A Machine Learning Approach to Boredom Detection using Smartphone's Sensors

Bruno Fernandes
Department of Informatics
ALGORITMI Centre
Braga, Portugal
bruno.fmf.8@gmail.com

Carlos Campos
Department of Informatics
ALGORITMI Centre
Braga, Portugal
a74745@alunos.uminho.pt

José Neves
Department of Informatics
ALGORITMI Centre
Braga, Portugal
jneves@di.uminho.pt

Cesar Analide
Department of Informatics
ALGORITMI Centre
Braga, Portugal
analide@di.uminho.pt

University of Minho

Abstract

Boredom is a subjective feeling that tends to emerge in situations where there is a lack of stimuli, being a state of relatively low arousal and dissatisfaction. With the massive use of smartphones worldwide, people found new ways to fight boredom. Indeed, it is safe to assert that people tend to use their smartphone, if available, in situations where boredom arises. Hence, the goal of this work is to detect such feeling when using the smartphone, without any kind of biometric data and using only data from smartphones' sensors. To achieve this goal, a mobile application was conceived and made available online, with the assembled dataset containing samples that categorised the presence, or not, of boredom using the experience sampling method. After exploring and pre-processing the dataset, a set of candidate Machine Learning models were conceived, tuned and evaluated, with the best one, a Random Forest, attaining a promising performance in terms of accuracy, precision and recall for boredom detection.

1 Introduction

Smartphones are globally used devices, regardless of age, region or status. These devices have so many practical advantages that listing all of them would be an endless exercise. Boredom, on the other hand, may be defined as a state of relatively low arousal and dissatisfaction, which tends to emerge in situations where there is a clear lack of stimuli [MV93]. In other words, boredom may be a sign that the current activity is not sufficiently satisfying

In: Rudolph S., Marreiros G. (eds.): Proceedings of STAIRS, Santiago de Compostela, Spain, august 2020, published at http://ceur-ws.org

neither motivating. Common sense tells us that, when bored, people tend to use their smartphone. Indeed, it has been reported that bored people tend to manifest consumerism habits, which makes them a better candidate to subscribe promotional campaigns and engage with new content [SKK09]. On the other hand, bored people could also make a more productive use of such idle moments, taking the opportunity to complete old tasks or cultivate knowledge.

The main goal of this work is to detect boredom when using the smartphone, using only data that is made available by smartphones' sensors, without considering any biometric neither intrusive data. This could open new ways to entertain people, to provide actionable information and to create more robust recommender systems, which may vary from recommending tasks from *to-do* lists to targeted advertising, among many others. A research question has been elicited and stands as follows, viz., *Is it possible to accurately detect boredom using Machine Learning models and the smartphone sensors as base data?*.

To achieve the proposed goal, it is of the utmost importance the existence of a dataset from where one could conceive and evaluate several candidate models. Therefore, the first step was to conceive a mobile application for data collection, with data being collected at specific intervals, with the user being asked how bored he felt during that same period - with this, we were classifying/labelling the collected data. This is also known as the Experience Sampling Method, a research method that probes participants to report on their thoughts or feelings on multiple occasions over time. Three distinct Machine Learning (ML) models were then conceived and evaluated on the processed data, in particular Decision Trees (DTs), Random Forests (RFs) and Support Vector Machines (SVMs).

The remaining of this paper is structured as follows, viz., the next section describes the state of the art regarding boredom prediction in mobile and non-mobile devices; the subsequent section contains a description of the used materials, the methods and the software developed for data collection; later, the performed experiments are explained, with results being gathered and discussed; finally, conclusions are drawn and future work is outlined.

2 Literature Review

Literature shows that researchers have already engaged on studying ways to measure and detect boredom as well as other feelings, such as stress [SP13], fatigue [PCNN16] or happiness [BLP13]. Initial attempts to detect boredom focused on specific data collected by specific sensors that must be carried at all times for continuous monitoring, representing a major limitation [BD13]. Recent times came, however, with promising results regarding the use of computers and smartphones to detect boredom.

