessments

by Goetz *et al.* [17], where participants considered themselves to be bored in one third of the responses. In contrast, the data set with *absolute* ground truth has 446 (10.1%) instances classified as *bored*, hence it is less balanced.

Second, after having noticed that our scores might be affected by our participants' general proneness to boredom, and since psychological traits can be beneficial when detecting affect [3], we launched a post-hoc survey were we asked the participants to fill out the 28-item Boredom Proneness Scale (BPS) [13]. Since the study had been officially over, participants didn't have to gain anything from doing this extra work. Still, 22 of the 54 participants completed the survey and allowed us to optionally add a boredom *proneness* score as new feature to each of their self-reports. Our recent related work shows that it may not even be required to obtain the *proneness* from self-reports, since it can be estimated from average daily phone usage [24].

Our *primary data set* uses *normalized* ground truth and *no boredom proneness* information. The rationale is that this modeling choice increases the applicability of the model: it does not require to obtain boredom proneness scores before its deployment and it is tailored to detect deviations in boredom even for people who are not prone to be bored.

Classifier Selection

We used a variety of classification methods to empirically evaluate their performance in our problem setting. In particular, we compared three widely used classifiers: 1) L2-regularized Logistic Regression (LR) [19], as an example of a linear classifier; 2) Support Vector Machines with Radial Basis Functions kernel (SVM) [33], as an example of a non-linear classifier; and 3) Random Forests (RF) [6], as an example of ensemble learning.

We applied the same model-building methodology for the three classifiers. In particular, we used a nested cross-validation approach [8] in which an inner loop performs a grid search over the space of model hyper-parameters to select the best performing values, and an outer loop measures the performance of the model found in the inner loop. This strategy guarantees that in each step of the outer loop the fold being evaluated is not used during the training phase at any point, avoiding any positive bias when measuring the performance. In our implementation, we chose 10-folds for the outer loop and 5-folds for the inner loop.

The classifier that yielded the best performance was RF. Similarly to other ensemble learning methods, *e.g.* as Boosting and Bagging, Random Forests make use of multiple weak-learners and aggregate their results, looking at optimizing the bias-variance tradeoff. RF uses decision trees as the base classifiers, where randomization is introduced as several stages. First, each tree in the forest is constructed using a different bootstrapped sample of the training data. Second, each node of a tree is greedily split considering the best feature for only a random subset of all the variables. This process aims at removing correlations between the different trees in the forest, and helps reducing over-fitting. The resulting models are inherently non-linear, tolerate the presence of outliers and have implicit support for categorical variables.

Classification Performance

Figure 3 depicts the classification performance results for the best performing approach (RF). The performance metric we report is AUCROC, area under the ROC (Receiver Operating Characteristic) curve, typically used to replace the standard classification accuracy metric for unbalanced datasets, as it is our case.

Figure 3 shows the classification performance for the 4 data sets. The *absolute* ground truth yields consistently higher performance than using the *normalized* ground truth. Boredom *Proneness* slightly reduces the variance.

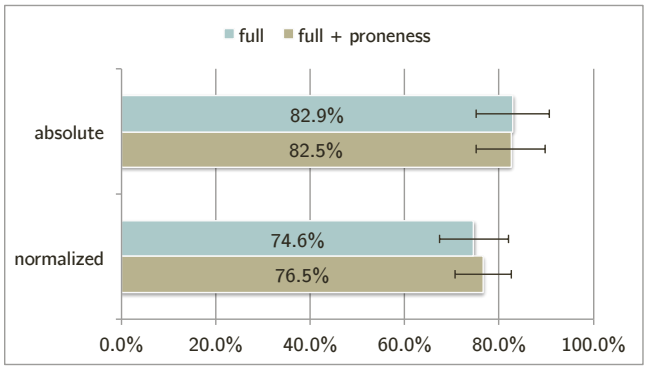

Figure 3. Area Under the ROC Curve (AUCROC) performance of the RF classifier with the different data sets.

Figure 4 shows the precision-recall curve for the *primary data set*. We observe that this model offers a high level of flexibility in choosing different classification thresholds to trade precision for recall, depending on the characteristics of the application setting. In general, for scenarios in which boredom detection is used to actively probe users, it is convenient to prioritize precision as to minimize the number of false positives (which may annoy users). In this sense, we can tune the model to get precision levels of 70.1% (for over 30% recall), or 62.4% (for 50% recall) in less restrictive scenarios.

