D4.8: Evaluation of the Holistic Behaviour Analysis Framework

Dissemination level: Public
Document type: Report
Version: 1.0.0
Date: August 28th, 2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement #769553. This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains.

Document Details

Project Number	769553
Project title	Council of Coaches
Title of deliverable	Evaluation of the Holistic Behaviour Analysis Framework
Due date of deliverable	August 31st, 2020
Work package	WP4
Author(s)	Oresti Banos (CMC), Kostas Konsolakis (CMC)
Reviewer(s)	Tessa Beinema (RRD), Harm op den Akker (RRD), Jorien van Loon (CMC)
Approved by	Coordinator
Dissemination level	Public
Document type	Report
Total number of pages	57

Partners

- University of Twente – Centre for Monitoring and Coaching (CMC)
- Roessingh Research and Development (RRD)
- Danish Board of Technology Foundation (DBT)
- Sorbonne University (SU)
- University of Dundee (UDun)
- Universitat Politècnica de València, Grupo SABIEN (UPV)
- Innovation Sprint (iSPRINT)

Abstract

This deliverable (D4.8) evaluates the functional operation of the Holistic Behaviour Analysis Framework (HBAF). This evaluation, which is executed under Task 4.4, investigates the resilience and reliability of the HBAF models developed in previous tasks for the inference of short-term behaviours (T4.1), long-term behaviours (T4.2) and behaviour changes (T4.3) over time. Based on these models, raw sensor data are processed into meaningful representations of human behaviour which are communicated to the Shared Knowledge Base in order to nurture some coaching actions and trigger possible interventions of the virtual council. Hence, this deliverable aims to describe the evaluation study that has been conducted in order to check the validity of the developed framework.

Table of Contents

List of figures

List of tables

Symbols, abbreviations and acronyms

CaaS	Coach-as-a-Sensor
CMC	Centre for Monitoring and Coaching
COUCH	Council of Coaches
COVID-19	Coronavirus Disease 2019
D	Deliverable
DBT	Danish Board of Technology Foundation
EC	European Commission
EEMCS	Electrical Engineering, Mathematics and Computer Science
ESM	Experience Sampling Method
GPS	Global Positioning System
HBAF	Holistic Behaviour Analysis Framework
iSPRINT	Innovation Sprint
M	Month
MS	Milestone
Q	Question
RRD	Roessingh Research and Development
SMS	Short Message Service
SU	Sorbonne University
T	Task
UDun	University of Dundee
UPV	Universitat Politècnica de València
UT	University of Twente
WP	Work Package

1 Introduction

After presenting the models for inferring short-term, long-term and behaviour changes from sensor data in previous deliverables, this deliverable assesses the potential of the three main core components of the Holistic Behaviour Analysis Framework (HBAF) when tested in the wild. The COVID-19 pandemic has posed important obstacles to conduct experiments involving humans worldwide. This is also the case for the experiment originally envisioned for T4.4. However, the particular impact that lockdowns and de-escalations has had in the daily behaviour of most people, where physical, social, emotional and cognitive activities have varied in general more rapidly and dramatically than usual, this scenario has been considered here as a great opportunity to assess the possibilities of the developed framework.

The evaluation study, which is used to collect the necessary data in order to evaluate the HBAF core components, is thoroughly presented and explained in Section 3. This includes the description of the protocol for the data collection, the sensing devices, and the methodology for the assessment of the framework. A description of the recruited participants, the different types of collected data, and privacy aspects are presented in Section 4. The results of the evaluation study are shown in Section 5, while Section 6 reports the main findings of this deliverable.

2 Objectives

The main objective of this deliverable (D4.8) is to describe the results of the assessment of the functional operation of the Holistic Behaviour Analysis Framework. Accordingly, this document aims to explain and describe the evaluation study that was performed in real-world scenarios in order to assess the reliability and resilience of each of the three components of the framework; including the components that were presented in D4.2 (Banos, Konsolakis, op den Akker, Pelachaud, & Bangalore, 2018) and D4.3 (Banos, Konsolakis, Bangalore Kantharaju, Pelachaud, & op den Akker, 2018) for detecting the short-term behaviours (T4.1), the components that were presented in D4.4 (Banos & Konsolakis, 2019) and D4.5 (Banos & Konsolakis, 2019) for detecting the long-term behaviours (T4.2), and finally the components that were presented in D4.6 (Banos & Konsolakis, 2019) and D4.7 (Banos & Konsolakis, 2019) for detecting relevant estimators of behaviour changes (T4.3).

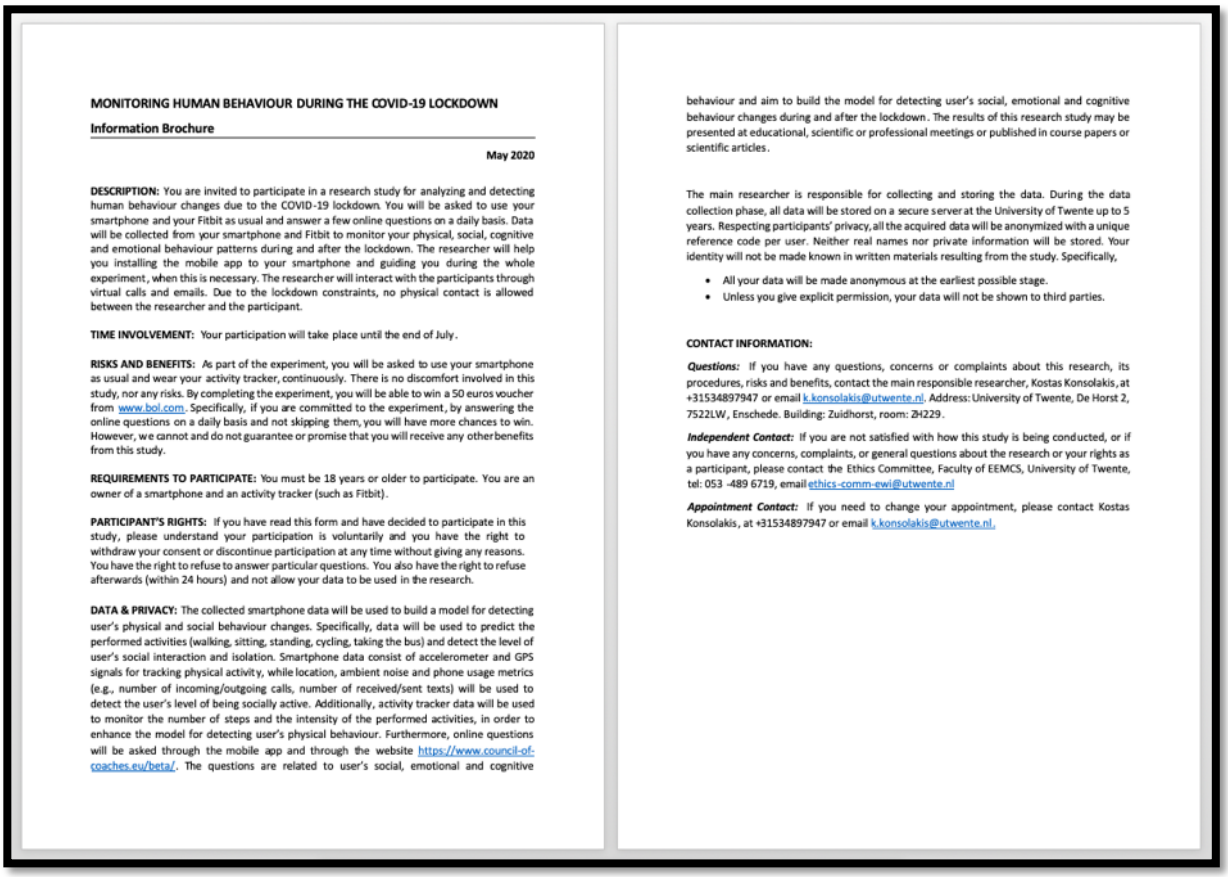
3 Methods

3.1 Evaluation Study

For T4.4, we decided to conduct a study in order to assess the functional operation of the HBAF. Initially, the study was designed to be performed in real-world scenarios, aiming to evaluate the three core components of the framework by focusing on COUCH's target groups (older adults, patients with diabetes type 2 and patients with chronic pain). However, the COVID-19 pandemic crisis has limited severely the options for recruiting participants from the target group, especially elderly people or those individuals at risk, since national regulations prohibited in general close contact (social distancing rules). This situation played a major role on setting up the experiment (such as distributing and configuring the sensing devices for the study participants). For this reason, and yet to ensure a proper evaluation of the developed framework, we decided to recruit healthy participants who already owned the necessary devices and were able to follow the instructions of the study remotely.

Participants were asked to use their own smartphone and activity tracker device, as usual, and answer a few online questions on a daily basis. The study was performed in an uncontrolled setting, and thus, no explicit instructions were given to the participants on what tasks to perform and how often. The study was carefully designed to take place during the COVID-19 lockdown, where we would have the opportunity to monitor human behaviour during and after the lockdown, and presumably changes in between. After submitting the investigation protocol, the recruitment phase started.

The recruitment phase took place in two different rounds. During the first round, we distributed the information brochure (see Figure 1) and we invited subjects who owned a smartphone and a Fitbit device to participate in the evaluation study. No restrictions were given with respect to gender, age, or health conditions of the participants. However, due to the relatively low number of responders, we decided to extend the protocol on the second round, by recruiting participants who owned any activity tracker that monitors steps (including pedometer applications installed on users' mobile phones). Overall, 23 healthy subjects from the Netherlands and from Greece participated in the data collection phase.



MONITORING HUMAN BEHAVIOUR DURING THE COVID-19 LOCKDOWN

Information Brochure

May 2020

DESCRIPTION: You are invited to participate in a research study for analyzing and detecting human behaviour changes due to the COVID-19 lockdown. You will be asked to use your smartphone and your Fitbit as usual and answer a few online questions on a daily basis. Data will be collected from your smartphone and Fitbit to monitor your physical, social, cognitive and emotional behaviour patterns during and after the lockdown. The researcher will help you installing the mobile app to your smartphone and guiding you during the whole experiment, when this is necessary. The researcher will interact with the participants through virtual calls and emails. Due to the lockdown constraints, no physical contact is allowed between the researcher and the participant.

TIME INVOLVEMENT: Your participation will take place until the end of July.

RISKS AND BENEFITS: As part of the experiment, you will be asked to use your smartphone as usual and wear your activity tracker, continuously. There is no discomfort involved in this study, nor any risks. By completing the experiment, you will be able to win a 50 euros voucher from www.bol.com. Specifically, if you are committed to the experiment, by answering the online questions on a daily basis and not skipping them, you will have more chances to win. However, we cannot and do not guarantee or promise that you will receive any other benefits from this study.

REQUIREMENTS TO PARTICIPATE: You must be 18 years or older to participate. You are an owner of a smartphone and an activity tracker (such as Fitbit).

PARTICIPANT'S RIGHTS: If you have read this form and have decided to participate in this study, please understand your participation is voluntarily and you have the right to withdraw your consent or discontinue participation at any time without giving any reasons. You have the right to refuse to answer particular questions. You also have the right to refuse afterwards (within 24 hours) and not allow your data to be used in the research.

DATA & PRIVACY: The collected smartphone data will be used to build a model for detecting user's physical and social behaviour changes. Specifically, data will be used to predict the performed activities (walking, sitting, standing, cycling, taking the bus) and detect the level of user's social interaction and isolation. Smartphone data consist of accelerometer and GPS signals for tracking physical activity, while location, ambient noise and phone usage metrics (e.g., number of incoming/outgoing calls, number of received/sent texts) will be used to detect the user's level of being socially active. Additionally, activity tracker data will be used to monitor the number of steps and the intensity of the performed activities, in order to enhance the model for detecting user's physical behaviour. Furthermore, online questions will be asked through the mobile app and through the website https://www.council-of-coaches.eu/beta/. The questions are related to user's social, emotional and cognitive behaviour and aim to build the model for detecting user's social, emotional and cognitive behaviour changes during and after the lockdown. The results of this research study may be presented at educational, scientific or professional meetings or published in course papers or scientific articles.

The main researcher is responsible for collecting and storing the data. During the data collection phase, all data will be stored on a secure server at the University of Twente up to 5 years. Respecting participants' privacy, all the acquired data will be anonymized with a unique reference code per user. Neither real names nor private information will be stored. Your identity will not be made known in written materials resulting from the study. Specifically,

- All your data will be made anonymous at the earliest possible stage.
- Unless you give explicit permission, your data will not be shown to third parties.

CONTACT INFORMATION:

Questions: If you have any questions, concerns or complaints about this research, its procedures, risks and benefits, contact the main responsible researcher, Kostas Konsolakis, at +31534897947 or email k.konsolakis@utwente.nl. Address: University of Twente, De Horst 2, 7522LW, Enschede. Building: Zuidhorst, room: ZH229.

Independent Contact: If you are not satisfied with how this study is being conducted, or if you have any concerns, complaints, or general questions about the research or your rights as a participant, please contact the Ethics Committee, Faculty of EEMCS, University of Twente, tel: 053 -489 6719, email ethics-comm-ewi@utwente.nl

Appointment Contact: If you need to change your appointment, please contact Kostas Konsolakis, at +31534897947 or email k.konsolakis@utwente.nl.

Figure 1: Information brochure for the study "Monitoring human behaviour during the COVID-19 lockdown".

The data collection phase started once the study was approved by the Ethics Committee EEMCS (EC, 2020) of the University of Twente. Specifically, the study with the title "Monitoring human behaviour during the COVID-19 lockdown" was approved on 19 May 2020 with the registration number 'RP 2020-43'. The data collection started on Wednesday, May 20th, 2020 (the actual date varies per participant, since some subjects started later depending on their availability) and it ended on Sunday, July 5th, 2020.

In Figure 2, the timeline of the Evaluation Study is depicted. On Friday, May 1st, 2020 (recruitment round 1), the information brochure, including a short description of the study, was distributed through the BSS department at UT and was also uploaded on social media (Facebook, Twitter, LinkedIn). Initially, 30 candidates were interested to participate in the research study. On Tuesday, May 19th, 2020, the investigation protocol for the evaluation study was approved by the Ethics Committee EEMCS with the registration number 'RP 2020-43'. On Wednesday, May 20th, 2020 (recruitment round 2), further explanations of the research study were given to the candidates, including participants' tasks and the data privacy issues. The recruited participants had to confirm their participation by signing the informed consent form, while final instructions were given explaining how to set up the sensing devices in order to participate in the study. In total, 23 subjects confirmed to participate in the data collection phase. The data collection phase started on May 20th, and ended on July 5th (varying per participant). Some days later, a compensation (voucher for online shopping) was given to the participants who managed to complete the experiment successfully.

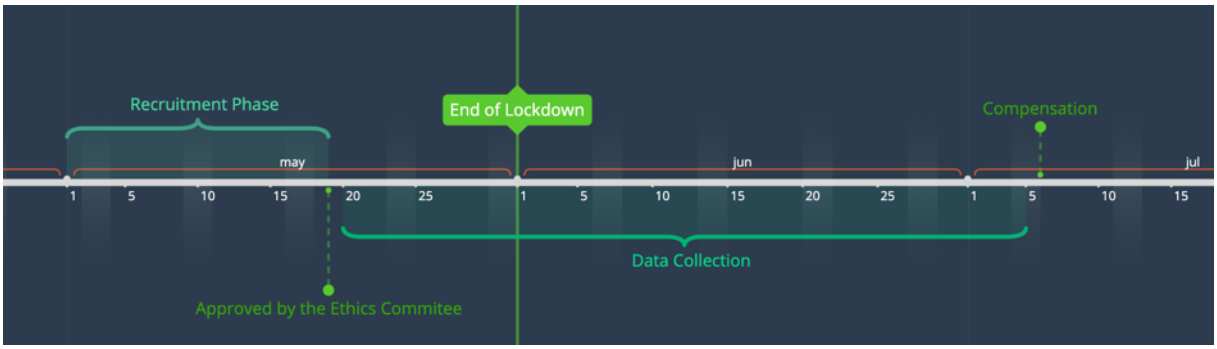


Figure 2: Timeline of the Evaluation Study.

During this phase, three main types of data were acquired; including raw smartphone data, processed data from activity trackers, and users' answers from the ESM questions. Additionally, users' demographics, including age, gender and occupation, were collected. The data types are presented and described in Section 4 (specifically sub-section 4.2)

The participants received two tutorials with detailed instructions on a) how to install the free AWARE app (AWARE-Framework, 2020) on their mobile device and what smartphone sensors to enable/disable for the data acquisition, and b) how to register and use the COUCH website in order to connect their activity tracker (if they owned a Fitbit) and answer the Coach-as-a-Sensor (CaaS) online questions (for further information see the deliverable D4.6). An overview of these instructions can be seen in Figure 3 and Figure 4, respectively.

Initially, the AWARE app was installed on the user's smartphone, which allowed the smartphone data acquisition. The smartphone collected data are used to build a model for detecting user's physical and social behaviour changes. Additionally, data from activity trackers were used to enhance the model for detecting user's physical behaviour. Finally, ESM and CaaS questions were asked through the mobile app and through the COUCH website (Council-of-Coaches, 2020). The questions are related to user's social, emotional and cognitive behaviour and were used to build the model for detecting user's social, emotional and cognitive behaviour changes during and after the lockdown. An overview of these questions is presented on Table 1 (mobile app) and Table 2 (COUCH website).

This study aims to evaluate the models for detecting users' short-term, long-term and behaviour changes that occur during the lockdown and how these differentiate when the users return back to their normal lifestyle. For this reason, two extra questions were asked at the end of each week (see questions Q8 and Q9 on Table 2) and were used as the ground truth for monitoring trends and changes related to physical, social, emotional and cognitive behaviour.

Overall Instructions:
You are asked to use the AWARE app. You will install the app on your Android smartphone and you will answer a few questions that are triggered automatically at different periods of time, every day. The AWARE app is also used to collect your smartphone data (number of phone calls/SMS, the ambient noise and your location). The collected data will be anonymized in order to not raise any privacy concerns. You can find more information on the Consent Form Document. If you experience any issue, please contact the researcher Kostas Konsolakis (k.konsolakis@utwente.nl).

More information about the AWARE app can be found here:
https://awareframework.com/

1. **Synchronize Google Calendar**: The researcher has invited you to join a calendar for this study. If not, please contact the researcher (you cannot continue to the next steps). The shared calendar will be used to send you some questions and notifications. Make sure that your Google Calendar is synchronized and that the 'AWAREStudy' is visible on the calendar app on your smartphone.

2. **Install app**: At first, you have to click on the link from your Android smartphone and install the app. The app file is also sent to you by email. Keep in mind that the AWARE app is not available on the Google Play Store (you cannot search/install it on the Play Store).

3. **Open app:** Once you install it, you open it and you give permission to any notifications that you get related to the accessibility option (see Figure 1). The accessibility option refers to stop optimizing device battery usage and allowing the AWARE app to always run in the background. Depending on the type of your Android smartphone, this step differentiates a bit (you might not get this type of notification). If you experience any issue, please contact the researcher.

4. **Join study:** Once you open the AWARE app, you have the option to scan a QR code (top right of the screen). Please scan the QR code in Figure 2. This QR code allows you to connect to a specific study. Once you scan the picture, you will be able to see the description of this study. Please use a name on the 'Onboarding' section, so we can link your smartphone with a specific id. This might be added automatically from your phone. If not, you can type anything that you want, like your name, or your email address or anything else that you prefer. Finally, you click on "SIGN UP!" and you give access to the app to read your smartphone data (see Figure 3).

5. **Do not kill the app**: Since the app collects data in a continuous way, it is important to give the necessary permission regarding the battery optimization. Depending on the smartphone brand, the smartphone will try to kill and stop the app since it is running on a silence mode. Specifically, your smartphone will try to put the app into 'a sleeping mode' and it will not be possible anymore to run on the background and collect data.

   Every smartphone brand has a different battery optimization policy. For example, for Samsung smartphones you have to go to the 'Settings' and then 'Device care' (or Device Maintenance) and then 'Battery' and then 'Unmonitored apps', where you will allow the AWARE app to use as much power as possible (see Figure 4). The power monitor system of your device should always allow the AWARE app to run and not try to stop it. This differentiates based on your smartphone device. In the link: https://dontkillmyapp.com, you can find how to disable the sleeping mode for the AWARE app. If you experience any troubles, please contact Kostas Konsolakis.

6. **Final remark**: It is important to keep the app and not close it while the experiment takes place. When the experiment is over you can uninstall the app or quit the study and the data collection will stop automatically.

Visual Instructions:

Figure 1: AWARE Interface & notification for Battery Optimization (the notification depends on the smartphone device).

Figure 2: Scan QR code.

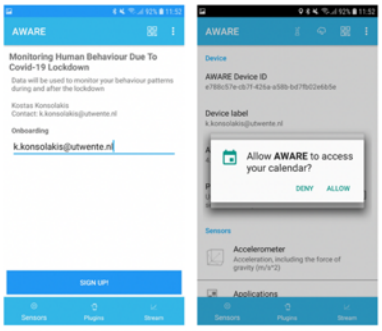


Figure 3: Join Study & Give access to smartphone data.