In 2013, Bixler and D'Mello focused on the writing periods at a computer to predict boredom, using DTs and Naive-Bayes classifiers, achieving satisfactory results [BD13]. Predictions were based on a set of values obtained with the help of a keystroke analysis, task appraisals and personality traits. It should be noted that the naive-bayes classifier obtained an accuracy of 82%, and the RF obtained a precision of 87%, when trying to predict boredom through the interaction of the person with the machine. Still considering computers, Guo et al. (2009) found, with the help of SVMs, that web interaction events, such as mouse movements or the number of clicks on a page, allow the prediction of whether a person is willing to be distracted, which may be indicative of boredom [GACA09]. In 2019, Seo et al. developed an Artificial Neuronal Network (ANN) that was able to classify boredom using data from electroencephalography and galvanic skin response exams, achieving an interesting accuracy of 79.98% [SLS19a]. A very similar study of the same authors focused only in data from electroencephalography exams, obtaining greater precision with a K-Nearest Neighbor based-model [SLS19b].

Considering smartphones, one important attempt has been made to use smartphone's data to predict user boredom. In 2015, Pielot et al. focused, among others, on what aspects of mobile phone usage are the most indicative of boredom. The authors developed a mobile application for data collection which run during 14 days [PDPO15]. The collected data included, among others, user's demographic information. The problem was framed as a binary classification one, i.e., whether the user is bored or not. Three distinct classifiers were used, with RFs being the one showing the best accuracy (when compared to L2-regularized Logistic Regression and SVMs). On the other hand, the overhead between precision and recall is significant, since precision levels limited the recall to 30% [PDPO15].

In addition to boredom, studies were also carried out in relation to other feelings, such as the level of stress [SP13], happiness [BLP13] or fatigue [PCNN16]. Sano and Picard (2013) have reported good results for stress recognition using a combination of mobile devices and clothing-associated sensors, despite some restrictions, such as the limited number of subjects and data [SP13]. Another interesting research, conducted by Bogomolov

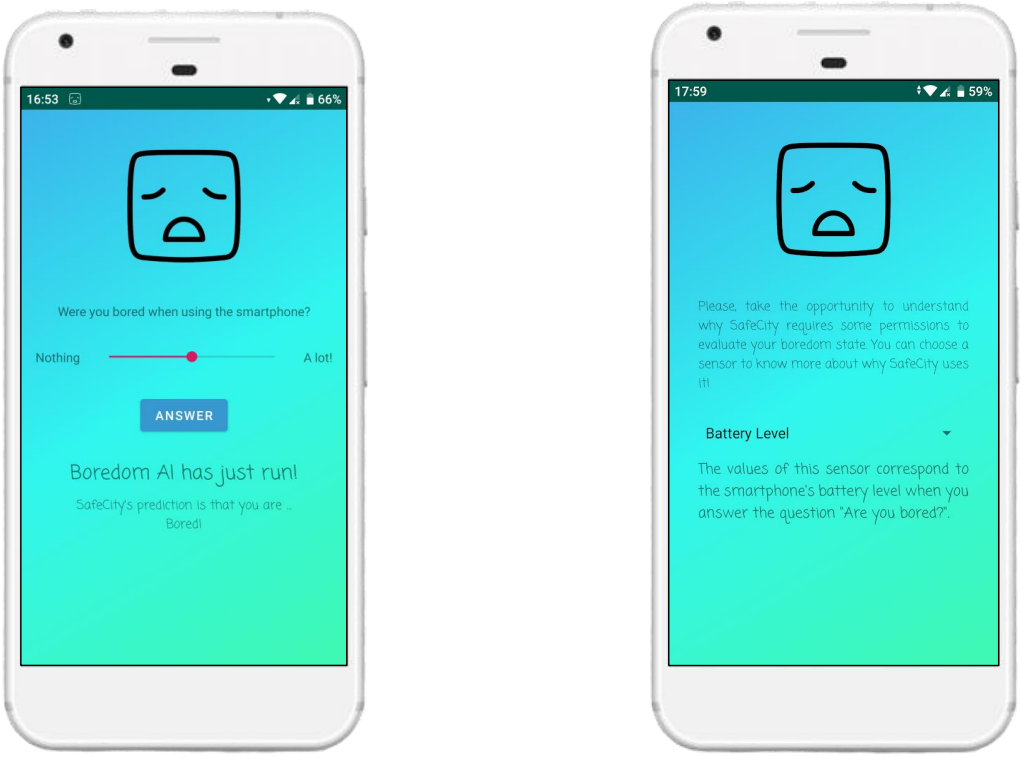
et al. (2013), concluded that it was possible to accurately recognise the happiness of an individual during is daily affairs, using a large set of features, obtained with the help of smartphones, such as communication data, proximity sensors information, bluetooth connection, weather data and personal characteristics [BLP13]. In 2016, Pimenta et al., developed an ANN to detect mental fatigue based on the user's interaction patterns with the computer (mouse and keyboard), showing that when users claim to feel mentally fatigued, they use these peripherals differently [PCNN16].

