Towards Hybrid Profiling: combining digital phenotyping with validated survey questions to balance data entry effort with predictive power

Citation for published version (APA):

DOI:
10.1145/3459104.3459199

Document status and date:
Published: 19/02/2021

Document Version:
Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 07. Apr. 2022
Towards Hybrid Profiling: combining digital phenotyping with validated survey questions to balance data entry effort with predictive power

Ehsan Hadian Haghighi
Department of Industrial Engineering
Eindhoven University of Technology, The Netherlands
e.hadian.haghighi@tue.nl

Raoul C.Y. Nuijten
Department of Industrial Engineering
Eindhoven University of Technology, The Netherlands
r.c.y.nuijten@tue.nl

Pieter M.E. van Gorp
Department of Industrial Engineering
Eindhoven University of Technology, The Netherlands
p.m.e.v.gorp@tue.nl

ABSTRACT
Tailoring apps based on user traits has attracted tremendous interest in developing mHealth apps, and understanding a user’s personality is a key challenge in this context. This challenge is typically addressed via classic surveys, which pose a regrettable high burden on app users. This study aims to reduce the response burden of personality tests by introducing a model for predicting the user personality based on digital footprints of app usage. At the same time, skipping surveys completely turns out to undermine prediction accuracy. Therefore, this paper conceptualizes a hybrid framework that utilizes user event data in combination with surveys that have fewer questions than conventionally. The proposed method demonstrates a promising trade-off between the simplicity of using user event data and the accuracy of the validated survey methods when applying the hybrid method to a retrospective case study, the accuracy is higher than when using the event data exclusively. Also, the number of survey questions needed is significantly lower. Since this is a novel method, we expect that results will strengthen as larger data sets available over time. To facilitate that process, we also present a mathematical model for rationalizing the design process of hybrid profiling systems. This may boost adoption by developers who aim to implement the method in their specific app.

Keywords
digital phenotype; mHealth; big-five; personality prediction; health interventions; multi-layer perceptron; neural network; hybrid system

1. INTRODUCTION
Recently, mHealth interventions have shown significant potential to empower its users to change their daily routines and adopt healthier lifestyles [10]. However, these digital support tools often fail to retain their impact in the long-term due to their users losing interest [12].

A promising strategy to retain users is by tailoring your intervention to the users’ individual wants and needs [7]. Particularly, it was found that personalized mHealth interventions are more effective than interventions that adopt one-size-fits-all approaches [7].

User preferences are commonly derived from knowledge on a user’s personality [11]. Therefore, to personalize an mHealth intervention, the tool first needs to understand the preferences a particular user has. Historically, personality traits are derived using questionnaires [22, 24]. However, users often perceive questionnaires as cumbersome [24], and the quality of their answers decreases with every next question [2]. Hence, lengthy personality tests increases the risk of users dropping out, and as such, the aim to reduce dropouts by applying trait-based personalization may fail a priori.

Although efforts have been invested in shortening popular personality tests, and user experience design techniques have been employed to reduce the perceived burden of questionnaires (e.g., using blocks and progress bars) [13], it was found that filling out popular personality tests still requires between five to fifteen minutes [15].

When the primary aim of such surveys is to personalize an mHealth intervention then they are typically part of a digital onboarding process. In digital onboarding processes, dropout is already a challenging issue so adding five to fifteen minutes there is probably unacceptable from the dropout point of view.

Another strategy to gain understanding of a user’s personality is to apply unobtrusive learning methods from the domain of Artificial Intelligence (AI). For example, user event data is automatically recorded by users performing digital actions on mHealth tools. These actions have a certain meaning that may, or may not be, iconic for people with a certain personality. Hence, these digital traces can be tracked to construct an image of one’s personality. For example, AI has been deployed to automatically determine a user’s personality based on: the apps someone has installed [26], someone’s
behavior on social media platforms \[3\], and someone’s call and messages logs \[6\]. These unobtrusive approaches are typically categorized as “Digital phenotyping, behavioral sensing, or personal sensing” \[17\]. Unfortunately, the current state-of-the-art in digital phenotyping in mHealth is insufficiently accurate to predict one’s personality reliably \[18\].