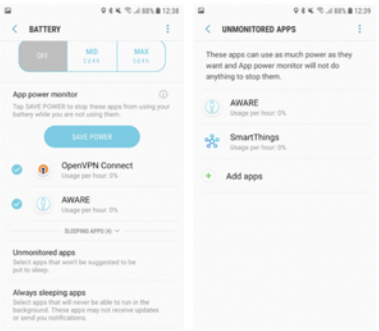


Figure 4: Making the AWARE app unmonitored in the battery settings of your device. This differentiates per device (in the above picture we present the settings for a Samsung S7 device).

Figure 3: Instructions for installing the AWARE app.

Figure 4: Instructions for using the COUCH website.

Table 1: An overview of the ESM questions asked through the mobile app.

Measured Behaviour	Frequency	Question	Answer
Emotional (morning, afternoon & evening hours)	Three times per day: at 12pm, 5pm and 9pm	**Q1**: How are you feeling right now?	5-Scale ranging from very sad to very happy
Social (morning hours)	Once per day at 12pm	**Q2**: Have you interacted with other people (physically or digitally) during the morning hours (between 7am and 12pm)? If so, please indicate the total duration in hours (rounding up).	5-Scale ranging from zero hours to five hours
Cognitive (morning hours)	Once per day at 12pm	**Q3**: Have you participated in any cognitive activities during the morning hours (between 7am and 12pm)? If so, please indicate the total duration in hours (rounding up). Cognitive activities are considered tasks such as reading a book or newspaper, learning a new skill, watching an educational TV-show or playing a board game (e.g., chess, sudoku, etc.).	5-Scale ranging from zero hours to five hours
Social (afternoon hours)	Once per day at 5pm	**Q4**: Have you interacted with other people (physically or digitally) during the afternoon hours (between 12pm and 5pm)? If so, please indicate the total duration in hours (rounding up).	5-Scale ranging from zero hours to five hours
Cognitive (afternoon hours)	Once per day at 5pm	**Q5**: Have you participated in any cognitive activities during the afternoon hours (between 12pm and 5pm)? If so, please indicate the total duration in hours (rounding up). Cognitive activities are considered tasks such as reading a book or newspaper, learning a new skill, watching an educational TV-show or playing a board game (e.g., chess, sudoku, etc.).	5-Scale ranging from zero hours to five hours
Social (evening hours)	Once per day at 9pm	**Q6**: Have you interacted with other people (physically or digitally) during the evening hours (between 5pm and 12am)? If so, please indicate the total duration in hours (rounding up).	5-Scale ranging from zero hours to five hours

Cognitive (evening hours)	Once per day at 9pm	**Q7**: Have you participated in any cognitive activities during the evening hours (between 5pm and 12am)? If so, please indicate the total duration in hours (rounding up). Cognitive activities are considered tasks such as reading a book or newspaper, learning a new skill, watching an educational TV-show or playing a board game (e.g., chess, sudoku, etc.).	5-Scale ranging from zero hours to five hours
User's feedback for the last 7days	Every Sunday at 4pm	**Q8**: Overall, how do you feel about the last 7 days?	5-Scale ranging from very dissatisfied to very satisfied
User's feedback for the last 7days	Every Sunday at 4pm	**Q9**: Try to compare this week (last 7 days) with the previous week. Did you experience any significant difference? If so, please specify. For example, you can type "more/less physically active", "more/less socially active". Furthermore, you can type anything that you think is relevant for this week; such as "no difference", "this week I was tired/sick", "this week I met more/less people", etc.	Free-text

Table 2: An overview of the CaaS questions asked through the Council of Coaches web application.

Measured Behaviour	Frequency	Question	Answer
Social	Once per day	**Q10**: Are you satisfied with your social interactions today?	Multiple choice including very dissatisfied, dissatisfied, satisfied, very satisfied
	Once per day	**Q11**: Have you interacted physically or digitally with other people today?	Yes[1] or No
	Once per day	**Q12**: Could you tell me the total duration (in minutes) of interacting with family members? So, for example write "30" if you spent half an hour with family members, or "0" if you haven't seen them.	Numeric input
	Once per day	**Q13**: Could you tell me the total duration (in minutes)	Numeric input

[1] If the answer is Yes then questions Q12-Q14 will be asked. Otherwise, they will be skipped.

		of interacting with friends? Again, write "0" if you haven't seen any.	
	Once per day	**Q14**: Finally, could you tell me the total duration (in minutes) of interacting with other acquaintances (like colleagues)? Again, write "0" if you haven't had any interaction.	Numeric input
Emotional	Once per day	**Q15**: Please describe your overall mood, from negative to positive, during the morning hours, let's say between 8am and 12pm.	Multiple choice including very negative, negative, neutral, positive, very positive
	Once per day	**Q16**: Now, please describe your overall mood, from negative to positive, during the afternoon hours, so between 12pm and 5pm.	Multiple choice including very negative, negative, neutral, positive, very positive
	Once per day	**Q17**: And finally, please describe your overall mood, from negative to positive, during the evening hours, so between 5pm and 12am.	Multiple choice including very negative, negative, neutral, positive, very positive
Cognitive	Once per day	**Q18**: Have you participated in any of the following tasks today, such as reading a book or newspaper, learning a new skill, watching an educational TV-show or playing a board game?	Yes[2] or No
	Once per day	**Q19**: That's very good. Now, could you tell me for how long did you spend reading today in minutes? Just type "0" if you didn't read at all.	Numeric input
	Once per day	**Q20**: Okay, and how much time did you spend on learning a new skill, such as studying a foreign language, in minutes? Again, just type "0" if you did not do this today.	Numeric input
	Once per day	**Q21**: Very good! Now, how much time did you spend watching an educational	Numeric input

[2] If the answer is Yes then questions Q19-Q22 will be asked. Otherwise, they will be skipped.

		TV-show? Again, just type "0" if you did not do this.	
	Once per day	**Q22**: And finally, how much time did you spend either playing sudoku, crossword games, puzzles, or other similar games in minutes?	Numeric input

3.2 HBAF assessment

Data from the evaluation study will be used in order to assess the reliability and resilience of each of the three core components of the HBAF framework (T4.1, T4.2 and T4.3). Thus, the main focus will be on evaluating the accuracy and robustness of detecting users' short-term, long-term and behaviour changes over time.

The **accuracy** metric will be used to evaluate the reliability long-term and behaviour change components of the HBAF and can be tested by contrasting what the system detects and the "ground-truth" given by the users. The ground-truth can be found on users' answers to the questions Q8 and Q9 (see Table 1). Briefly, we asked the users at the end of every week to comment retrospectively (last 7 days) whether they noticed any relative difference in terms of behaviour compared to the previous week. A sample of these answers can be seen in Figure 5.

Figure 5: A sample of users' answers based on the question Q9 asked on Sunday 14 June.

The **robustness** metric will be used to assess the resilience of the HBAF components by forcing the system to handle abnormal/erroneous situations which, although unlikely, could occur at some time and affect the overall performance of the system. The most common way to simulate such situations is to add some level of noise or absence of signal, which is then injected into the real data:

1. Noisy raw sensor data: we purposely introduce some level of noise or absence of signal into the collected data in order to validate the detected short-term behaviours. Our aim is to answer questions such as "What is the tolerance of modelling the short-term social behaviour to noise in the ambient sound?".

2. Noisy short-term behaviour: similar approach but now we alter some short-term behaviour values in order to see the impact on the detected long-term behaviours. Our aim is to answer

questions such as “What is the tolerance of the sedentary behaviour trend detection when outliers are present in the step count for some days?”.

3. Noisy long-term behaviours: similar to the previous approach, but now focused on assessing the impact that errors on the short-term behaviour has in the detection of behaviour changes. Our aim is to answer questions such as “What is the tolerance of a negative physical change detection when outliers are introduced in the step count for some days?”.

4 Experimental Setup

4.1 Participants

Initially, more than 30 participants showed interest to participate in the evaluation study. However, only 23 confirmed to participate in the data collection phase, while two of them decided to withdraw in a later phase. In the end, 21 subjects managed to complete the experiment successfully. Among the participants, 15 were females with an average age 31.7 years old and 6 were men with an average age 34 years old (the total average age is 32.4). Five participants were Greek residents, while the rest lived in the Netherlands. It is worth mentioning that the participants consist of one pensioner with low-back chronic pain, two healthy students, while the rest of the subjects are healthy employees.

Some participants were not able to install the mobile app on their smartphones, while others did not give consent to collect their own smartphone data. Thus, these participants only used the COUCH website to answer the CaaS related questions (see Table 2) and their activity tracker for monitoring steps. Consequently, the participants were divided into two groups. The Group A consists of 12 participants (with an average age: 34.58 years old), including sensor data from the mobile app, the COUCH website and the activity tracker. Group B consists of 9 participants (with an average age: 29.44), including data from the COUCH website and the activity tracker. An overview of the Group A and Group B can be seen in Table 3 and

Table 4, respectively. It is worth mentioning that the total duration for the Group B does not represent the actual days that data was collected for these participants. For instance, the total duration for Subject 7 is 40 days (starting on 22/05/2020 and ending on 01/07/2020), while the actual days that data were collected are 25. This happens because the subjects did not answer the CaaS questions on the COUCH website on a daily basis. Another important remark is that the participants who installed the mobile app on iOS devices could not answer the ESM questions that are presented on Table 1.

Table 3: Overview of the collected data for the Group A.

Subject ID	Sex	Age	Mobile App	COUCH App	Activity Tracker	Duration in days
Subject01	female	58	Android	yes	Fitbit	53
Subject04	male	29	Android	yes	Fitbit	47
Subject09	female	30	Android	yes	Pedometer app	43
Subject12	male	29	Android	yes	Fitbit	41
Subject13	female	30	Android	yes	Fitbit	42
Subject14	female	37	Android	yes	Fitbit	43
Subject15	female	30	Android	yes	Fitbit	39
Subject17	female	33	Android	-	Other tracker	27
Subject18	female	25	Android	-	Garmin tracker	38
Subject19	female	30	Android	yes	Pedometer app	45
Subject20	male	54	Android	yes	Fitbit	40
Subject21	male	30	Android	yes	Fitbit	53

Table 4: Overview of the collected data for the Group B.

Subject ID	Sex	Age	Mobile App	COUCH App	Activity Tracker	Duration in days
Subject02	female	30	-	yes	Fitbit	46
Subject03	female	28	iOS	yes	iOS pedometer	45
Subject05	female	29	-	yes	Pedometer app	42
Subject06	male	32	iOS	yes	iOS pedometer	44
Subject07	female	29	iOS	yes	iOS pedometer	40
Subject08	female	30	-	yes	Pedometer app	44
Subject10	female	28	-	yes	Pedometer app	42
Subject11	male	30	-	yes	Pedometer app	44
Subject16	female	29	iOS	yes	iOS pedometer	39

4.2 Data types

During the data collection phase, four different types of data were acquired. These include raw and processed smartphone data through the mobile app, processed data from the activity trackers, users' answers from the questions asked through the mobile app and through the COUCH website, and users' demographics (including age, gender and occupation).

The smartphone collected data are used to build a model for detecting user's physical and social behaviour changes. Specifically, sensor data are used to predict the performed activities (walking, sitting, standing, cycling, taking the bus) and detect the level of user's social interaction and isolation. Data from activity trackers are used to enhance the model for detecting user's physical behaviour, while user's answers are used to build the model for detecting user's social, emotional and cognitive behaviour changes during and after the lockdown.

Smartphone data consist of accelerometer and GPS signals for tracking physical activity, while ambient noise, location, the number of incoming/outgoing calls and the number of received/sent text messages are used to detect the user's level of being socially active. The smartphone data types were thoroughly presented and described in D4.2 (Banos, Konsolakis, op den Akker, Pelachaud, & Bangalore, 2018). Further information can be also found here (Ferreira, 2020). It is worth mentioning that no raw audio or text content were stored. Concerning the smartphone audio (e.g., ambient noise and phone calls) and text messages (SMS), these were only used for annotating if a user was socially active, without examining the audio or text content. For instance, the audio decibels were used for modelling user's social behaviour instead of the audio content. Similarly, the GPS location was only used to cluster users' locations respecting the privacy concerns that might raise.

Data from activity trackers include the number of steps and the intensity of the performed activities. It is worthwhile to mention that Fitbit data are uploaded on the HBAF server, while steps data from other activity trackers are provided through excel files. Another data type is related to the ESM and CaaS questions, including users' answers with numeric and free text values. Further description of these questions can be found on Table 1 (mobile app) and Table 2 (COUCH website).

The aforementioned types of data are presented on Table 5.

Table 5: Overview of the collected data types.

Data Types	Sensing Device	Description
physical activity (collected through the plugin_google-_activity_recognition)	AWARE android	it describes user's movement and the mode of transportation, which are detected through the Google Location API (combining accelerometer and GPS data). The plugin can monitor the following activities: walking, cycling, using a vehicle (e.g., taking the bus), tilting and being still.
physical activity (collected through the plugin_ios-_activity_recognition)	AWARE iOS	it describes user's movement and the mode of transportation, which are detected through the iOS API (combining accelerometer and GPS data). The plugin can monitor the following activities: walking, cycling, using a vehicle (e.g., taking the bus), tilting and being still.
step counts (collected through the plugin_ios_pedometer)	AWARE iOS	it counts the number of steps, based on accelerometer data through the iOS API.
locations	AWARE android, iOS	it estimates user's location based on the GPS coordinates (longitude and latitude).
messages	AWARE android, iOS	it logs received and sent SMS texts, performed by or received by the user; it does not record personal information, such as phone numbers or contact information, but uses a unique ID based on SHA-1 encryption.
calls	AWARE android, iOS	it logs incoming and outcoming call events, performed by or received by the user; it does not record personal information, such as phone numbers or contact information, but uses a unique ID based on SHA-1 encryption.
conversations (collected through the plugin_studentlife_audio)	AWARE iOS	it identifies if there is an ongoing conversation based on the ambient noise (0 = voice/noise and volume, 1 = audio features, 2 = conversations). The plugin uses audio from the microphone sensor on iOS devices and detects if the user is engaged in a conversation or not. It follows a duty cycle of 1-minute audio collection, 3 minutes pause and it does not store the raw audio (it's disabled).
conversations (collected through the plugin_studentlife_audio_android)	AWARE android	it identifies if there is an ongoing conversation based on the ambient noise (0 = voice/noise and volume, 1 = audio features, 2 = conversations). The plugin uses audio from the microphone sensor on android devices and detects if the user is engaged in a conversation or not. It follows a duty cycle of 1 minute audio collection, 3 minutes pause

		and it does not store the raw audio (it's disabled).
users' answers (collected through the ESM questions)	AWARE android	it provides users' answers for the questions Q1-Q9 asked on android devices, including integer and text values. The data were collected based on the plugin_esm_scheduler which reads the events of a shared Google calendar (with the study participants) and triggers the ESM questions.
users' answers (collected through the CaaS questions)	COUCH website	it provides users' answers for the questions Q10-Q22 asked through the COUCH website, including integer and text values.
demographics	COUCH website	it provides users' answers for their sex, age and occupation.
step counts	Activity tracker	it counts the number of steps.

4.3 Privacy

The protocol for the evaluation study with the title "Monitoring human behaviour during the COVID-19 lockdown" was approved by the Ethics Committee EEMCS on 19 May 2020 with the registration number 'RP 2020-43'. Respecting participants' privacy, the collected data were anonymized and stored to a secure server at the University of Twente. All data were de-identified, and thus, neither real names nor private information were stored. Furthermore, data access has been limited only to researchers with authorized credentials and passwords. The acquired data will be stored on the server up to five years, in compliance with the UT data management plan.

It is also worth mentioning that no raw audio or text content were stored whatsoever. Concerning the data related to smartphone audio (e.g., ambient noise and phone calls) and text messages (SMS), these were only used for annotating if a user was socially active, without examining the audio or text content. For instance, the audio decibels were used for modelling user's social behaviour (instead of the audio content) and detecting if a conversation was taking place. Similarly, the GPS location was only used to cluster users' locations to indoors or outdoors (without recording user's address), respecting the privacy concerns that might raise. Thus, the processed GPS data provide information for the frequency of the visited locations and not the identification of these locations.

5 Results

5.1 Dataset: Collected Data

The dataset consists of raw data collected from 21 subjects. In particular, the acquired data types and duration vary per participant (see participant's group on Table 3 and Table 4). Regarding the physical behaviour, data related to the number of steps, the performed activities and the location were collected. An overview of the total number of steps for all the participants can be seen in Figure 6. An overview of the different types of performed activities can be seen in Figure 7, while the activities have been clustered to sedentary (including vehicle and stationary activities) and vigorous (including cycling, walking and running activities) in Figure 8. An overview of the collected GPS data can be seen in Figure 9, which varies per users' location (participants are located either in the Netherlands or in Greece). It is worthwhile mentioning that the GPS data will be clustered to indoors and outdoors locations (during the processing phase for detecting short-term behaviours) and will be used to extract the necessary features in order to describe the long-term physical, social, emotional and cognitive behaviours as was described on D4.4 (Banos & Konsolakis, 2019) and D4.5 (Banos & Konsolakis, 2019).

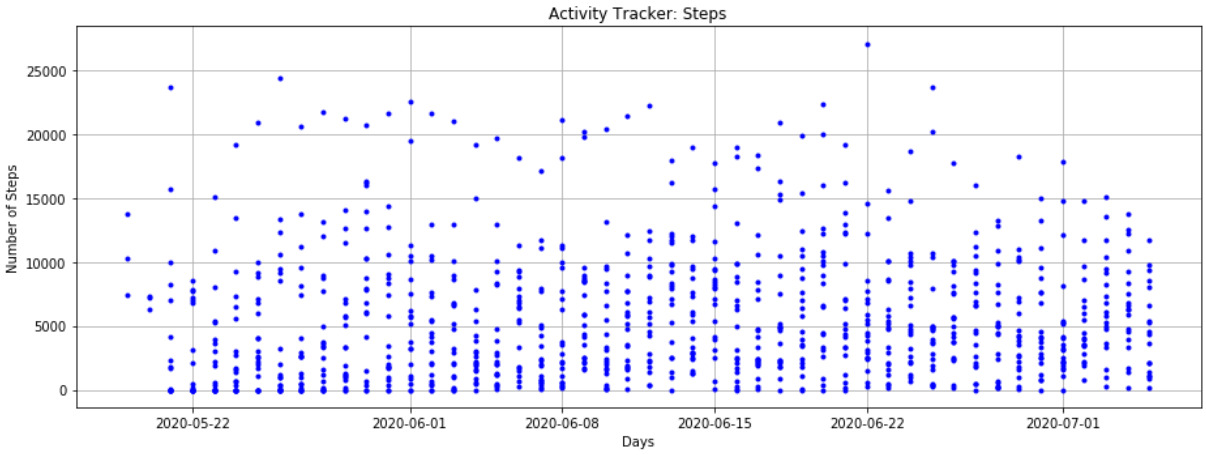


Figure 6: Total number of steps per day and subject, collected via the activity trackers.

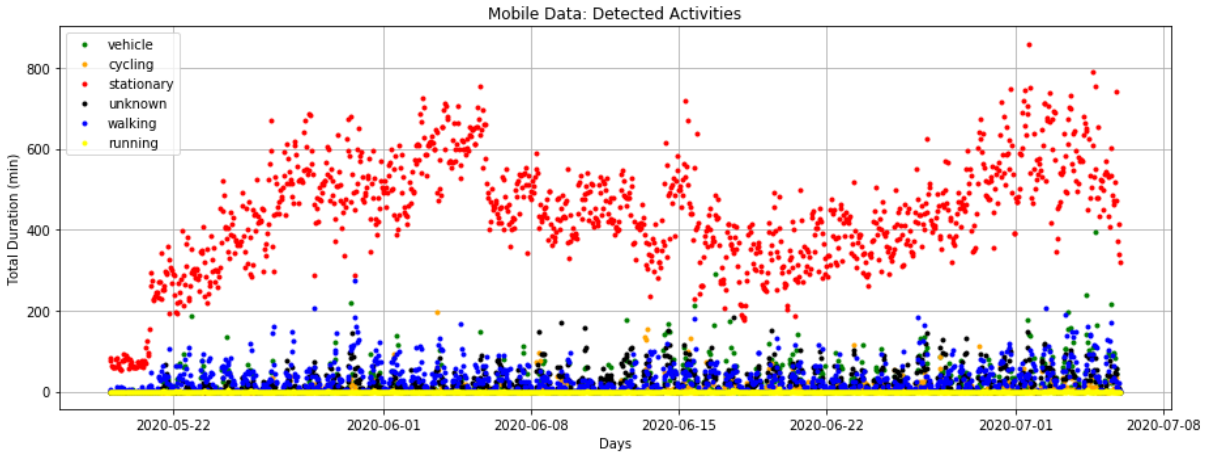


Figure 7: Total duration for each detected activity per day, collected via the mobile devices. The duration refers to the activities performed on an hourly basis for all subjects.