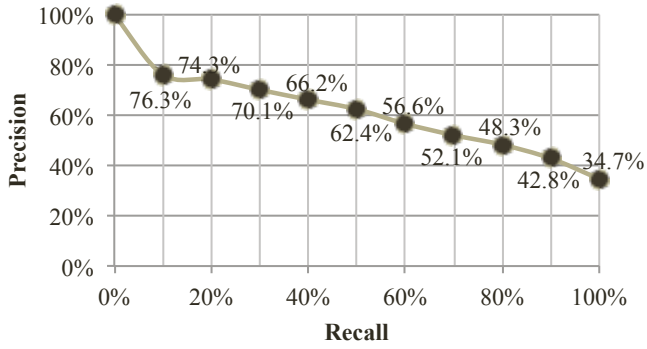

Figure 4. Precision-recall plot for the primary data set (normalized ground truth, no proneness).

Feature Analysis

Random forests can be used to rank features by their importance in the classifier, which provides useful insights about

the discriminative power of the features in the considered problem setting. *Mean Impurity Decrease* is the most common way to obtain feature importance from random trees [6]. It is computed by averaging across all the trees in the forest the amount of impurity removed by each feature while traversing down the tree, weighted by the proportion of samples that reached that node during training.

Using this method, we obtained the feature importances for all features in the primary data set. Table 5 depicts the top 20 features and their importance (Column: *Import*). Grouping these features in different categories –as depicted in Figures 3 and 4, we find that the most important categories of features are:

- **Recency of communication activity** expressed by the features regarding the last time that the user communicated via phone or SMS, and the last time since notifications arrived, as notifications were largely generated by applications from the communication category;
- **Intensity of recent usage** reflected by features such as the volume of internet traffic, # phone unlocks, and level of interaction with applications in the last five minutes;
- **General usage intensity** captured by *e.g.* battery drain, state of the proximity sensor (*i.e.*, whether the phone's screen is covered), or time since last phone use;
- **Context / time of the day** reflected by the hour of the day and the values of the light sensor[5]; and,
- **Demographics**, *i.e.* the participants' gender and age.

Feature	*Import*	*Correlation*	*The more bored, the ..*
time_last_outgoing_call	0.0607	-0.143	less time passed
time_last_incoming_call	0.0580	0.088	more time passed
time_last_notif	0.0564	0.091	more time passed
time_last_SMS_received	0.0483	0.053	more time passed
time_last_SMS_sent	0.0405	-0.090	less time passed
time_last_SMS_read	0.0388	-0.013	less time passed
light	0.0537	-0.010	darker
hour_of_day	0.0411	0.038	later
proximity	0.0153	-0.186	less covered
gender (0=f, 1=m)	0.0128	0.099	more male (1)
age	0.0093	n.a.	+20s/40s, -30s
num_notifs	0.0123	0.061	more notifications
time_last_notif_cntr_acc	0.0486	-0.015	less time passed
time_last_unlock	0.0400	-0.007	less time passed
apps_per_min	0.0199	0.024	more apps per minute
num_apps	0.0124	0.049	more apps
bytes_received	0.0546	-0.012	less bytes received
bytes_transmitted	0.0500	0.039	more bytes sent
battery_level	0.0268	0.012	the higher
battery_drain	0.0249	-0.014	the lower

Table 5. Most important features in the primary data set sorted by their Mean Impurity Decrease score. More positive (blue) correlation values should be interpreted as "the higher the value the more bored".

Relation of Boredom and Top Features

To understand which usage patterns are related to boredom, we computed the direct, global relationship between the most important features and boredom. We trained a Linear-Regression Model and analyzed the sign (positive or negative) that it assigned to each of the top-20 features. Table 5 visualizes the relation of the top-20 features in the *Correlation* column.

Our participants tended to be more bored the **more time** had passed since **receiving phone calls, SMS, or notifications**, and the **less time** had passed since **making phone calls** and **sending SMS**. However, the **volume of notifications** received in the last 5 minutes is likely to be **higher** when being bored.

Being bored is also correlated with more phone use: the **screen** was **less** likely to be **covered** (which, for example, happens when the phone is stowed away), **more apps** were used, the **last unlocking** and **checking** for new notifications happened **more recently**, and the **volume of data uploaded** was **higher** when our participants were bored. Interestingly, the amount of **data download** and **battery drain** were **lower** when people were bored.

