Decoding User Behaviour from Smartphone Interaction Event Streams

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Björn Friedrichs

March 2023

Cardiff University

School of Computer Science & Informatics

Abstract

The smartphone has become an everyday device for many people around the world and has led to an evolution in the way we use these devices. This has led to increased research interest in the effects of smartphone use on psychological traits, which could have a positive impact in clinical or self-help settings by identifying positively influencing variables.

In this thesis, a new model to extract behaviour information from a stream of usage is presented. The model aligns with previous methods in the research area but focuses on establishing a generalisable three-step process of processing user interaction to extract new user behaviour knowledge. This introduces a structured approach to smartphone usage evaluation and enables the implementation of customisable applications. It also creates a baseline to compare previously defined metrics which describe smartphone usage. Usage derived from metrics which could be considered high-level such as screen-on time is self-evident and therefore are common measure to distinguish usage between users. However, within usage sessions, they suffer from limitations such as a strong skew towards short bursts of usage because of how smartphones are often used. By utilising direct interactions with the user interface (such as taps and scrolls), usage at a lower level can be considered which can carry more elemental characteristics of behaviour. Thus, they can be used to model behaviour more accurately, which can be aligned with the user's mental state to identify habits which are caused by problematic use patterns. This enables the isolation of user trait classes reflecting smartphone addiction and impulsivity.

Contents

List of Publications

- [35] - Friedrichs, B, Turner L. D., and Allen S. M. Discovering Types of Smartphone Usage Sessions from User-App Interactions. In *International Conference on Pervasive Computing and Communications (PerCom)*, pages 459-64, IEEE 2021.
- [36] - Friedrichs, B, Turner L. D., and Allen S. M. Utilising the co-occurrence of user interface interactions as a risk indicator for smartphone addiction. In *Pervasive and Mobile Computing*, Volume 86, Elsevier 2022.
- Friedrichs, B, Turner L. D., and Allen S. M., Whitaker R. M., Linden D. E., Smartphone User Interface Interactions reflect Trait Impulsivity. To be submitted to *Addictive Behaviors*, Elsevier

List of Figures

List of Tables

List of Acronyms

UI User Interface

SA Smartphone Addiction

SAS Smartphone Addiction Scale

MCQ Monetary Choice Questionnaire

MWU Mann-Whitney U

NLP Natural language processing

TF-IDF Term Frequency-Inverse Document Frequency

BFUS Behaviour-From-Usage-Stream

Acknowledgements

Without the guidance and support of my supervisors, Liam Turner and Stuart Allen, I would not have been able to complete my thesis and write these words. They not only provided crucial guidance but also created an environment of trust and friendliness. I would like to extend a special thank you to Liam for going above and beyond what I could have expected. His sense of duty, responsiveness, and a seemingly endless supply of good ideas were truly remarkable. I wish him and his family all the best.

I would also like to thank my co-authors, Roger Whitaker and David Linden, for their valuable suggestions and help.

Lastly, I would like to thank all of my friends and family who stood by me and provided me with the support I needed. Danke!

To Ella and my friends;
My family Regina & Ernst, Thorben
Tanja & Timo, Ulrich & Karin
Josephine & Rudolf

For their endless support

Chapter 1

Introduction

The smartphone has become a personal, everyday device for many around the world. With this kind of saturation, innovations have resulted in an evolution in the way we use these devices [13]. A rise in processing power allows running more powerful applications, games, or multiple applications at the same time. Larger and brighter screens create surfaces that can hold more information and interactive elements. Extended battery life means devices can be used almost all day without a need for charging. Most smartphones also operate with a constant internet connection which given the infrastructure of many urban areas enables the possibility of information exchange at any time. Collectively, this has resulted in a highly stimulating device being embedded in our lives which enables not only research into how we use these devices, but also how usage may correlate with broader latent factors such as mood, anxiety, boredom, or stress.

The increased research interest in these effects and the types or habits of use behind them has led to an improved understanding of how they are linked to certain behavioural patterns (e.g., predictors of addictive smartphone behaviour [114] or the effect of social media on impulsive behaviour [155]). This understanding of these issues linked to smartphone use could have a positive impact in clinical or self-help settings, for example with the discovery of positively influencing variables [123]. Mapping this connection between smartphones and psychological traits is especially interesting as this enables a degree of passively monitoring correlating factors between usage and psychological states without specialised equipment (e.g., [22]) or other intervening

methods such as the collection of survey data (e.g., [165, 114, 72, 28]). However, because of the multitude of combinations in the definitions of behaviour, input options, and evolution of device capabilities available with these devices, the various techniques used to establish these connections are fractured.

1.1 Smartphones and Behaviour

Human behaviour is a complex system which can generally be described as the conscious or subconscious actions and mannerisms taken towards stimuli. The form of these responses is shaped by many factors such as personality or past experiences and can evolve throughout one's lifespan. Because of the variability of all these influencing factors, capturing behaviour is a difficult task.

While it is still difficult to encapsulate behaviour, the benefit of a device like a smartphone is that it presents a user with a limited number of input options. This limiting factor acts as an interface between overall behaviour and processable information to enable a generalisation of the highly diversified use cases a smartphone can provide. From this, is possible to look for certain psychological traits and relate them to specific patterns in a user's behaviour. Examples include boredom [89, 4], anxiety or stress [30, 31], mood or emotion changes (including their impact on the user's behaviour) [156, 37, 58, 143, 99] and problematic use patterns [130, 114, 32].

Smartphones are interacted with for a variety of reasons and can be used to fulfil a multitude of tasks. On the highest level, the user will have some goal (even if subconscious) when using their phone. Reasons can include examples such as viewing a notification, responding to a text message, browsing social media or taking a phone call. Some of these interactions can be intentional and planned (e.g., drafting an email for a scheduled meeting), while others can be circumstantial (e.g., receiving a text message from a distant relative). All these interactions contribute towards the overall behaviour of the user and the more hidden aspects of smartphone usage.

1.1.1 User Interaction Events

The touchscreen of smartphones acts as an intuitive dual-purpose interface by reflecting the information of the current device's state but also allowing direct inputs with high responsiveness. The most basic case (touch of the screen triggers event) of this however does not describe the structure of interaction in terms of temporal and contextual dependencies. It is missing semantic information in the form of device state, which is defined by factors such as the currently launched application, the target of the event (e.g., the on-screen keyboard or a button in the application) or any preceding events. Relying completely on non-contextual UI events would make extraction of behaviour challenging, therefore for a meaningful interpretation of behaviour, the device state has an important role.

At any given point there is a multitude of interactions that a user can choose to undertake with a smartphone (e.g., a tap, a long press, a swipe). These interactions become more contextual events when combined with on-screen information. Pressing a button which opens an application is physically the same as a press on the on-screen keyboard. Semantically, they achieve different goals and therefore are logically separate actions. Other influencing information can be temporal, converting a simple press to a long press or a swiping motion that transforms a touch into a drag can represent a scroll or unlock event based on the device state and location of the touch. And just like interactions are a key part of behaviour so is the lack thereof, not choosing to interact with a device for certain amounts of time (to take a break or consider options) can be just as relevant as opting to do so.

Lastly, there are system-specific effects that can be caused by the user or the device itself as a reaction to either direct interaction or externally triggered events. For example, the screen can turn on from a screen touch, a button press or a received notification. An application may be start from the home screen or by a redirection from another application. While these would not be considered direct inputs of a user, they are either a consequence of previous behaviour or an influencing factor for coming interactions.

Definition 1. *A user interaction occurs when a user physically handles the device and changes the devices state. Examples of this are the user unlocking the phone or interacting with the screen (e.g., to type on the on-screen keyboard or to change applications).*

Definition 2. *An event is the consequence of a state change in the smartphone as a system. This change can be triggered in multiple ways, such as a change of it's own state (e.g., battery draining), an external update (e.g., received notification) or manually via a user interaction. Each event is captured at a discrete time.*

As a user interacts with their phone, any singular event is part of a sequence which can be considered a stream:

Definition 3. *An event stream is a sequence of events generated by a user. Within the stream the events are inherently ordered by the time they occurred in.*

1.1.2 Session and Task-completion Boundaries

These singular interactions become more behaviour-defining when they can be retraced as a 'stream of events' with a defined start and end point, forming a bounded string of usage interactions which is grouped. These groups and their context-dependent effects are the observable representation of a user's undertaking towards a shared aim or task. This can also be understood as a session of usage, where it compromises a set of interactions to form one or multiple tasks.

In turn, tasks can be described as the 'operationalisation' of a user's goals [168]. These goals can be a conscious choice but do not have to be. As such, they can have a clearly defined result such as replying to an email or taking a picture but can also be less result oriented and rather a subconscious decision to use a smartphone out of habit, boredom or other factors (e.g. [89, 85, 110]). For example, a user might pick up their phone to browse social media to pass time without any specific motivation other than to

pass time. This means designing methodologies with clearly defined task boundaries is difficult for reasons like multi-tasking behaviour, shared responsibilities of applications for the same task, and rest periods that do not conclude the previous action. Furthermore, they encapsulate multiple wants and needs from the user combined with often unknown external factors which are hard to capture. Because of these difficulties in modelling tasks and the limitations of contextual information, which is capturable from interactions, tasks are not further explicitly considered. Instead, it is possible to utilise the perceived barriers between distinct periods of usage, as previously mentioned in the form of sessions. Sessions offer the benefit of defined start and end points without needing more contextual information such as needed to model tasks. This has been previously addressed in one of two ways:

Time-based session boundaries are a way of grouping UI events in a predefined timeframe. This enables analysing user behaviour over different spans of time such as minutes, hours or a whole day. In the case of longitudinal studies, this approach allows analysing per-application metrics for screen-on time or similar metrics in a highly condensed form. However, by discarding the task boundaries valuable usage information in the form of the differences in behaviour based on a user's potential wants, needs and motivation is lost. This shortcoming can be addressed by using more dynamic pause-separation boundaries which infer the borders of a task by idle time. This approach (explored by [152]) identified a time threshold of 45 seconds of a user being non-interactive (regardless of events) as the optimal point to separate task boundaries (regardless of success). A similar idea was proposed by Ferreira et al. who noted an additional in-between breakpoint for 'micro-usage' for usage which occurred in just the first 15 seconds [34]. With this approach, there is no arbitrary separation along time or application borders which enables more diverse and cross-application capture of usage behaviour.

Event-based session boundaries use specific logical points of UI events to identify task boundaries (e.g. [148, 172, 23, 52]). A focused approach can utilise the applica-

tion state boundaries (opening and closing of an application) to capture the behaviour specific to it. The benefits of this approach are strong encapsulation and offer very good information when comparing behaviour between specific applications. The drawback is that it is not possible to capture tasks that are cross-application, so it is not suitable for a generalised model. So, instead of relying on applications or timeouts the platform of smartphones allows capturing very specific start and end points of usage sessions by observing the possible interactibility with the screen. Interactions start when the screen is turned on and continue until it is turned off again. These breakpoints encapsulate any cross-application task behaviour and also inherently include some of the time-based boundaries in terms of longer breaks between sessions.

Those boundaries between tasks embed some of the nuances within the complex nature of user behaviour. Utilising them adds to the improved general understanding of usage and also can be useful for models that attempt to meaningfully distinguish between different kinds of users. These boundaries can also help form the basis for individual usage sessions within a constant stream of usage data from users.

The direct capture of user behaviour is difficult, a focus on capturing interactions and their boundaries as they occur allows to store an accurate representation of a user's journey in form of sessions. This representation of usage does not infer any actual behaviour patterns or habits but instead lends itself to explore them more easily. While various forms of processing have been utilised for multiple use cases, one step to understanding the actual behaviour behind usage lies within the ability to inform of a user's latent mental states reflected by their smartphone use.

Overall this leads to the following definition of sessions:

Definition 4. *A smartphone (usage) session is the part of an event stream which occurs within confined boundaries of a cognitively connected sequence of interactions. This connection can be established through boundaries of time (e.g., elapsed time without events) or specific event (e.g., screen turned on or off). All interactions that occur within those boundaries are part of the session. Via this definition, it is possible to split*

an event stream into distinct chunks of behaviour. In this thesis, sessions are considered to be event-bounded (screen on to screen off).

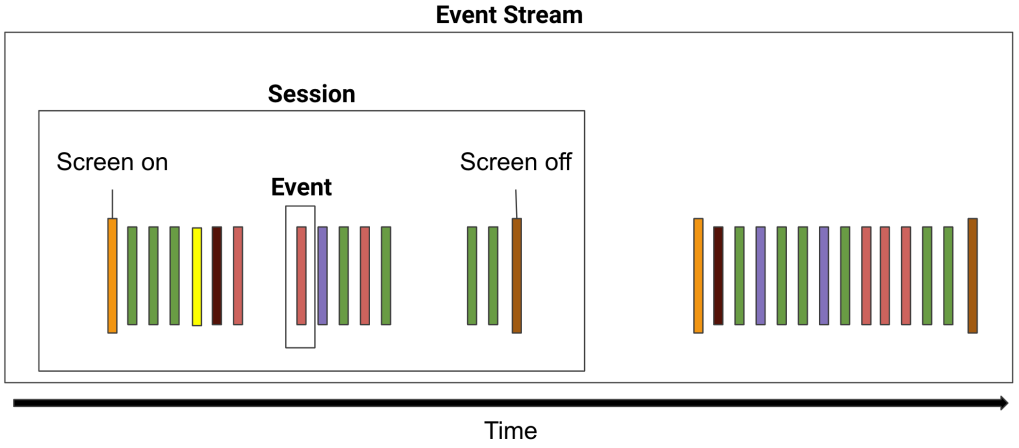


Figure 1.1: An example of the relationship between an event (Definition 2) within a session (Definition 4) within an event stream (Definition 3).

1.2 User Traits Reflected by Technology

The way that humans enact on their needs, wants and expectations is complex. Two people can react completely differently to the same stimulus, or the same for completely different reasons. This observable behaviour is a continuous stream of rapidly made conscious and subconscious decisions which does not necessarily inform intent. While some parts are unique traits of one's personality, there are patterns in behaviour that are shared between individuals and can hint at certain conditions or backgrounds. For example, Jesdabodi et al. identified 13 distinct usage states from 24 users which cover focused topics such as communication, shopping or photography [49]. On a large scale, Zhao et al. found 382 types of users from 106,762 users by clustering application usage over time and identified specific types such as "Night communicators"

or “Evening learners” [169]. Those states of usage are found between multiple users and indicate parallels in the modes of operation of mobile devices.