In this study, we aim to combine classical personality tests with digital phenotyping techniques in order to explore the personalities of users of an mHealth tool. Particularly, we aim to increase the accuracy of digital phenotyping techniques, by enriching these unobtrusive learning methods with user reported survey responses of highly trimmed down personality tests.

Our findings demonstrate the relevance of investigating more subtly how hybrid approaches can be developed which balance well the user efforts with the predictive power. Since specific recommendations are largely dependent also on the quality of the data set that one has available to train a predictive model, our study aims to contribute also a mathematical model for rationalizing the design process of this new category of hybrid systems.

We showcase and validate our method using the GameBus platform but the techniques are applicable more broadly: our proposed method can be implemented in any mobile application which (1) offers pages that have a strong link to psychological traits, and (2) which offers functionalities for prompting the user with short surveys. In this case study, we use user event data and a multilayer perceptron (MLP) neural network to have an estimation of two probability trait profiles of each user. Then, based on these predictions a limited number of questions will be asked to predict the user’s personality. Results show that the proposed hybrid approach is promising and deserves more research attention.

2. BACKGROUND

2.1 GameBus

Recently, mHealth tools have been changing the face of healthcare. The rapidly increasing number of mHealth apps available in the market, clearly reveals that mHealth plays an important role in increasing health literacy, health management and patient engagement. It does remain a scientifically challenging question how one can design mHealth tools in a manner that users remain engaged over long periods of time.

GameBus (www.gamebus.eu) is an mHealth platform that uses gamification techniques as classified by Hamari \[16\]: it provides elements such as ‘challenges’, ‘points’, ‘goals’, ‘progress’, ‘leader-boards’, and ‘rewards’ in non-game contexts for promoting health activities in a playful manner. The GameBus platform is designed to persuade its end-users to a healthier lifestyle. Depending on the challenges chosen to take, different types of healthy activities -from social, to mental, to physical-, are rewarded with virtual points. Such points can be linked to individual or team-based competitions, but also non-competitive goal tracking is supported. Moreover, Social support elements such as newsfeeds, chats and likes and comments can be employed for behavior change. The academic platform has already been used to support various controlled experiments that aim to show statistically the effect of certain platform configurations on user engagement \[19\] -20\]. It turns out that results often depend on personality traits of the study participants \[21\]. Personalization based on users’ traits in GameBus can be applied to persuasive messages, often called nudges \[9\]. It is also possible to personalize the content, offered challenges and design of the App in accordance with their personality. Therefore, we are preparing a dynamic variant that tailors the platform configuration based on user traits. Such plasticity in user experiences based on personal traits has been described in the context of marketing \[22\] but it seems not explored for health promotion well and the current state-of-art approaches are not accurate enough in mHealth \[18\]. Trait modeling can be done via a variety of models. In a gamification context, for example, one could use the HEXAD model \[25\]. In the following section, we clarify a more established model from general psychology.

2.2 Big-Five framework for modeling user preferences

The most common personality framework is the Big-Five \[14\], and there are several mappings framework between users’ personality trait and actual (game) content \[11\] -25\].

Personality testing in accordance with the Big-Five personality traits consists of five traits including extraversion, agreeableness, conscientiousness, neuroticism, and intellect representing personality at the broadest level of abstraction. Mini-IPIP \[8\] is one of the most commonly used short inventories measuring an individual on the Big Five dimensions of personality. Although Mini-IPIP with 20 questions provides a short-form IPIP survey delivering an acceptable proxy for a person’s traits, it may still be felt like a burden since it typically takes 5 minutes to fill out. Although we contextualize the research using the GameBus app primarily because we have full data access for scientific research in compliance with the GDPR, this proposed method can be used by other mHealth platforms that collect similar data. Moreover, this hybrid approach is applicable to any psychological model.