3 Materials and Methods

The next lines describe the materials and methods used in this work, including the collected dataset, which was created from scratch and contains real world data, as well as all the applied treatments.

3.1 Data Collection

To build the dataset, an Android mobile application was developed and made available at Google's Play Store[1] (Figure 1). It targets Android users with, at least, Android 6.0 (API 23). It is available in both Portuguese and English languages. In its essence, it consists of a set of services, broadcast receivers and sensors listeners that query both physical and soft sensors of the user's smartphone, periodically. It then asks the user to classify his boredom level in order to label the collected data. If the user does not answer, or if new measurements are recorded, previous unlabelled measurements are discarded. The user may decide, at any time, to turn on, or off, the boredom service, having also available an activity that describes all used sensors. The user may also choose which permissions he grants to the app. The mobile app will collect all possible data, respecting all permissions.



(a) Answering to "How Bored Do You Feel?".

(b) Rationale of all used sensors.

Figure 1: Mobile application for data collection.

The problem was initially framed as a regression one, i.e., the user defines his state of boredom to be of a value between 0 and 100, where 0 means "Absolutely Not Bored" and 100 means "Absolutely Bored". Figure 1 depicts

[1] https://play.google.com/store/apps/details?id=com.plugable.safecity

the main activities of the mobile app for data collection, including the question "How bored do you feel?" (Figure 1a) as well as the explanation of the collected data (Figure 1b). It is important to emphasise that no biometric neither demographic data is collected, neither it is possible to link a record of the dataset to the person who answered it. In other words, only the sensors' values and the boredom level are stored, there is no user identifier. In addition, in order to be as less intrusive as possible, we opt not to implement the *process_outgoing_calls* and *read_sms* android permissions.

3.2 Data Exploration

The collected dataset consists of a total of 1511 observations, ranging from December 03rd, 2019 to February 16th, 2020. Table 1 depicts all available features. No missing values are present. There are, however, failed sensors' readings filled with the value -1. Indeed, both the *ambientTemperatureSensor* and the *humiditySensor* features are completely filled with -1, meaning that none of the used smartphones had available such information.

Table 1: Features available in the collected dataset.

#	**Feature**	#	**Feature**	#	**Feature**
1	accelerometerSensor	16	gyroscopeSensor	31	orientation
2	ambientTemperatureSensor	17	homeButtonPress	32	otherNotifications
3	audioJack	18	hourFirst	33	outgoingCalls
4	batteryLevelFirst	19	hourSecond	34	pressureSensor
5	batteryLevelSecond	20	humiditySensor	35	proximitySensor
6	bluetoothConnection	21	incomingCalls	36	recentButtonPress
7	bored	22	isCharging	37	ringerMode
8	chattingNotifications	23	lightnessSensor	38	screenActivations
9	currentNotifications	24	magneticSensor	39	smsReceived
10	dayFirst	25	minuteFirst	40	socialNotifications
11	dayOfWeekFirst	26	minuteSecond	41	timestamp
12	dayOfWeekSecond	27	mobileDataSensor	42	weekend
13	daySecond	28	monthFirst	43	wifiSensor
14	flightMode	29	monthSecond	44	yearFirst
15	gravitySensor	30	notificationsRemoved	45	yearSecond

With all features assuming a non-Gaussian distribution (under the Shapiro-Wilk test with $p < 0.05$), the non-parametric Spearman's rank correlation coefficient was used. In this dataset there are a few pairs of highly-correlated features. In particular, the *wifiSensor* and *mobileDataSensor* features are negatively correlated, meaning that when a person has wireless connection enabled he tends to have his mobile data connection disabled. The pairs *hourSecond* and *hourFirst*, *daySecond* and *dayFirst*, *dayOfWeekSecond* and *dayOfWeekFirst*, *batteryLevelSecond* and *batteryLevelFirst*, and *year* and *month* have, as expected, an almost perfect correlation.