Related to demographics, **male** participants tended to be more bored than females, and boredom was **higher** for participants in their **20s** and **40s** and lower in their **30s**.

Boredom was more likely the **later** it was in **the day** and the **darker** the ambient lighting conditions.

Finally, apps that most strongly correlated with being bored were Instagram, email, settings, the built-in browser, and apps in the 'other' category. Apps that correlated most strongly with not being bored were communication apps, Facebook, SMS, and Google Chrome.

Please note that this analysis only reflects direct correlations between the features and boredom. Some features may not have a direct relationship with boredom levels, but may become indicative when combined with other features. Further, due to our observational approach, causal interpretations, such as incoming phone calls mitigating boredom, would only be speculative, hence we omitted them.

ONLINE-VALIDATION PILOT STUDY

To validate our third hypothesis that mobile phone users are more open to consume suggested content on their phones when they are bored (**RQ3**), we conducted an in-the-wild pilot study. We released a new version of *Borapp* called *Borapp2* that suggested to visit a popular "news" website that offers typically short, entertaining articles and which has been defined by its founder as designed to help people to combat boredom. Our hypotheses were that:

- **H1**: Mobile phone users are more likely to interact with suggested content when bored (as inferred by our machine learning algorithm), and
- **H2**: Mobile phone users spend more time interacting with suggested content when bored (as inferred by our algorithm).

[5]The same physical sensor returns the ambient light levels and whether the phone screen is covered

Methodology

Participants

For this validation study, we recruited 16 participants who were different from the participants in the first study. They were recruited through a large mailing list of volunteers to participate in research studies. They installed *Borapp2*, an updated version of *Borapp*, on their Android phones. Their ages ranged from 16 to 51 (Mdn = 39, M = 36.31, SD = 9.37). According to their locales and the time zones reported by the phones, the large majority of the participants were Spanish speakers living in Central Europe. Note that none of these participants had participated in the first *Borapp* study.

Apparatus

To infer boredom, *Borapp2* implements an online boredom detection module. It computes the previously described features on-the-fly and feeds them into an online instance of the RF classifier. The classifier is trained with the data obtained in the first *Borapp* study, described in the previous section.

Instead of experience-sampling probes, *Borapp2* creates notifications that suggests to open the *Buzzfeed*[6] news app. Buzzfeed describes itself as providing "the most shareable breaking news, original reporting, entertainment, and video across the social web"[7]. As shown in Figure 5, the notification showed the title of the most recent article and suggested the user to click-to-read. We chose the Buzzfeed app as suggested content, because (1) the app caches articles, so that the study did not rely on permanent availability of an internet connection, and (2) its content is designed to be interesting to a broad audience.

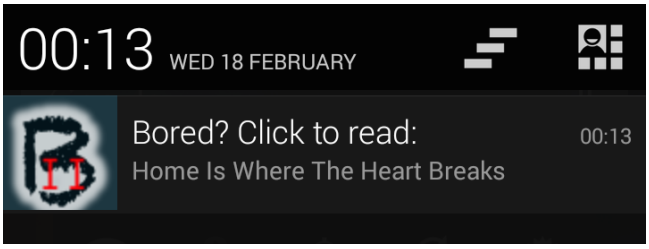

Figure 5. Example of a notification suggesting to visit the Buzzfeed app.

Whenever the user turned on the screen, *Borapp2* tested a number of conditions, such as whether a notification was already scheduled or whether enough time (30 minutes) had passed since the last notification. When those conditions were satisfied, the app scheduled a notification with a delay ranged from 10 sec - 5 min. At the end of this interval, the classifier was run to estimate whether the user was bored or not. When the user was detected to be bored, the notification was posted. If the classification result was the opposite, the notification was posted in only 1 out of 9 cases. Through empirical analysis during the testing phase, we found that this roughly yielded the same amount of notifications for each of the two boredom conditions (bored and non-bored). The notifications disappeared if they remained ignored by the user for more than 5 minutes.

[6]www.buzzfeed.com
[7]http://www.buzzfeed.com/about

Design

The online detection of boredom served as independent variable with two values: *bored* and *normal*. The *bored* value served as experimental condition, whereas the *normal* value was the control condition. We used repeated-measures design, *i.e.*, participants would be exposed to notifications in both conditions. Since the inferred emotional state cannot be specified by the researcher, the study has to be considered as quasi-experiment rather than a truly randomized controlled trial.