The idea of this mesh of individuality and shared traits has also been applied to on-device usage [67]. Specifically, the assumption that any action but also hesitation [61] is important to the characteristics of a person’s behaviour.

1.2.1 Identity of User Behaviour

Identity, in this case, refers to the idea that a user’s behaviour is unique enough to identify them among a group of others. Compared to user types, where it is desirable to find differences between groups but identify the patterns that are shared between users, the differences that are recorded for identity are usually fine-tuned and low-level interactions. For example, to answer questions such as “how fast does the user type” or “how quickly did they pick up their phone”, which are the result of physical interactions, their inputs are timed precisely so they can be used to extract an individual’s habits and mannerisms [126]. Other markers such as their physical location can be used to further verify them [158].

On top of physical traits, there are additional virtual markers that identify one’s use. These are more high-level traits such as application or website visits, how long or engaged they are, or what the outcome of any given interactions was (e.g., a purchase). This can also be understood as one’s digital footprint and is frequently utilised for purposes of personalisation such as targeted advertising [73].

1.2.2 Shared Types of User Behaviour

While the nuances of usage can be used to uniquely identify a user, certain patterns of use can identify mental states or user traits across groups of people. These overlaps of behavioural patterns create the basis for understanding the relationship of the user

to their devices. The emotional state or mood of a user (e.g., boredom, anger) can noticeably impact the factors of behaviour, for example, Visuri et al. [156] found that application selection is influenced by a user's mood state.

This connection between a user's mental state and their behaviour can further be observed with more extreme cases or disorders that mediate a certain change compared to the rest of use. Problematic behaviours such as over- (addiction) or uncontrolled use (impulsivity) are showing patterns that permeate throughout their behaviour (e.g., [104, 76, 39]). This has shown that not only issues that are connected with device usage per se but also stable user traits such as impulsivity have been found to be connected with how users interact with their phones.

The underlying factors which influence the patterns of behaviour have been explored by utilising various methods of data capture and processing. In most cases, they focus on establishing relationships between features of usage (frequently on a high level such as screen-on time) and a user's behaviour, which has resulted in a vast pool of methods and connections. This motivates a comparison to deepen the understanding of the relationship between these captured features and user mental states. Furthermore, it hints towards a gap in the formality of processing captured smartphone data and its application.

The following terms form a basis for this:

Definition 5. *Low-level features are events which occur at the closest border between user and device. They occur at a discrete time and do not envelop other events. These include direct interactions with the user interface (e.g., taps or scrolls), the device itself (e.g., a change in battery state) or other semantically atomic events (e.g., application switches or push notifications).*

Definition 6. *High-level features summarise behaviour over a bounded amount of time. This means they are inherently just the passed time between two points in time (e.g., the screen-on time bounded by timeouts or events) or an aggregation of events that occur*

within that bound. For example, summative features such as the count of application switches that occurred within a session.

1.3 Research Direction

Previously, smartphone behaviour research has mainly focused on derived features such as recorded screen-on time, but these features generalise the intricacies of human behaviour to a single dimension such as time. These features create an important baseline that allows understanding how they are influenced by a user's traits, however, they are limited in their ability to capture the specific nuances of how the user is interacting with the device.

In this, there is a high volume of information encapsulated in this stream of direct interaction data. This approach has been used in isolated cases before, such as tracking scrolling [146]. However, this motivates the exploration of this space and finding new methods to pervasively model and utilise all of the inputs of users towards a better understanding of types and habits of usage.

Smartphones are used every day by many, but the way they are handled can span a plethora of use cases. These and other underlying contextual factors of smartphone use contribute to a net of mental dependencies. This complexity makes it difficult to create a general understanding of the area.

The culmination of these key concepts in the literature, the limitations of those current methods and the missing links with direct user-interface interactions lead to the following aim of this thesis: *Modelling smartphone user-interface interactions to capture unseen characteristics of behaviour in usage sessions and utilise those signals that correlate with latent user mental states.*

1.4 Contributions

In the pursuit of this goal, this thesis makes the following contributions:

C1 **Identification of issues with current common methods for representing user behavior, which tend to focus on single, isolated features and high-level characteristics.** Explaining the variance in user behaviour by focusing on isolated features fails to capture usage complexity. Screen-on time and others have been commonly used to establish links between usage and behaviour. However, they are too simple and symptomatic instead of being reflective of actual intricacies in behaviour.

C2 **The proposal of the Behaviour-From-Usage-Stream model which represents a formal framework to process and evaluate user behaviour data.** The model demonstrates how to encode the stream of input events generated by a user into a vectorised format representing a rich data space which can be used for information gain such as inferring stable user traits. This is meant to fill in the gaps of the literature which make it difficult to compare multiple studies because of diverging data collection methods and evaluation methodologies. Through its adaptability the BFUS model can be used with any input features (e.g., high-level, low-level) or evaluation methods and because steps can be adjusted or exchanged while keeping the rest of the model fixed it enables the possibility of comparing features or methods for their individual parts.

C3 **A case study of UI event based user behaviour capture being powerful enough to distinguish users based on psychological traits such as addiction or impulsivity.** Given an independent variable, the BFUS model is validated by predicting user traits from just the transformed event stream of a given user. In this, a novel approach utilises NLP-embedded weighting and vectorisation techniques to capture the nuances of user behaviour. Low-level features are evaluated against more

commonly used high-level features to show how they can improve prediction accuracy for user labels.

1.5 Thesis Structure

The remainder of this thesis follows the Contributions outlined in Section 1.4:

An introduction to the key concepts and supporting literature follows in Chapter 2. This includes a review of the various themes and applications of user behaviour research, a brief discussion of statistical powers and an overview of the dataset which is used as the main source.

In Chapter 3 common behaviour metrics that can be used to distinguish types of user behaviour sessions are presented. These, often summative, features are critically assessed on their capabilities of effectively representing their respective sessions which contributes towards the C1. It then continues to establish the BFUS model with its three-step process. This is expanded on by applying the model with a novel vectorisation methodology for behaviour capture by considering the co-occurrence and weight of physical and logical interface interactions. This contributes towards C2.

Chapter 4 applies the BFUS model with summative and isolated features to detect a user's predisposition to psychological traits and how they influence user behaviour. This chapter shows that these features are not apt for separating groups of users properly. Additionally, it motivates the interest in finding methods that can separate groups of users with certain traits. This contributes towards C1 and C3.

Following, Chapter 5 demonstrates how given a correctly tuned methodology user traits can be detected using a transformed usage event stream. It highlights how user-interface interactions are more strongly influenced by these traits compared to summative features. This contributes towards C3.

This leads to Chapter 6 which discusses how the user-session relationship can play a

big role in any real-world evaluation. Additionally, it discusses the feasibility of a psychological interpretation of results from the previous chapters. This further contributes to the BFUS model's potential adaptability (C2) and shows additional perspectives on the effects of user traits on usage on a low-level, contributing to C3.

The thesis concludes in Chapter 7, where a summary of all findings and contributions is presented. It also outlines potential future work that could extend the findings of this thesis. Finally, the thesis ends with an overall critical assessment and conclusion.

Parts of this thesis have previously been peer-reviewed and published (assuming minor alterations to ensure consistency and fit the overall narrative). In particular, parts of Chapter 3 have previously been published in © 2021 IEEE, *International Conference on Pervasive Computing and Communications (PerCom)* [35]. SA-related parts have largely been published in © 2022 Elsevier, *Pervasive and Mobile Computing* [36]. Impulsivity-related parts are largely included in a research paper currently posed to be submitted and peer-reviewed.

Chapter 2

Background

Smartphone usage is a broad field of research which targets different aspects of the device, its users and their relationship. Li et al. [77] defined the different main contexts of the literature as:

- "App", includes applications and their categories as a whole, and how they are connected to entire ecosystems, e.g. their evolution within app stores.
- "Smartphone", is the domain with the device's capabilities and requirements at its centre, e.g. how sensor data influences usage with influences such as battery drain or network traffic.
- "User", describes the domain which explores the interplay of user characteristics and usage, which can be used for profiling purposes on a group (e.g., attributes such as gender or age) or individual (e.g., tracing of a digital footprint) level.

In combination, these observe how shifts in the usage of smartphones affect the ecosystem at scale. This encompasses trend data such as application and network use in the total population of smartphone users and also includes how influencing factors such as battery drain, network traffic or discoverability (through categories) of applications change their perception and desirability of use from an economic point of view. Following this definition usage may be collected via cell towers or logged by an external authority. For example, call logs from a network provider may be analysed to identify

trends for entire regions [134]. For the most part, the focus of the survey is on monitoring these macro-trends of smartphone usage and how the market is changed through them, therefore it is recommended for those cases of interest.

However, as the survey is mostly focused on usage in the context of those macro trends, it only touches on some aspects of user behaviour and its relationship with different demographics (user profiling) and on the individual (user identification) level but misses some key components of what constitutes behaviour for individuals. User behaviour in this context is dictated by a variety of features that affect people directly (e.g., how long they use their phone) and how this is influencing (or influenced by) their decision-making process. So, while smartphone usage on this level is mostly encapsulated in the definition of "User" by Li et al. there is no strict separation between it and "App" or "Smartphone" when considering what constitutes behaviour overall. For example, some information such as application switching patterns or sensor data of the smartphone can be contributing influences on the behaviour of users.

This chapter will focus on this link between the smartphone and user-centric aspects of behaviour. It will expand Li et al.'s survey with additional literature concerning the various possibilities of capturing and encapsulating device behaviour and how it can be utilised to inform of personal and intrinsic user habits, patterns, or attributes. Furthermore, it will discuss how smartphones are reflective of users' mental states beyond personality and show how users can be affected by their mood, boredom or cases of problematic use such as addiction and impulsivity. Finally, it presents necessary background information on effect sizes, natural language processing, and datasets, which will be repeatedly used throughout this thesis.

2.1 Related Work

Figure 2.1 shows the relationships between the parts of the structure proposed for app usage research by Li et al. The most related parts and the focus in this thesis are

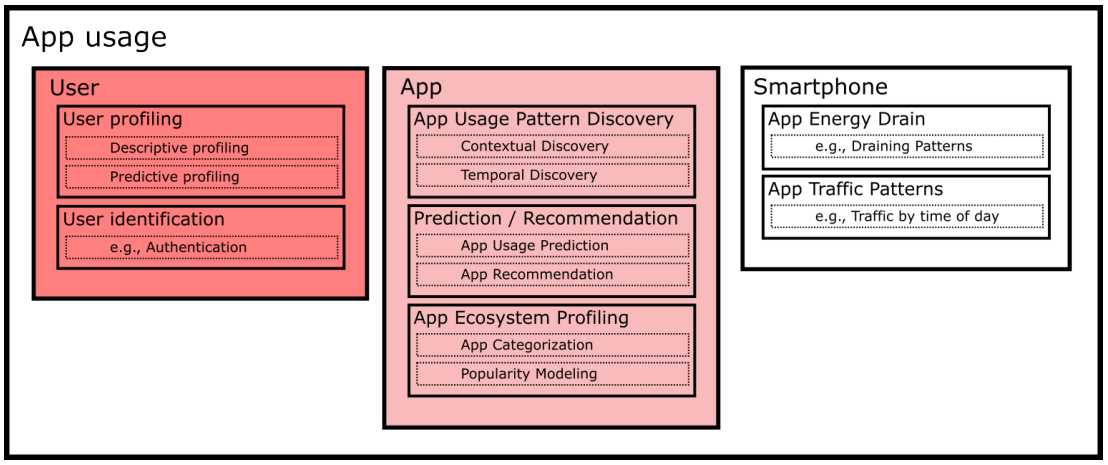


Figure 2.1: The structure of app usage research as identified by Li et al [77]. Parts which are relevant to this thesis are highlighted, the intensity of the highlight corresponds to how strongly this thesis ties into the respective research area.

on the 'User' and 'App' dimensions. The Smartphone dimension which focuses on energy drain and large scale application traffic is not relevant to user-centric behaviour research of user groups or individuals and will not be covered in detail.

Behaviour in smartphones has been encapsulated by many different approaches. This is partly due to the evolution of the smartphone as a device but also because of different interpretations of what constitutes "use". For behaviour research it is interesting to consider the effect of apps on the user and how a user expresses themselves in different kinds of applications. Contexts from sensors and social surroundings have been studied and found to be influential, however we want to investigate behaviour and its effect on usage at a lower level.

2.1.1 Usage Contexts

One aspect of defining usage has been the identification of how the changes in the smartphone's state but also the user's surrounding environment influences their decision making processes. For this reason, researchers often collect data from many dif-

ferent sensors to find correlations between a user's behaviour and their context. While methodologies frequently overlap, this can be split into the following three categories:

Usage or Application Context Nowadays the use of a smartphone has far exceeded the original use cases of mobile phones which were mostly just communication. They can be used for a large range of use cases now, including productivity apps for tasks such as document drafting or task management, entertainment such as playing games or browsing social media and many more. Therefore, which applications are used is very commonly collected as part of the full user behaviour assessment since their content and interactions are often considered distinct. This usage context describes the users engagement with the content of their device. The focus is specifically on which application they are using and how they have interacted with them. Since they can offer vastly different functionality grouping the behaviour of all applications together can discard a lot of the available nuance. For example, Deng et al. found that time spent in applications can be vastly different across categories, over a 7-day period the daily use of social applications (e.g., facebook or instagram, 44 minutes) was almost double that of using web browsers (25 minutes) [25].

Phone calls and text messaging used to be the main use case for mobile phones before smartphones became ubiquitous and enabled a wider breadth of functionality through various kinds of applications. However, those base functions are still an important feature of smartphones. Their frequency of occurrence and length of use can give insight into the patterns and habits of users (e.g. [90, 68]).

In general, applications (and their categories) have become more and more important with the rise of mobile device ubiquity. The emergence of smartphones means that non-voice applications have become more and more accessible as. In 2010, as phones had started to be capable for multi-purpose tasks, Rahmati and Zhong equipped participants with the same mobile phones over a 4 month period. They found that users were interacting with a large mix of phone applications regardless of location or vicinity to PCs. The also found that the devices would be used for many purposes including

accessing information but also socialising [122]. Rahmati et al. followed this up by showing that the previous usage context of a user influences app selection heavily, but only the directly proceeding app used may be of actual importance [121].