The target, i.e., the bored feature, is, in its original form, imbalanced. Figure 2a depicts the cardinality of observations per bored values in 20 bins of equal width. The great majority of observations, 388, fall into the first bin, which ranges from 0 to 5 (no boredom). The bin comprising bored values from 50 to 55 (somewhat bored) contain 274 observations. The third most frequent bin is the last one, from 95 to 100 (absolutely bored), with 193 observations. On the other hand, if to divide the bored feature into two bins of equal width (Figure 2b), both ones would contain approximately the same amount of observations (≈ 750).

The correlation between the independent features and the dependent one was analysed using the F-test, which can assess multiple coefficients simultaneously. In this test we aim to find the independent features that allow us to reject the null hypothesis that the fit of an intercept-only model and a linear model is equal, having $p < 0.05$. 11 features were found to be statistically significant, allowing us to reject the null hypothesis for each one of them. Among such features, the ones with most importance are *bluetoothConnection*, *mobileDataSensor*, *wifiSensor*, *weekend*, *currentNotifications* and *isCharging*. Since these features can, in someway, influence boredom, they are are good candidates to be used by the ML candidate models.

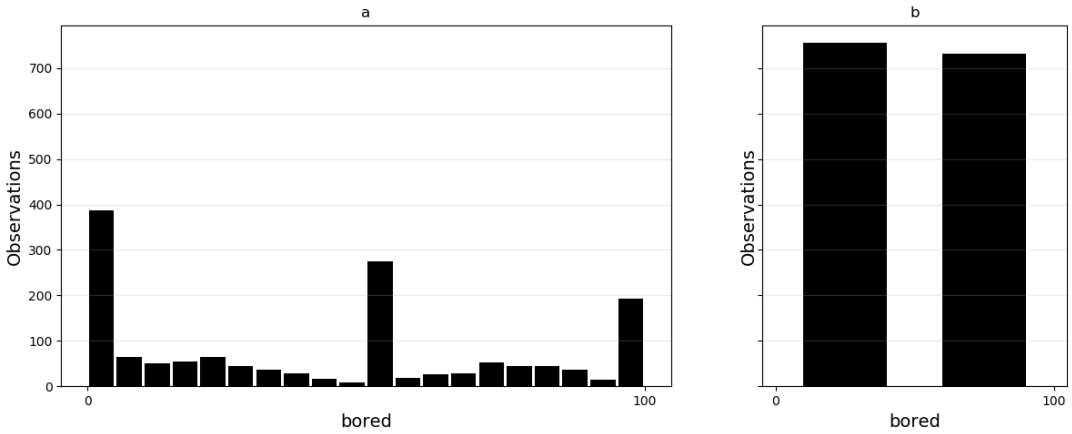


Figure 2: Histogram for the bored feature, using (a) twenty bins of equal width and (b) two bins of equal width.

3.3 Data Pre-processing

A set of methods were applied to clean and prepare the input data. The first step was to remove both the *ambientTemperatureSensor* and *humiditySensor* features. Afterwards, since the number of rows with -1 as year was reduced (23 rows), they were also removed. Since there were no variations between the pairs *yearFirst* and *yearSecond*, and *monthFirst* and *monthSecond*, the first were removed and the second were renamed to *year* and *month*, respectively. The month feature was then incremented by one since it was in the interval [0, 11]. An index, based on the timestamp up to the minutes, was then created, being sorted in ascending order by date.

As feature engineering, three new features were created. The first one, entitled as *batteryVariation*, was created based on the variation of the battery level values. It allows one to understand if the level of battery dropped, remained stable or even if the smartphone was charging. The second and third ones, *minutesInterval* and *hourInterval*, allows one to understand how many minutes and hours that particular observation encompasses, respectively.

Based on the correlation matrix analysis, eight features were dropped, in particular, *dayFirst*, *hourFirst*, *minuteFirst*, *year*, *month*, *otherNotifications*, *batteryLevelSecond* and *dayOfWeekFirst*. Then, some features were renamed, in particular, *dayOfWeekSecond* to *dayOfWeek*, *daySecond* to *day*, *hourSecond* to *hour* and *minuteSecond* to *minute*.

Even though DTs are not affected by any monotonic transformation to the input data, SVMs work better with normalized data. Data was, therefore, normalized, i.e., scaled, to the range [0, 1]. Since no categorical feature was present in the dataset, no feature encoding method was applied. Finally, the problem was framed as a binary classification one, i.e., the target feature, bored, was binned into "Not Bored" and "Bored" bins. As shown previously, with a binary classification problem the dataset gets well balanced, with 732 "Bored" observations and 756 "Not Bored" ones. The final dataset is made of 1488 observations with 35 features, including the target.