The participants' reaction to the notification was used as dependent measure. We computed two scores:

- **Click-ratio**: which is defined as the number of notifications clicked in a condition divided by the total number of notifications presented in this condition. This score was designed to validate hypothesis 1.
- **Engagement-ratio**: which is defined as *click-ratio*, but in addition to clicking the notification, Buzzfeed also had to be kept open for at least 30 seconds. This score was designed to validate hypothesis 2.

Procedure

The procedure to join and set up the study was similar to the procedure in the first study. Participants installed *Borapp2* from Google Play. The app then walked the participants through the informed consent and the setup steps, including the installation of the Buzzfeed app. Each participant had to keep *Borapp2* actively running for at least 14 days. Afterwards, we collected the participants' basic demographics with a questionnaire and debriefed participants. As reward, we conducted a raffle of one 300 EUR gift certificate of a large online retailer.

Results

Many of our participants had not used Buzzfeed before. To make sure that no novelty bias would affect our results, we removed the first 2 days of the data.

Over the remaining 12 days, our participants received 941 notifications (M = 60.81, SD = 38.27) suggesting to open a piece of content from Buzzfeed. Our participants were predicted to be bored by our algorithm in 48.0% of the cases.

A Shapiro-Wilk test showed that neither of the scores were normally distributed, hence results were analyzed using non-parametric statistics. We used the median as measure of central tendency, the median absolute deviation (MAD) as replacement for the standard deviation, and the 4th and 13th rank as approximation of the 95% confidence interval, following the procedure described in [35]. Significance was tested by using the Wilcoxon-Signed Rank Test.

Click-ratio (H1)

Figure 6 shows the average *click-ratio* per condition. In the *bored* condition, individual *click-ratio* scores ranged from 0% to 71% (**Mdn = 20.5**, MAD = 13, Q = 8.25 – 31.5, CI_{95} = 6 – 42). In the *normal* condition, individual *click-ratio* scores ranged from 0% to 45% (**Mdn = 8**, MAD = 8, Q = 1.5 – 21, CI_{95} = 0 – 30). The difference is statistically significant (z = −2.102, **p = .018**) and the effect size large (**r = -.543**).

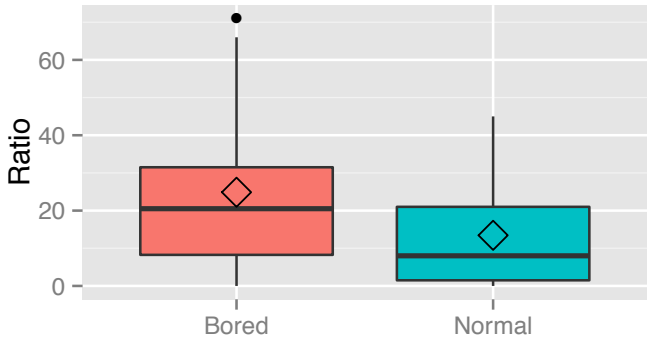

Figure 6. Click-ratio per condition.

Engagement-ratio (H2)

Figure 7 shows the average *engagement-ratio* per condition. In the *bored* condition, individual *engagement-ratio* scores ranged from 0% to 66% (**Mdn = 15**, $MAD = 12$, $Q = 3.5 - 21.25$, $CI_{95} = 2 - 31$). In the *normal* condition, individual *engagement-ratio* scores ranged from 0% to 34% (**Mdn = 4**, $MAD = 4$, $Q = 0 - 10$, $CI_{95} = 0 - 10$). Again, the difference is statistically significant ($z = -2.102$, **p = .018**) and the effect size large (**r = -.511**).

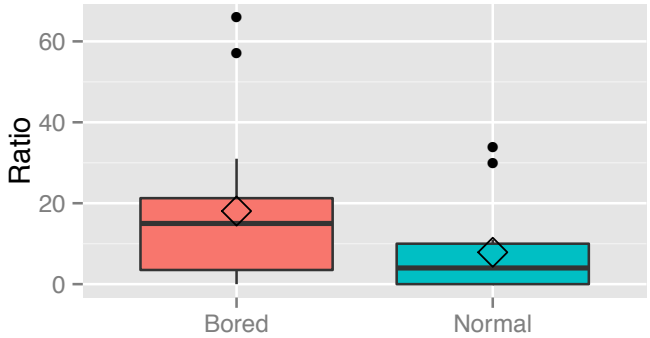

Figure 7. Engagement-ratio per condition.