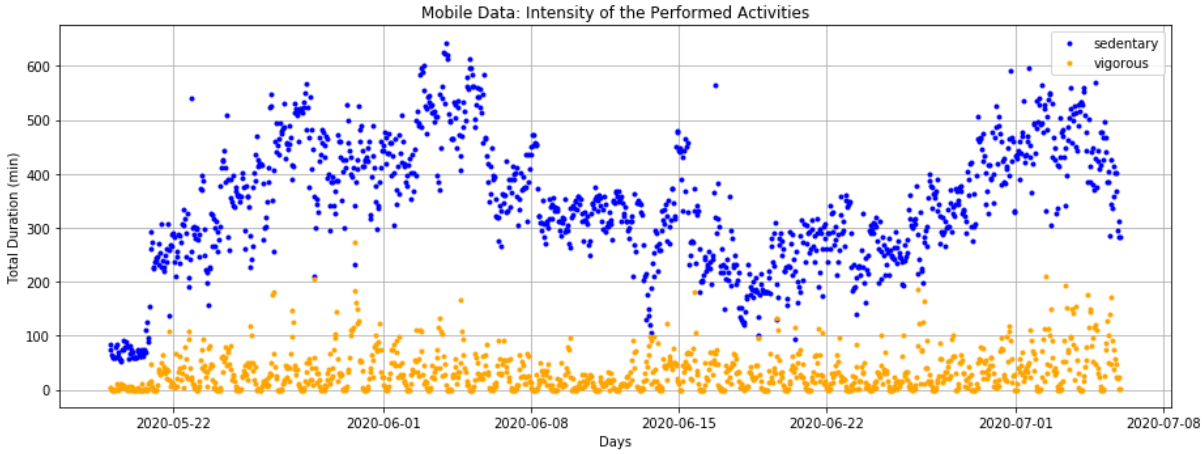


Figure 8: Total duration for the sedentary and vigorous activities per day, collected via the mobile devices. The duration refers to the activities performed on an hourly basis for all subjects.

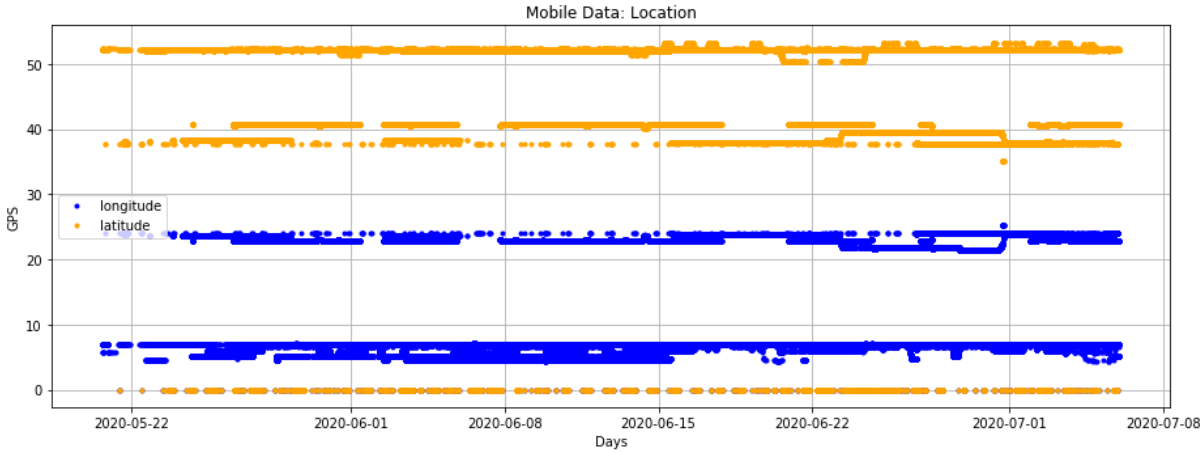


Figure 9: GPS (longitude and latitude) values per day and subject, collected via the mobile devices.

Regarding the social behaviour, mobile data related to the incoming/outgoing phone calls (duration of the conversations), received/sent text messages and ambient noise (duration of the conversations based on audio) were collected. An overview of all the phone calls can be seen in Figure 10. An overview of the text messages (SMS) sampled per hour can be seen in Figure 11. An overview of all the detected conversations (based on audio voice through the mobile microphone) can be seen in Figure 12.

In addition to the aforementioned mobile data for detecting social behaviour, users' answers were collected through mobile (ESM) and online questionnaires (CaaS). An overview of the users' answers for the ESM social questions can be seen in Figure 13, while the answers for the CaaS social questions can be seen in Figure 16.

Regarding the cognitive and emotional behaviour, data from mobile and online questions were collected. An overview of the users' answers for the ESM cognitive questions can be seen in Figure 15, while the answers for the CaaS cognitive questions can be seen in Figure 16. Similarly, an overview of the users' answers for the ESM emotional questions can be seen in Figure 14, while the answers for the CaaS emotional questions can be seen in Figure 17.

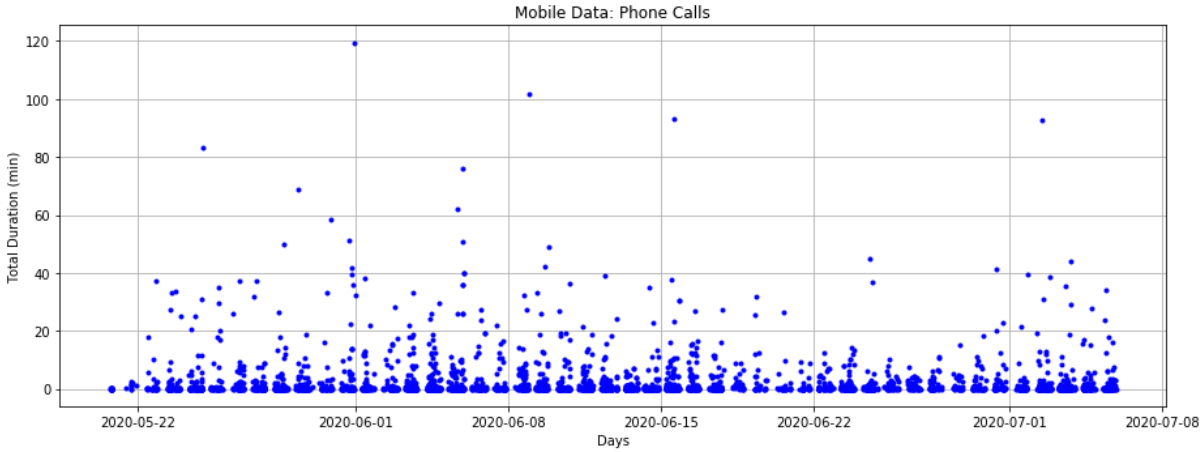


Figure 10: Total duration for each phone call per day and subject, collected via the mobile devices.

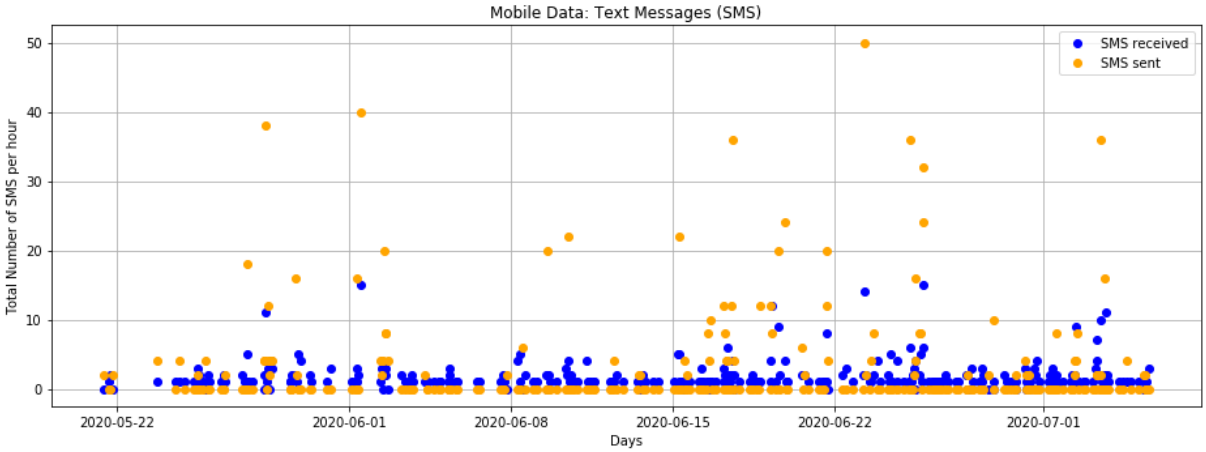


Figure 11: Total number of received/sent text messages (SMS) per day, collected via the mobile devices. This number refers to all the SMS received/sent on an hourly basis for all subjects.

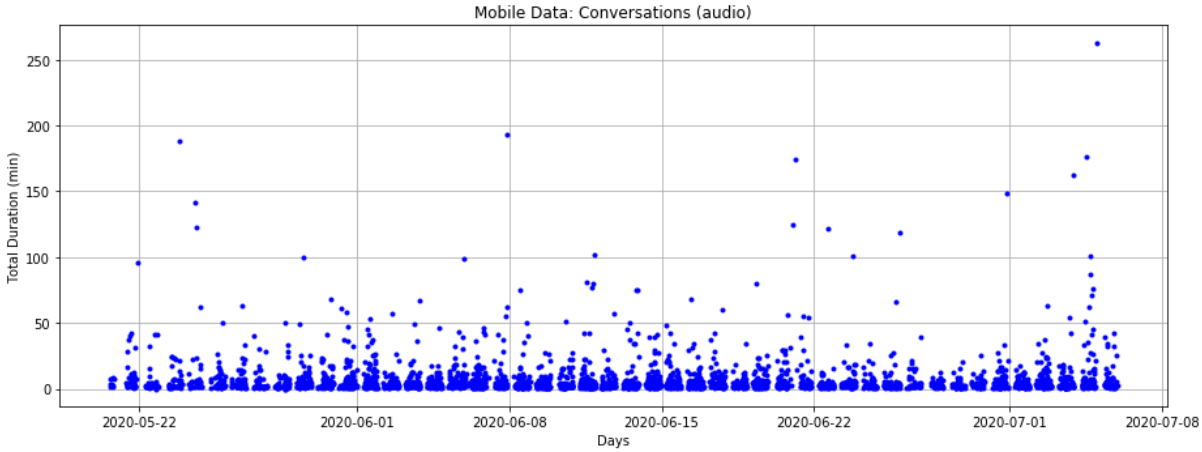


Figure 12: Total duration for each audio conversation per day and subject, collected via the mobile devices.

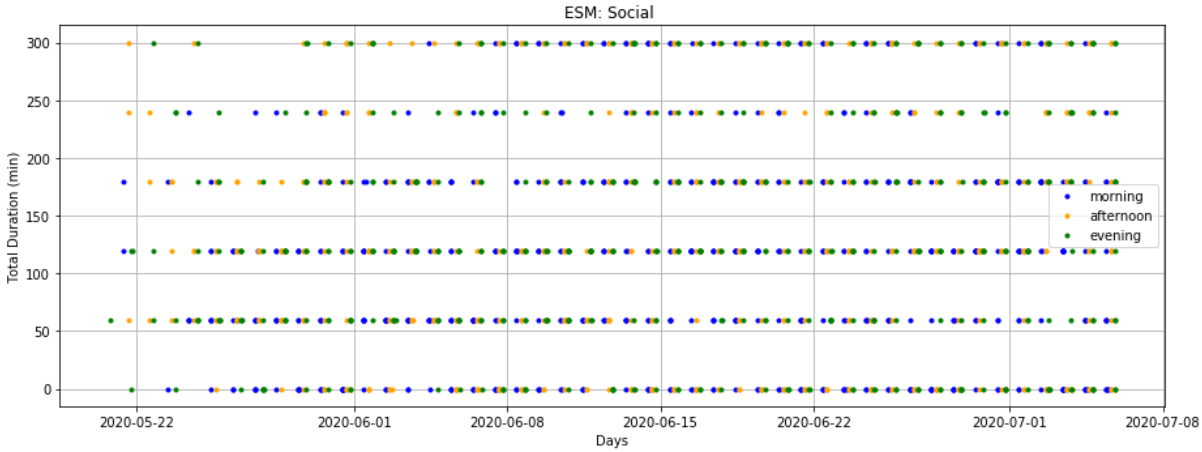


Figure 13: ESM - Social answers per day and subject for being socially active, collected via the mobile devices.

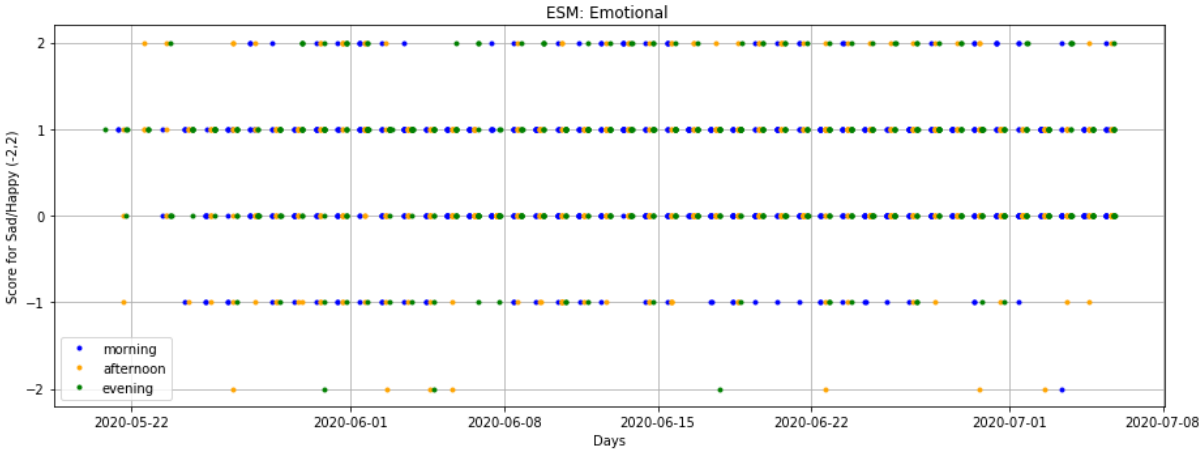


Figure 14: ESM - Emotional answers per day and subject for being sad or happy, collected via the mobile devices; -2 score refers to very sad, -1 refers to sad, 0 refers to neutral, 1 refers to happy and 2 refers to very happy.

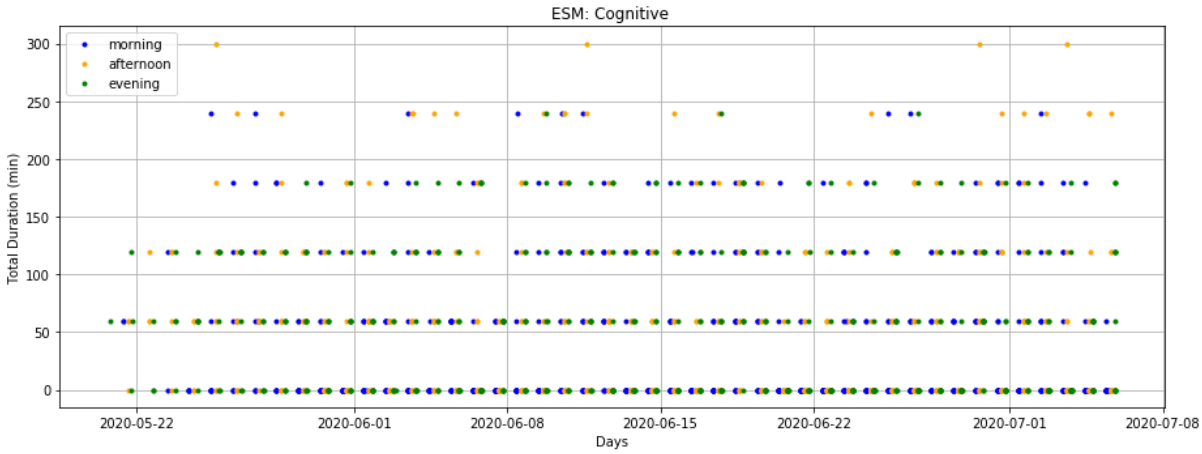


Figure 15: ESM - Cognitive answers per day and subject for being involved into cognitive tasks, collected via the mobile devices.

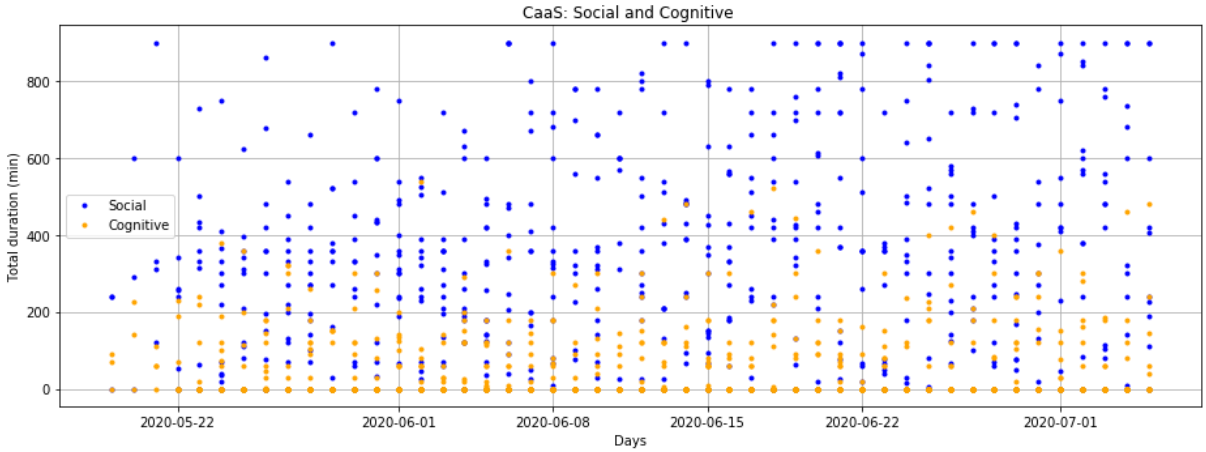


Figure 16: CaaS data with the Social and Cognitive answers per day and subject, collected via the website.

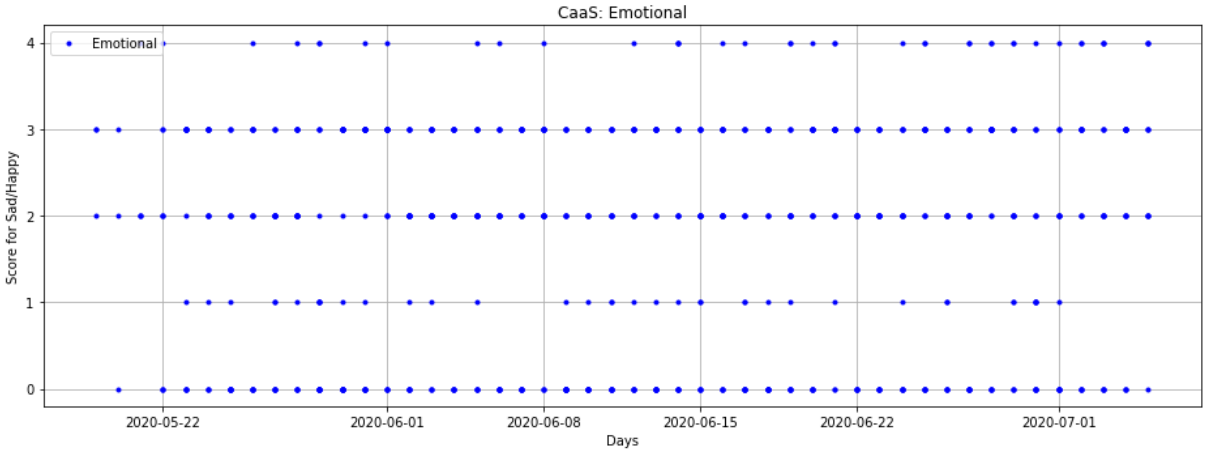


Figure 17: CaaS data with the Emotional answers per day and subject for being sad or happy, collected via the website; 0 score refers to very negative, 1 refers to negative, 2 refers to neutral, 3 refers to positive and 4 refers to very positive.

5.2 Dataset: Processed Data

5.2.1 Short-term Behaviours

After the data collection phase, data are processed by the HBAF in order to transform the raw sensor data into physical, social, emotional and cognitive short-term behaviours for each participant. Depending on the participants' group, different types of data are used. For the Group A, data from users' smartphones, activity trackers and questionnaires (both ESM and CaaS) are processed. For the Group B, data from activity trackers and CaaS questions are processed.

It is worth mentioning that data related to activity trackers on Group B are sampled every day, while mobile and activity trackers data on Group A are sampled every minute. The minute-sampling gives the opportunity to extract some additional features with more meaningful information, such as the number of steps during the morning/afternoon/evening hours, during the working hours, etc. (see D4.4 for further explanation), which can be later used for the inference of long-term behaviours. This is only possible for Group A and only for the subjects who have an activity tracker. A day-sampling is rather applied to Group B.

The **short-term physical behaviour** can be described by the number of steps in the range of minutes per day. For instance, the total number of steps per day is depicted for Subject01 in Figure 18. As can be seen in the time-series plot, there is a significant increase on the performed number of steps during the first week of June (3rd week of data collection) which can be an estimator of a positive trend for the physical behaviour. For the sake of simplicity, we have considered Subject01 (Group A) to showcase the short-term behaviour inference results of the HBAF. It is important to mention that Subject01 is the only participant who adheres to the COUCH original inclusion criteria for the HBAF evaluation (elderly person with chronic pain).