3. METHODOLOGY

Construction of user’s trait profile is considered as one of the main challenges in personalization \[3\]. User preference as one of the most important aspects of user satisfaction should map into personality characteristics. Therefore, the user’s personality as the users’ most representative of personality trait makes this as a classification problem. The main goal of this study is to evaluate users’ preferences from their user event data in order to reduce their effort by answering fewer survey questions. The user event database is a collection of behavioral and analytics data that represents how users act in the app. Big-five survey results are considered as the ground truth for one’s actual preferences. We collected such data for pilots that enable us to train and test machine learning models such that future users do not have to fill out such long surveys. Specifically, we start from a 20-item Mini-IPIP survey (Appendix \[B\] and aim to further reduce the number of survey items without significantly compromising the predictive power of the questions that remain. Figure \[1\] is the graphical representation of the study design and general user journey when using the digital phenotype platform within mHealth applications.

3.1 Data collection

Data for this study was reused from a previous study \[19\] which evaluated the impact of competitiveness in a mHealth setting for 12 weeks among first-year high-school students. In total, 313 unique participants – including 290 students, and 23 teachers – have joined the study. Consent was collected to use the pseudonymized data for scientific studies on mHealth behavior change. Participants also were invited to complete a Mini-IPIP survey personality test in accordance with the Big-Five personality traits leaving a total of 94 participants in the dataset. Regarding event data, a total of 35 input parameters were derived. These parameters are fully listed in Appendix \[A\] and some of them are mentioned below: \(x_1\): number of days a user has been online in the given period (excluding the automatically tracked events)
3.2 Data preparation

As clarified in the introduction, our proposed method aims to classify users in psychological profiles in order to personalize the user experience and optimize user engagement. When aiming to reduce dropout rates, that classification should happen as fast as possible, with adequate accuracy. Therefore, we analyzed empirically different scenario’s (ranging from classifying already after one day to classifying only after multiple weeks). The study took place in 12 weeks and datasets as it is expressed in the form Eq. (1) they were prepared based on daily intervals \( T \) from users’ data. Datasets include both the values of input parameters \( X_{i,j} \) (user event data) and the actual degree of belonging to the Big-Five classes \( y_{i,l} \) for all participants that completed the survey.

\[
\text{Dataset}_T = [X_{i,j}, y_{i,l}]_T
\]  

\[
X_{i,j} = \begin{bmatrix}
x_{1,1} & \cdots & x_{1,j} \\
\vdots & \ddots & \vdots \\
x_{i,1} & \cdots & x_{i,j}
\end{bmatrix}
\]  

\[
y_{i,j} = \begin{bmatrix}
y_{1,1} & \cdots & y_{1,l} \\
\vdots & \ddots & \vdots \\
y_{i,1} & \cdots & y_{i,l}
\end{bmatrix}
\]  

\[
x'_{i,j} = \frac{x_{i,j}}{\frac{1}{n} \sum_i x_{i,j}} \cdot \frac{1}{\frac{1}{n} \sum_i x_{i,j}}^2
\]

To normalize the Big-Five survey data against data that represents Dutch society and categorize each user into one of the five trait groups, the score of each individual participant is compared with the average \( \mu_{nl} \) and standard deviation \( \sigma_{nl} \) score of people in the Netherlands (see Fig. 1 and Table 1 i.e., average scores and standard deviations of personality traits for the Dutch populations, we made use of data of the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata (Tilburg University, The Netherlands) to the Big-Five personality test by computing the z-score for each trait group \( l \) (Eq. 3). Consequently, user \( i \) belongs to trait group \( Y_i \) which has higher value among other trait groups (Eq. 6).

\[
y_{i,l} = \frac{y_{i,l} - \mu_{nl,l}}{\sigma_{nl,l}}
\]
\[ Y_i = \left[ \max_l(y_i,l) \right] \] (6)