This evidence supports both hypotheses. Participants were more likely to open Buzzfeed (**H1**) and engage with it (**H2**) when they were predicted to be bored by the model. We found strong effect sizes for both dependent measures.

DISCUSSION

Our two field studies provide empirical evidence that (1) it is possible to infer boredom from phone usage patterns with acceptable accuracy; (2) boredom is related to communication activity, usage intensity, hour of the day, and demographics; and (3) mobile phone users are more likely to engage with suggested content when they are predicted to be bored.

RQ1: Inferring Boredom

The data from our first field study with 54 participants shows that it is possible to create a machine-learning model to automatically detect when people are bored while using their phone.

We tested the impact of two additional factors. First, the ground truth being *normalized* vs. *absolute* self-report scores. Second, including a *boredom proneness* score obtained via a validated questionnaire.

The performances of the machine learning models range from **74.6%** AUCROC for the *primary data set* with *normalized* ground truth and *no boredom proneness* to **82.9%** AUCROC for the data set with *absolute* ground truth and *boredom proneness*. When tuning to a recall of 50%, a model trained with the *primary data set* achieves a precision of 62.4% for the *bored* class.

Performance was consistently higher for the *absolute* ground truth. Our interpretation is that the definition of absolute boredom is more agreed upon and it seems that the level of boredom needs to be higher in order to agree with a statement of feeling bored. Normalized boredom, while harder to infer, reflects the fact that everybody has a different boredom proneness, hence, a different reference point for boredom. While using normalized boredom makes boredom detection less accurate, it allows to also detect when people slightly deviate from their baseline state towards boredom, even if they would not wholeheartedly agree to the statement 'I feel bored'.

Including boredom proneness scores had only slight effects on the models' accuracies. It decreased the variance for both ground truths, which means that it helped to make the models more stable. However, the effect was not as pronounced as by Biller and D'Mello [3]. A reason might be that our analysis relies on scores from only 22 (40.7%) participants. However, recent related work shows that boredom proneness does not necessarily have to be collected via self-reports, but that it can be estimated from daily mobile phone usage patterns [24].

Limitations arise from the fact that the study took place *in-the-wild* in an unsupervised setting. These studies trade the ability to control for ecological validity. Our participants were free to dismiss probes as long as they responded to 84 in total. This may create some bias, as we had no control over when our participants dismissed probes. They might have dismissed more probes when they were deeply engaged or busy with other things. Hence, in particular the results with the absolute ground truth might emphasize on events where the user is bored.

Yet, the accuracy of the models are significant. Previous works that used a similar approach to detect daily happiness [5] or daily stress levels [4] from mobile phone use, where not able to achieve acceptable prediction accuracy from mobile phone use data alone. They had to include personality traits and weather information. This comparison indicates how closely interwoven boredom and mobile phone use have become.

RQ2: Strong Indicators of Boredom

The features identified as the strongest indicators of boredom were related to five aspects: the recency of communication activity, the intensity of recent usage, general usage intensity, the context (hour of the day and proximity sensor), and basic demographics.

Boredom correlated with more time having passed since the last incoming communication, and less time passed since the

user last initiated outgoing communication via calls, SMS, and messages. This finding suggests that being contacted by others is generally correlated with being less bored. Contacting others, however, is more likely to happen while being bored.

Boredom further correlated with the intensity of mobile phone use. In general, we found that the higher the usage intensity, the higher the boredom. This confirms observations in previous work [7, 25] that people use their mobile phones when bored to kill time. Our work advances these previous findings by providing empirical evidence that this increase in usage contributes to the detection of boredom or phases of killing time.

Furthermore, boredom positively correlated with the time of the day and darker ambient lighting conditions. This finding means that there are boredom levels vary throughout the day, as shown in [23]. Moreover, in contrast to Mark *et al.* [23] who found that boredom is lower during late working hours, our results include after-work hours, and indicate that boredom tends to increase as the day progresses.

Finally, boredom correlated with demographics. Boredom tended to be higher for male participants, and higher for participants in their 20s and 40s and lower in their 30s. This findings are in line with previous work which found that age [36] and gender [2] are significant predictors of experiencing boredom in leisure time.

One limitation here is that our analysis exclusively yields correlational results. While these allow to learn which usage patterns co-occur with boredom and allow inferences, we cannot establish a causal relationship between predictive usage patterns and boredom.