Prediction and recommendation of applications specifically based on the usage context is a common use case [118, 169, 142, 29, 131, 47, 11]. Parate et al. trained a model for predicting and preloading applications based on 7,630 devices in the wild and validated it with a 22-user control group. By utilising a sequence of prior used applications they were able to predict a user's next app within 5 choices with an 81.89% accuracy [113]. Furthermore, based on a 2019 study Turner et al. suggest that application switching occurs as a part of a larger network and while there are overlaps in frequency for commonly used applications, that these switching networks are highly dependent on the individual since the networks show high variability in their edit-distance (M=0.89, SD=0.06) [149].

Zhao et al. conducted a study using the application history of 106,672 Android users and noticed that applications usage is influenced by factors such as gender, age or income. By analysing the application descriptions they were able to infer differences between groups of application. For example, they found that male users (p=0.0049) were 10 times more likely to use 'gaming' applications than female users (p=0.0005) or that applications which allow customisation of the lock screen and launcher were a lot more used by younger users (age 0-17, p=0.012) than older users (age 35+, p=0.0003) [171].

As such and continuing within this thesis they are defined as follows:

Definition 7. *Applications in the context of a smartphone are individual pieces of software which usually cover one specific purpose. Every action performed on a smartphone occurs within an application, which means that even standard or pre-installed functionality like the 'phone' or 'messages' are applications. Even the home screen (often called "launcher" within Android) counts as an application.*

Definition 8. *The category (of an application) is a tool to label and group it with other*

applications of similar or adjacent functionality. An application category is usually required to be defined by the application store (e.g., the Google Play Store) where the application can be downloaded.

Sensor Context Another aspect which can influence usage is the surroundings of the smartphones itself. This can also be referred to as the sensor context. Over time smartphones have been equipped with more ways to understand their environment and how they are being used within it, including their orientation, location, time and more. Therefore the sensor context includes any data which can be sensed by the smartphone itself and is then frequently used to establish classes of usage and compare them as an influence on a user's behaviour.

The location of a user can be retrieved using GPS or location via Wi-Fi networks and tracking the physical location enables the retrieval of movement throughout their life, establishing typical locations such as work, home and other common locations. Soikkeli et al. found that users are using their phones for longer per session at home (M=321s) than in the office (M=234s) or 'Elsewhere' (M=209s). However, they actually generate the least sessions per hour when at home compared to the others, showing that there is clear differences in how often and much the devices are used in different location contexts. Additionally, analysing a user's location in relation to other usage data such as use duration or frequency can inform about habits that are bound to specific locations. Using data from 140 devices over a 3 week period, Karkoski et al. discovered that users used communication tools such as email and SMS with higher intensity (more frequently and for longer) than in any other location context (such as at 'Home' or 'Elsewhere') [55].

Similarly, movement sensors such as the accelerometer, gyroscope or magnetometer are used to register a user's movement which can have direct effects on them using their phone. From these sensors Zhuo et al. were able to to detect a users activity (typing, scrolling or watching videos) and the users activity status (walking, standing or sitting down) with a 75% accuracy [173]. In conjunction with other sensors such as

GPS, it can also be used to detect if the user is being transported and which routes are being taken. Nawaz and Masocolo were able to detect whether they were driving a car and enrich this with the specific routes they were taking [102]. The battery can also be used to infer some characteristics about their usage. Higher efforts in computation cause the battery to deplete faster, meaning that in certain situations users can become conscious of their battery usage. This means user behaviour changes in certain states of the battery such as when it is at low power or plugged in and charging. Kang et al. found that by analysing time spent in calls or using data they were able to observe a user trying to conserve battery influence their usual habits [54].

Network status can be relevant for apps that require an active internet connection. The differences between being connected to a cell or Wi-Fi network, or having strong compared to low signal strength impacts waiting times and responsiveness. With data often being limited by mobile network providers, these changes can influence a user's decision-making regarding which applications to use (and to which extent). Srinivasan et al. derived usage rules for individual users based on multiple factors such as time spent, app patterns and signal strength. They found that some of their rules only applied when connected to WiFi or cellular [137].

Social Context Since smartphones are devices that can be carried anywhere by a person, the social environment it is used in can become a relevant datapoint when considering how a user interacts with it. Ferreira et al. showed that based on the four different social settings 'Alone', 'With friends', 'With strangers' and 'Other', over 80% of micro-usage sessions (15 seconds or less) occurred when the user was alone [34]. Furthermore, Papapetrou and Roussos detected three distinct clusters of usage which they attributed to be 'professional', 'family' and 'leisure' activity [112]. Overlapping with detecting whether a user is 'at home' or 'at work' based on their location, Srinivasan et al. have shown that users have diverging usage patterns in those different social contexts [138].

The dependence of usage on a user's social (or non-social) context has also been ex-

plored by studies surrounding problematic or addictive use. For example, Salehan and Negahban found a link between social networking applications and phone addiction [125]. However, later Elhai et al. investigated the link between 'social use' and 'process use' (e.g., "news consumption, entertainment, relaxation") again and found that process use is actually more indicative of addiction than social use [31]. This expands on their previous work that smartphone addiction is caused by a need for social fulfilment but might not directly be correlated with the usage of those application [30]. This shows that the user's social context can have an impact in how or why a smartphone is used. However, it is difficult to interpret the combination of signals which is introduced through the complexities behind user behaviour and the network of social needs.

Time represents a special case as a context. It can technically be captured as distinct datapoints with a sensor, but that oversimplifies its impact. While it can be used for distinct sets such as 'Morning' or 'Evening' it can also be used as a separate factor to observe continuous change over time for any sensor. Therefore, it may be considered a whole separate dimension of usage discovery, Li et al. classifies temporal patterns as different to contextual patterns as the there may never be an exact distinction between context [77]. In either case, the time of day or day of the week are metrics frequently used to establish differences in behaviour for temporally distinct periods. Böhmer et al. found that session length and frequency of app use were different at different points of the day such as morning, evening or night [18]. Further, using multiple metrics including app duration and frequency, messages, website visits and location data LiKamWa et al. were able to build a predictive mood model on a smartphone which included findings such as a worse mood on weekdays compared to weekends [79].

This section detailed the different contexts that are relevant for smartphone usage and how they have been used previously for understanding user behaviour. Of note are particularly that contexts can be used to meaningfully distinguish different modes of usage. However, for many of these contexts the underlying features are high-level

features such as screen-on time or time spent in applications or their use frequency. Low-level features are only used rarely, perhaps because it's harder to capture data in such detail. This means using these high-level features is currently often considered the baseline for how usage is defined. It may be worth reconsidering this usage baseline in the first place even without added contexts such as sensors or social constructs.

2.1.2 User Types And Use Patterns

These context of smartphones have improved the understanding of how smartphones and their application are used on a general level. Those differences in usage stem from the individual users that handle their phones, there is a large variety of people with different characters, living in different environments and with different backgrounds. The identity of any individual user mixed with shared attributes between specific types of users contribute to a large variety of how how behaviour is expressed in the form of usage. These user types can be understood as groups within all smartphone users that show collective patterns and habits in their usage. One way to identify these groups is through their descriptive usage demographics, which are generally linked with digital traces [67]. This includes attributes such as gender (e.g., [7, 65, 170, 87]), age (e.g., [7, 87]), income [170] or marital status [87]. Users with different cultural backgrounds also have been found to have differentiating usage characteristics surrounding smartphone use [117, 80].

Another concept for user types is behaviour-driven and shares characteristics with those of "power users" using desktop computers. In a parallel with smartphones, Kang and Shin found that users with higher intensity ('power user') are less likely to share personal information than 'nonpower' users [53]. This separation into two specific groups is otherwise less common, rather, given the diversity of platforms, users, and phones there have been a few to dozens or hundreds of types of users found (e.g. [169, 49, 52]). Zhao et al. found up to 382 distinct types of users based on application use duration and frequency [169]. These groups of users do not necessarily all have a

label to describe them directly but from their behaviour, it is possible to extract types of usage that might differ in where, when, how often, for how long and which applications they interact with. This type of identification is a more general approach to understanding behaviour and often helps in highlighting the diversity that users might show not just between each other but even between their own sessions in day-to-day use.

A more directed approach is when very specific patterns of behaviour are extracted. This is also called "rule mining" (e.g., [101]) because the resulting sequences (rules) follow a conditional "if-then" pattern as described by Pinder et al. [119]. In these rules factors such as location, weekday and time, battery state and more are used to construct a conditional for a given outcome. For instance, Srinivasan et al. found that a user being home at night on a weekday had a high likelihood of opening Facebook [137].

The detection of types of users or at least types of sessions has been of interest to understand how groups of users interact with their devices. This has been largely successful with high-level features and has shown that those features are capable to effectively separate different kinds of users and modes of usage. This means that if features are generally able to distinguish between different types of users they are likely to be useful for other tasks such as predictions or classifications.

2.1.3 Reflection on Usage Definition

One of the classic metrics to observe usage is monitoring the time which is spent using the device (e.g. [155, 123, 84, 60, 25, 115, 23]). This can be understood as the time that the device is not off or sleeping [33], but also includes intricacies such as time spent on the lock screen [43], time spent in certain applications (or categories thereof) [33, 104] or time with a minimum amount of attentiveness [152]. Aside from the length of use, a different way to consider usage is by how often actions are made. For example,

how often an application is opened in a specific time frame. Both of these concepts, frequency and duration of use can be applied to many contexts, such as user location, phone calls or internet usage (e.g., [55]).

It is used a frequently because it has some nice properties in that it's one of the easier metrics to capture and it is relatively straightforward to interpret. Especially when paired with other high-level features such as application launch frequency it delivers very interpretable results. However, Oulasvirta et al. found that smartphone sessions are often very short bursts of interactions. When investigating sessions they found that those which are predominantly 'touching' focused are only 1 minute or shorter 92% of the time and when 'scrolling' focused shorter than 35 seconds 90% of the time [111]. Also, Ferreira et al. found that a lot of applications are used with only very small sequences of 'micro-usage'. They identified that 41.5% of applications are only used for less than 15 seconds [34]. In another study, Banovic et al. found that out of all sessions 95% are shorter than 360 seconds [12]. This shows that interactions with smartphones are dense and that differences in how they are used may be small. This means that not only time may be compressed to only a few seconds of usage at a time, but also that application switches and other high-level features have to be arranged within those tiny bubbles of use.

Furthermore, using 30 devices and 82,620 application usages per device, Gouin-Vallerand et al. showed that the grouping of usage based on transitions created by switching applications is not the same as those found when considering time. By analysing switching behaviour in the form of Markov-Chains is distinct from usage described by just time spent in applications per-day (F=0.00156) and also per-hour (F=0.00117) [41]. This is reinforced by the findings for common application networks by Turner et al. discussed earlier [149].

User behaviour may frequently be viewed through a potentially distorted lens by relying on high-level features. It is possible that while they certainly represent usage at some level, they compress important nuances. This may also be further exacerbated by

issues such as data sourcing and evaluation while aspects of smartphones keep changing rapidly.

2.1.3.1 Diversity of Data Sources

Some of the variances in the definition of behaviour can be attributed to the different means of acquiring information from users. Between surveying users and recording interactions directly on the device, the latter has seen a rise in popularity recently [97]. However, there are some real challenges to recording this data directly from a user. Since smartphones have evolved to be a companion in many people's lives, carrying a lot of identifying or private information, privacy and ethical related problems need to be addressed. Additionally, there is substantial work required to design an accurate infrastructure to capture the high volume of data generated by users. These automatic collections might either happen as part of an application with another use case, as a result of a direct recruitment effort or sometimes even in a fully controlled environment to establish a specific context.

A significant contrast in the collection might be presented by the difference between popular operating systems. Given the market, nowadays the most popular systems are Google's Android and Apple's iOS. Because of system restrictions of the latter, it usually has data collected by survey only (e.g., [155]). Studies that do have generated user data either did so on old operating system versions when the system was still less restricted (e.g. [145]) or completely focus on the collection of in-application data (e.g., through the help of a library that developers have to add to their individual applications [153]). While focusing on single applications can be an interesting approach for certain cases such as marketing, it loses a lot of the general use information, similar to constraining sessions to application sessions as discussed in or these reasons, the most common platform to collect data from is Android, which imposes none of these restrictions.

With a field moving as fast as the development of smartphones a consideration of how

devices might have changed over time is also needed. What qualified as a smartphone a decade or two ago is a completely different user experience from what is used nowadays. This is a motivating factor for an adaptive model of smartphone usage analysis which is not locked to specific collection and processing methods. These show vast differences between devices and influences when it comes to the collection of the high-level metrics surrounding usage. Combined with the indecisiveness in the literature of which features accurately encapsulate behaviour leads to the question of whether these features characterise the complexity of user behaviour accurately.

This is further exacerbated by the differences between survey-only behaviour capture (e.g. [161, 4]) to on-device capture (e.g. [104]). On-device logs or monitoring apps provide detailed and accurate information about how a smartphone is being used, while surveys rely on self-reported information from participants, which can be prone to bias and inaccuracy. Additionally, survey responses are often based on a limited scope (such as a 6-point Likert scale, e.g. [161]) and may not capture the full range of influences present when evaluating smartphone usage behaviour. Accurately comparing the data gathered from these approaches is difficult, and it hinders the ability to draw meaningful conclusions about smartphone usage patterns.

The diversity in collection and definition of usage behaviour was partly sparked due to the vast differences in how this information can be processed. Research has unfolded surrounding modelling smartphone behaviour as a data complexity problem. Therefore, making these actions interpretable and comparable as part of understanding a user's behaviour relies on a formal description of what behaviour is. Multiple concepts and techniques have been discovered to quantify smartphone usage but because of the complexities inherent in human behaviour, there is no unifying solution that is generally accepted.

2.1.3.2 The Case for Low-level Features

If high-level features may cause issues with compression and perhaps obscure the root complexities of behaviour there may be better ways to capture usage. Previously, some actions or events have been observed specifically because smartphones offer unique ways of interacting with them. Drag-and-drop or swipe actions on a touch screen differ strongly from interactions that would be possible with buttons or switches. Alqarni et al. was able to identify common swiping patterns and gestures with a 74.97% accuracy and identify users based on their keystroke pattern with a 63.72% accuracy [6]. Gooding et al. used scrolling data to distinguish text readability between languages and found that users interact differently with 'advanced' or 'elementary' texts and also that when scoring their comprehension of these texts faster scrolling generally correlated with worse results [40]. Also utilising touch and scrolling events, Yu et al. identified their impact on energy as a consequenceand proposed changes to adaptable frame rate models that would not impact the user but improve energy efficiency [167].

Mehrotra et al. included events such as tap and long tap and found that how physically active a user is was not a statistically significant influence on how long they used their phone, but was significant for how many applications switches they had made and on how many times they tapped the screen [90]. By focusing on just the keyboard, Ghosh et al. were able to predict a users emotional state from taps and swipes with an accuracy of 70% [37].