Table 1: Average scores of Big-Five personalities test in the Netherlands

<table>
<thead>
<tr>
<th>Trait (l)</th>
<th>extra.</th>
<th>agree.</th>
<th>consc.</th>
<th>neuro.</th>
<th>intel.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. (µ_{nl})</td>
<td>3.12</td>
<td>3.84</td>
<td>3.72</td>
<td>3.48</td>
<td>3.48</td>
</tr>
<tr>
<td>Std. (σ_{nl})</td>
<td>0.66</td>
<td>0.52</td>
<td>0.52</td>
<td>0.71</td>
<td>0.5</td>
</tr>
</tbody>
</table>

3.3 Statistical Analysis

3.3.1 Optimum time frame and key features

One reason why user might abandon an app is we were not able to deliver personalized experiences that fit into their needs. The previous study shows most of the users abandon the app after only two weeks [19]. Also, analysis show users’ activities reached their peak on the third day, but then started to decrease significantly. Hence, the decision came down to a trade-off between the accuracy and observation time. Three days was considered as the optimum minimum time span required to have a good balance of the estimation accuracy and time. This time span can be considered as the golden time to engage people and make a personalized intervention.

Feature selection is one of the most important concepts in Machine Learning techniques to eliminate redundant variables and keeps only the best subset of predictors in the model. In the first phase, after categorizing users in their trait groups, average and standard deviation values of the input variables \( x_{i,1} \) to \( x_{i,35} \) (features) for users in the same group were calculated. Based on the standard deviation and dispersion of features, the algorithm selects three features with a higher value of standard deviation between the average of that variable within groups. In another word, this filter method follows the reasoning scheme to choose variables which their value being close within their group but have a significant distance from the other group.

3.3.2 Inactive users

The second set of statistical analyses was focused on the evaluation of the users’ activities. A respondent labeled as an inactive user if their key user event data have not changed during the study period of 12 weeks. It means these features have zero standard deviation during this period.

3.4 Model building

3.4.1 Artificial neural networks

Artificial neural networks (ANNs) are a soft computing technique inspired by simulating the behavior of the human brain. This study utilizes the power of Keras library in Python for this purpose and a multi-layer perceptron (MLP) with 1 input layer, 2 hidden layers, and 1 output layer are used to build the model. Each hidden layer consists of 10 nodes as depicted in Figure 2. The hyperbolic tangent activation function in each hidden layer and softmax function for the output layer are supplied along with the number of nodes. Since this problem is a type of single-label multi-class classification, categorical cross-entropy was applied in its settings rather than binary cross-entropy as the loss function of the model. Therefore, the proposed model gives a dispersion of probabilities between all five traits groups. Two categories that receive the highest probability will be the output for the model for the next step.

3.4.2 Selective questions

By predicting two possible personality groups for each user from the previous step, users’ mini-IPIP test scores on those two personality groups are considered. In practice, it is equivalent to just 8 questions out of 20 questions of the Mini-IPIP test that are asked from a user according to the predicted probable personality traits. In the next experiment with the aim to decrease the burdensome of the test, even more, we reduced the number of questions to 6 questions which includes 1 question with a positive keyed direction and 2 questions with negative keyed directions for each group. These questions were selected based on the confirmatory factor analysis (CFA) of the Mini-IPIP by choosing the highest values [5]. Confirmatory factor analysis is a popular powerful statistical tool in social research for evaluating latent variables. This analysis measures whether a prespecified factor model (hypothetical constructs) provides a good fit to the data or not. Consequently, questions number 6, 11 and 16 for extraversion, 2, 7 and 17 for agreeableness, 3, 8 and 18 for conscientiousness, 4, 9 and 19 for neuroticism, 5, 15 and 20 for intelligence were selected (Table 2). As can be seen from Table 3 in the next step, the number of questions is reduced to 4 questions by using confirmatory factors and choosing the highest positive and negative keyed directions in each group.