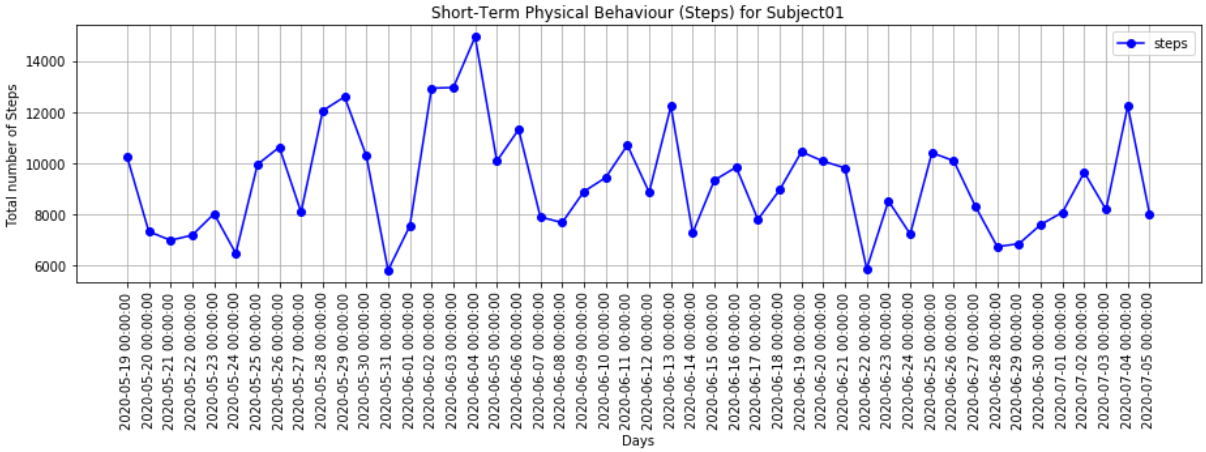


Figure 18: The Short-Term Physical Behaviour for Subject01 based on the steps.

In addition to the steps data, short-term physical behaviour can be described by the intensity of the detected activities (clustered to sedentary or vigorous) and by the location clustering to indoors/outdoors (based on the GPS data). Indoors location refers to the time when the user is located inside the house (in our scenario normally due to the lockdown restrictions), while outdoors location refers to elsewhere. The short-term physical behaviour for the Subject01, based on the activity's intensity and user's location, can be seen in Figure 19. The picture showcases the large difference of being indoors and sedentary compared to being outdoors and vigorously active.

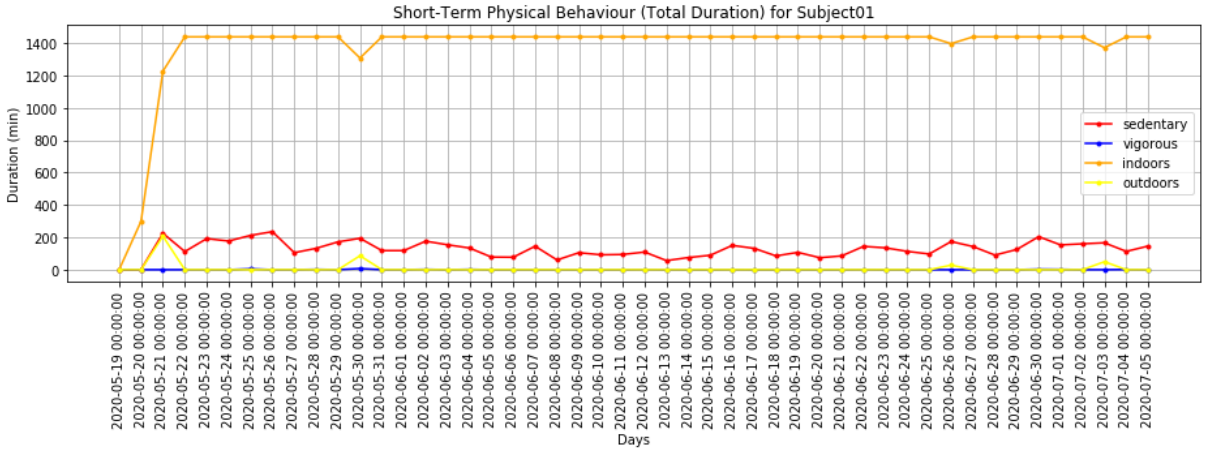


Figure 19: The Short-Term Physical Behaviour for Subject01 based on the activity intensity and location.

The **short-term social behaviour** can be described by the mobile sensors data that were presented in D4.2 (which includes phone usage parameters such as the duration of incoming/outgoing phone calls, the number of received/sent text messages, the duration of detected conversations based on ambient audio), and the users' answers to the ESM and CaaS social questions. Thus, we decided to follow a data fusion approach by averaging the detected duration of being socially active based on social sensors, ESM and CaaS data when these are available (see Figure 20). However, this approach can be only used for the Group A participants. The short-term social behaviour for the Group B participants can be described based on the CaaS social questions.

In Figure 20, the social behaviour time-series data for the Subject01 show that the user had been more socially active at the end of June, which can be a clear indicator that the subject was more socially active when the lockdown ended. Another important remark is that the total duration of being socially active via the CaaS, ESM and mobile sensors show similar patterns for most of the days, while the data fusion approach suffices for the days that this is not feasible. It is also important to mention that the CaaS zero values refer to days with missing data; the subject did not answer the CaaS questions through the website for these days.

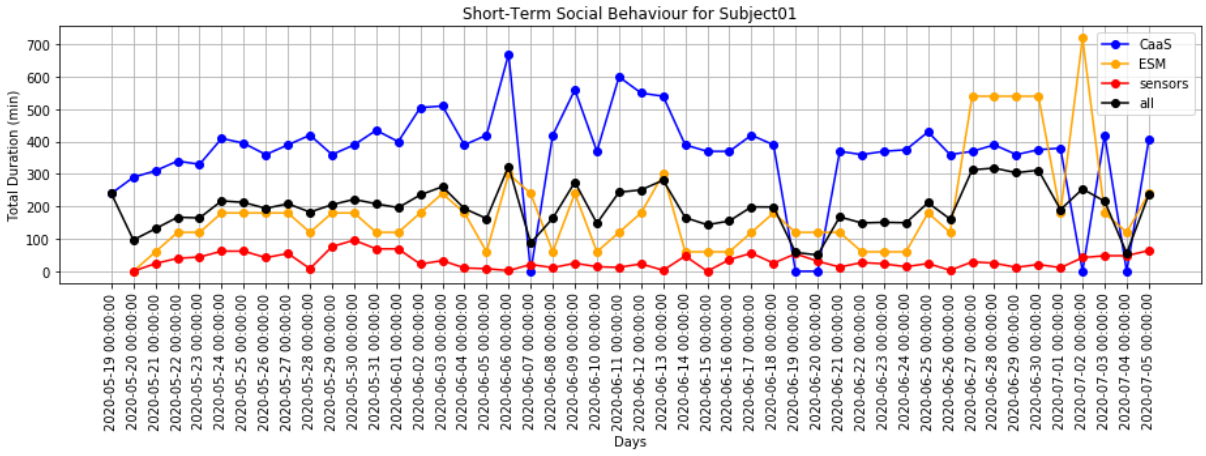


Figure 20: The Short-Term Social Behaviour for Subject01 based on the total duration of being socially active; including the data fusion approach (all), mobile data (sensors), ESM and CaaS data.

The **short-term emotional behaviour** can be described by the total duration of being happy or sad through the users' answers to the ESM and CaaS emotional questions (following a similar data fusion approach). More specifically, participants were asked to estimate the emotional score of being happy or sad. If the score was negative (-2 or -1), this information was converted to being sad for 5 hours (which represent the total time during the morning, afternoon and evening hours). If the score was positive (1 or 2), we estimated that the participant was happy for 5 hours, while we did not consider the

neutral case that the score was zero. For instance, the short-term emotional behaviour for Subject01 can be seen in Figure 21 (total duration of being happy) and Figure 22 (total duration of being sad). Even though there are some deviations on the emotional behaviour of the user, both figures show that the Subject01 was progressively happier when the lockdown ended (last days of June).

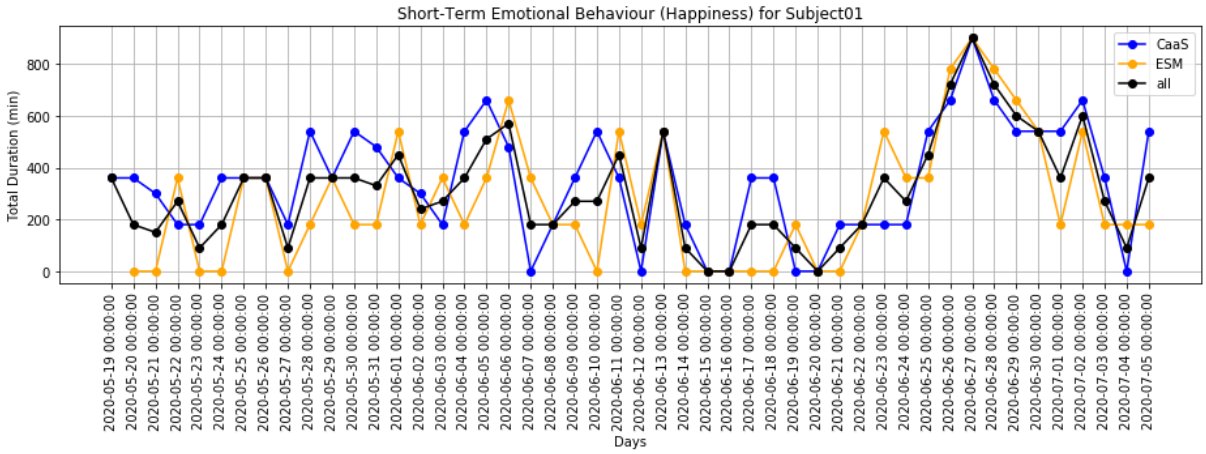


Figure 21: The Short-Term Emotional Behaviour for Subject01 based on the total duration of being happy; including the data fusion approach, CaaS and ESM data.

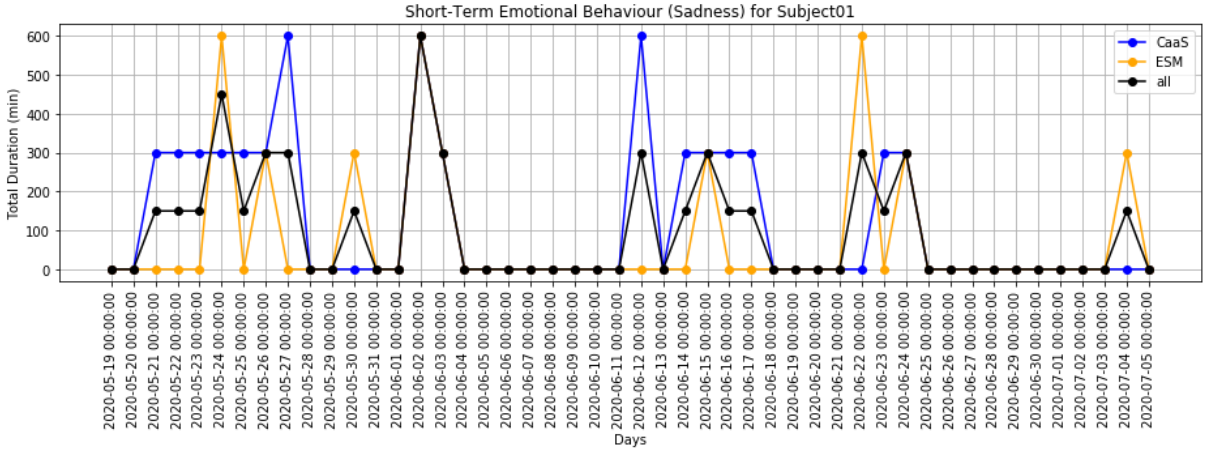


Figure 22: The Short-Term Emotional Behaviour for Subject01 based on the total duration of being sad, including the data fusion approach (all), CaaS and ESM data.

Similar to the emotional behaviour model, the **short-term cognitive behaviour** can be described by the total duration of being involved into cognitive tasks and based on users' answers on the ESM and CaaS cognitive questions. Cognitive tasks include activities such as reading a book, playing a board game, etc. For instance, the short-term cognitive behaviour for Subject01 can be seen in Figure 23. As figure shows, Subject01 spent more time into cognitive tasks during the last days of June, which overlaps with the previous figures related to the social and emotional behaviour model.

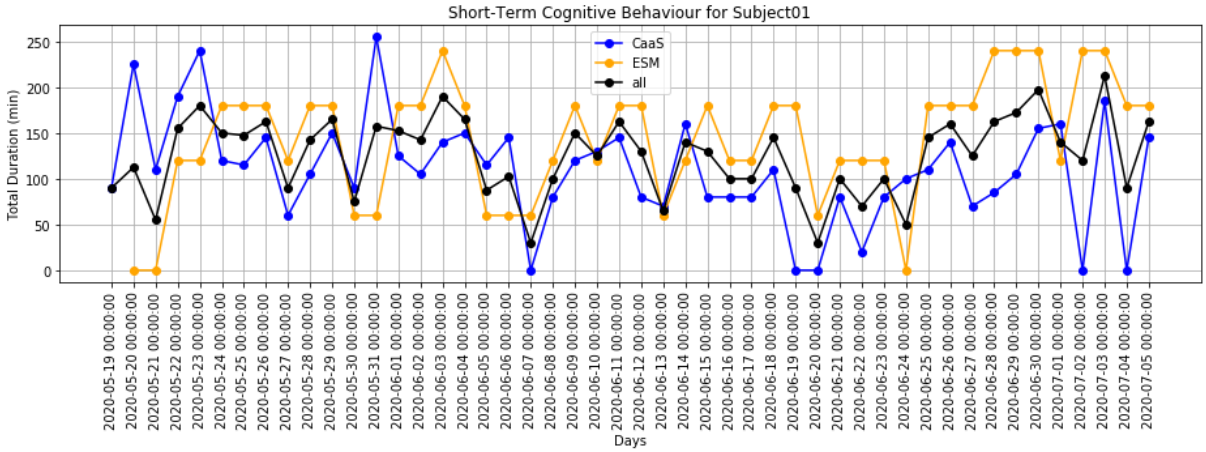


Figure 23: The Short-Term Cognitive Behaviour for Subject01 based on the total duration of being involved into cognitive tasks, including the data fusion approach (all), CaaS and ESM data.

5.2.2 Long-term Behaviours

After the inference of short-term behaviours, data are further processed in order to detect physical, social, emotional and cognitive long-term behaviours. For further information see Tables 3-9 in the deliverable D4.4 (Banos & Konsolakis, 2019). The long-term behaviours model uses the short-term behaviours and extracts additional features when the behaviours are performed during a certain timeframe. For instance, for the long-term physical behaviour based on steps, we extract the following features:

a. The total number of steps per week;
b. The total number of steps during the morning hours (8am-12pm) per week;
c. The total number of steps during the afternoon hours (12pm-5pm) per week;
d. The total number of steps during the evening hours (5pm-12am) per week;
e. The total number of steps during the daytime hours (7am-12am) per week;
f. The total number of steps during the working hours (8am-4pm) per week;
g. The total number of steps during the weekdays (Monday-Friday) per week;
h. The total number of steps during the weekends (Saturday-Sunday) per week;
i. The total number of steps performed indoors per week and;
j. The total number of steps performed outdoors per week.

As mentioned previously (see Section 5.2.1 above), all the aforementioned features can be extracted for the participants in Group A, while only a few features can be calculated for the participants in Group B due to missing information (such as data related to the morning/afternoon/evening hours and the indoors/outdoors location).

After the feature extraction, we calculate the differences of each value compared to the same value from the previous week. For instance, the difference of the total steps for a week, compared to the steps from the previous week. Then, we cluster the long-term behaviours based on the average value of the aforementioned features, and we estimate if there is a trend (positive or negative) or if there is no difference between two weeks.

The long-term physical behaviour can be described by the summary of the averaged features for the steps, the sedentary duration, the vigorous duration, the indoors duration and the outdoors duration. For example, if the subject is more physically active during the week 22 compared to the week 21, then there is a positive trend. A similar data fusion approach is followed for the long-term social, emotional and cognitive behaviours in order to detect the weekly trends. The predicted versus the actual trends (based on the ground truth given from the users through the questions Q8 and Q9) are depicted in Figure 24 for the long-term behaviours for Subject01.

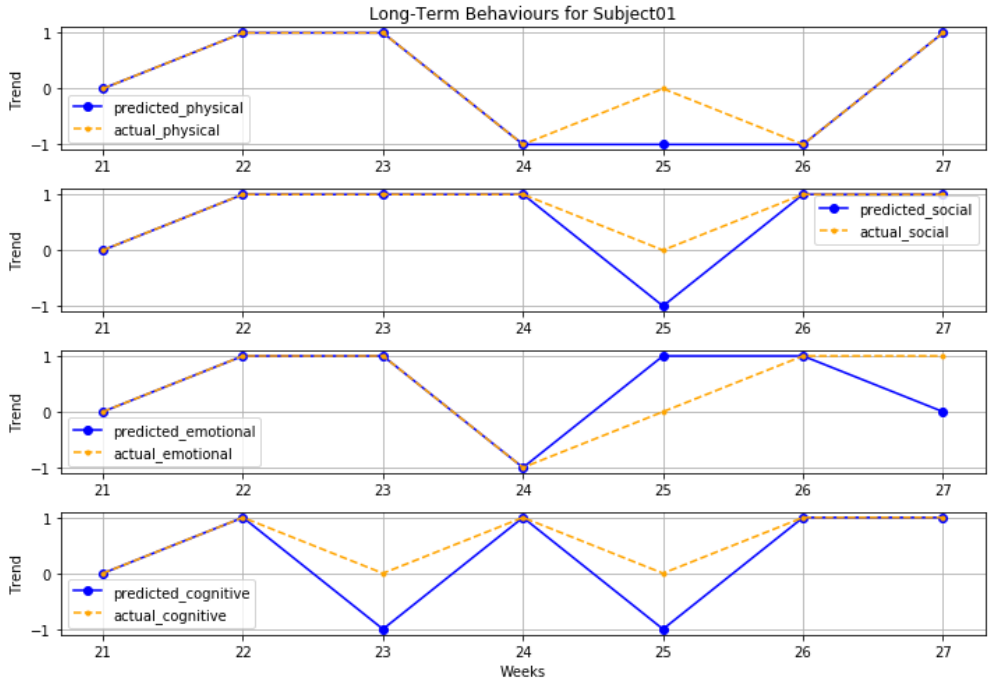


Figure 24: Long-Term Behaviours for Subject01, including positive, negative or no trends.

In the following figures Figure 25, Figure 29 and Figure 27, we also present the long-term behaviours (the actual versus the predicted trends) for the Subject04 , Subject14 and Subject19. The data of these subjects will be used for the HBAF evaluation in the following sections. According to these pictures, it is clear that the long-term behaviours deviate between each user's behaviours, but also among users. For instance, the period between week22 and week23 is marked with a positive trend for the Subject01, while a negative trend is noticed for the Subject04. However, it can be said that the physical, social, emotional long-term behaviours can be described by following similar patterns, while this is not the case for the cognitive long-term behaviours (see Figure 27 for Subject19).

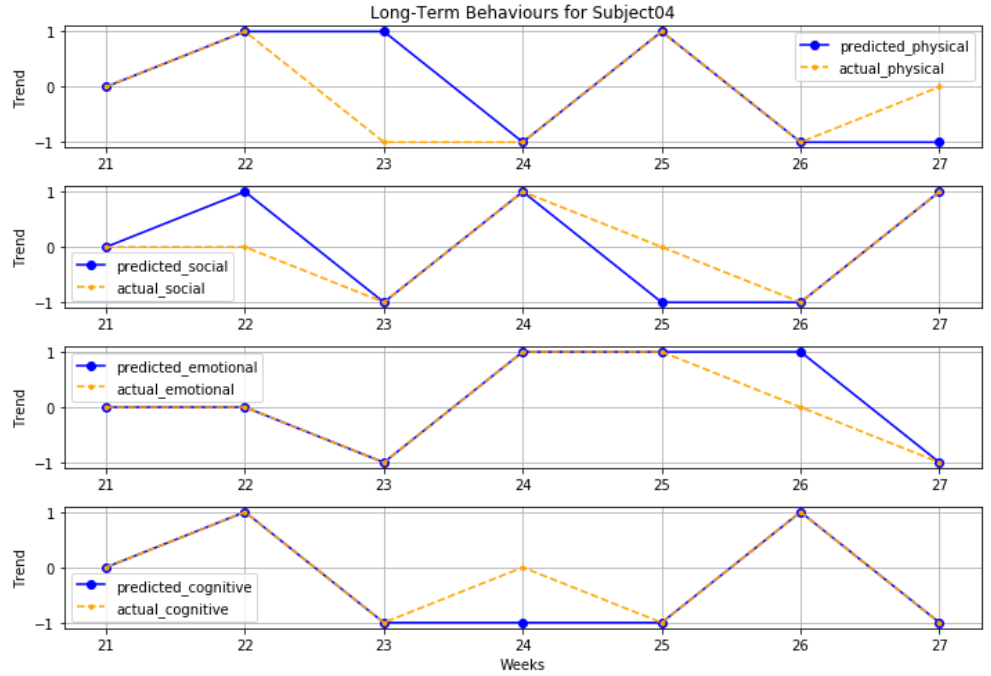


Figure 25: Long-Term Behaviours for Subject04, including positive, negative or no trends.