4 Experiments

The goal of this work is to develop and tune the best possible ML model to detect, i.e., predict, boredom when using the smartphone. Hence, to take advantage of the dataset's characteristics, three distinct ML models were experimented. Such models, besides behaving well in binary classification problems, also have the ability to generalise well in smaller datasets. All candidate models were tuned as described in the following lines.

4.1 Technologies and Libraries

Python, version 3.7, was the used programming language for data preparation and pre-processing as well as for model development and evaluation. Pandas, NumPy, scikit-learn, matplotlib and pickle were the used libraries. Knime, a free and open-source data analytics and ML platform, was also used. For increased performance, Tesla T4 GPUs were used. It is worth mentioning that this hardware is made available by Google's Colaboratory, a free python environment that requires minimal setup and runs entirely in the cloud.

4.2 Model Conception and Tuning

The first model, A (DT), was tuned in regard to used quality measure, the pruning method and the minimum number of records per node. The second one, Model B (RF), was tuned in regard to the used quality measure, the number of trees and tree depth. The last one, Model C (SVM), had its kernel, C value and gamma value tuned. Model A and C had their results validated under 10-fold cross validation, while Model B experienced 5-fold cross validation. Table 2 describes the search space for each hyperparameter of the candidate ML models.

Table 2: Search space for each hyperparameter of the candidate ML models.

Model	Hyperparameter	Search Space
(A) Decision Tree	Quality Measure	[Gain ratio, Gini index]
	Pruning	[No Pruning, MDL, Reduced Error]
	Minimum Number of Records	[1, 25] with a unitary step size
(B) Random Forest	Quality Measure	[Information Gain (Ratio), Gini index]
	Number of Trees	[100, 1000] with step size of 100
	Tree Depth	[-1] & [5, 25] with step size of 2
(C) SVM	Kernel	[Linear, Polynomial, RBF, Sigmoid]
	C	[0.001, 0.01, 0.1, 1]
	gamma	[0.001, 0.01, 0.1, 1, Auto, Scale]
	degree	[1, 7] with a unitary step size

5 Results and Discussion

The candidate models were evaluated in regard to their accuracy, precision and recall. The accuracy metric is quite straightforward, telling us the percentage of correctly classified observations. On the other hand, the precision and recall metrics allow an understanding, and measure, of relevance based on true and false positives and negatives. In other words, precision tells us how many of the observations classified as "bored" were, indeed, "bored" observations. Recall tells us, from all "bored" observations, how many did the model found (did the model found all the "bored" observations? How many observations did the model incorrectly classified as "not bored"?). Precision may be prioritised over recall if the goal is to reduce the number of false positives. Table 3 summarises the obtained results for the conceived models.

Model A had its best performance using Gini Index as quality measure, Minimum Description Length (MDL) as pruning method and 14 as stopping criteria. It achieved an accuracy of 0.653, a mean precision and a mean recall of 0.658. Model B, which took significantly more time to train than Model A, had its best performance with Information Gain as quality measure, 800 decision trees making the forest and a maximal tree depth of 25 levels. It achieved an accuracy of 0.684, an overall improvement of more than 3% when compared to Model A. In terms of mean precision and recall, it achieved a value of 0.709 and 0.706, respectively. An overall improvement of 5%. The fact that the dataset was well balanced, allowed the models to achieve interesting precision and recall values. In fact, the "Not Bored" class had a slightly better precision and recall than the "Bored" one. The better performance of Model B when compared to Model A can be explained by the fact that RF make use of several decision trees (800 in this case), making it an ensemble learning method that makes use of a simple yet powerful concept - the wisdom of crowds. The last Model, C, showed to have its best performance using a Polynomial Kernel, with a forth degree support function, a regularisation parameter of 1, and a scaled gamma of $1/(nrOfFeatures \times variance)$. The best candidate SVM model achieved an accuracy of 0.659, a mean precision of 0.615 and a mean recall of 0.661. Indeed, both SVMs and DTs had similar performances, with RFs outperforming both models.

Taking into account all data and model restrictions, the achieved results show a promising future with regard to the detection of boredom using only smartphone's sensor data, leaving aside all biometric and intrusive data, such as calls, cameras, location and all kinds of messages.