Table 2: Selected Mini-IPIP item number for 6 questions based on confirmatory analysis

<table>
<thead>
<tr>
<th>Item number</th>
<th>extra.</th>
<th>agree.</th>
<th>consc.</th>
<th>neuro.</th>
<th>intel.</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Selected Mini-IPIP item number for 4 questions based on confirmatory analysis

<table>
<thead>
<tr>
<th>Item number</th>
<th>extra.</th>
<th>agree.</th>
<th>consc.</th>
<th>neuro.</th>
<th>intel.</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>8</td>
<td>9</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

3.4.3 Cost function

Since this study uses the standard mini-IPIP questions to serve as ground-truth representing users’ scores on each of the Big Five dimensions, cognitive load is not contributing to our decision and the only factor that may have the greatest impact on the respondent burden is the length of the questionnaire. Therefore, to better visu-
alization, we defined users' effort and model accuracy as the cost function, to make a balance between users' effort and the accuracy of the model. In Eq.\(7\) \(n\) is the number of questions and \(a\) is the accuracy of the model.

\[
\text{Cost} = 0.55 - 0.55 \frac{a}{100} + 0.45 \frac{n}{20}
\]  

\(7\)

4. RESULTS

As previously mentioned, the main objective of this paper is to predict users’ digital phenotype as the first step of personalization to create an optimized mobile app user experience based on users’ trait. To ensure that the accuracy of the model persists, a 5-fold cross-validation method is applied for all the experiments.

4.1 Features selection

Three features are selected based on their scores in the first statistical analysis. These features, as presents in Table 4, are \(x_{i,3}\), \(x_{i,7}\), \(x_{i,32}\) which respectively are the number of times a user has viewed their personal activity log, the number of times a user has viewed their profile, and the number of points a user has collected. Table 5 presents the average and the standard deviation of these features in all trait groups.

<table>
<thead>
<tr>
<th>vactivities</th>
<th>vprofile</th>
<th>points</th>
</tr>
</thead>
<tbody>
<tr>
<td>extra.</td>
<td>7.81</td>
<td>6.89</td>
</tr>
<tr>
<td>agree.</td>
<td>10.90</td>
<td>9.76</td>
</tr>
<tr>
<td>consc.</td>
<td>19.70</td>
<td>32.15</td>
</tr>
<tr>
<td>neuro.</td>
<td>7.25</td>
<td>6.08</td>
</tr>
<tr>
<td>intel.</td>
<td>10.85</td>
<td>9.46</td>
</tr>
</tbody>
</table>

Table 5: Statistical analysis of key features for all trait groups

<table>
<thead>
<tr>
<th>vactivities</th>
<th>vprofile</th>
<th>points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. between trait groups</td>
<td>11.3</td>
<td>13.57</td>
</tr>
<tr>
<td>Std. between trait groups</td>
<td>4.46</td>
<td>5.64</td>
</tr>
</tbody>
</table>

4.2 Active users

Analyzing users based on the standard deviation of their key event data during these 12 weeks shows about 40% of participants who filled the survey (30 out of 94) were not active in that study at all. By comparing users in each trait group, it appears that in three out of five categories we have fewer (active) users (see Fig. 3 and Fig. 4).

4.3 MLP classifier

Initially, an MLP neural network was used to assess the feasibility of determining users’ personality traits just using user event data. Considering the highest probability in groups as the predicted personality trait, the balance average accuracy of 42.7\% (±7.7\%) and the F-score 32.6\% (±11.8\%) for all 5-folds was obtained. Other scores of using MLP in predicting each trait group is summarized in Table 6.

<table>
<thead>
<tr>
<th>Scores</th>
<th>accuracy</th>
<th>recall</th>
<th>precision</th>
<th>f-score</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>42.7%(±7.7%)</td>
<td>42.7(±7.7%)</td>
<td>34.5%(±16.7%)</td>
<td>32.6%(±11.8%)</td>
<td>0.49(±0.04)</td>
<td></td>
</tr>
</tbody>
</table>

As discussed earlier, this paper proposes a model to combine user event data and ask some selective questions with the aim of reducing the response burden of the Mini-IPP test and having a more accurate model for predicting user personalities. Therefore, by considering the two groups with the highest probability and rating those based on their scores to the questions for the predicted groups, the accuracy increases by 30% compared with using just MLP. Although the decline in the number of questions to 6 reduces the burden of the users by 10\% more, the accuracy of the model slightly affected. Consequently, by using the defined cost function (Eq. \(7\)) it is concluded asking 8 questions would be better. Table 7 shows the accuracy of proposed model for each trait group with

![Figure 3: Active and inactive users](image3.png)

![Figure 4: Dropout rate](image4.png)
8 and 6 questions.