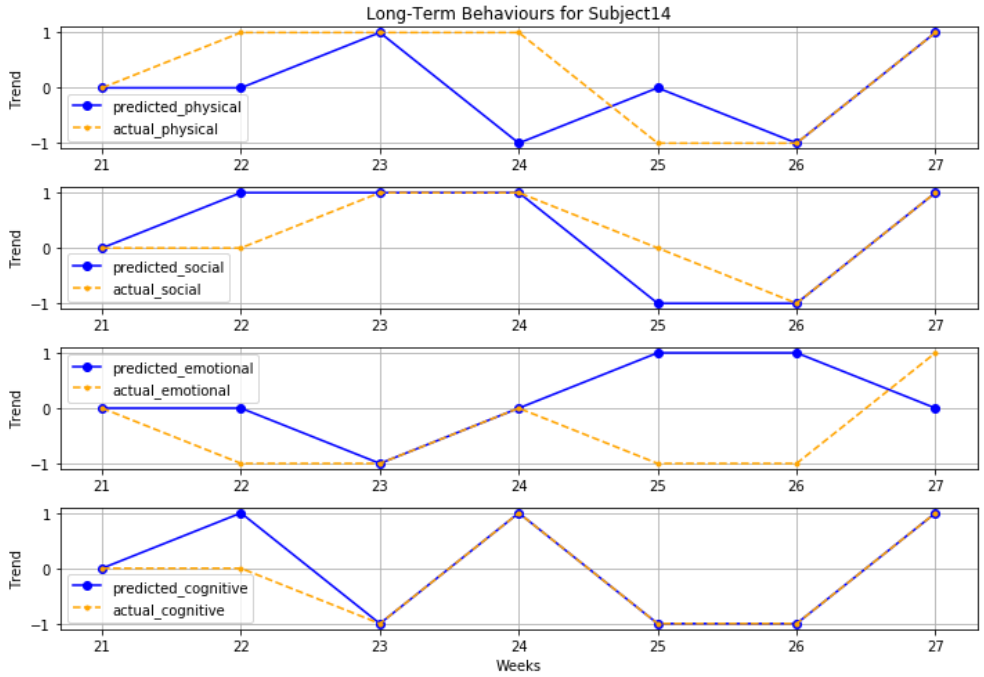


Figure 26: Long-Term Behaviours for Subject14, including positive, negative or no trends.

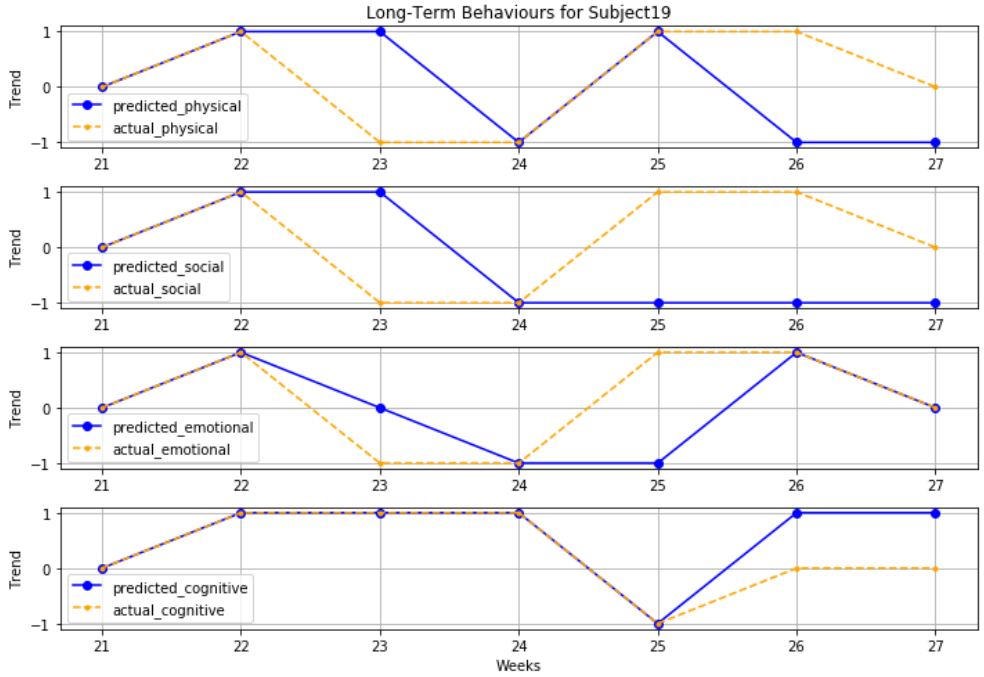


Figure 27: Long-Term Behaviours for Subject19, including positive, negative or no trends.

5.2.3 Behaviour Changes

After the inference of long-term behaviours and the estimation of trends, we apply the change point detection algorithm in order to detect if there is a significant change or not over the time series data. We thoroughly elaborated the model for detecting behaviour changes in the deliverable D4.6 (Banos & Konsolakis, 2019). In the following figures, see from Figure 28 to Figure 34, we present the detected physical, social, emotional, and cognitive behaviour changes for Subject01, where alternating colours designate segments with estimated change points. Specifically, the pink areas mark the estimated start and end point for a change, while the dashed line marks the time where a true change occurred (the actual end of lockdown which varies per participant). It is clear from these pictures, that the detected changes overlap for most of the behaviours on a weekly basis, where three noticeable changes at least occur; one change during the week before the end of the lockdown, one change during the week that the lockdown ended, and one change after the end of the lockdown.

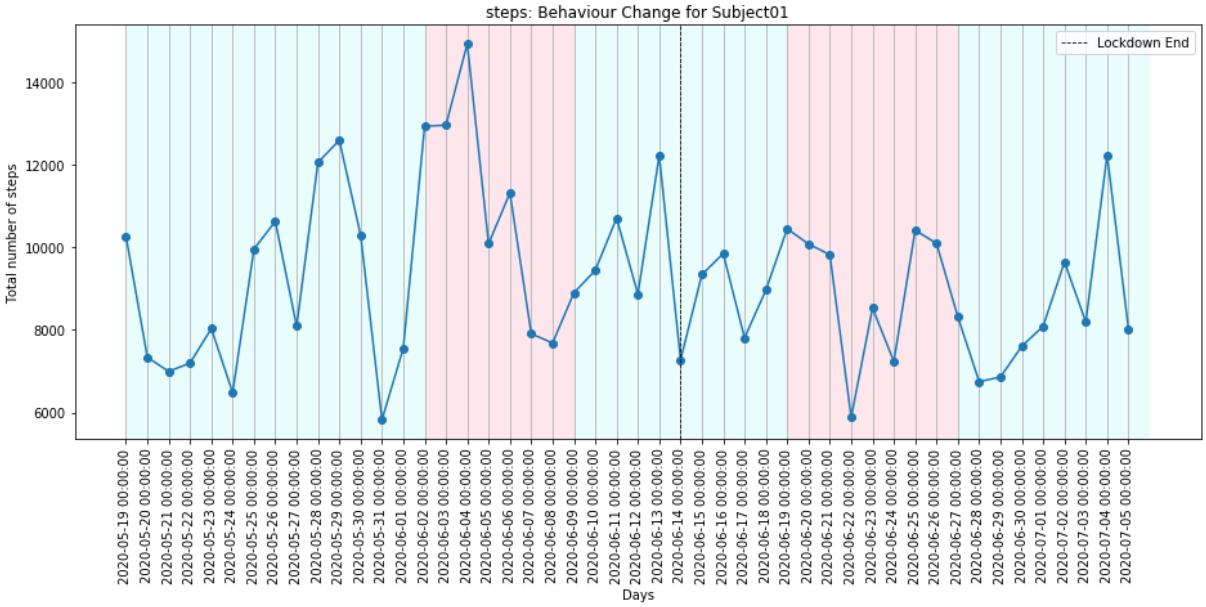


Figure 28: The Physical Behaviour Change for Subject01 based on the steps.

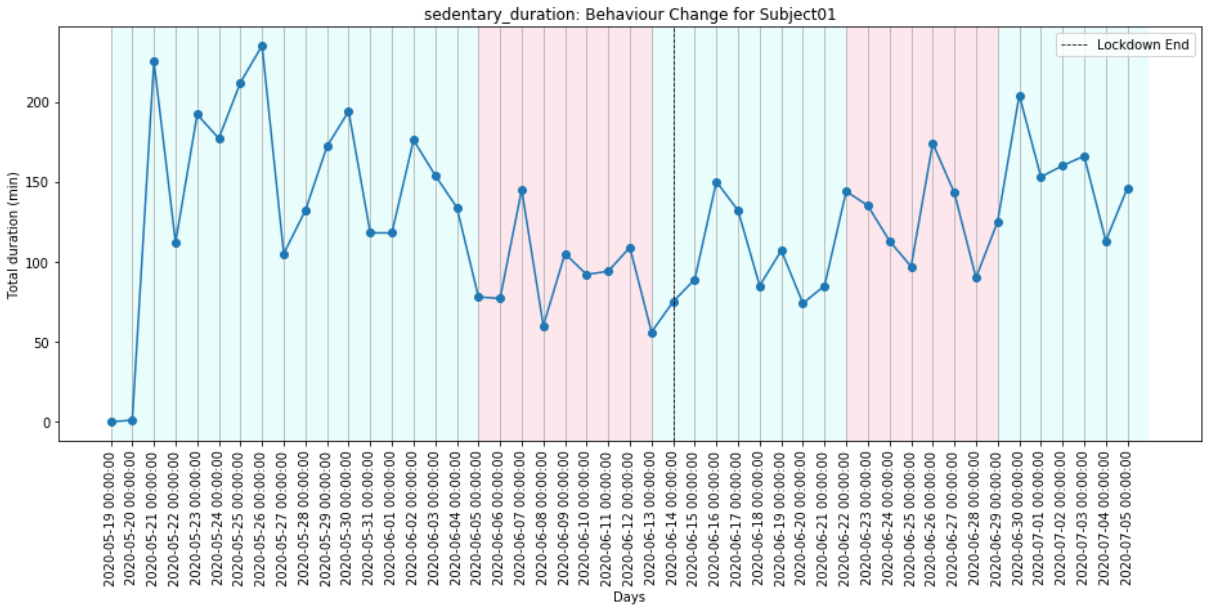


Figure 29: The Physical Behaviour Change for Subject01 based on the sedentary duration.

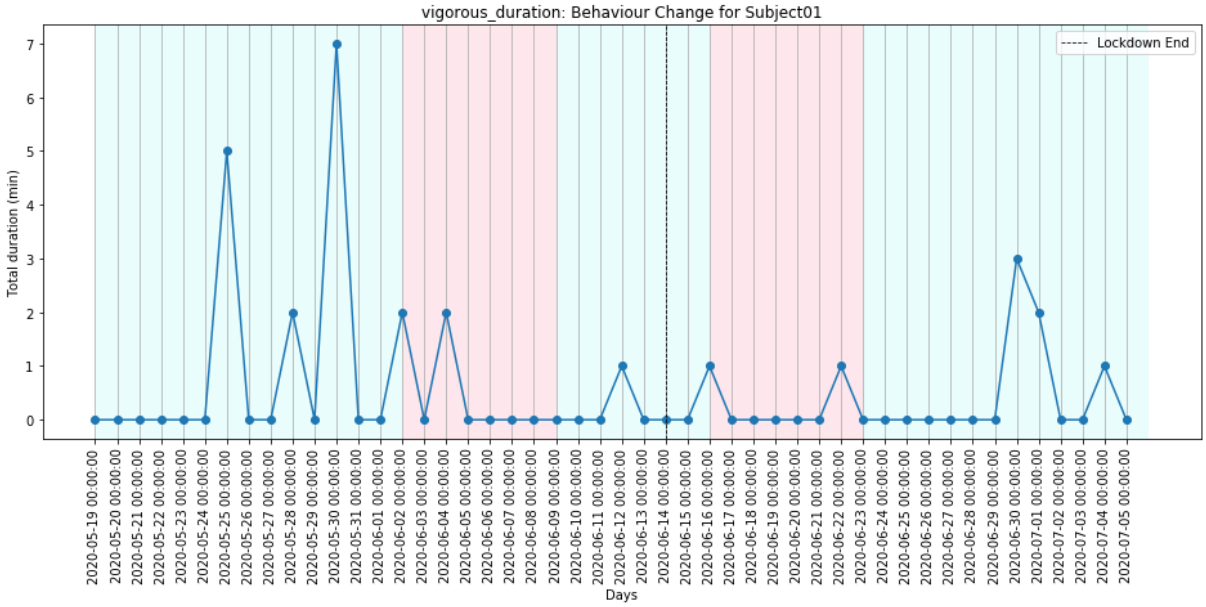


Figure 30: The Physical Behaviour Change for Subject01 based on the vigorous duration.

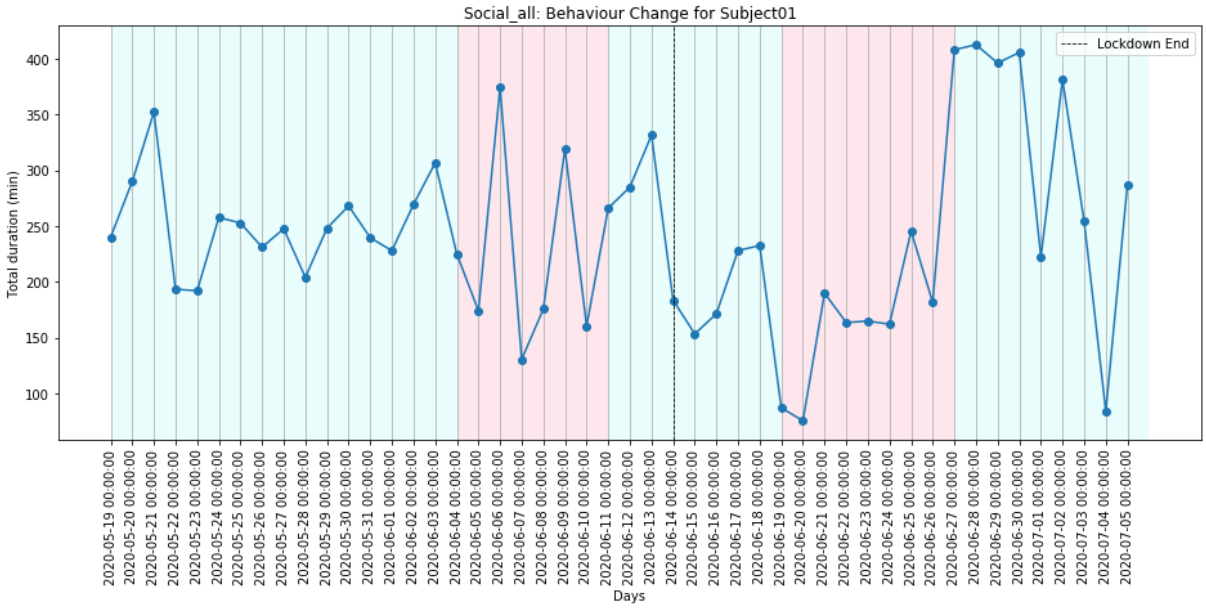


Figure 31: The Social Behaviour Change for Subject01 based on the total duration of being socially active (through the data fusion approach).

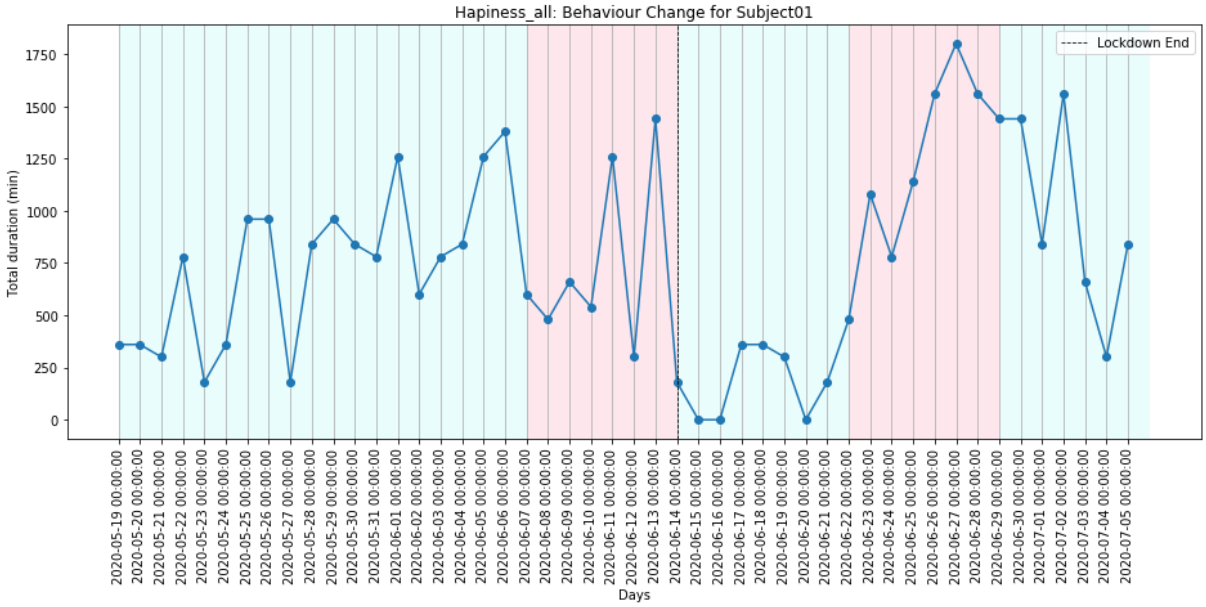


Figure 32: The Emotional Behaviour Change for Subject01 based on the total duration of being happy (through the data fusion approach).

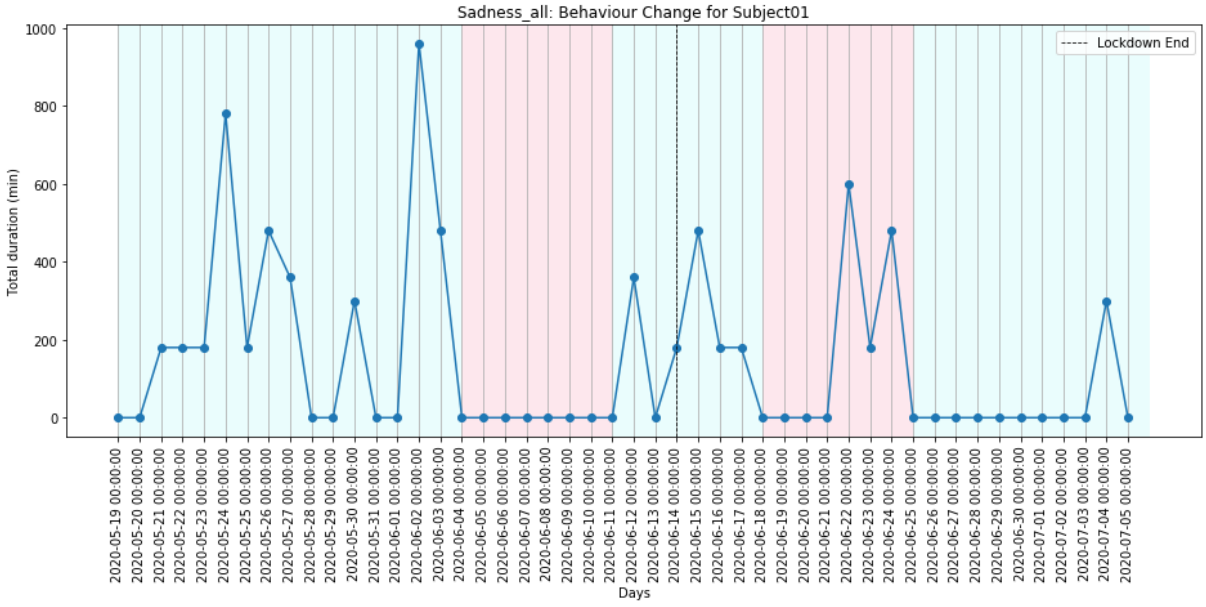


Figure 33: The Emotional Behaviour Change for Subject01 based on the total duration of being sad (through the data fusion approach).