Low-level events have seen consideration in the literature previously, but the default is often still their high-level counterparts. Partially this may be because there is no consensus on which one to use and because there is no direct comparison in how they perform next to each other. Introducing a model which can compare low-level and high-level features directly may help understand their respective differences and potential advantages.

2.1.4 Smartphone Behaviour Informing User Traits

Another aspect of smartphone behaviour research explores the interactions of smartphones and psychological traits (and their psychological effects in day-to-day use). The collection of users experiences and feelings during their usage allows us to establish links that help to deepen the understanding between the user and device. One of the common techniques which enables the collection of more personal user data points is the Experience Sampling Method (ESM) [151]. As the name suggests, the ESM is a method of sampling otherwise inaccessible information from users at various points in time. An external signal such as a notification or sound prompts the user to record their current state, usually in the form of a multiple choice or Likert scale (e.g. [105, 17, 163]), and thus allows to capture what a person is thinking, feeling or doing just by reminding them to record it. The method does not provide any survey or questions by itself; it just establishes the framework for how data can be sampled effectively from a user.

Through this information is possible to model how the smartphone affects certain mental states such as the user's mood reflected by emotions (happy, sad, angry) (e.g., [66, 156]), their level of stress (e.g., [23, 165]) or their anxiety levels in general and social situations [30, 165]. Often this is coupled with predictive approaches of user profiling to detect differentiating aspects within usage based on a user's traits, or also emotional or mental state.

While mood and emotions can steer usage, they are not necessarily indicative of problematic behaviour. It is difficult to distinguish when and how the effects of problematic usage occur, however markers for them have been identified before. This is one of the focus points of behaviour research, which includes the identification of factors which may cause or lead up to uncontrolled, impulsive and problematic behaviour. This also overlaps with the behaviour research surrounding user types, for example, addicted users could be considered a group of users with specific aspects in their habitual usage that can be detected.

2.1.4.1 Digital Phenotyping and Psychoinformatics

A phenotype can be described as the observable traits or characteristics of an individual, which are induced by a combination of their environment and genes. Digital phenotyping describes the ability to determine a user's phenotype from just their interactions with digital devices. The term was first introduced by Torous et al. to describe a framework for the pervasive capture, summation and processing of user data to infer their psychological traits [144]. In this, data is captured on a (mobile) device and then sent and stored on a remote server for further analysis.

Since digital devices have taken over such a large chunk of people's lives, inferring user types from their characteristics in usage has become more readily available. While the ubiquity of devices is not a necessity for digital phenotyping, it has pushed methods to more popularity. With computing capabilities and sensors improving steadily, the options for a wide variety of possible data capture (as discussed in Section 2.1.3.1) enable the collection of precise information about a user's status. This information enhances the understanding of user behaviour and personality from what is possible by completely relying on surveys or in-person assessments.

Psychoinformatics is a related emerging field with a focus on utilising the vast amounts of data available through personal digital devices. The basis for it lies within the ability to precisely record and measure information of a user in an almost constant stream. The distinction (or addition) to digital phenotyping is in the granularity and volume of recorded data. Where digital phenotyping refers to the general concept of using digital devices to infer psychological features, psychoinformatics focuses specifically on collecting as much data as possible for future rule mining [88, 97]. In this, there have been suggestions of how this recorded data can be applied, for example in diagnostics of problematic use patterns [96]. Various applications and studies have employed these strategies, including Tymer [104], Menthal [8], mPulse [159], and AWARE-Light [150].

These areas show the potential of processing usage events, especially in large amounts. This further motivates exploration in how effective these low-level features are at uncovering the complexities behind usage.

2.1.4.2 Smartphone Addiction

With smartphones having such a constant impact on people's day-to-day lives they are faced with problematic habits which might inhibit their normal behaviour. Generally, addiction is a case of compulsive or obsessive behaviour that continues even when faced with the negative consequences (financially, socially, etc) of those actions. For example, Van den Bulck has found links between smartphone addiction (SA) and its impact on sleep and how use late at night can affect long-term tiredness in users [154].

There are different kinds of addiction and SA is often considered a 'behavioural addiction' [21] where habits get enforced from gratification instead of as a result of e.g., substances. Additionally, Liu et al. found that there exist parallels of the behaviourial issues between gaming [81] and smartphone addiction and that users with gaming groups had a higher chance to show signs of smartphone addiction. Similarly, Beranuy et al. has investigated potential links between internet addiction and problematic phone use [15]. This was also reflected by Jin Jeong et al. where they found multiple correlations between internet addiction SA, but they also suggested that SA's contributing factors are harder to differentiate from those of non-addicted users compared to the same factors when considering internet addiction [51]. This kind of addiction has also been linked to various personality and identity traits such as anti-social behaviours by Pivetta et al. [120].

SA is often linked with certain application categories such as social networking and communication [135, 26, 104]. This is because there are strong assumptions about the role of "social-seeking" behaviours that are either a result of or caused by depression, anxiety and stress [120]. Elhai et al. also related this conceptually to the idea of "fear of missing out" and feeling a strong obligation to be updated with their social circle at

all times [30]. Deng et al. mirrored this in a study which found that, while there can be many reasons why one is unable to stop using their phone, many times it is down to keeping up their status in their relevant social circles [25].

The relationships that have been found between application categories and SA also extend to non-social ones, Bae et al. discovered that there are correlations between SA and entertainment applications or video games [10]. Additionally, there have been reports of categories of applications or individual applications being identified as a contributor to SA that do not seem immediately obvious, for example, while investigating links between addiction and application categories Roberts et al. found that for male users reading in a bible application was correlated with their likelihood of addiction [124]). Park et al. discovered that there are patterns which overlap in casually habitual and addictive use behaviours, but that when actual addictive use takes place it has additional effects on a user's life such as their sleep duration [114].

The Smartphone Addiction Scale (SAS) developed by Kwon et al. is a measure of an individual's proneness for addictive smartphone behaviour. The SAS is a self-report questionnaire based on six factors: "daily-life disturbance, positive anticipation, withdrawal, cyberspace-oriented relationship, overuse, and tolerance" [71]. It consists of 33 questions on a 1 ("strongly disagree") to 6 ("strongly agree") Likert scale, resulting in a point range of 33-198. Its continued development led to a short version (SAS-SV) [72] which reduced questions to 10 and defined cut-off values to identify individuals as addicted or non-addicted. With a reduced point range of 10-60, for male participants, this cut-off is 31, while for female participants it is 33. The exact questions can be found in Appendix E.

2.1.4.3 Impulsivity

Impulsivity is described as a personality trait that surfaces by an uncontrolled reaction to stimuli which has roots in thrill or novelty seeking behaviour but also potentially self-harming disorders such as kleptomania, pyromania, borderline personality

disorder or hyperactivity [160]. There have been multiple variants of detecting impulsivity using self-reported surveys. For example, the Barratt impulsivity scale (BIS-11) divides impulsivity into three subcategories: attentional (e.g., loss of concentration), motor (e.g., action without thought), and non-planning impulsivity [116], while the UPPS-P as proposed by Whiteside and Lynam identified five personality-based traits [160]. There is no complete consensus surrounding the classification of impulsivity but a general acceptance of different factors exists. These surveys use questions in all categories to probe the variously identified sub traits of impulsivity.

Impulsivity has been linked with many personality disorders such as drinking [140], gambling [27], and other substance addictions [62, 64], but has also been applied in other fields. For example, adoption of security measures for physical health concerns such as Covid-19 regulations has been studied in relation to impulsivity [83, 163]. Additionally, increased impulsivity has been linked to posing a greater risk to cybersecurity via e.g. private information disclosure [2, 3].

Moreira and Barbosa suggest delay discounting as an initial assessment for impulsive behaviour [98]. Instead of probing factors individually, delay discounting employs a more generalised attempt at identifying impulsivity as a function of real-world rewards. Delay discounting is a measure of whether a sooner, but smaller reward is preferred to a larger, delayed reward. It has been linked with impulsive behaviour since it reflects multiple aspects of impulsivity such as acting without thought and non-planning.

Kirby et al. designed the Monetary Choice Questionnaire (MCQ) as a series of 27 questions designed to measure a person's delay discounting based on monetary choices [62]. Rewards are grouped into three separate magnitudes (small, medium, and large) as the delay discounting decreases for larger rewards [63]. Each group contains 9 levels of discounting defined according to a hyperbolic function. The geometric mean of the most consistent choice for each group results in the final discounting level k for each user. This result usually falls in the range of $0 <= k <= 0.5$ where a smaller value indicates low discounting and therefore a preference for a higher later reward, whereas

a higher value indicates high discounting and a preference for an immediate smaller reward. Appendix F lists the questions of the MCQ.

Given these issues of problematic and impulsive use, it would be beneficial to unlock the patterns which coincide with these traits by only monitoring user behaviour on their smartphone without having to impede the user's life otherwise. Long-term, such a system could also avoid side effects from re-prompting their condition caused by direct interactions surrounding the user's problematic behaviour as would be the case with ESM methods or surveying.

2.1.5 Emerging Research Questions

In summary, smartphone usage can be quantified by a wide variety of metrics including the time spent on apps, the number of apps used, the frequency of app usage, the amount of data downloaded, the number of calls and messages sent, or the number of searches conducted. They are influenced by factors such as time of day, location of the user, the device's battery level or signal strength. Generally, as smartphones have evolved to adapt to the various use cases of their users, the vectors of capturable data have increased in parallel. This is only intensified by the variety and limitations of available methods to capture data from smartphone users.

Though all of these methods have been repeated multiple times, they commonly rely on high-level features such as screen-on time to characterise usage. While these features are easy to capture and understand they potentially introduce a stark layer of simplification on the actual behaviour behind the usage. This is because especially screen time obscures all the intensity and frequency of interactions that may be taken while the screen is on. Furthermore, features are often used in isolation (e.g., just screen-on time or count of applications used) to infer usage behaviour which may further contribute to a loss of nuance when evaluating a user's behaviour. **RQ1** In what ways are isolated and high-level features such as screen-on time mischaracterising the actual complexity

of user behaviour?

Furthermore, it may be possible to characterise usage on a more personal level by utilising low-level features in the form of user-interaction events instead. Low-level interactions such as taps and scrolls hold a lot of value to distinguish patterns of usage specific to individual users. It may be possible to utilise them to identify types of behaviour expressed through their usage. **RQ2** How can low-level user-interface interactions be used to effectively infer user behaviour through usage?

Finally, if low-level features offer this capability it leads to another assumption surrounding behaviour being influenced by a user's traits. Traits are considered stable within each user. This means they are a constant part of the user and should exert influence on their decision making. Traits such as addiction or impulsivity have been found to influence usage on a high level (e.g., screen-on time). If this is the case, they may also embed themselves in those low-level user interactions. **RQ3** To what extent are stable user traits such as SA or impulsivity represented by a user's events on a smartphone?

These questions lead back to the overall aim of this thesis to capture the influences on usage which may only be encoded in interactions on a low-level. It follows a discussion of the requirements for the shape and kind of data which is mandatory to work with low-level features.

2.2 Dataset Requirements and Availability

Prior research has collected data in many forms, but most data is not publicly available. Depending on the hypotheses of the research these datasets were captured only certain data points are available. This combined with the fast progression of the technology means that datasets are often not granular enough.

RQ2 and RQ3 aim to model user behaviour using low-level events. Therefore, if those

interactions were not captured in a dataset they are not suitable for processing. Additionally, since hesitation on this level might be an important factor it also is important to have the individual capture windows as small as possible. Ideally no binning at all takes place and all events are timestamped individually. Furthermore, semantic information is required to establish borders between different actions such as screen events and application launches. With this the following requirements for usable datasets are established:

- A unique user identification (can be anonymised)
- System-wide capture (not limited to specific applications)
- Records of a representative range of low-level interactions (e.g., taps and scrolls)
- Application and category data
- Screen state (on or off)
- No or small temporal bins (at most one-second intervals)

Table 2.1 shows a list of publicly available datasets and their overlap with the requirements for the work in this thesis. Most of the datasets had a different focus from this thesis and added recording of certain data (e.g., application launches) as part of their collection for thoroughness. Only the Tymer dataset was set up with user interactions in mind and collected events at an appropriate level. Additionally, only the Tymer and LiveLab datasets recorded the display state and therefore could inform of usage sessions between turning the screen on or off. While not directly relevant to smartphone research, the table includes two datasets that recorded desktop usage. Those include recorded data of low-level events such as clicks, text input and scrolling which could be seen as a parallel to similar low-level events on a smartphone. This shows that while rare for smartphone datasets, this level of detail has been of interest in related use cases

Mobile datasets	**Year**	**U**	**SW**	**AD**	**ET**	**SS**	**LL**
PhoneStudy* [139]	2014-2018	X		X	X		
Carat [107]	2014-2018	X	X	X	X		
Tymer* [103]	2017	X	X	X	X	X	X
TalkingData [141]	2016	X		X			
LiveLab [129]	2010-2011	X	X	X	X	X	
MDC [75]	2009-2011	X	X	X	X		
Desktop datasets	**Year**	**U**	**SW**	**AD**	**ET**	**SS**	**LL**
Behacom [127]	2019-2020	X	X	X	~		X
Four HCI Tasks* [92]	2012	X		X			X

* Available by request

Table 2.1: Public datasets and captured data in relation to the requirements for this thesis. U=Unique user, SW=System-wide capture, AD=Application data, ET=Exact time capture, SS=Screen state (on/off), LL=Low-level capture (e.g., taps, scrolls).

for user behaviour research. As the Tymer dataset offers all the required data and therefore will be the basis for most of the analysis in this it is explored in more detail in the next section.

2.3 The Tymer Dataset

This dataset is a collection of data collected by the similarly named Android application “Tymer”. The app was designed to collect a multitude of device events in the background of a user’s regular usage. This was emphasized by the general availability of the application on Android devices of version 4.4 or higher (the collection on iOS was not possible because of platform restrictions). Distributing the app to users’ personal devices instead of provided, homogenous ones meant that the usage data would

reflect their behaviour in a real-world setting.

Over a period of 8 weeks, it collected usage data from 64 users (see Appendix A). Written informed consent was provided by all participants and the study was approved by the ethics committee of the School of Psychology, Cardiff University. The dataset was collected in 2015 and has been utilised for previous studies of a related nature as this thesis [104, 105, 149].