Table 7: Prediction scores of using Hybrid method (user event data and asking selective related questions)

<table>
<thead>
<tr>
<th></th>
<th>Scores for 6q</th>
<th>Scores for 8q</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>63% (±7.2%)</td>
<td>72% (±4.7%)</td>
</tr>
<tr>
<td>recall</td>
<td>63% (±7.2%)</td>
<td>72% (±4.7%)</td>
</tr>
<tr>
<td>precision</td>
<td>53.3% (±10%)</td>
<td>60.6% (±11.7%)</td>
</tr>
<tr>
<td>f-score</td>
<td>56% (±8.1%)</td>
<td>63.7% (±7.7%)</td>
</tr>
<tr>
<td>cost</td>
<td>0.33 (±0.04)</td>
<td>0.33% (±0.02%)</td>
</tr>
</tbody>
</table>

4.4 Evaluation of model performance

The values of accuracy are quite stable in all five folds of cross-validation. Hence, promising results are achieved for the prediction of the digital phenotype by the proposed framework. To test the effect of asking smart selective questions, the model was tested by random questions. It was ensured that at least one question from each group was asked about that type of personality. The accuracy in this scenario drops to 33.7% (±5.6%).

4.5 Performance analysis

To analyze the performance of the proposed model, it was tested by a set of 14 users’ data representing in Table 8.

Table 8: Test group dataset

<table>
<thead>
<tr>
<th></th>
<th>points</th>
<th>profile</th>
<th>activities</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>29</td>
<td>10</td>
<td>11</td>
<td>extra.</td>
</tr>
<tr>
<td>User 2</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>extra.</td>
</tr>
<tr>
<td>User 3</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>intel.</td>
</tr>
<tr>
<td>User 4</td>
<td>29</td>
<td>11</td>
<td>13</td>
<td>neuro.</td>
</tr>
<tr>
<td>User 5</td>
<td>143</td>
<td>51</td>
<td>49</td>
<td>consc.</td>
</tr>
<tr>
<td>User 6</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>extra.</td>
</tr>
<tr>
<td>User 7</td>
<td>24</td>
<td>12</td>
<td>12</td>
<td>extra.</td>
</tr>
<tr>
<td>User 8</td>
<td>34</td>
<td>20</td>
<td>21</td>
<td>agree.</td>
</tr>
<tr>
<td>User 9</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>extra.</td>
</tr>
<tr>
<td>User 10</td>
<td>13</td>
<td>7</td>
<td>9</td>
<td>intel.</td>
</tr>
<tr>
<td>User 11</td>
<td>150</td>
<td>133</td>
<td>57</td>
<td>extra.</td>
</tr>
<tr>
<td>User 12</td>
<td>4</td>
<td>10</td>
<td>4</td>
<td>intel.</td>
</tr>
<tr>
<td>User 13</td>
<td>49</td>
<td>10</td>
<td>22</td>
<td>intel.</td>
</tr>
<tr>
<td>User 14</td>
<td>0</td>
<td>7</td>
<td>2</td>
<td>extra.</td>
</tr>
</tbody>
</table>

Moreover, users’ personality was defined by the Mini-IPIP test as the ground truth table. The results of this test appear in Tables 8 as well. In the first step, an MLP classifier is used to predict the test groups’ traits. Figure 5 shows confusion matrix of using MLP classifier and as can be seen the accuracy of using user event data without asking any question is 43%. Also, precision, recall and F-score, in this case, are 25%, 43%, 31% respectively.