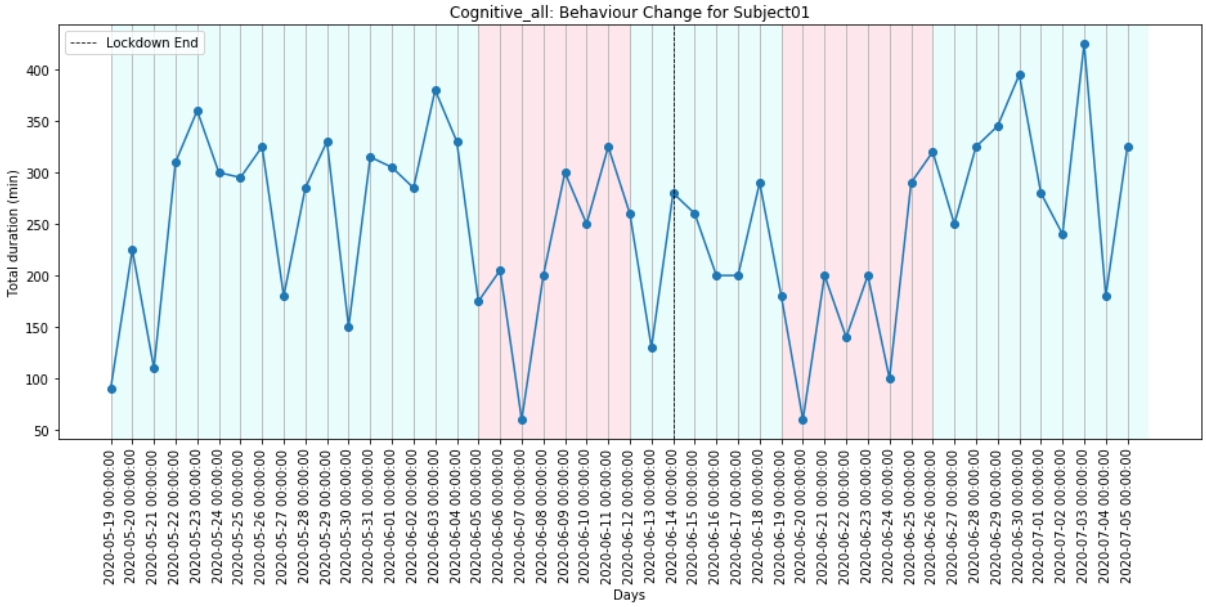


Figure 34: The Cognitive Behaviour Change for Subject01 based on the total duration of being involved into cognitive tasks (through the data fusion approach).

Furthermore, we present the results for detecting behaviour changes for the Subject04, Subject14 and Subject19 (see Figure 34 - Figure 46). These are the subjects who will be used for the HBAF evaluation in the following sections. For the sake of simplicity, only the time series with steps are shown for the physical behaviour changes and only the happiness time series for the emotional behaviour changes.

Regarding the Subject04, it can be seen from the figures that the period that a change occurs is similar for both physical, social, emotional and cognitive behaviours. In particular, the first big noticeable change takes place on 02/06/2020. However, the duration that a change takes place differentiates among the behaviours. Similar to Subject01 figures, the period when the lockdown ended can be described by a significant change.

Regarding the Subject14, the first change overlaps with the time that the lockdown ended, while the other changes can be easily seen as deviations in the time series plots. Similarly, the detected changes for Subject19 can be easily noticed. It is worth mentioning that the lockdown for this user ended on 21/05/2020 (three days after the data collection started), where the collected data up to this date were not enough to define a change. Consequently, it is clear that the historical data play a major role in defining when there is an anomaly over the time-series, resulting to accurate predictions of behaviour changes.

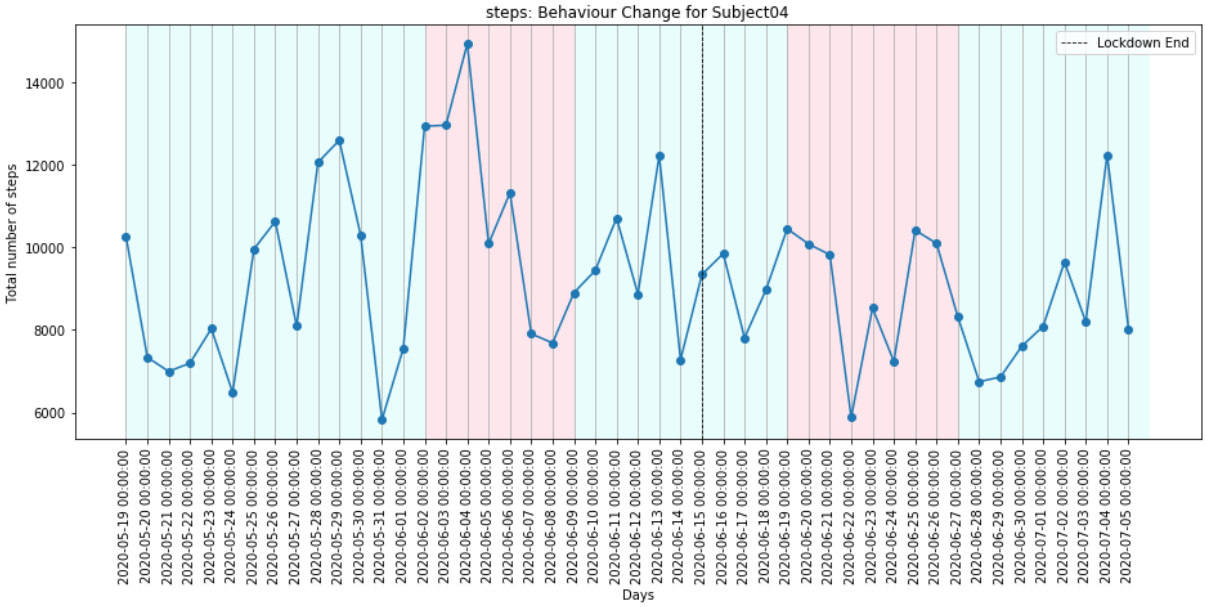


Figure 35: The Physical Behaviour Change for Subject04 based on the steps.

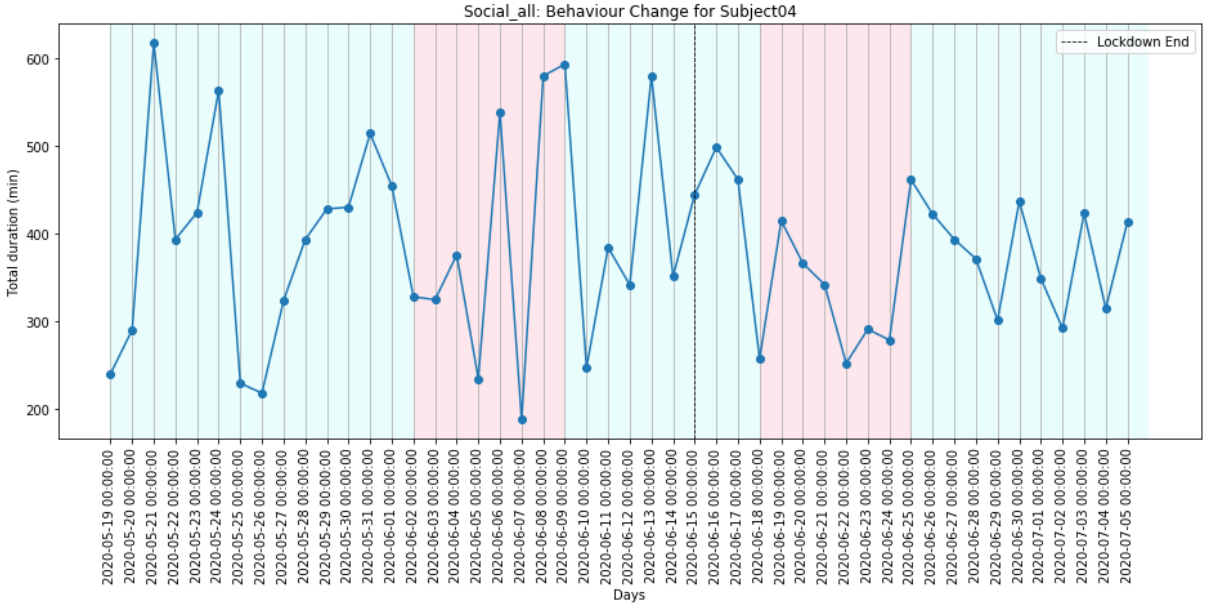


Figure 36: The Social Behaviour Change for Subject04 based on the total duration of being socially active (through the data fusion approach).

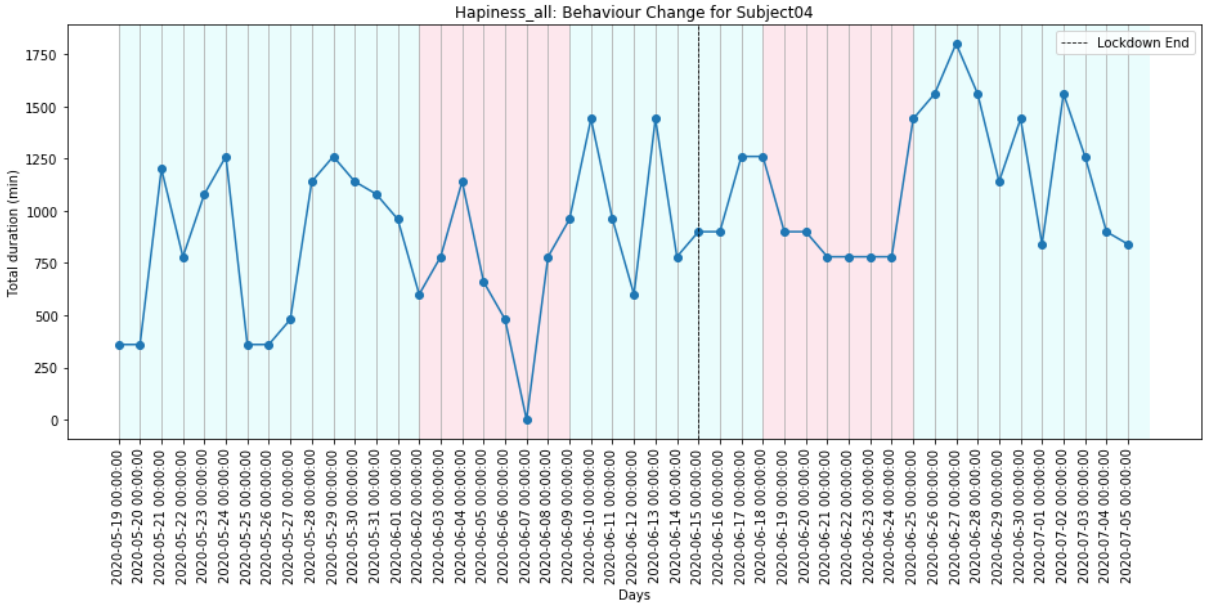


Figure 37: The Emotional Behaviour Change for Subject04 based on the total duration of being happy (through the data fusion approach).

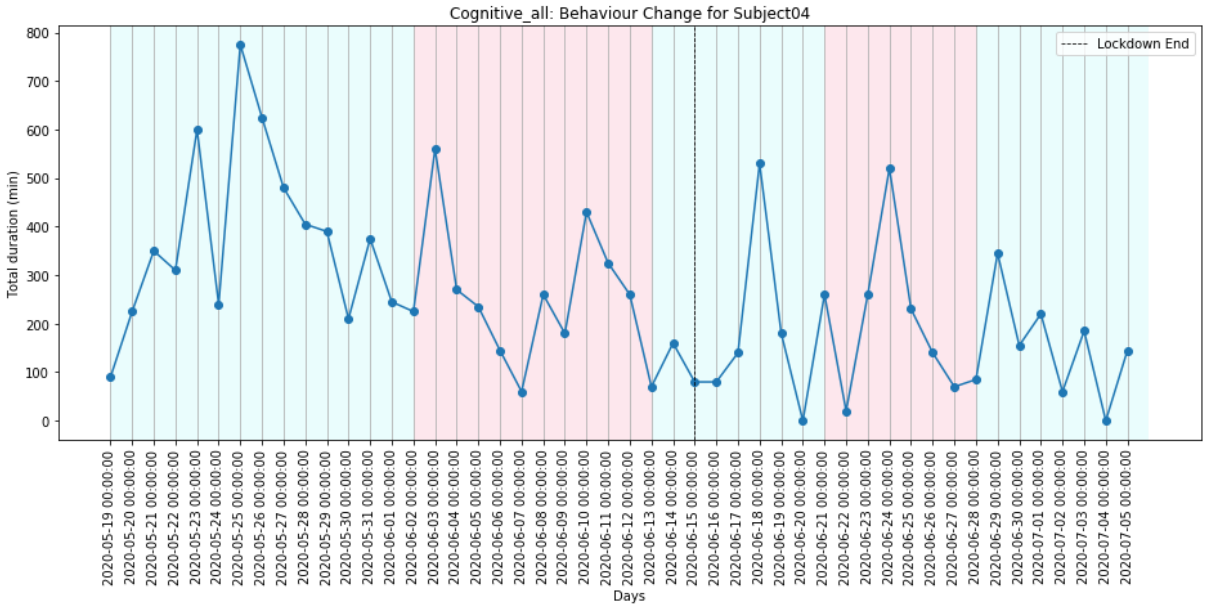


Figure 38: The Cognitive Behaviour Change for Subject04 based on the total duration of being involved into cognitive tasks (through the data fusion approach).

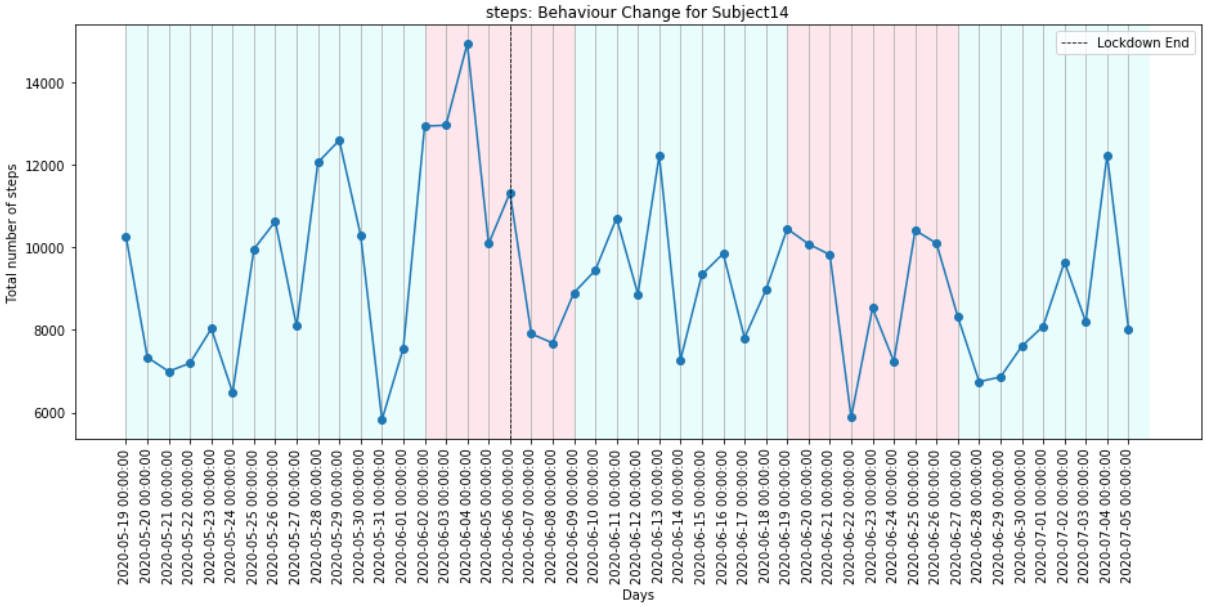


Figure 39: The Physical Behaviour Change for Subject14 based on the steps.

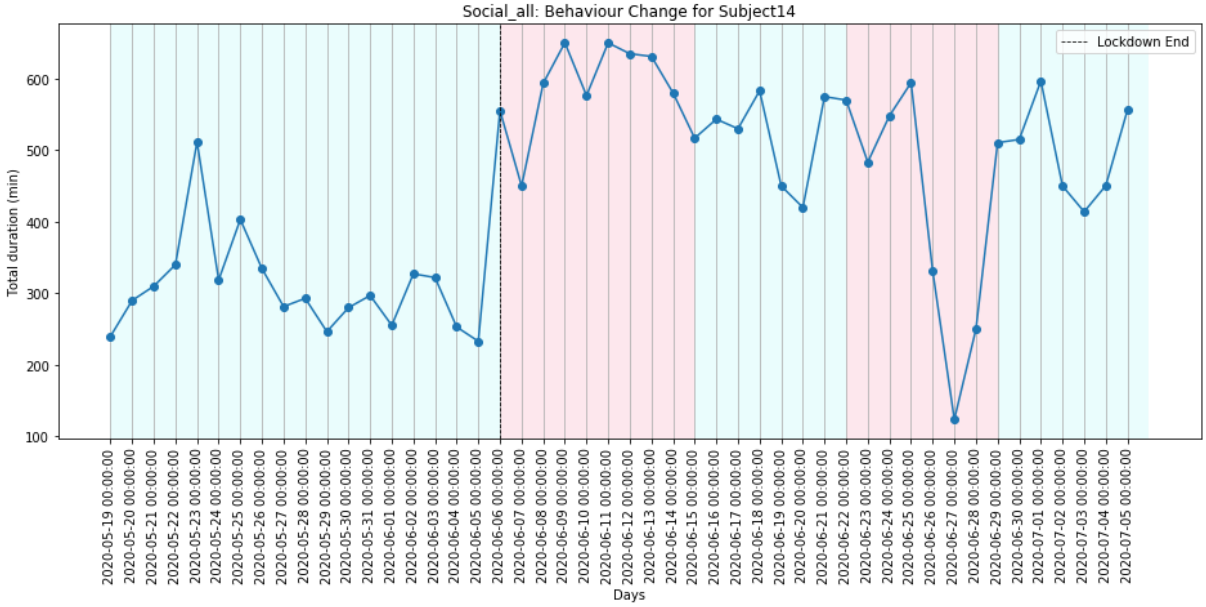


Figure 40: The Social Behaviour Change for Subject14 based on the total duration of being socially active (through the data fusion approach).

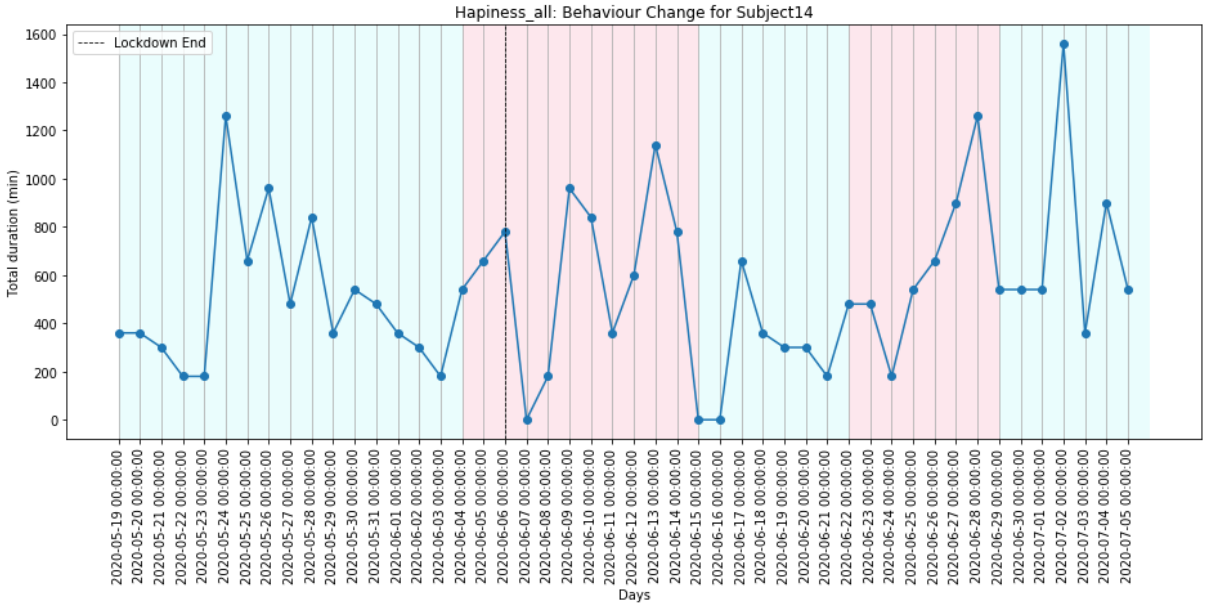


Figure 41: The Emotional Behaviour Change for Subject14 based on the total duration of being happy (through the data fusion approach).