Yet, not all of the most important features are related to actually using the phone. We also learn that contextual factors, such as the time of the day, and demographics play a role.

RQ3: Boredom and Consumption of Suggested Content

The main motivator for our second field study was to answer our third research question: are people more open to suggested activities and content when bored? The study provides evidence to answer this question affirmatively: our participants were significantly more likely to open and engage with suggested content on their mobile phones when our algorithms predicted them to be bored.

The interpretation of these findings has to take important factors into account. First, the user study was not a true experiment, but a quasi-experiment, since the conditions of being bored or not could not be randomly assigned but occurred naturally. Second, the tested sample was rather small and somewhat biased (by self-selection). Third, the conditions were not pure given the error rate of the boredom inference algorithm (40% as per the results of the first user study). Hence, the findings should be regarded as preliminary.

The interesting aspect is that the notification we posted was not related to any communication activity. Our previous work [29, 32] shows that mobile phone users find notifications from communication apps (messengers, email, social networks) to be important, while notifications from other types of apps are largely being ignored and at times even perceived as annoying.

Thus, these findings are significant, as they show that automatically-detected boredom may be an ideal way to deal with peoples' increasingly scarce attention. We envision it's application in *boredom-triggered proactive recommendations* [28], an approach to increase the success rates of proactive recommendations regardless of the content.

CONCLUSIONS

In this paper, we have proposed a machine learning method to automatically infer boredom from mobile phone usage, context and demographics. In an in-the-wild study with 54 participants, our models have reached accuracies ranging from 74.6 to 82.9% AUCROC. We have also studied the most predictive features and found that recency of communication, usage intensity, time of the day, and demographics are the categories of features with the highest discrimination power. Furthermore, in a second in-the-wild study we have found that users are more likely to engage with suggested content on their phones when they are bored.

We believe that boredom-triggered proactive recommendations open the way towards the design mobile recommender systems that have a better understanding of when and how engage with their users. Potential application scenarios we envision are:

(1) Engage with their users by providing interesting suggestions (*e.g.* videos, activities, contacting friends) in moments of boredom;

(2) suggest useful but not necessarily boredom-curing activities (*e.g.* clear the backlog of a todo or read-later list) instead of lackluster *killing time* activities; and

(3) help to make positive use of boredom by fosting introspection, reflection, and creativity.

While this work has provided first evidence that notifications delivered during phases of inferred boredom can drive engagement with an entertaining news website, future work needs to be carried out in order to provide stronger statistical proof that this effect can be observed in different settings as well. In addition, we plan to carry out future work on how to intervene *in-situ* when mobile phone users are detected to be bored, including the exploration of points (2) and (3) above.

ACKNOWLEDGMENTS

We thank the people who participated in this study.