To follow our goals, instead of using binary values in classification with MLP, we use predicted probabilities that of to what extent a user belongs to each trait group. Therefore, Z-score for related questions to the two groups having the highest probabilities is calculated. In this method, the accuracy of using hybrid technique, as it is shown in Table 9 and Figure 6, is 71.43% which is about 30% higher than using the MLP neural network.

Table 9: Prediction scores of using the proposed hybrid method by asking 8 questions

<table>
<thead>
<tr>
<th></th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>71.43%</td>
</tr>
<tr>
<td>recall</td>
<td>71.43%</td>
</tr>
<tr>
<td>precision</td>
<td>56.55%</td>
</tr>
<tr>
<td>f-score</td>
<td>62.86%</td>
</tr>
<tr>
<td>cost</td>
<td>0.33%</td>
</tr>
</tbody>
</table>

By reducing the number of questions to 6, the performance of the model dropped. Figure 7 shows the confusion matrix of that and Table 10 clearly shows that when users are asked 6 questions, the result is worse than when they are asked 8 questions.
in our prediction. Even when users do not actively participate in the survey, by using user event data exclusively, we would reach only about 50% accuracy. Although diminishing the number of questions to 6 reduces the burden of the survey by 10%, it reduces the performance of the model by the same level as well.

The retrospective case study was based on a health gamification app. In that setting, digital traces included variables such as the number of times a user has viewed their personal activity log, the number of times a user has viewed their profile, and the number of points a user has collected. These variables aided to predict significant differences in users with different personality traits. On one hand, users in the Conscientiousness group are more loyal to the platform; on the other hand, the app has not engaged the interest of users with agreeableness and extraversion trait to use the app and they have a higher rate of dropout.

5.1 Future works

Running a trial on the impact of personalization of challenges and users’ experience would be an opportunity to have a better understanding of the effect of it in the context of mHealth tools. More broadly, having a good proportion of samples in all traits’ groups could improve the accuracy of the proposed model.

6. REFERENCES


APPENDIX

A. EVENT DATA

\[ x_1 \]: number of days a user has been online in the given period (excluding the automatically tracked events)

\[ x_2 \]: number of days a user has been tracked using automatic trackers

\[ x_3 \]: number of times a user has viewed their personal activity log

\[ x_4 \]: number of times a user has viewed their friends’ activity log

\[ x_5 \]: number of times a user has viewed the details of an activity

\[ x_6 \]: number of times a user has viewed their newsfeed

\[ x_7 \]: number of times a user has viewed their profile

\[ x_8 \]: number of times a user has viewed their friends’ profile

\[ x_9 \]: number of times a user has viewed the details of (one of their) circles (= teams)

\[ x_{10} \]: number of times a user has viewed their overview of challenges

\[ x_{11} \]: number of times a user has viewed the leaderboard of a particular challenge

\[ x_{12} \]: number of times a user has viewed the leaderboard of a particular circle (= team) within a particular challenge

\[ x_{13} \]: number of times a user has given support to others

\[ x_{14} \]: number of unique players a user has supported

\[ x_{15} \]: number of times a user has received support from other players (i.e., not including the number of times a user has supported herself)

\[ x_{16} \]: number of unique players that have supported the user (i.e., not including the user herself, in case she has supported herself)

\[ x_{17} \]: number of reactions on activities a user has made

\[ x_{18} \]: number of unique players a user has reacted on (i.e., a reaction on their activities)

\[ x_{19} \]: number of reactions a user has made on her own activities

\[ x_{20} \]: number of reactions a user has received on her activities, not including the reactions she has made on her own activities

\[ x_{21} \]: number of unique players a user has received reactions from on her reactions, not including herself (if she has made a reaction on her own activities)

\[ x_{22} \]: number of circle (= team) chat messages sent by a user

\[ x_{23} \]: number of unique circles (= teams) a user has sent a chat message

\[ x_{24} \]: number of circle chat messages a user has received, to including the chat messages she has received from herself