A total of 82,242,309 individual events were recorded. The dataset comprises multiple types of events that encompass a wide variety of physical interactions a user can take with their phone. As shown in Table 2.2, some events are interactions that are always invoked by the user such as taps, scrolls, or typing while some can occur after external events such as received notifications or the battery falling to low levels. The dataset also captured the participants scores for the SAS and MCQ in briefing and debriefing sessions.

Figure 2.2 is an example of usage from a single user throughout the collection period. Each ring is the equivalent of one day of usage. This reflects patterns of non-activity during night-time (circa 1am - 8am) and most activity during the late evening (circa 9pm - 11pm). This demonstrates how some habits and patterns of a user are imprinted in the usage history of their smartphone, but also that a lot of noise exists in the data.

2.3.1 Pre-processing and Inferred Data

The data is mostly left untouched for analysis as part of the thesis. This is to ensure that the model works on the actual inputs of a user. By doing so there is minimal decisions being made about the input data from any given user. However, because the original data collection was run on user's personal devices and had to be uploaded to an online server some issues with data duplication and inconsistent event capture did occur in some cases.

Firstly, any duplicate events in the dataset were removed - defined as events with the

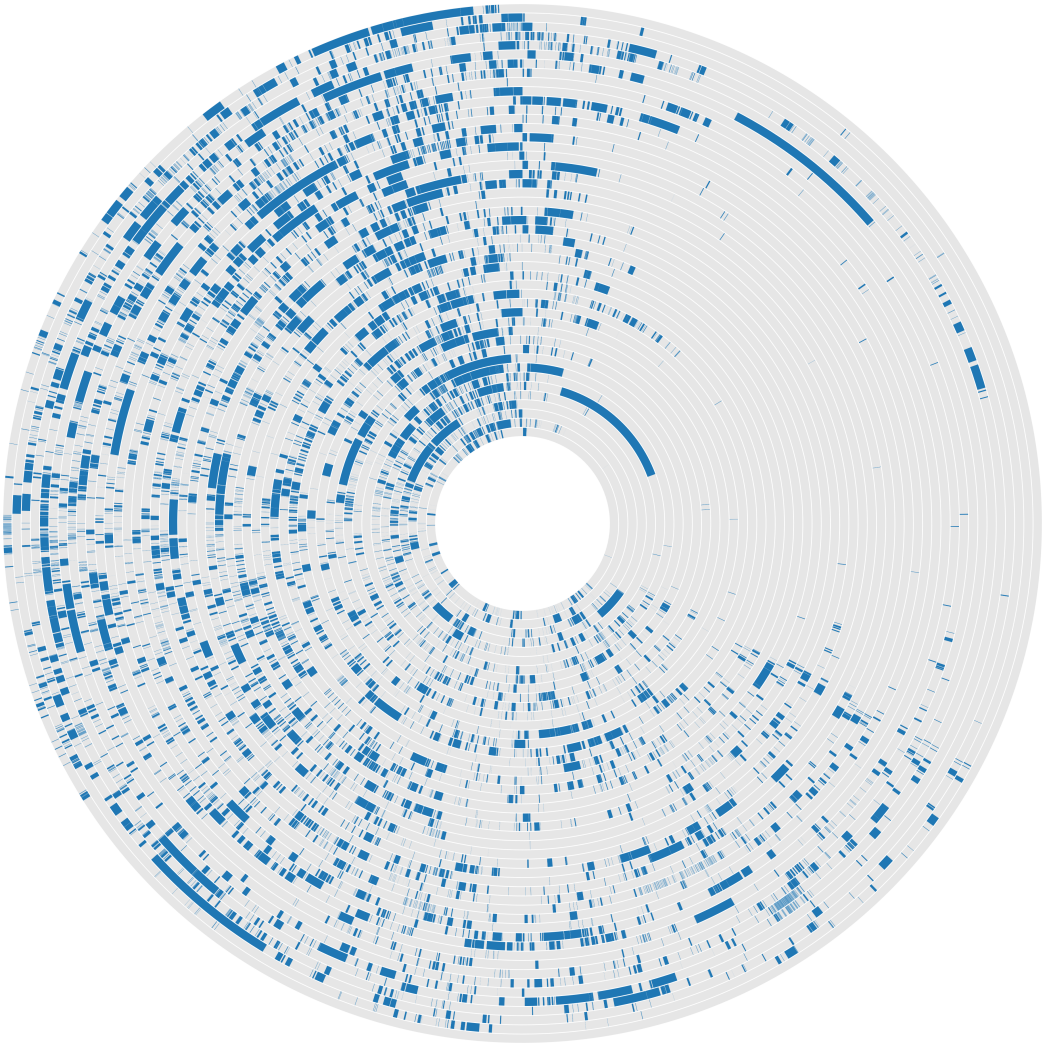


Figure 2.2: A sample of smartphone usage for a single user throughout the collection period. Each ring represents one day, this user totalled 47 days. Periods of screen-on time are highlighted. The data is arranged to reflect a 24-hour analogue clock.

same user, timestamp, and event type. Screen-off endpoints were inferred for any session that missed this event naturally. Application switches were not captured directly but only general system level window state changes. From these it is possible to compute application switches by comparing the change in application package the event is related to, for example only when the package between two window events changes

was the app changed. This enables to infer the application for every event that occurred between these two.

Scrolling was captured via the detection of UI events scrolling off-screen, while this does reflect the action of scrolling it massively overstates the amount of scrolling a user has initiated. To align this with a more natural idea of scrolling events triggered by a user we remove all scroll interactions that are preceded by another scrolling event in a time window of 200ms. This corresponds with a lower bound to what a user would commonly be able to process and react to [50].

2.3.2 Statistical Analysis

A statistical test is a method of posing a hypothesis and a null hypothesis which, if rejected, leads to acceptance of the original hypothesis. Statistical tests commonly are defined to report statistical significance based on an alpha threshold for their reported p-value. The normal threshold for alpha is 0.05 and this (given appropriately constructed tests) results in formally acceptable results.

However, when working with large sample sizes, statistical tests tend to show even small differences as statistically significant. At this point the p-value can become deceiving as statistical significance does not necessarily reflect how strong the effect is. The effect size can supplement the mere existence of significance with a rating of how prevalent the effect is. Therefore, the effect size is frequently reported alongside the p-value to enable comparison of magnitude between multiple significant results. Additionally, when multiple tests can enable to measure the correlation (or separation) of two samples, the effect size can be used to evaluate method effectiveness beyond significance.

In this thesis we report the results of non-parametric tests (such as the Kruskal-Wallis or Mann-Whitney U tests) to compare the effectiveness of different usage features. This is achieved using the effect size of these tests.

2.3.2.1 Effect Sizes of Non-parametric Tests

While the conversion of effect sizes between tests is not strictly necessary to compare results it has advantages to know how the statistical power between tests compares. This is not to probe the many tests for the best results but rather can inform slightly different methodologies compare to each other. For example, how different sampling techniques which derive their data from the same source can influence results.

Given z as a z-test statistic it is possible to calculate Pearson's r statistic result and d as defined by Cohen [24], which is a common effect size figure:

$$r = \frac{|z|}{\sqrt{N}} \tag{2.1}$$

$$d = \frac{2 \times r}{\sqrt{1 - r^2}} \tag{2.2}$$

Some tests such as the Kruskal-Wallis test by ranks calculate the z statistic by default for further internal calculation. Other tests require to derive z from their own test statistic manually. Such a case is the Mann-Whitney U test, here given n_x is the size of a sample and U is the test statistic it is possible to derive z as follows:

$$z = \frac{U - \frac{n_1 \times n_2}{2} - 0.5}{\sqrt{\frac{n_1 \times n_2 \times (N+1)}{12}}} \tag{2.3}$$

To compare effect sizes between tests we add the area under curve (AUC) effect size to the usually reported statistics and p-values. The AUC, as a common effect size metric [91], will allow us to not only compare performance between similar tests but also between tests with similar causal relationships.

The AUC score ranges from 0 to 1, where given two sets of data it describes the predictive capabilities of a chosen variable or model. The bounding values 0 and 1 correspond to a strong (negative or positive) diagnostic ability and 0.5 to no diagnostic ability. While AUC values have no strict boundaries they can be categorised by rule of thumb [1]: poor for 0.5<=AUC<0.7, acceptable for 0.7<=AUC<0.8, excellent for 0.8<=AUC<0.9 and outstanding for AUC>=0.9.

Given d the AUC can be formally derived [42, 19]. This can be achieved by letting ϕ be the normal cumulative distribution function so that the score can be calculated as:

$$AUC = \phi\frac{d}{\sqrt{2}} \tag{2.4}$$

The AUC will be used for comparisons in multiple tables of this thesis, it will mark red for poor (0.5<=AUC<0.7), yellow for acceptable (0.7<=AUC<0.8), light green for excellent (0.8<=AUC<0.9) and neon green for outstanding (AUC>=0.9). To make distinctions more visible the colour scale applies to the values as if they were rounded to their next single digit.

2.4 Conclusions

Understanding a user's smartphone behaviour allows us to explore how they respond to different stimuli and how those responses can be used to improve their lives in a variety of fields. It enables the possibility of making more informed decisions and developing more effective strategies for tackling issues such as problematic smartphone use.

The literature describes many ways to utilise the input, sensor and context data of smartphones to capture and process behaviour information. However, many parts of it are either different or incomplete which reduces confidence in any individual approach. This is intensified by literature that challenges or conflicts with previous approaches. The cause of this is partly due to the natural progression of improved methodologies but a large part is also due to the rapid evolution of smartphone capabilities themselves. The improved processing and networking features allow for capturing large amounts of time-accurate data. Identifying the potential issues with outdated techniques and also the potential application of a new approach is going to be explored as part of RQ1 and RQ2.

This thesis will align some of the loose ends in the current literature and iterates on the common themes to improve the accuracy of decoding behaviour. The aim is to create a basis for smartphone behaviour research which can be used to identify the various types, habits and patterns of users. It also aims to utilise this information to make accurate predictions about a user's mental state and traits to answer RQ3.

In the following chapter, the limitations of current approaches are discussed and evaluated. Based on those limits a new model for behaviour capture is formed. This model is then used to dissect the Tymer dataset and identify various types of users.

Event type	N	Description
Scrolling	48,479,672	Triggered when UI elements were scrolled on or off the screen following user input.
Typing	11,841,853	Text input into input fields.
Text selection	11,596,910	Selection of text.
Tap	4,048,481	A single tap on a user interface view element such as a Button.
Window state change	3,081,043	A change in the internal window state.
Notification	2,480,356	A push notification was received.
Screen events	412,366	The screen turned on or off. The screen can turn on from interactions or notifications. The screen can turn off via timeout or a manual lock.
Unlock	221,617	A manual invocation of the screen getting unlocked. This refers to navigating from the lock screen to the device's home screen or last-used application. Typically, this requires the input of a passcode.
Power connection	42,173	The phone started charging or was disconnected from power.
Long tap	28,530	A press that continues for a long enough time to invoke an alternative event.
Battery state	6,344	The battery charge depleted to low levels or charged back up.
Device power	2,964	The device was turned on or off.

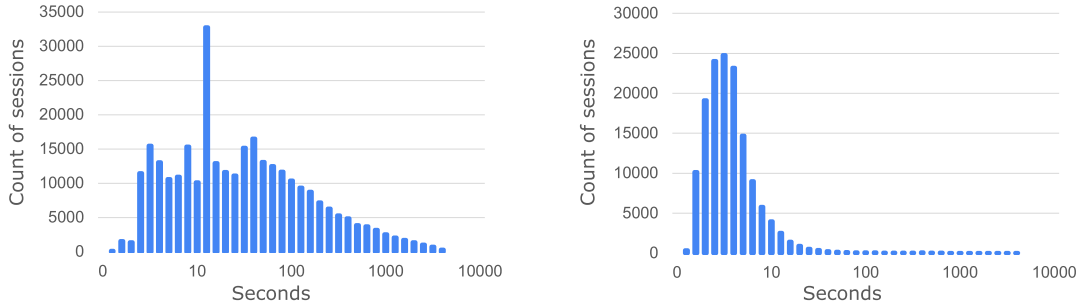
Table 2.2: The low-level event types that were collected by the Tymer application and the number of times each event occurred. Each event has a precise timestamp of when it occurred and from the window state it is possible to infer which application it occurred in.

Chapter 3

Measuring the Relevance of UI events

One of the challenges of defining user behaviour is in reducing the complexity of the wide range of functionality that is behind smartphone use to traceable levels. The smartphone allows capturing both high-level information about usage as well as every low-level interaction. This exposes a lot of collectable detail in the available data but presents a challenge of determining what information is useful for supporting different tasks. For some use cases it may be useful to focus on the higher level information, such as timings of specific device interactions (e.g. prompt timing [8, 103] or notification delays [147]) or which applications were in use (e.g. to infer the impact of specific applications on mood [156]). Whereas for other cases, the nuances of low-level interactions (such as UI events) may be of interest (e.g. when inferring general user traits or personality [95]). This motivates a review of previous methods to extract usage patterns and a deeper investigation into how usage can be defined by user behaviour. For example, while some users might choose to interact with their devices more frequently or for longer, usage has been discussed to be more complex than just timings [93, 94]. This chapter proposes a novel methodology to extract the relevance of specific UI events in order to retain the variance in the encoded data while transforming it into a more usable shape. Furthermore, it demonstrates the success of this methodology in distinguishing various groups of usage from these events.

Part of the current landscape for smartphone behaviour research is the identification of features which are expected to contribute towards an independent variable (e.g. [130]). As stated before, these include features such as screen-on time or app-switching be-



(a) Log screen-on time

(b) Log unlock time

Figure 3.1: Screen-on time and unlock time in seconds.

haviour. The effect of these features is then observed in isolation to understand how a user's context might influence them. Limiting the scope to just a single influence, rather than a combination of metrics, heavily reduces the variance that can be captured using modern methods of monitoring. This shows a discrepancy considering that the complexities of usage behaviour are understood to be dependent on a whole net of interconnecting mental dependencies. Thus, it motivates the identification and formalisation of the limits of behaviour modelling using a constrained feature scope.

3.1 Limitations of Isolated Features to Represent Usage

Certain aspects of usage frequently reappear as features when the identification of similarities in smartphone behaviour is desired. These are commonly high-level summative features that have been the focus of research efforts. This section details these key features and how they can be used to represent usage. Some of their properties reveal how they have inherently limiting attributes which make their use as features problematic. These issues with isolated key summative features then mark the entry point for a discussion and reconsideration of isolated features in general.