REFERENCES

1. Avrahami, D., and Hudson, S. E. Responsiveness in instant messaging: predictive models supporting inter-personal communication. In *Proc. CHI '06*, ACM (2006).
2. Barnett, L. A., and Klitzing, S. W. Boredom in free time: Relationships with personality, affect, and motivation for different gender, racial and ethnic student groups. *Leisure Sciences 28*, 3 (2007), 233–244.
3. Bixler, R., and D'Mello, S. Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. In *Proc. IUI '13*, ACM (2013).
4. Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., and Pentland, A. S. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proc. MM '14*, ACM (2014).
5. Bogomolov, A., Lepri, B., and Pianesi, F. Happiness recognition from mobile phone data. In *Proc SOCIALCOM '13*, IEEE Computer Society (2013).
6. Breiman, L. Random forests. *Machine Learning 45*, 1 (2001), 5–32.
7. Brown, B., McGregor, M., and McMillan, D. 100 days of iphone use: Understanding the details of mobile device use. In *Proc MobileHCI '14*, ACM (2014).
8. Cawley, G. C., and Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res. 11* (Aug. 2010), 2079–2107.
9. Cherubini, M., and Oliver, N. A refined experience sampling method to capture mobile user experience. In *Proc. Workshop of Mobile User Experience part of CHI '09* (2009).
10. Consolvo, S., and Walker, M. Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing 2*, 2 (Apr. 2003), 24–31.
11. Davenport, T. H., and Beck, J. C. *The Attention Economy: Understanding the New Currency of Business*. Harvard Business Press, 2002.
12. Eastwood, J. D., Frischen, A., Fenske, M. J., and Smilek, D. The unengaged mind: Defining boredom in terms of attention. *Perspectives on Psychological Science 7*, 5 (2012), 482–495.
13. Farmer, R., and Sundberg, N. D. Boredom proneness: The development and correlates of a new scale. *J. Pers. Asses 50*, 1 (1986), 4–17.
14. Fenichel, O. *On the psychology of boredom*. Columbia University Press, 1951.
15. Fisher, C. Boredom at work: A neglected concept. *Human Relations 46*, 3 (1993), 395417.
16. Fogarty, J., Hudson, S. E., Atkeson, C. G., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J. C., and Yang, J. Predicting human interruptibility with sensors. *ACM Trans. Comput.-Hum. Interact. 12*, 1 (Mar 2005), 119–146.
17. Goetz, T., Frenzel, A. C., Hall, N. C., Nett, U. E., Pekrun, R., and Lipnevich, A. A. Types of boredom: An experience sampling approach. *Motivation and Emotion 38*, 3 (2014), 401–419.
18. Guo, Q., Agichtein, E., Clarke, C. L. A., and Ashkan, A. In the mood to click? towards inferring receptiveness to search advertising. In *Proc. WI-IAT '09*, IEEE Computer Society (2009).
19. Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
20. Horvitz, E., Koch, P., Sarin, R., Apacible, J., and Subramani1, M. Bayesphone: Precomputation of context-sensitive policies for inquiry and action in mobile devices. In *Proc UM '05* (2005).
21. Iqbal, S. T., and Bailey, B. P. Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Trans. Comput.-Hum. Interact. 17*, 4 (Dec 2010), 15:1–15:28.
22. LiKamWa, R., Liu, Y., Lane, N. D., and Zhong, L. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proc. MobiSys '13*, ACM (2013).
23. Mark, G., Iqbal, S. T., Czerwinski, M., and Johns, P. Bored mondays and focused afternoons: The rhythm of attention and online activity in the workplace. In *Proc. CHI' 14*, ACM (2014).
24. Matic, A., Pielot, M., and Oliver, N. Boredom-computer interaction: Boredom proneness and the use of smartphone. In *Proc. UbiComp '15*, ACM (2015).
25. Oulasvirta, A., Rattenbury, T., Ma, L., and Raita, E. Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing 16*, 1 (2012), 105–114.
26. Picard, R. W., Vyzas, E., and Healey, J. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell. 23*, 10 (Oct. 2001), 1175–1191.
27. Pielot, M. Large-scale evaluation of call availability prediction. In *Proc. UbiComp '14* (2014).
28. Pielot, M., Baltrunas, L., and Oliver, N. Boredom-triggered proactive recommendations. In *Smarttention, Please! Intelligent Attention Management on Mobile Devices Workshop @ MobileHCI '15*, ACM (2015).
29. Pielot, M., Church, K., and de Oliveira, R. An in-situ study of mobile phone notifications. In *Proc. MobileHCI '14* (2014).
30. Pielot, M., de Oliveira, R., Kwak, H., and Oliver, N. Didn't you see my message? predicting attentiveness to mobile instant messages. In *Proc. CHI '14*, ACM (2014).

31. Rosenthal, S., Dey, A. K., and Veloso, M. Using decision-theoretic experience sampling to build personalized mobile phone interruption models. In *Proc. Pervasive '11*, Springer-Verlag (2011).

32. Sahami Shirazi, A., Henze, N., Dingler, T., Pielot, M., Weber, D., and Schmidt, A. Large-scale assessment of mobile notifications. In *Proc. CHI '14*, ACM (2014).

33. Vapnik, V. N. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

34. Vodanovich, S. J. On the possible benefits of boredom: A neglected area in personality research. *Psychology and Education: An Interdisciplinary Journal* (2003).

35. Wade, A., and Koutoumanou, E. Statistics & research methodology. https://epilab.ich.ucl.ac.uk/coursematerial/statistics/non_parametric/confidence_interval.html, 2010.

36. Weissinger, E., Caldwell, L. L., and Bandalos, D. L. Relation between intrinsic motivation and boredom in leisure time. *Leisure Sciences 14*, 4 (1992), 317–325.

37. Wilson, T. D., Reinhard, D. A., Westgate, E. C., Gilbert, D. T., Ellerbeck, N., Hahn, C., and C. L. Brown, A. S. Just think: The challenges of the disengaged mind. *Science 345*, 6192 (2014), 75–77.