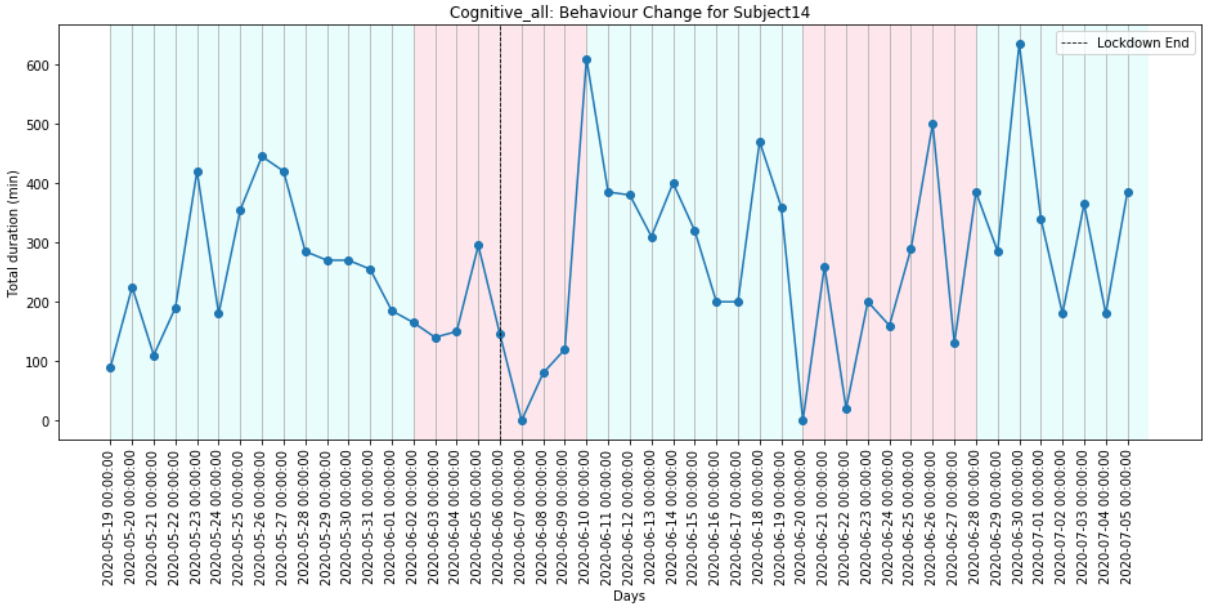


Figure 42: The Cognitive Behaviour Change for Subject14 based on the total duration of being involved into cognitive tasks (through the data fusion approach).

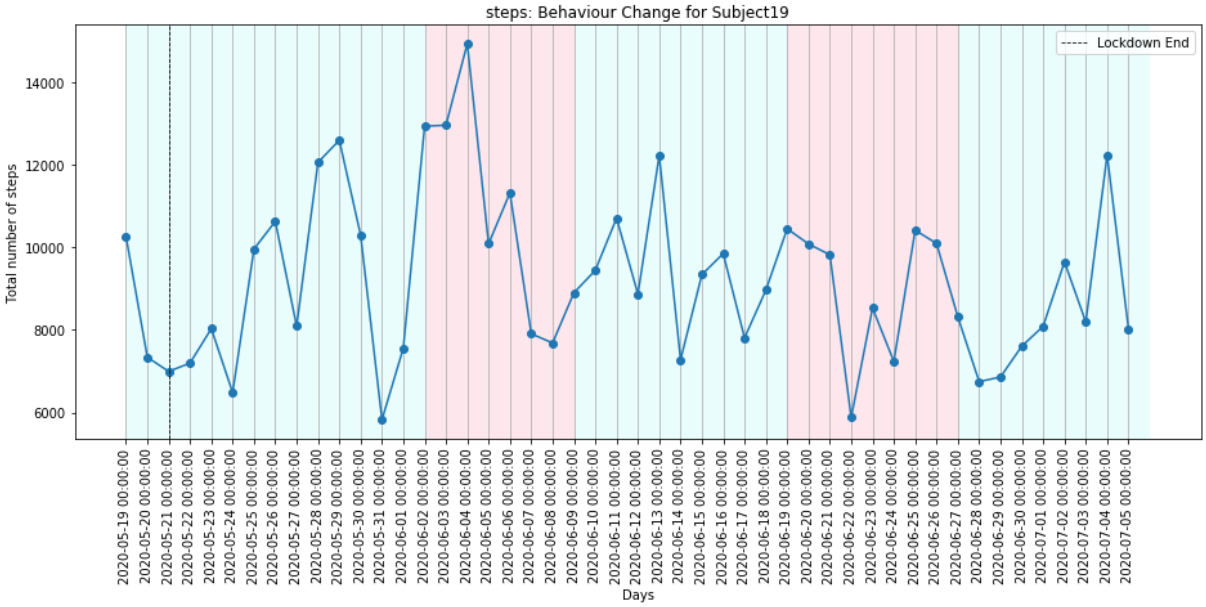


Figure 43: The Physical Behaviour Change for Subject19 based on the steps.

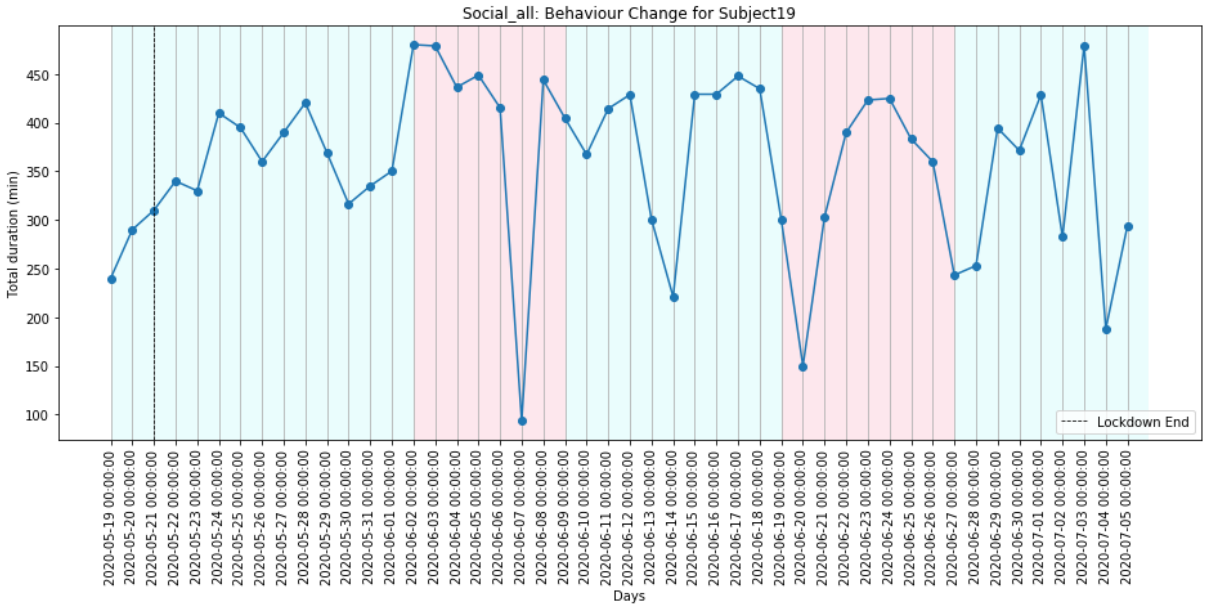


Figure 44: The Social Behaviour Change for Subject19 based on the total duration of being socially active (through the data fusion approach).

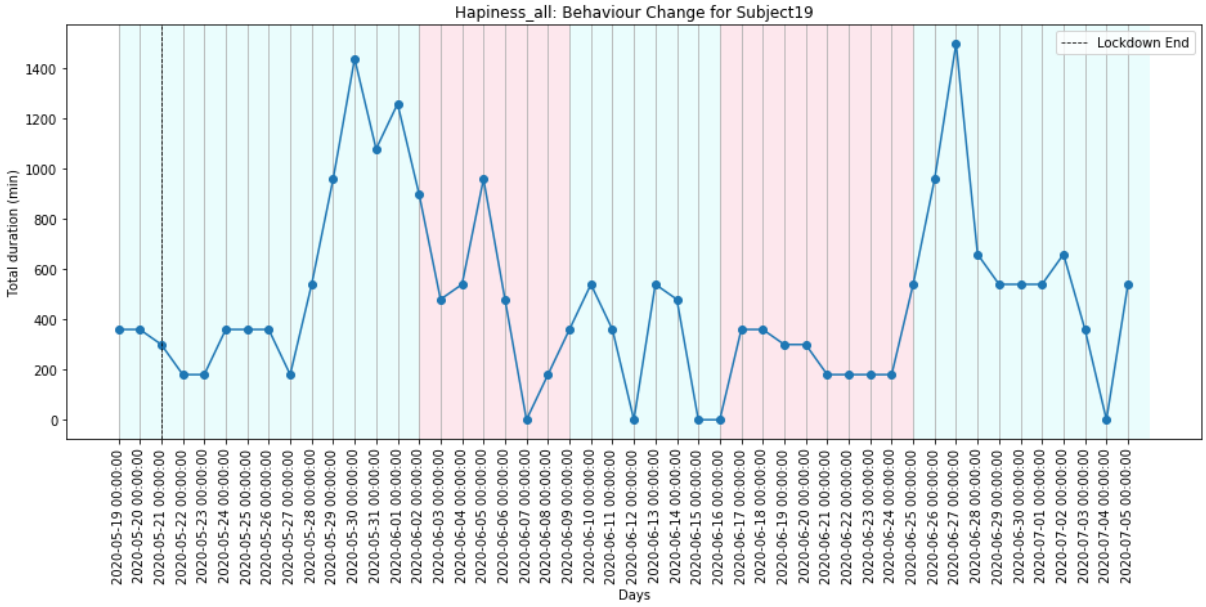


Figure 45: The Emotional Behaviour Change for Subject19 based on the total duration of being happy (through the data fusion approach).

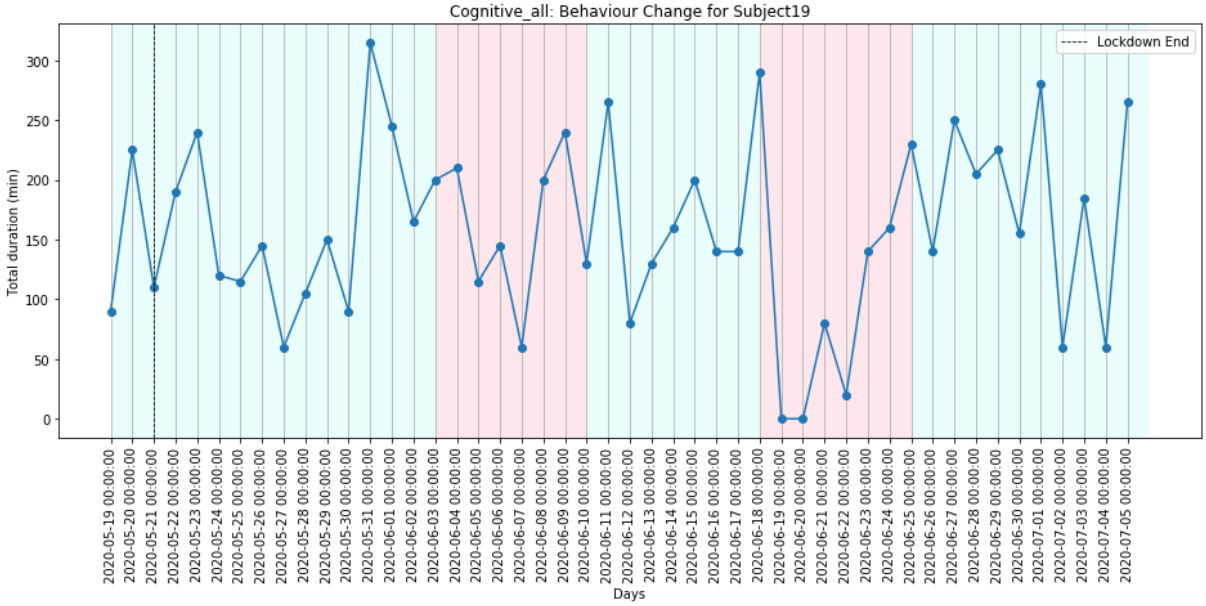


Figure 46: The Cognitive Behaviour Change for Subject19 based on the total duration of being involved into cognitive tasks (through the data fusion approach).

5.3 HBAF accuracy

After processing the collected data in order to infer the short-term, long-term and behaviour changes, the evaluation phase takes place. The evaluation phase consists of two parts; the first part validates the accuracy of the HBAF models while the second part validates the HBAF robustness dealing with missing data.

Initially, the HBAF framework will be evaluated in order to validate the models' accuracy for detecting long-term and behaviour changes. This part is based on users' answers on the weekly ESM questions Q8 and Q9 which contain the 'ground-truth' information related to users' physical, social, emotional and cognitive long-term behaviours.

Among the subjects who participated in the evaluation study, only Subject01, Subject04, Subject14 and Subject19 managed to answer almost all the Q8 and Q9 questions. Even though many more subjects participated in the evaluation study, their answers on the questions Q8 and Q9 were deemed after analysis not sufficient or relevant to describe possible weekly trends for their physical, social, emotional and cognitive behaviours. Specifically, most of the subjects did not manage to answer these questions, while others kept answering that there was not a difference on their behaviours. For instance, they kept answering "no difference" on Q9, which was translated to no difference between two weeks for both physical, social, emotional, and cognitive behaviours. However, due to the significant difference observed during the analysis, the reliability of the answers was deemed inadequate. Consequently, in order to have reliable data for the HBAF evaluation we decided to perform the evaluation based on these 4 subjects.

It is important to mention that these 4 Subjects belong to Group A, even though there are a few participants in Group B (Subject02, Subject03, Subjects06, Subject08 and Subject10) who gave sufficient answers on the Q8 and Q9 questions. However, these participants in Group B did not manage to answer regularly the CaaS questions (which are the only collected data to describe users' social, emotional and cognitive behaviours), and thus they were excluded from the evaluation phase.

5.3.1 Accuracy of the Model for Inferring Long-Term Behaviours

The overall evaluation score for the model is depicted in Figure 47. It can be seen that the performance score varies per participant, while the average accuracy score for all the subjects (Subject01, Subject04, Subject14 and Subject19) approximates 70%. Specifically, the average score for predicting physical long-term behaviours is 67.85%. A similar average score is achieved for predicting social and emotional long-term behaviours, while for predicting cognitive long-term behaviours the accuracy is 78.56%.

It is worth mentioning, that Subject14 and Subject19 did not give a complete answer on the question Q9 for multiple times. For instance, they answered "this week I was more tired, I had to work a lot of hours" which can roughly be converted to a negative emotional trend and to a positive cognitive trend. However, no other information is provided for the physical or social behaviour of the user for this specific week. Thus, the ground truth for the physical and social behaviour is zero (we assume that there is no difference between the two weeks). Consequently, incomplete answers had a negative impact on the performance score of the prediction model for detecting the long-term behaviours.

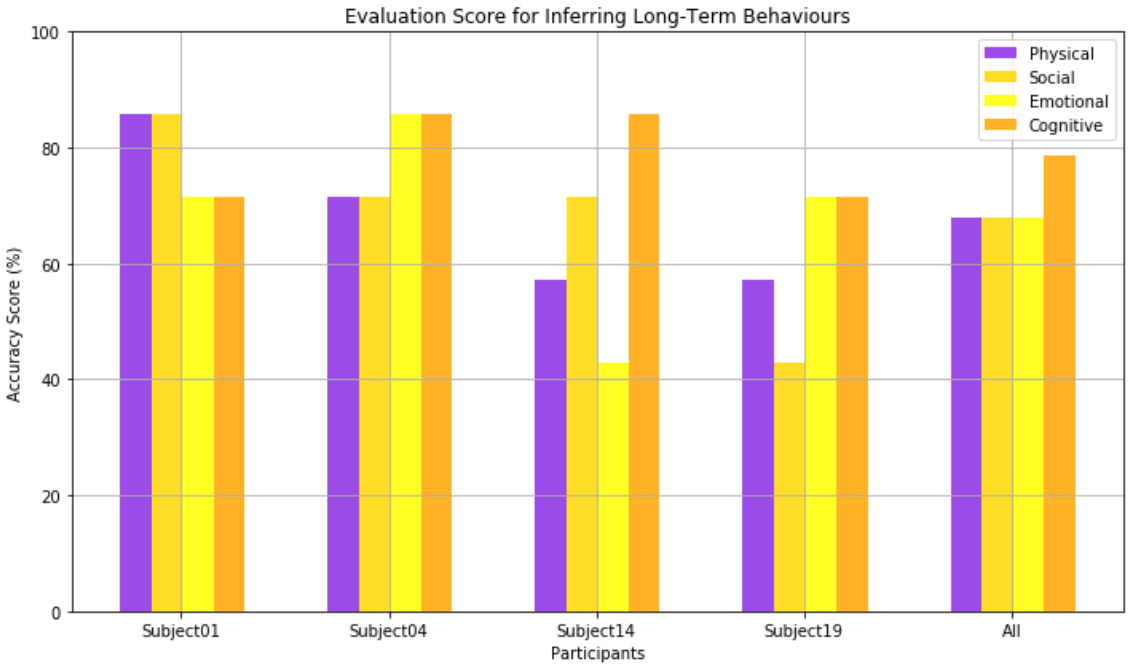


Figure 47: Evaluation score for inferring long-term behaviours comparing the performance for each subject.

5.3.2 Accuracy of the Model for Inferring Behaviour Changes

Even though the 'ground truth' provided through the questions Q8 and Q9 was sufficient enough to evaluate the prediction model for the long-term behaviours, it does not suffice to annotate the actual changes in order to evaluate the model for the behaviour changes. Thus, we decided to ask all the subjects to answer a few extra questions at the end of the data collection phase (through Google forms) aiming to gain the necessary information, retrospectively, in order to evaluate the accuracy of the behaviour changes model. In Table 6, the users' answers on the questions Q23 – Q27 can be seen. These five questions are the following:

Q23 (lockdown information): *"Depending on each country, the government announced some rules/suggestions for the end of the COVID-19 lockdown. You can consider that the lockdown was over for you when you started going out more frequently (going shopping, going to work and not working from home, going out with friends, etc.). Please estimate the date that the lockdown was over for you."*

Q24 (physical behaviour changes): *"Try to compare the period during and after the lockdown. Did you experience any significant changes for your physical behaviour? If so, please specify. You can type anything that you think is relevant. For example, you can type "During the lockdown, I was walking every day because I had more free time", "After the lockdown, I started going to the gym more frequently", etc."*

Q25 (social behaviour changes): *"Try to compare the period during and after the lockdown. Did you experience any significant changes for your social behaviour? If so, please specify. You can type anything that you think is relevant. For example, you can type "During the lockdown, I wasn't satisfied with my social life and I felt lonely", "After the lockdown, I try to go out more often and meet more people", etc. "*

Q26 (emotional behaviour changes): *"Try to compare the period during and after the lockdown. Did you experience any significant changes for your emotional behaviour? If so, please specify. You can type anything that you think is relevant. For example, you can type "During the lockdown, I was sad and stressed due to the virus pandemic", "After the lockdown, I feel happier because I can travel again", etc. "*

Q27 (cognitive behaviour changes): *"Try to compare the period during and after the lockdown. Did you experience any significant changes for your cognitive behaviour? If so, please specify. You can type anything that you think is relevant. For example, you can type "During the lockdown, I was reading more books", "After the lockdown, I play board games with friends more often", etc. "*

Table 6: An overview of the questions asked in order to get the ground-truth for the behaviour changes.

userID	Q23	Q24	Q25	Q26	Q27
Subject01	14/06/2020	During the lockdown, I was going only to the supermarket and pharmacy every fifteen or more days. Rarely I was going for walking. After lock down I am yet very careful, I avoid going to places with many people and generally I have not return yet to my daily routine	During the lock-down, I wasn't satisfied with my social life, I felt lonely. After the lock down I try to meet more people but I haven't managed it.	During the lock down, I was sad, stressed and afraid of the virus pandemic. I was afraid and very anxious about my children's health, especially about my younger son. After the lock down I feel a little more freedom about going out of my house, but I don't feel that I can go wherever and whenever I want	I think there is not any difference
Subject04	15/06/2020	During the lockdown I was more physically active. In particular, at the beginning I had more time to experience and learn new things and topics, while being in a good mood, but close to the end of the lock-down was really difficult to concentrate and find the motivation to work or just spend time on something new. Therefore, I got excited to start working again. On the other hand, I lost all the bits and pieces of time where I could have exercise, or reading something or simply go out for a walk.	During the lockdown I was surely in contact with more people (via social network and phone) than even before and after the lockdown.	At the beginning of the lock down, I was a bit stressed for the situation and for the uncertain future. Later on, the fear and stress reduced to just common sense and the necessity of going out/back to work was pungent. However, I would not mind to have a situation in between where I could have more free time once in a while, for example per week.	At the beginning of the lockdown, I was more prone to read and learn new things. I learned some new topics and I was really excited. However, I think that the situation of being all the time at home didn't help and I slowly lost the motivation of committing myself to expand my cognitive horizons.
Subject14	06/06/2020	I was less physically active despite going out for 30 mins every day and online yoga once a week, I was missing daily mobility to work. Today I keep working from home and looking for a gym because I gained a few kilos.	I was spending all the time with my boyfriend and wasn't feeling lonely. My social communication only increased due to regular skyping with friends and colleagues. The only	At the beginning of the lockdown I was stressed and felt something like panic attacks when I had to go shopping, was afraid of people, but then I got used to self-isolation. Now I am not happy I can travel again - I have to get used to traveling. Also, I was stressed because my parents had Corona, and my mother was in	During lockdown, I was reading more books and played board games sometimes, I was also learning Dutch and learning to paint. I am continuing

			downside was that I could not travel to meet my parents.	intensive care, I could not sleep well during that time. Now I am happy they have recovered, but I am still afraid of the second wave of Corona in the fall/winter.	with reading and learning the language, now it is 3 times a week instead of 1, but otherwise no difference.
Subject19	21/05/2020	During the lockdown I was taking long walks in the parks to compensate for not doing any other exercise. Now that lockdown is over I stopped that habit but I am still scared to return to the gym.	During the lockdown I used a lot internet to communicate. I felt on with it. I missed the real-life interaction. So now i am trying to meet people more outside face to face.	During lockdown I was sad and stress because i felt Trapped and waiting. I am used to very fast rhythm un my daily life and the change of staying at home was a big one for me. Until now i still feel weird from the aspect that i don't feel carefree.	During lockdown I read a lot of books. I tried to get new skills also like foreign language and online courses. After the lockdown I continue with the book reading but not so much with gaining new skills

Even though the provided answers for these questions are quite elaborated, the subjects did not eventually specify the exact date or time that a real change occurred. Consequently, the ground truth for the behaviour change prediction model was manually generated (as a consequence of not having the necessary information directly through the Q24-Q27 questions) by combining all the aforementioned answers for each participant, including the questions Q8 and Q9. Specifically, if a subject indicated on Q9 that there is "a significant change" for a certain type of behaviour that overlaps with user's answers on Q24-Q27, then the weekly change can be considered as an actual change for this type of behaviour. For example, Subjec19 mentioned that the lockdown was over three days after the data collection started, which is not sufficient enough to conclude if there were any real changes during the 7 weeks of the data collection. However, by combining user's answers on the Q9 and Q24-Q27 questions, we estimated that an actual change took place in week 24 and week 26 for the physical and social behaviours, in week 23 and week 26 for the emotional behaviour and in week 24 and week 25 for the cognitive behaviour.