One of the most common features used for distinguishing session behaviour is the

screen-on time (e.g. [25, 172, 169, 26, 49]), which presents a convenient and obvious point to analyse user behaviour for several reasons. Not only does it reflect the temporal usage directly in a very understandable format (i.e., seconds/minutes), but it can also be split in a multitude of ways to extract other factors such as time spent in specific applications or categories thereof. Additionally, it is consistent and not selective for certain users, i.e. it will be produced by every user, every time they use their device. Because of the short bursts of interactions common for smartphones [111] the distribution of screen-on time per session is heavily weighted towards very short sessions as seen in Figure 3.1. This and all the following results from here are based on the Tymer dataset.

The context of whether the device is in a locked state has been discussed as an influence on user behaviour, for example, the length of a usage session is influenced by the devices lock state [45, 46]. Users may opt to just glance at information from a device or it could have been turned on from a received notification. This can be a valuable piece of information when trying to understand smartphone-specific use patterns. Derived from the moment of an unlock is the time taken to unlock (the time from the point of the screen turning on until a manual unlock event), which is like screen-on time in that it shares many of those properties apart from being present in every use. Unlocks typically only occur at the very start of any usage session (as reflected by Figure 3.1). Given this distribution, the hesitation of a slow unlock could be interpreted differently than the intent behind an immediate interaction [43].

Application transitions [68, 41], networks and re-visitation patterns [52] represent a different aspect of usage. These can be used to unravel the way different categories of applications interplay and how smartphones are used as a whole to accomplish different tasks. The simplest way of capturing application transitions is by counting how many applications were used throughout a session of use.

While the differences in these features mean to distinguish groups of behaviour from each other, they only partly capture the nuance present in the patterns that make up be-

haviour. These features can misrepresent the variance in actual usage. In an example scenario, a user opens a document-based application to write an essay, another does the same to read an article and a third opens the application but just leaves the application open without any input or interaction. Although the same application was used for similar lengths of time in each scenario, the actual interactions were completely different. When viewing these interactions from the lens of isolated features such as screen-on time or application switches they would show similar signals. Constructing alternate features such as 'count of keystrokes' might enable to describe the differences in this scenario but this will show issues in cross-category or non-typing based scenarios.

From this the following limitations are identified:

L1 The variability of usage is misrepresented by single features because they can only capture specific aspects of the entirety which make up behaviour on a smartphone.

L2 Sessions are skewed towards short bursts of interactions [111] which means that features which are inherently bound to the length and interactiveness of a usage session experience the same skew. This results in it not being possible to properly split single features into appropriate groups of usage, e.g. by considering same-length cut-off points.

Some approaches in the literature have already engaged with these issues by gathering and processing large amounts of user-generated data. This has also included various approaches of models predicting external variables based on multiple features [9, 70]. However, an adaptive, general-use model which makes it possible to address these limitations and accurately encapsulate usage has yet to be defined in the literature.

3.2 The Behaviour-From-Usage-Stream (BFUS) Model

The *Behaviour-From-Usage-Stream* model is proposed in which finer-grained behaviour in the form of a stream of user interactions can be used to model the complexities of usage. Previous approaches have been scattered between correlations with time or count-based features (e.g., with features such as screen-on [25, 172], unlock state [45] or event counts [104]) without formalising a method that can evolve with the capabilities of smartphones and thus available data.

One of the core assumptions of the model is that it is required to be adaptable for multiple modes of current and future usage. Smartphones have evolved quickly over the past decade, and it is likely that they will continue to increase their processing and sensing capabilities in the future. Formalising a static model built around specific input (and output) capabilities that are normal at the current point in time would likely be superseded quickly given the rapid rate of progression of the domain. Thus, we aim to abstract the information retrieval process away from a distinct definition of usage and instead formalise the structure of methods that have been discovered in the literature. From there the task for future researchers is closer to parameter selection rather than requiring rebuilding a model of how to approach user behaviour evaluation. It also offers flexibility towards the devices to which the model can be applied, even if the empirically validated evidence is focused on smartphones of the current time. This follows the fundamentals of comparative emerging research in which user types and labels are assumed to be encoded in usage data [88, 97].

To do so the parameters of the model need to be adaptive so the model will be able to encompass previous approaches within the literature, but restrictive enough that individual parameters can be changed which would enable a comparison between the methods. Instead of having difficult to replicate methods for data transformation and analysis this aims to add a more rigid (step-by-step) structure overall while maintaining customisability for every step itself. This would enable to directly compare the results of those modifications and iterate for the best possible selection of steps (i.e., which

features to select, how to transform them or what method including hyper parameters to use to evaluate them).

For example, a method could then be more easily tested with multiple alternatives by changing just one parameter (e.g., the feature selection) while keeping the data transformation and analysis exactly the same. In Section 3.4 this is used in a cluster analysis for types of usage sessions between users and enables a direct comparison between potential usage clusters found by utilising screen-on time and low-level features.

3.2.1 BFUS Model Concepts

The basis to apply this model for information extraction lies within the assumption that a stream of well-defined events exists, generated by users during any specified usage period. These event types (e.g., taps, scrolls or key presses) are not prescribed by the model itself but have to be defined before the application of the model. The model aims to allow to extract factors such as relationships and relevance from these streams, either associated with their own user but also potentially between users.

A range of n users $u_1, u_2, ..., u_n$ have a one-to-one relationship with event streams so that every user produces exactly one stream of usage $S_{u_1}, S_{u_2}, ..., S_{u_n}$. For each user, given a collection of event types $E = \{e_1, e_2, ..., e_m\}$ where m is the count of types, a full interaction event I_{et} can be described as an interaction with a type e and time of occurrence t such that $e \in E, t \in \mathbb{R}, I_{et} = (e, t)$. In this, each stream is then comprised of a list of events so that $S_{u_x} = \{I_{et_1}, I_{et_2}, ...\}$ where $t_i <= t_{i+1}$, which means that any event $I_{et_{i+1}}$ always directly follows I_{et_i}. For example, Figure 3.2 demonstrates what such an event stream could look like when a single user generates events by interacting with their phone.

These event streams represent the continuous interactions of a user with their device and form the basis of the BFUS model. A three-step process is proposed to handle their transformation and evaluation. The transformation specifically addresses previ-

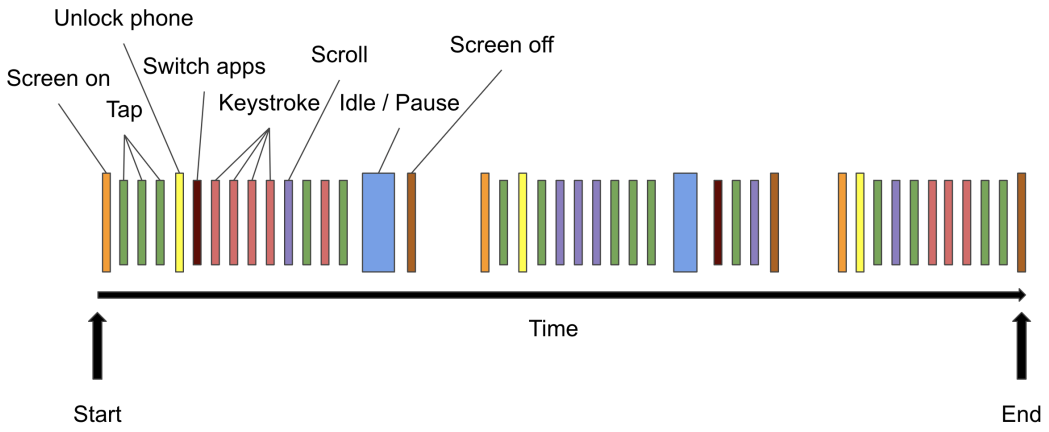


Figure 3.2: An example of a stream of events generated by a single user. The events are labelled and marked with colours based on their types. Some specific events that are contextually important to bounding (i.e., screen-on and unlock time) and also pseudo events (idle) are highlighted.

ous discoveries in usage capture which introduced splitting event streams of each user into multiple smaller sessions of usage such as events per day, within a timeframe or between session boundaries. The second step addresses feature selection, this could be isolated features (such as screen-on time) but addressing the limitations of Section 3.1 allows to create a fixed-size vector space for each session which can include multiple features. The final step handles the application of any chosen method such that knowledge can be extracted from the data.

1 Bounding. While a stream of interactions from any user does not have to have a defined start and end point it is possible to identify behavioural boundaries in them. Task switching and attentiveness are related fields to this within which boundaries have been discussed [110]. Any stream S_x can be split into multiple groups or sessions via cut-off points that align with behaviour boundaries such as screen events, time of day or cognitive timeouts. These groups lay the foundation of a one-to-many relationship between users and usage sessions. This relationship enables to view behaviour as frag-

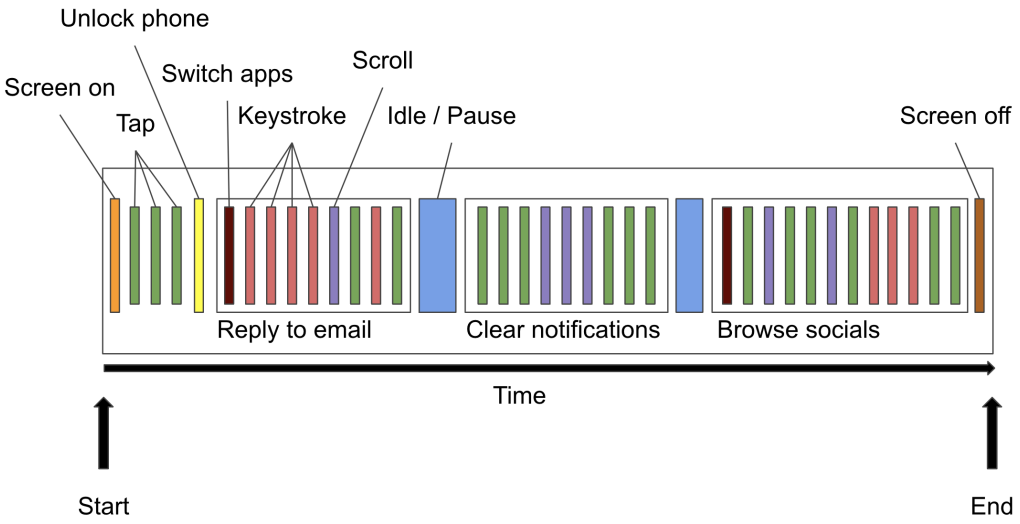


Figure 3.3: An example of how events in a session are arranged. Includes imagined task boundaries which demonstrate how events between apps can be similar or different.

mented pieces which can be evaluated, instead of having to understand it in its entirety. This means that when all groups $G_a \subseteq S_a$ are ordered and added together are equivalent to the original stream S so that $\cup_{G_a} = S_a$. For example, Figure 3.3 demonstrates how events from a stream as introduced by Figure 3.2 could be bounded by screen events to form individual sessions.

2 Vectorisation. For the purposes of analysis, multiple strings of events in the form of sessions can be hard to interpret because the encoded information must be extracted first. Therefore, BFUS proposes a system which transforms the events of each group G_{a_x} into a fixed vector space to align the data to more standard formats which can be processed in the last step. The space itself is not defined by BFUS but is constrained by two conditions.

1. The vector space must be the same size for each group.

2. Every position in the vector describes a distinct feature.

Constraining the vector space in this way introduces a homogeneous structure into the possibly diverse and scattered series of captured events. Simultaneously, it allows to separate information in the data to highlight specific aspects of usage that would otherwise be difficult to summarise. For example, counting the occurrences of events in each group constructs a vector space of length m for each group where every value in the vector reflects the frequency of exactly one type in the data.

3 Application. Once the vector space has been established it can be used to extract useful information. Generally, two modes of operation are applicable when using the BFUS model.

- *Internal disambiguation*, which can be used to infer differences in the usage itself. The vectors can be grouped or clustered using various methods to further understand or separate usage. This requires no further variables and follows more exploratory approaches (e.g., [169, 49]).
- *Relationship estimation* is a method which relates the data in vector groups with one or more independent variables of their respective user. This requires that the variables are captured from each user additionally and independently from the event capture.

This is achievable since the previous two steps have prepared the data in such a way that common methods of hypothesis testing are available. The BFUS model does not enforce any of the statistical tests because the literature has shown that usage data has so much variance based on multiple factors (such as gender, age, location, and more) that tests are not necessarily applicable for all possible combinations.

By following these steps, the Behaviour-From-Usage-Stream model is able to transform a time-sorted usage event stream in such a way to enable extracting information that would otherwise be hidden. The model structures the approach of decoding without enforcing specific methods. However, the literature has shown that some parameters show stronger responses than others (e.g. low-level features compared to

summative features such as screen-on time). The following section presents a few suggestions for model parameters in each of the steps which have been identified for decoding smartphone usage.

3.3 BFUS Suggestions for Smartphones

As discussed previously, because the nuances in behaviour are difficult to capture, the area of research, in general, has not yet established a consensus on how to extract behaviour information. The BFUS model itself does not explicitly define which methods should be used either. Instead, markers identified previously (such as screen-on time, event count and others) have shown varying levels of effect when it comes to identifying the difference in usage. Based on those, bounding and vectorisation methods can be chosen to reflect behaviour following current research methods.

3.3.1 Utilising Screen-Event Boundaries

While group boundaries for the BFUS can be defined in multiple ways in the context of smartphones by using application start and endpoints such as screen on and off events [12] or cognitive timeouts [152]. Between those, both can be argued to be more applicable to separate sessions and the choice has been in contention in the literature [44]. However, since both can also be considered events (screen events literally, timeouts as a 'pseudo-event'), making them equals in theory, either approach is suitable. For ease of computation, a screen-to-screen approach with added timeout events is suggested.

3.3.2 Utilising Methods from Natural Language Processing

Natural language processing (NLP) is a broad field of computer science and artificial intelligence which aims to teach computers to understand and repeat human language.

It spans many areas such as optical or speech recognition, learning syntax and grammar or understanding the semantic relations of words and sentences.

NLP is challenging due to the complexities of human languages, not only are there many spoken and written languages which have completely different expressions and rules, but even internally many languages show variance based on location, dialect, era and other influences. It is also difficult because the same words can have different meanings depending on context. The ambiguous nature of language makes it hard for machines to understand natural language in the same way as humans do. This creates an amount of variation that is impossible to encode in rules written manually. Instead, using large datasets and strong computational capabilities it is possible to analyse text automatically for statistical patterns.