\[ x_{25} \]: number of unique players that have sent the user a circle (= team) chat message, excluding the user herself (if she has sent any messages herself)

\[ x_{26} \]: number of unique circles (= teams) that have sent the user a circle chat message

\[ x_{27} \]: number of activities a user has performed (both manually and automatically registered)

\[ x_{28} \]: number of activities that were automatically registered for a user (for example via Google Fit, Fitbit, or any other automatic tracker)

\[ x_{29} \]: number of activities (both annually and automatically registered) that were rewarded by a challenge

\[ x_{30} \]: number of automatically registered activities that were rewarded by a challenge

\[ x_{31} \]: number of challenges a user was enrolled in

\[ x_{32} \]: number of points a user has collected

\[ x_{33} \]: number of unique challenge rules a user has performed

\[ x_{34} \]: number of unique activity types a user has performed

\[ x_{35} \]: number of unique activity types a user has performed that were rewarded

B. BIG-FIVE MINI-IPIP SURVEY

Table 11: Mini-IPIP survey

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor</th>
<th>Text</th>
<th>Keyed direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Extra.</td>
<td>Am the life of the party.</td>
<td>Positive</td>
</tr>
<tr>
<td>2</td>
<td>Agree.</td>
<td>Sympathize with others’ feelings.</td>
<td>Positive</td>
</tr>
<tr>
<td>3</td>
<td>Consc.</td>
<td>Get chores done right away.</td>
<td>Positive</td>
</tr>
<tr>
<td>4</td>
<td>Neuro.</td>
<td>Have frequent mood swings.</td>
<td>Positive</td>
</tr>
<tr>
<td>5</td>
<td>Intel.</td>
<td>Have a vivid imagination.</td>
<td>Positive</td>
</tr>
<tr>
<td>6</td>
<td>Extra.</td>
<td>Don’t talk a lot.</td>
<td>Negative</td>
</tr>
<tr>
<td>7</td>
<td>Agree.</td>
<td>Am not interested in other people’s problems.</td>
<td>Negative</td>
</tr>
<tr>
<td>8</td>
<td>Consc.</td>
<td>Often forget to put things back in their proper place.</td>
<td>Negative</td>
</tr>
<tr>
<td>9</td>
<td>Neuro.</td>
<td>Am relaxed most of the time.</td>
<td>Negative</td>
</tr>
<tr>
<td>10</td>
<td>Intel.</td>
<td>Am not interested in abstract ideas.</td>
<td>Negative</td>
</tr>
<tr>
<td>11</td>
<td>Extra.</td>
<td>Talk to a lot of different people at parties.</td>
<td>Positive</td>
</tr>
<tr>
<td>12</td>
<td>Agree.</td>
<td>Feel others’ emotions.</td>
<td>Positive</td>
</tr>
<tr>
<td>13</td>
<td>Consc.</td>
<td>Like order.</td>
<td>Positive</td>
</tr>
<tr>
<td>14</td>
<td>Neuro.</td>
<td>Get upset easily.</td>
<td>Positive</td>
</tr>
<tr>
<td>15</td>
<td>Intel.</td>
<td>Have difficulty understanding abstract ideas.</td>
<td>Negative</td>
</tr>
<tr>
<td>16</td>
<td>Extra.</td>
<td>Keep in the background.</td>
<td>Negative</td>
</tr>
<tr>
<td>17</td>
<td>Agree.</td>
<td>Am not really interested in others.</td>
<td>Negative</td>
</tr>
<tr>
<td>18</td>
<td>Consc.</td>
<td>Make a mess of things.</td>
<td>Negative</td>
</tr>
<tr>
<td>19</td>
<td>Neuro.</td>
<td>Seldom feel blue.</td>
<td>Negative</td>
</tr>
<tr>
<td>20</td>
<td>Intel.</td>
<td>Do not have a good imagination.</td>
<td>Negative</td>
</tr>
</tbody>
</table>