Trying to estimate when a real change happened is quite challenging. For this reason, we decided to estimate the week that a change occurred instead of the day. Depending on the participant and the total duration of the collected data, the prediction model detects at least two changes. Our assumption is that the period during and after the lockdown could be linked into two changes, respectively. Based on this assumption, which might be different for each subject, the overall accuracy for the prediction model for detecting physical behaviour changes is 62.5%, 87.5% for detecting social and emotional behaviour changes, and 75% for cognitive behaviour changes. For instance, Subject19 achieves 50% for the physical behaviour (predicted changes in week 23 and week 26), 100% for the social and emotional behaviours, and 50% for the cognitive behaviour (predicted changes in week 23 and week 25). The best prediction score is achieved for Subject04, where the model estimated that there is a change in week 23 and in week 25, for both the physical, social and emotional. For the cognitive behaviour, the model

estimates two changes in week 23 and week 26. The overall accuracy for inferring behaviour changes can be seen in Figure 48.

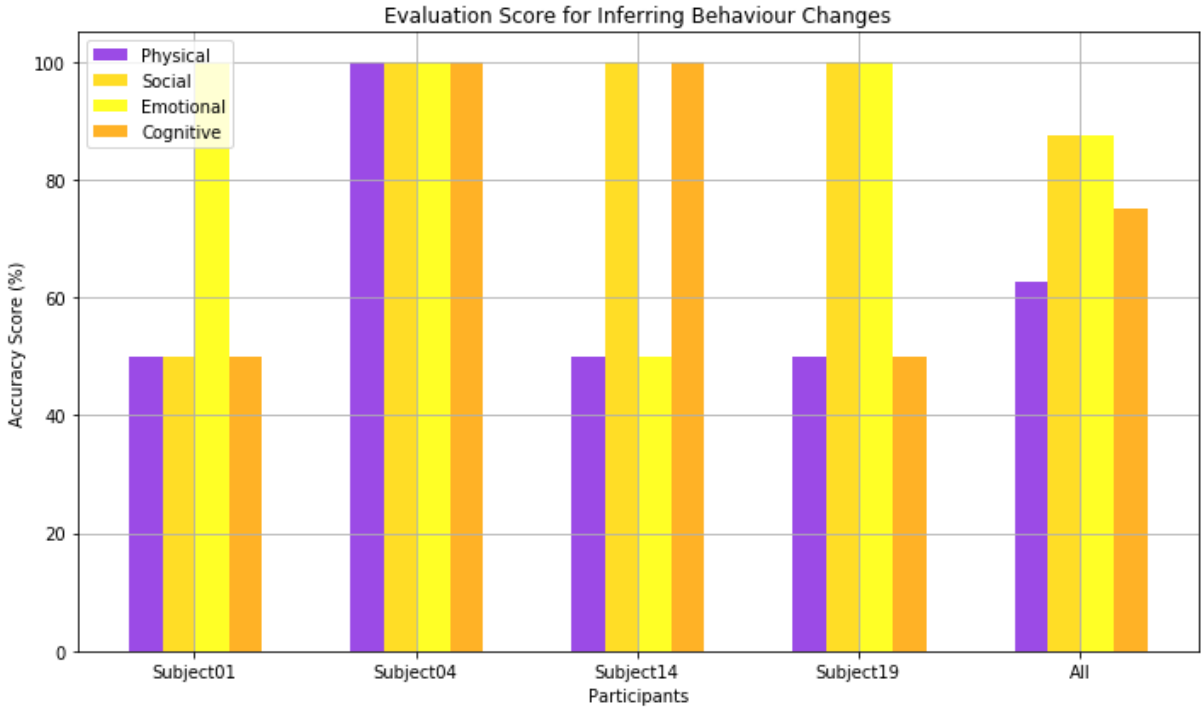


Figure 48: Evaluation score for inferring behaviour changes comparing the performance for each subject.

5.4 HBAF robustness

In order to validate the robustness of the HBAF, some level of noise has been added to the collected data aiming to evaluate the prediction models while the system tries to handle erroneous situations. In order to inject some level of noise into the dataset, the data were under-sampled by progressively removing random rows from the initial dataset. For instance, the size of the collected data for the short-term physical behaviour of Subject01 is 68.262, which can progressively be reduced to 3.413 (-95%) rows. The rows were removed for different percentages of the initial dataset, starting from -0.5% and ending to -95%. In order to ensure statistical robustness, the rows were randomly removed for each percentage ten times in total and for each iteration the accuracy of the model was computed. Eventually, all the accuracies for each percentage were averaged. It is worth mentioning that the initial size of the dataset for each behaviour comes from the data fusion of the available collected data. That means that some rows might contain NaN values, while other rows contain important information for describing users' behaviour.

In order to evaluate the robustness of the prediction model for the inference of long-term behaviours, we calculated the error rate (1 – accuracy) for the different scenarios of handling missing data for the Subject01. As it can be seen in Figure 49, the prediction model is extremely robust for handling missing data and especially for the physical behaviour. Thus, we can assume that if a user does not enable the data collection for a few hours or days during the data acquisition (while using the COUCH system), then the HBAF will still have the necessary input to make accurate predictions for the user's long-term behaviours.

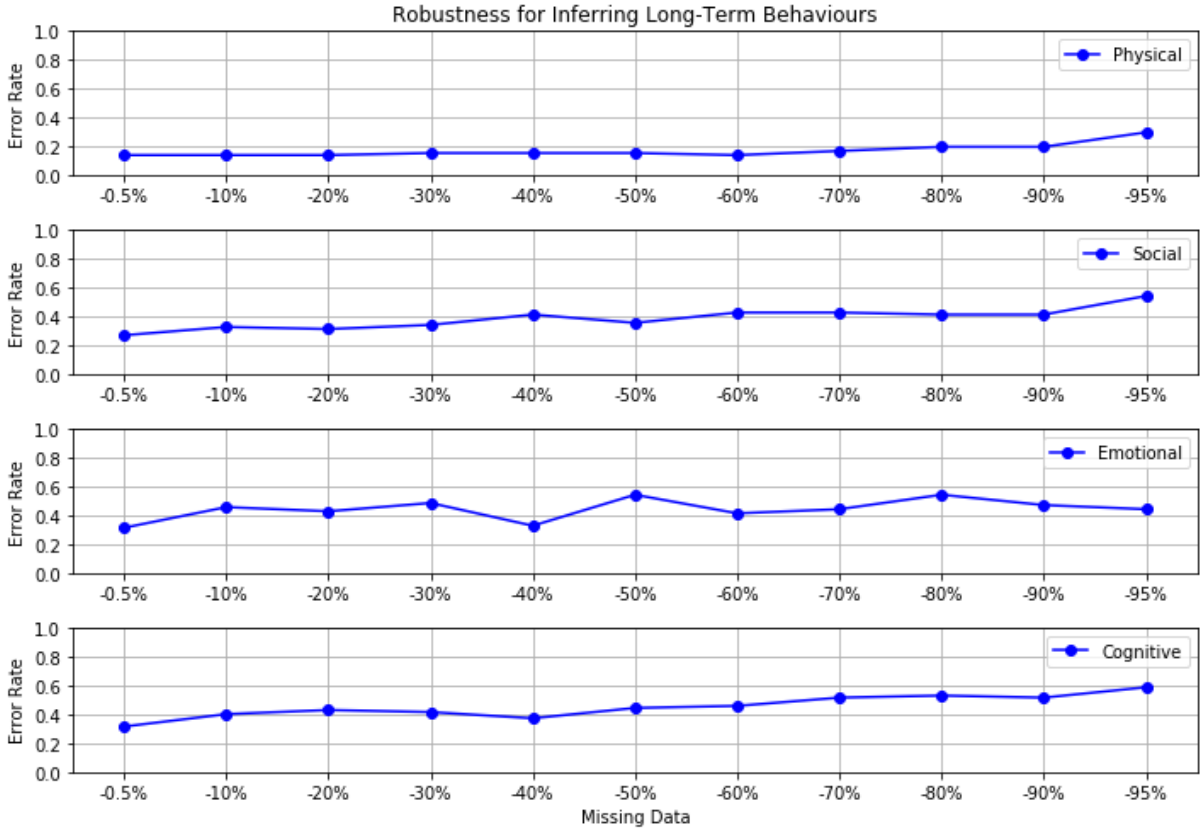


Figure 49: Robustness for Inferring Long-Term Behaviours.

Similarly, we calculated the error rate for the different scenarios of handling missing data in order to detect behaviour changes. In Figure 50, the HBAF robustness for inferring behaviour changes is depicted. The physical, social and emotional prediction model can be considered robust by handling up to -10% of missing data, while the cognitive model looks to perform the best by being able to handle up to -50% of missing data. However, it is important to mention that the error rate highly depends on the accuracy of the ground truth. The limitations that were presented in subsection 5.3.2 could possibly explain the reason that the error rate does not get linear, since it should be increasing while more rows are missing.

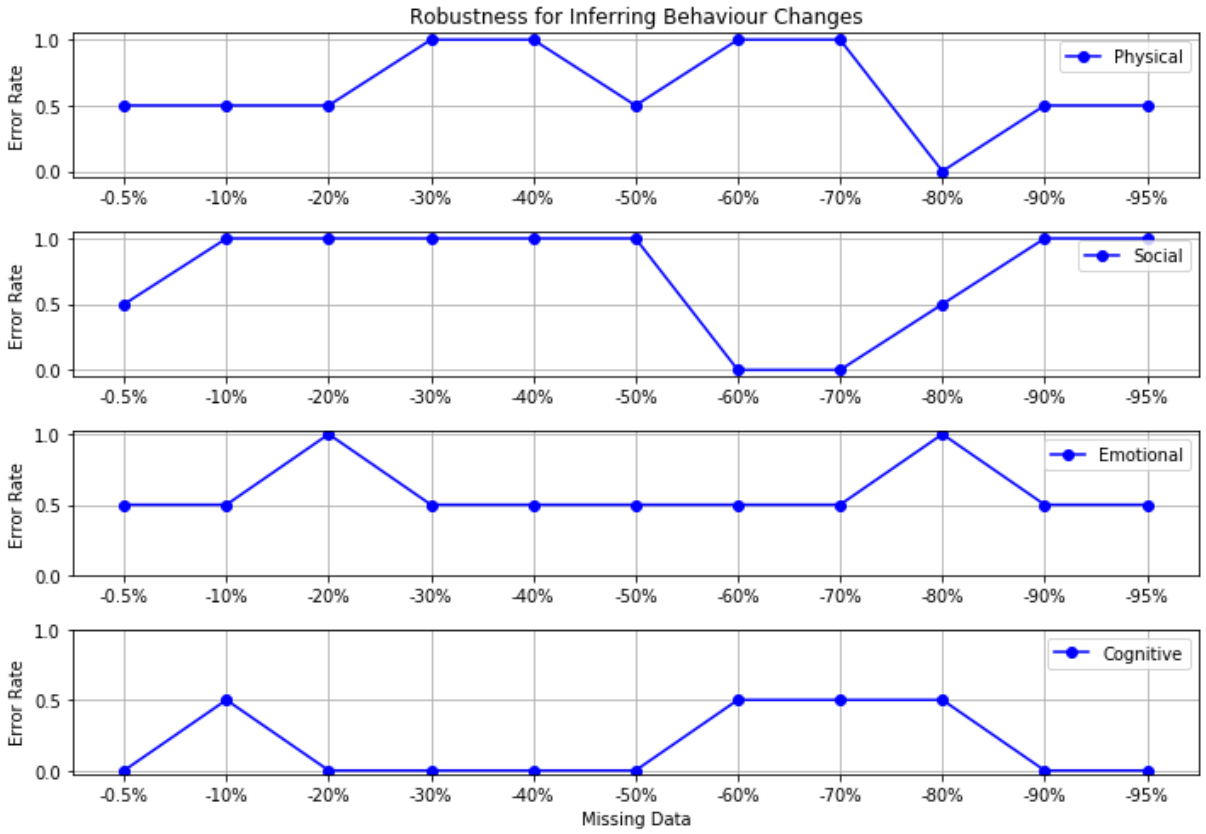


Figure 50: Robustness for Inferring Behaviour Changes.

6 Discussion

6.1 Main Findings

This deliverable aims to evaluate the main components of the HBAF framework for detecting long-term and behaviour changes. While the COVID-19 pandemic limited our actions initially, we successfully managed to set up remotely the experiment and conduct an alternate evaluation study which lasted around 7 weeks. In total, 21 subjects completed the data collection phase; the duration and the types of collected data vary per participant. Overall, the participants found the devices for collecting data (including installing the mobile app and answering the ESM and CaaS questions) easy to use and they did not experience any problem. In particular, we noticed that the participants showed a preference on using the mobile app and answering the ESM questions on a daily basis. Only one recruited participant was afraid of the privacy concerns that might raise from the mobile data collection; such concerns forced him to withdraw from the study. However, during the data analysis phase we noticed a lot of missing data (due to the decision of the user to intentionally not answer the questions and not due to a technical issue of the HBAF); either the subjects did not answer all the questions (both ESM and CaaS) on a daily basis, or their answers were insufficient. This limitation played a major role in our decision on evaluating the HBAF framework, and thus, we decided to use data only from 4 subjects mainly: Subject01, Subject04, Subject14 and Subject19.

Due to the COVID-19 lockdown, there were some irregularities on the users' habits and trends that were difficult to be determined. For instance, the lockdown started and ended on a different date for each participant (depending also on the country of residence), which forced the participants to leave their comfort zone and behave in an unforeseen way. While time passed by, the participants got used to the new situation (e.g., working from home, being socially isolated, being afraid and stressed of the pandemic, etc) and they tried to follow an ordinary lifestyle. However, during this time their actions were progressively changing, followed by the end of the lockdown which forced them to get used to another situation. Even though there were a few participants who got back to their usual daily life after around 1 week of the lockdown end, most of the participants were sceptical and afraid of this change (they kept staying indoors).

Based on the evaluation study and the acquired data, we showed that the suggested data types are sufficient to detect users' short-term, long-term and behaviour changes in a holistic approach. We found that a data fusion approach is highly recommended in order to overcome the problem of missing data and infer short-term behaviours successfully. For instance, combining data from mobile devices and activity trackers in order to describe short-term physical behaviour when data from one device are missing, or combining users' answers when either ESM or CaaS questions are missing. It is important to mention that the short-term core component has been evaluated through the evaluation of the long-term and behaviour change components of the HBAF, since the short-term behaviours model feeds the other two components. Overall, we proved that the model for inferring short-term behaviours can sufficiently be used to describe physical, social, emotional and cognitive behaviours, and thus, no further evaluation of this core component is needed.

Furthermore, we observed that the long-term and behaviour changes might deviate between each user's behaviours, but also among users. Which proves the urgent need to apply a personalized coaching system through the virtual system of the 'Council of Coaches'. It is worth mentioning that the physical, social and emotional long-term behaviours seem to follow similar patterns, in contrast to the cognitive long-term behaviours. This assumption is valid for certain users, such as Subject19 (see Figure 27), where many of the weekly trends that describe the physical, social and emotional long-term behaviours overlap. Thus, a correlation between physical, social and emotional behaviours for certain days or weeks can be noticed. Overall, the accuracy for predicting physical, social and emotional long-term behaviours is 67.85%, while the accuracy for predicting cognitive long-term behaviours is 78.56%.

Another prominent outcome of this study is the importance of initially collecting data during a learning phase, so the HBAF can learn and identify any unusual trends over time in order to accurately detect a behaviour change when the user deviates from a normal lifestyle. For instance, Subject19 mentioned that a big change occurred on the 3rd day of the data collection (the end of the lockdown), which could

not be detected due to the insufficient amount of data before that date. It is worth mentioning that the lockdown for this user ended on 21/05/2020 (three days after the data collection started), where the collected data up to this date were not enough to define a change. Consequently, it is clear that the historical data play a major role in defining when there is an anomaly over the time-series, resulting to accurate predictions of behaviour changes. The overall accuracy for the prediction model for detecting physical behaviour changes is 62.5%, 87.5% for detecting social and emotional behaviour changes, and 75% for cognitive behaviour changes.

Eventually, we concluded that the HBAF framework components are robust and can achieve an acceptable performance for feeding the coaching strategies during a COUCH session. In particular, we found that the HBAF will still have the necessary input to make accurate predictions for the user's long-term behaviours and behaviour changes, if data are missing for a few hours or days. More specifically, the model for inferring long-term behaviours is considered significantly robust achieving a good performance score even if the dataset is removed by 50%. On the other hand, the model for inferring behaviour changes can be considered robust for detecting users' cognitive behaviour or by handling up to -10% of missing data for detecting users' physical, social and emotional behaviour changes.

6.2 Open Issues

The HBAF framework was evaluated using data from the Subject01, Subject04, Subject14 and Subject19. These subjects managed to answer the ESM and CaaS questions almost every day. However, their answers were sometimes limited which caused problems on the ground truth annotation for the weekly changes. Especially, this issue was noticed while identifying the actual behaviour changes over time. Furthermore, we found that the data were collected over a situation where human behaviours could not be clearly detected and described due to the COVID-19 lockdown, while some participants did not manage to sufficiently annotate them.

Even though the final dataset proved to be adequate for evaluating the HBAF core components, since the data were collected for more than a month (7 weeks in total), it is important to mention that more time might need in order to accurately detect behaviour changes when no temporal changes due to the pandemic phenomenon (or other changes such as seasonality trends due to holidays) are present. Overall, the lockdown was a peculiar situation that may not take place anytime soon or ever again in order to further investigate human behaviours, while the participants of this study do not fully represent the target groups that were initially envisioned for the HBAF evaluation.

7 Bibliography

Ferreira, D. (2020). *AWARE Framework*. Retrieved from https://awareframework.com/sensors/

EC. (2020). *Ethics Committee EEMCS*. Retrieved from https://www.utwente.nl/en/eemcs/research/ethics/

AWARE-Framework. (2020). Retrieved from https://awareframework.com/

Council-of-Coaches. (2020). *COUCH*. Retrieved from https://www.council-of-coaches.eu/beta/

Banos, O., & Konsolakis, K. (2019). *D4.7: Behaviour change detection analysis prototype.* The Council of Coaches Consortium.

Banos, O., & Konsolakis, K. (2019). *D4.6: Methods for detecting behaviour changes from short-term behaviour.* The Council of Coaches Consortium.

Banos, O., & Konsolakis, K. (2019). *D4.5: Long-term behaviours analysis prototype.* The Council of Coaches Consortium.

Banos, O., & Konsolakis, K. (2019). *D4.4: Methods for inferring short-term behaviour from multimodal sensor data.* The Council of Coaches Consortium.

Banos, O., Konsolakis, K., Bangalore Kantharaju, R., Pelachaud, C., & op den Akker, H. (2018). *D4.3: Short-term behaviour analysis prototype.* The Council of Coaches Consortium.

Banos, O., Konsolakis, K., op den Akker, H., Pelachaud, C., & Bangalore, R. (2018). *D4.2: Methods for inferring short-term behaviour from multimodal sensor data.* The Council of Coaches Consortium.

Banos, O., Konsolakis, K., op den Akker, H., Pelachaud, C., & Bangalore, R. (2018). *D4.1: State-of-the art, requirement analysis and initial specification of the Holistic Behaviour Analysis Framework.* The Council of Coaches Consortium.

Acknowledgements

The Council of Coaches project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement #769553. This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains.

Headings and titles in this document, as well as the Council of Coaches logo use the Comfortaa font, designed by Johan Aakerlund and Cyreal and licensed under the Open Font License[3].

Additional text in this document uses the Roboto font, designed by Christian Robertson and licensed under the Apache License, Version 2.0[4].

The Council of Coaches logo and Blobmen graphics were drawn freely in Inkscape, licensed under the GNU General Public License[5].

[3] Open Font License: http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=OFL_web

[4] Apache License, Version 2.0: http://www.apache.org/licenses/LICENSE-2.0

[5] Inkscape License Information: https://inkscape.org/about/license/