In Section 3.1 the issues of screen-on time and similar features to distinguish user behaviour was discussed. This includes the attempts which utilise the counts of interactions to infer user behaviour and traits. However, this disregards the inherent complexity of usage by focusing on the summed-up number of events which do not necessarily co-exist equally in relation to each other (e.g. amount of scrolling vs long tap events generated). This has an impact on how these counts can be compared to each other.

As touched upon in Section 1.1.2, interactions in sessions can be considered a highly diverse sequence of entities making up a task (or goal). Given this, there are parallels that can be drawn between interactions and language. Both topics deal with highly diverse, sometimes non-logical sequences of entities (words or interactions). If a single interaction on a smartphone is considered a word, and a session a sentence we face a similar interest as NLP in how we want to extract how the "words" relate to other concepts such as meaning or psychological profiles.

While the field of NLP has evolved to use highly specific models that can extract meaning from written text, research revolving around smartphone behaviour rarely uses advanced models that compress high amounts of data. Attempting to build patterns and rules from specific data points is still common.

3.3.2.1 Re-purposing TF-IDF for Smartphone Usage

One key area of natural language information retrieval is terminology extraction. This refers to the idea that some terms in a corpus hold more weight than others in the context of what they are describing. The common application of these weighted terms would be to find the words which have the highest relevance. For example, after retrieving those words you could use them as keywords or to index searches. This can be achieved by picking the highest-scoring words or defining a cut-off point above which all words are relevant. However, the entire vector contains a lot of information about which words are deemed relevant in the supplied corpus.

This form of extraction can be achieved using the term frequency-inverse document frequency (TF-IDF). It is a method of proportionally scaling the occurrences of words in a sentence against their total occurrences in a corpus.

In detail, *term frequency (TF)* is the count of each word (term) in each sentence of a corpus. To adjust for varying lengths of sentences word counts can be logarithmically scaled to reduce the impact of very long sentences. This counting word method by itself would only function correctly if all words in a language would have the same relevance. However, in natural language words such as “a” or “the” are frequent because they appear in almost every sentence but do not describe the topic.

The relevance of these words can be adjusted by applying the *inverse document frequency (IDF)*. In this, every word is tested on how frequent (or rare) it is across all sentences. If a word appears in almost every sentence it is likely less relevant than a word that appears in just a few key sentences.

Together, TF-IDF formally defines the score for each word w in sentence d as:

$$\begin{aligned} \text{TF-IDF}_{wd} &= \text{tf}_{wd} \times \text{idf}_w \\ \text{tf}_{wd} &= 1 + log\left(freq(w,d)\right) \\ \text{idf}_w &= log\left(\frac{1+n}{1+\text{df}_w}\right) \end{aligned}$$

where $freq(w, d)$ is the number of times that word w occurred in sentence d, n is the total number of sentences and df_w is the number of sentences that contain word w. Each sentence d is then represented by a feature vector $f_w = (\text{TF-IDF}_{w_1 d}, \ldots, \text{TF-IDF}_{w_n d})$.

Through this process, the extraction of 'relevancy' for each event over time is proposed. Where relevancy is defined as the importance of an event being quantified in comparison to all other occurring events, it is possible to utilise this knowledge of co-occurrence for further processing. This concept of terminology extraction is already well-established in the NLP domain. While originally thought to be used in order to process this concept of relevancy for words in a document, TF-IDF is a general-use weighting algorithm which can be used in any use case if applicable. Therefore, to extract the relevance of each event in each session the input needs to be aligned to fit the TF-IDF format.

To apply TF-IDF, a corpus and multiple documents (sentences) consisting of words are required. A parallel to usage in the form of event streams can be drawn. A corpus is a sequence of words which is partitioned by punctuation. Each partition is one sentence. An event stream is a sequence of events, partitioned by cognitive boundaries (e.g. screen events or timeouts). Each of these partitions is a session. To summarise, in the context of TF-IDF every event stream is a corpus, every session is a sentence and every event is a word. Figure 3.3 demonstrates how a session between screen on and off events (compare punctuation) can be interpreted to utilise the events similar to words. The definition of TF-IDF_{wd} can therefore change so that w describes event types instead of words and d represents a session instead of a sentence.

The major difference between real languages and considering event types as the dictionary for this transformation is that a real language's dictionary consists of hundreds of thousands of words whereas event types from smartphone interactions would be significantly lower (e.g., 16 types as used in the next section). However, the size of the dictionary should not matter in this case because of multiple factors. Firstly, the formula of TF-IDF does not change because of a smaller dictionary, it is still possible

to apply it in exactly the same way and achieve results that are relevant to its relative dictionary (and corpus). It would certainly change how you would interpret the results of individual words between a sentence from a real language and the results from the impact of an event in an event stream. The frequencies of events would almost certainly be very different than what you would expect from a real language. However, the basis of TF-IDF is the relative impact of words compared to other words transformed within the same corpus. While the expectation within a smaller dictionary is that words are repeated much more frequently within documents they would also be more common within the entire corpus, which means the scaling of occurrences that TF-IDF produces is still valid. Therefore, this difference in dictionary size is not relevant as long as comparisons of TF-IDF frequency results are considered only within the same corpus.

Furthermore it has to be considered that compared to a real language the event stream of smartphone usage has no explicit grammar. Though, a parallel could be drawn between the flow of language (and its erratic and hard to capture nature) and the stream of events that are generated as part of a user's behaviour. For example, the way taps, keystrokes and scrolls combine in certain applications is assumed to not be completely random but instead have patterns of recurring usage throughout. While this assumption is not required to use TF-IDF for transforming events like words in a sentence, it does support the concept of it being possible to extract the relative impact event types within sessions.

To conclude, implementing TF-IDF as part of the vectorisation step in the Behaviour-From-Usage-Stream model enables extracting the 'relevance' of each event in every session. With this transformation, all low-level features can be included without causing issues by having large discrepancies in frequencies between them.

3.3.3 Considering Vector Space Compression

One of the complications of analysing usage is how to process the constant stream of data. Every user may produce different amounts of events, with changing densities of usage and patterns. Even when user interactions are split into sessions, they can drastically differ between users in terms of how much or for how long they interact with their device. This makes it difficult to find methods to analyse usage data directly from an event stream.

In most cases, some form of compression is necessary to analysis usage (and behaviour) since most regression, clustering and similar methods have requirements for the input data (e.g., having a fixed vector space). Compression of interactions to (often isolated) high-level features (e.g., screen-on time) is commonly done as it is simple to process and the results are easily understood. However, this is a pretty strong form of compression which reduces all the interactions to very simple forms, this could mean a lot of information is lost when subsequently inferring user behaviour.

The previous section detailed how TF-IDF can be adapted to process low-level features instead. This still represents a kind of compression in that it inherently discards the temporal relationship of individual events. It is able to compute how impactful certain events were to the overall session, but not how they interleave each other during usage. A possibility would be to expand the dictionary by using n-grams, these would allow to capture sequences of interactions rather than atomic actions.

For example, a simple session consisting of just a "tap scroll tap" could be represented as two sequence, "tap-scroll" and "scroll-tap" instead of three individual events. For the TF-IDF transformation it would be possible to consider these sequences as words instead, however this inflates the input vectors by a lot. For example, the Tymer dataset with 16 event types would result in 256 bigrams or 4096 trigrams which would all represent an individual feature. Computing this for hundreds of thousands of sessions is very resource intensive, and the resulting data is only useful in direct comparison

to this specific corpus. As already discussed, TF-IDF is only a relative metric which means its embeddings only make sense when compared within this exact corpus of n-grams. This makes it difficult to extract a benefit from using n-grams over the atomic events.

Other approaches such as transition matrices (markov chains) can fully represent the flow between events in a session. This can be useful to fully visualise how events commonly flow from one state to another. However, this can be computed overall or for every individual session but does not encode how common transitions are across all interactions. This is different to TF-IDF's ability to include the importance of events even between sessions and not just within them.

For these reasons TF-IDF seems to strike a good balance between providing the compression needed for further analysis but also allows more granularity than simpler compression methods by being able to transform low-level events.

3.4 Exploring Types of Usage Sessions with the BFUS Model

The Tymer dataset presents the opportunity to apply the model in a real context. In this section, the general pre-processing (Section 2.3.1) is followed up with some consideration about the inclusion of specific events. The exploration of habits, usage patterns and user types has propelled the understanding of how mobile devices are used (e.g. [157, 85, 162, 110]). This can be used not just to model the behaviour of a user (e.g. [82]) but also to detect various mental states of the user (e.g. [10, 79]). Applying the model to the Tymer dataset should yield information about some of the latent behaviour in the form of usage clusters.

3.4.1 Data Preparation

The screen-on-to-off bounding step generates N=415,505 sessions. This includes n= 114,485 sessions in which no further interaction occurred. These sessions are on average only 5.97 (SD=5.86) seconds long and are most likely short-glance (e.g. for incoming notifications) or accidental as they are corresponding with the 5-second automatic screen timeout. Since they do not contribute any additional value in terms of usage events they should be considered a type of usage session by themselves. In the context of the model, they do not offer any information for further analysis and therefore are excluded. The remaining sessions with usage behaviour are used as the focus to apply the BFUS approach.

Additionally, some very long-running sessions without any real inputs that seemed to arbitrarily end were encountered. These could be the result of issues with the collection method or part of very unusual usage patterns such as long-running navigation software or similar usage methods. This issue is addressed by removing all sessions that continued for longer than one hour (N=4), leaving a total of N=301,024 sessions.

Features	Min	Max	M	Mdn	SD
Screen-on time	0.1	3595.8	148.3	37.3	342.7
Unlock time	0.1	3576.8	5.4	2	56.3
App switches	0	1004	5.1	3	8.8
Event count	1	104 824	317.7	22	1320.3
Category count	0	16	1.9	2	1.5

Table 3.1: Descriptive statistics of features. Screen-on time and unlock time after screen-on in seconds [35].

Following this, all sessions are vectorised using TF-IDF as described in Section 3.3. Before this, some enhancements to event selection are required to maximise the utilisation of available information in the data. One aspect to consider is the temporal relationships between events. Events are only captured when a user interacts with

their device and there is no data when no explicit action was taken. Just as non-communication is a form of communication in itself, in the context of capturing usage, periods of inactivity may also provide useful information in characterising usage. This is supported by the concept of time introducing cognitive boundaries which have been found to have an influence on usage behaviour [152]. Therefore, a custom idle event for any 30 second[1] intervals of non-interaction was added. This also addresses the choice of smartphone bounding as introduced in Section 3.3. Additionally, screen-on and -off events by definition of splitting the stream of events into sessions had to occur in every single session and would therefore not contribute any meaningful data points for the vectorisation and were removed. The result is a 16-dimensional TF-IDF transformed vector (all features included from the Tymer dataset as outlined in Table 2.2) for each session in the dataset.

Experiment 1 summary

Aim: Explore types of users by applying the BFUS model with TF-IDF.

Input sessions: All available sessions after pre-processing (N=301,024).

Features: TF-IDF vectors created from UI events.

Output: Clusters of usage sessions which represent distinct types of usage.

3.4.2 Types of Smartphone Usage Sessions

One way of applying the BFUS model is by observing the similarities and differences in session behaviour, this constitutes an 'internal' comparison of behaviour encoded in the data. In the case of discovering 'types' of usage sessions, a type could be defined by natural clusters formed by the closeness of the data. While there are multiple clustering options, K-means enables to identify similar sessions around centre points with

[1] In Section 4.1 a discussion follows why this was amended with a short and long idle time of 1 second and 45 seconds. This was not amended for this section as the ability to detect types of sessions did not seem to be dependent on that difference.

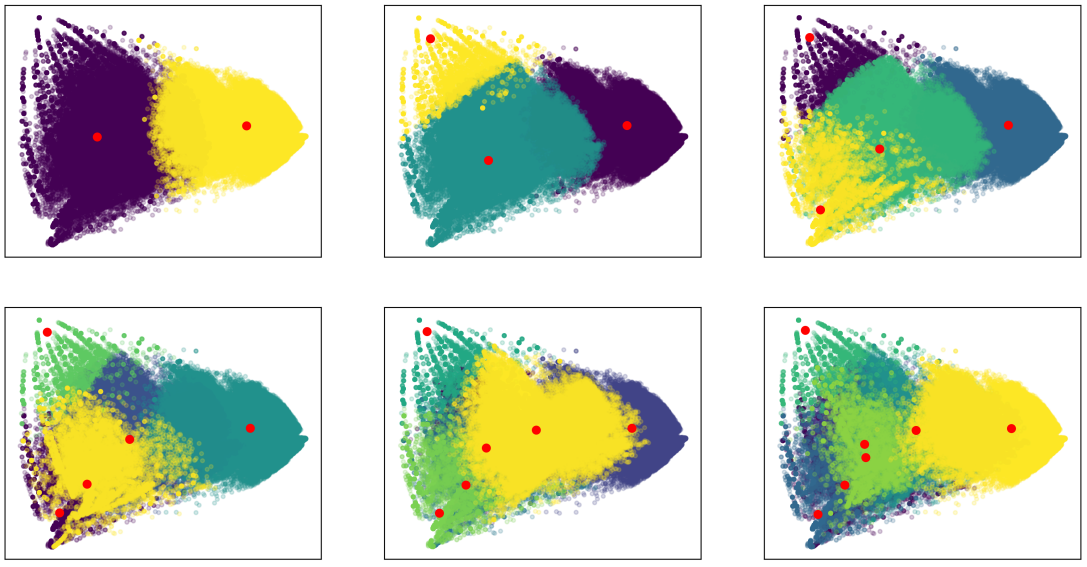


Figure 3.4: K-Means clusters of all sessions for k=2,...,7 formed by the 16-dimensional TF-IDF features after a PCA reduction to 2 dimensions. Each red dot represents a cluster's centre [35].

similar features. The distinction between those clusters reveals the various kinds of sessions and how they are related. The appropriate number of clusters can be chosen by observing the inertia (sum of squared error per cluster) elbow. The inertia elbow shows a dip at k=3 and k=5, for demonstration purposes k=5 is going to be used as the best approximation (see Appendix B). However, it should be noted that the elbow test in general is not a definitive test for how many specific clusters exist and rather offers an initial estimation point of where clusters might be less similar than with a different configuration [57].

Visualising the 16-dimensional cluster data can be achieved by utilising a principal component analysis (PCA) to compress the data points into a two-dimensional co-ordinate system. This is a commonly used technique to make multivariate data more interpretable [38]. Figure 3.4 shows the clusters and their centres when transformed in such a way. While the majority of the data in the clusters separate from the other data there is some overlap visible. This demonstrates how behaviour can not necessarily be

contained by a few specific classes.