6 Conclusions and Future Work

Boredom detection, using a non-intrusive approach, may change the way we interact with smartphones, opening new ways for the recommendation of to-do tasks, books or even ways to grow one's knowledge upon situations

Table 3: Summary results for the conceived models.

Model	**Hyperparameters**			**Accuracy**
	Quality Measure	*Pruning*	*Number of Records*	
(A) Decision Tree	Gain ratio	No pruning	19	0.627
	Gain ratio	Reduced Error	19	0.629
	Gain ratio	MDL	17	0.633
	Gini index	No pruning	11	0.649
	Gini index	MDL	14	0.653
	Quality Measure	*Nr. of Trees*	*Tree Depth*	
(B) Random Forest	Information Gain Ratio	400	13	0.642
	Information Gain Ratio	300	-1	0.658
	Gini	500	-1	0.680
	Information Gain	800	-1	0.683
	Gini	400	15	0.683
	Information Gain	800	25	0.684
	Kernel	*C*	*gamma*	*degree*	
(C) SVM	RBF	1	0.1	-	0.575
	RBF	1	0.001	-	0.603
	Polynomial	0.1	1	3	0.631
	Polynomial	1	Scale	4	0.659

where there is lack of stimuli.

To achieve the proposed goal, a mobile application was developed and made available online. Users were probed using the Experience Sampling Method in order to quantify, in a scale of 0 to 100, how bored they felt in the corresponding period. Observations were then binned and framed as a binary classification problem ("Not Bored" and "Bored"), with bins having approximately equal width. The collected data was cleaned and pre-processed taking into account the three candidate models. As expected, the best RF candidate model achieved a better performance, in terms of accuracy, precision and recall, than the best DT and SVM one. The SVM showed relatively high accuracy, although it showed lower mean precision compared to the other models, being more sensitive to false positives. The overall results are in line with our expectations, given the fact that the dataset is still in an early stage and leaves aside all kinds of biometric and intrusive data. In addition, the conceived approach only considers smartphone sensors', eliminating the need for subjects to carry specific sensors or use computers. As direct answer to the elicited research question, it can be said that it is possible to conceive ML models that are able to detect, with promising accuracy and precision, user boredom using only the smartphone's sensors as base data. This can open new avenues to improve targeted advertising (in particular, its timing) and to improve time management.

As future work, the plan is to apply deep learning models to the collected dataset, which is growing daily. We also aim to re-frame the problem as a regression or as a multi-class one in order to evaluate the performance of the conceived models in such scenarios.

6.0.1 Acknowledgements

This work has been supported by FCT – *Fundação para a Ciência e Tecnologia* within the R&D Units Project Scope: UIDB/00319/2020. It was also partially supported by a Portuguese doctoral grant, SFRH/BD/130125/2017, issued by FCT in Portugal.

References

[BD13] R. Bixler and S. D'Mello. Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 225–234, 2013.

[BLP13] A. Bogomolov, B. Lepri, and F. Pianesi. Happiness recognition from mobile phone data. In *Proceedings of the 2013 International Conference on Social Computing*, pages 790–795, 2013.

[GACA09] Q. Guo, E. Agichtein, C.L.A. Clarke, and A. Ashkan. In the mood to click? towards inferring receptiveness to search advertising. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 319–324, 2009.

[MV93] W.L. Mikulas and S.J. Vodanovich. The essence of boredom. *The Psychological Record*, 43(1):3–12, 1993.

[PCNN16] A. Pimenta, D. Carneiro, J. Neves, and P. Novais. A neural network to classify fatigue from human–computer interaction. *Neurocomputing*, 172:413–426, 2016.

[PDPO15] M. Pielot, T. Dingler, J.S. Pedro, and N. Oliver. 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*, pages 825–836, 2015.

[SKK09] A.C. Scheinbaum and M. Kukar-Kinney. Beyond buying: Motivations behind consumers' online shopping cart use. *Journal of Business Research*, 63(90):986–992, 2009.

[SLS19a] J. Seo, T.H. Laine, and K.A. Sohn. An exploration of machine learning methods for robust boredom classification using eeg and gsr data. *Sensors*, 19(20):20, 2019.

[SLS19b] J. Seo, T.H. Laine, and K.A. Sohn. Machine learning approaches for boredom classification using eeg. *Journal of Ambient Intelligence and Humanized Computing*, 10:3831–3846, 2019.

[SP13] A. Sano and R.W. Picard. Stress recognition using wearable sensors and mobile phones. In *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 671–676, 2013.