A first observation is that while users are very diverse in their overall usage, they are consistent in their types of session, with 98.84% of users having at least one session of each of the five types. This shows that while smartphone usage in the literature has previously been shown to be driven by individual patterns (e.g. [136]), that additional commonalities exist through these types of sessions.

Comparing the most important events to high-level features (Tables 3.2 and 3.3) identifies coherent patterns in each cluster:

1. Sessions whose main events are focused around text input and editing, also with a high screen-on time, unlock time and multiple app switches (i.e. not just text messaging).

2. Comparatively long sessions with a focus on scrolls, taps and switches between applications, which may imply high activity for a prolonged time.

3. Very short sessions (median event count of one), with a heavy focus on app switching, taps and notifications, consistent with glances after receiving a notification.

4. Sessions of 3-4 minutes, with low interaction/app switching, and large idle time between events, potentially being sessions that end up idling until the screen locks by itself.

5. Sessions with a strong focus on notifications but short on all high-level features could be indicative of a session that is changing media and triggering an internal notification.

Despite the process not including high-level features, the typical sessions within each cluster also have a defining set of typical high-level features. Kruskal-Wallis H-tests show that the distributions for each high-level feature do vary significantly across

	Event type	**TF-IDF**		**Count**	
		M	SD	M	SD
Cluster 1	Text Box	0.50	0.15	98.07	240.03
n=98267; 34.08%	Text Selections	0.45	0.21	94.26	217.91
	Scrolls	0.29	0.16	168.31	817.38
	Taps	0.26	0.12	204.90	462.40
	App Switches	0.17	0.11	0	0
Cluster 2	Scrolls	0.31	0.29	78.75	452.7
n=105571;36.61%	Taps	0.23	0.21	8.03	37.22
	App Switches	0.17	0.17	0	0
	Unlocks	0.17	0.24	0.76	0.56
	View Selections	0.16	0.28	58.14	412.82
Cluster 3	App Switches	0.93	0.13	0	0
n=27115; 9.4%	Taps	0.06	0.17	0.20	0.86
	Unlocks	0.06	0.15	0.15	0.37
	Notifications	0.02	0.11	0.04	0.21
	App Switches	0.02	0.10	0	0
Cluster 4	Idles	0.83	0.17	6.9	12.81
n=29945; 10.38%	Unlocks	0.19	0.20	0.68	0.65
	App Switches	0.07	0.14	0	0
	Notifications	0.05	0.14	0.33	1.61
	Taps	0.05	0.12	0.67	4.11
Cluster 5	Notifications	0.81	0.19	16.30	101.70
n=27470; 9.53%	App Switches	0.15	0.20	0	0
	Taps	0.11	0.18	1.20	3.97
	Scrolls	0.06	0.15	1.96	31.66
	View Selections	0.05	0.17	6.03	67.38

Table 3.2: Frequency statistics of the top 5 TF-IDF features for each k=5 K-Means cluster [35].

	Screen-on time	Event count	Switches	Categories
Cluster 1	229.6 (427.4)	748 (1997.6)	8.6 (11.3)	2.55 (1.4)
Cluster 2	114.7 (290.8)	161.8 (877.5)	4.8 (7.9)	2.06 (1.5)
Cluster 3	9.3 (10.5)	2.3 (3)	1.6 (1.7)	1.1 (0.3)
Cluster 4	227.7 (396.2)	10.4 (18.8)	0.9 (2.2)	0.5 (0.9)
Cluster 5	36.7 (127.9)	28 (163)	1.5 (2.8)	0.8 (1)

Table 3.3: Mean (SD) of the high-level features in each of the k=5 K-Means cluster [35].

clusters, (screen-on time: $H = 86111.4$, event count: $H = 147386.8$, app switches: $H = 86678.0$, category count: $H = 79491.7$, all $p < 0.01$), with Dunn posthoc tests also showing significance between all pairs for all features except cluster 4 and 5 for event count.

3.4.3 Comparing Clusters Against Session Features

The previous section detailed how low-level events can be used to cluster towards groups that reflect usage. But this does not yet show how they compare to high-level features of sessions commonly used in the literature to summarise usage.

3.4.3.1 The Role of Lock-state

A way to distinguish sessions is to consider the time it takes a user to unlock (the time from screen-on to unlock event) or if a user unlocks their phone at all [52]. A fast unlock could be considered to be attached to a different kind of interaction, user or session than a slow one. In our records, approximately a third of all sessions do not have an unlock event attached to them (95,432 compared to 162,259 sessions). While this indicates a decently sized split, when mapping the sessions across the impact data of the TF-IDF results it seems that unlock time actually plays a much smaller, or even

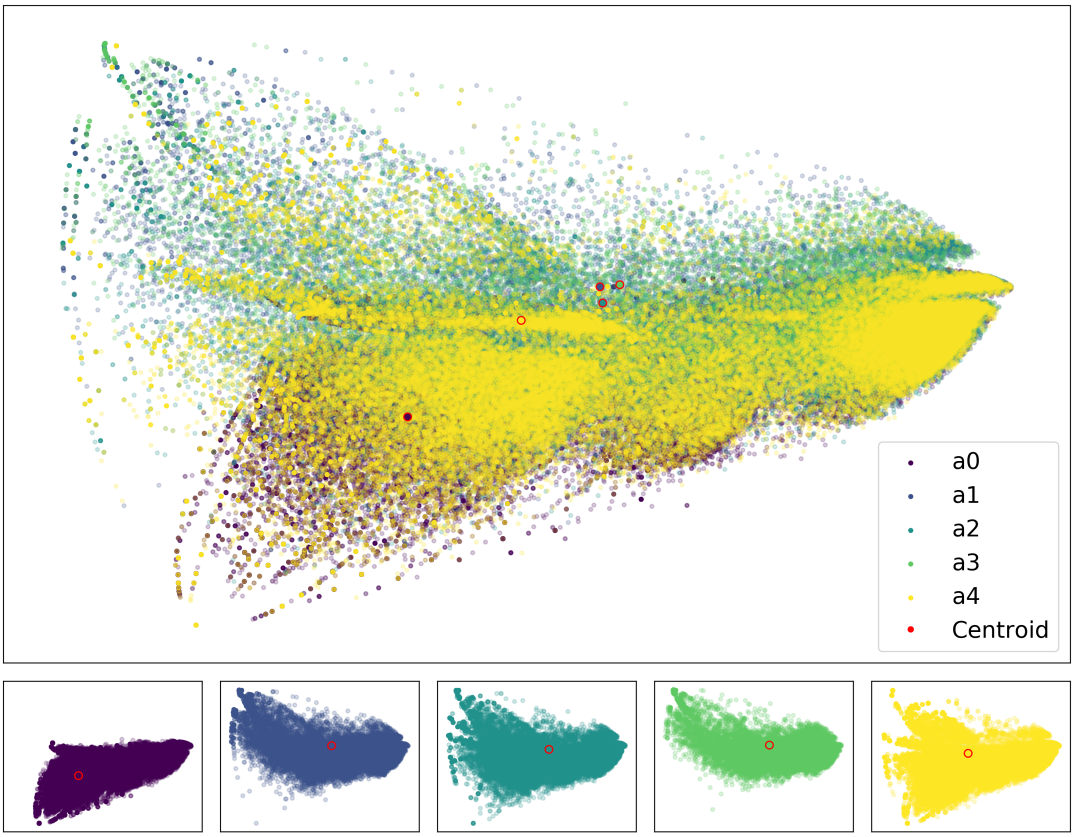


Figure 3.5: All sessions clustered by only using their high-level features (screen-on time, event count, unlock present, application switches and count of app categories) instead of TF-IDF features, plotted using PCA and showing the individual layers [35].

counterproductive role in grouping sessions. In the clusters the ratio of sessions having an unlock event are as follows in order: ~84.3%, ~71.7%, ~15%, ~62.8% and ~25.7%. In comparison, when sorting sessions by unlock time and then splitting them into 5 groups, 99.95% of all sessions and 100% of all sessions without an unlock end up in the first of those slices. Instead, the BFUS model shows that the unlock times are diverse in each cluster by applying a non-parametric pairwise comparison test. The Kruskal-Wallis H-test followed by Dunn posthoc comparisons is applied to the unlock state data, using the clusters as sample groups. A 100% null hypothesis rejection rate (i.e. none of the samples varies significantly from any of the others) with $H = 67097.67$

and $p < 0.01$ suggests that the lock-states are not from the same population for each cluster.

3.4.3.2 Comparing Against High-level Features

To show the additional utility of considering user-app interaction behaviour to characterize sessions, an alternative to how usage sessions could be grouped from the distributions of high-level features is examined. Firstly, Table 3.1 shows that the distributions of individual features have long tails with similar means and medians for most features. Splitting the distribution into group sessions using the range of the distribution results in most sessions being contained within a single group. For example, ~96% of sessions are placed within the same group for screen-on time. Equally, splitting the distributions into tertiles, quartiles, or quintiles results in a high degree of similarity between most groups of sessions. This suggests that additional granularity is necessary to capture notable characteristics of usage and that the high-level features are not a suitable proxy for user-app interaction behaviour.

To examine this further, Figure 3.6 shows how high-level features correlate to one another and the TF-IDF cluster each session is assigned to. Importantly, it shows that the TF-IDF clusters overlap and span across the distributions of high-level features, both individually and in pairs. This highlights that high-level features do not provide a suitable proxy for user-app interaction activity and that observing this granularity of behaviour is useful. This is demonstrated further by repeating the clustering process discussed using a vector of all high-level features to represent a session, rather than TF-IDF scores of the lower-level features. Figure 3.5 shows how the clusters created by high-level features overlap poorly with the TF-IDF clusters by fixing the individual sessions in the same position as Figure 3.4.

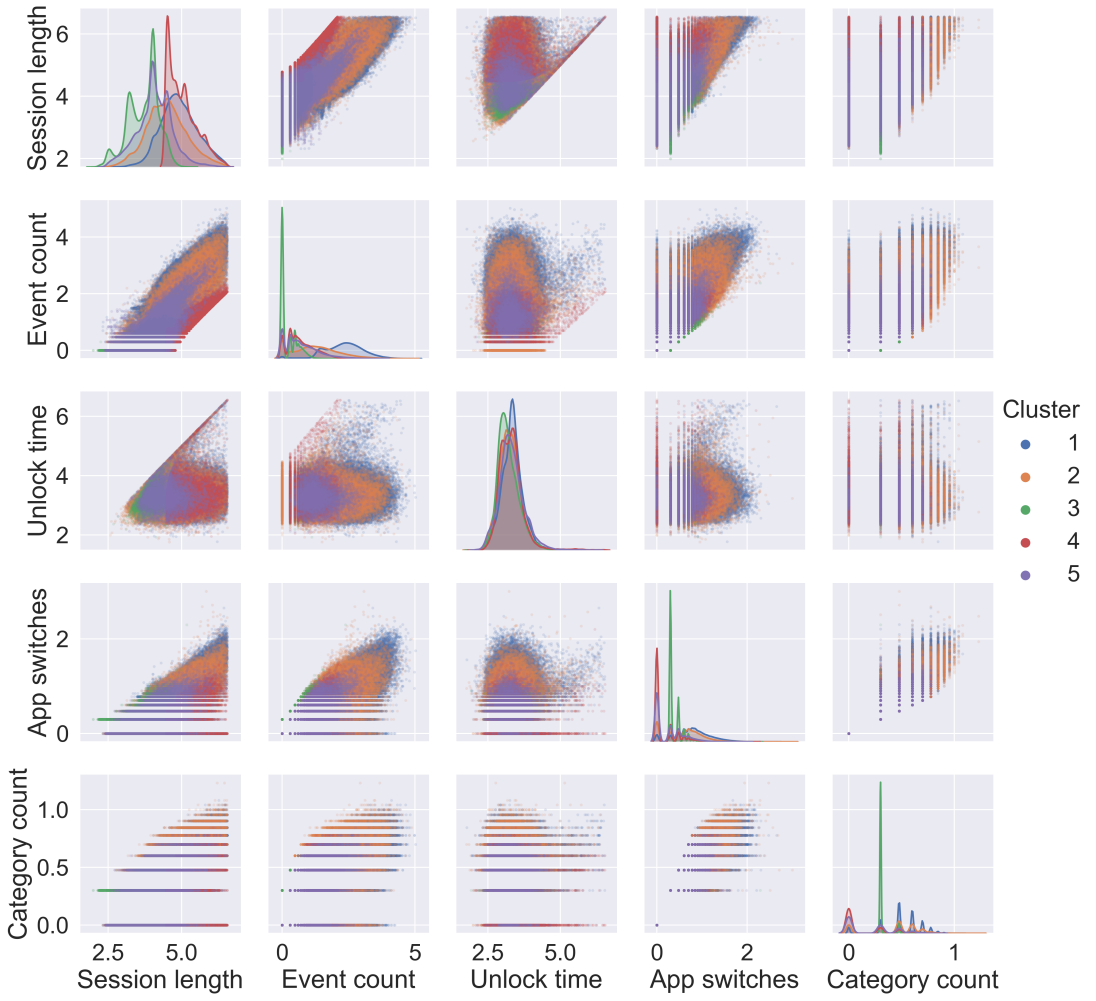


Figure 3.6: Pairwise comparison of correlations between high-level features of sessions and TF-IDF clusters (Log-Log) [35].

3.4.4 Discussion of Session Types

Applying the BFUS with K-means clustering shows that clusters with similar high-level features can be substantially different at the event level, indicating the diversity of smartphone usage. It avoids potential issues of bias from using raw counts by utilising TF-IDF to increase the impact of significant events. By analysing a rich dataset, which is unique in the level of user events that have been captured.

These results show that high-level features on their own are not sufficient to accurately

group sessions that are indicative of the user's cognitive goals. This is an important result, as previous analysis in the literature largely uses sessions defined by time between screen or application events [49, 110].

While it was possible to find an initial k via the elbow method it is unlikely that the 5 clusters in this section actually describe all types of usage that exist. There are likely more types with more nuanced differences hidden within these. Also the inherently spherical nature of K-means clusters means that unless session usage occurs in such an arrangement this form of clustering may not be ideal. Alternative methods of clustering such as DBScan or hierarchical clustering do not have the same limitation with the benefit of it being that they are able to detect clusters by distance between sessions by itself. However, those alternatives may also not be perfect. Firstly, DBScan is much more expensive to compute and comes with a much higher time investment because generally it requires its parameters to be tuned for accurate results. Even after tuning its parameters, it struggled to pick up more than one cluster in the variation of session usage. It is possible that this is due to sessions being arranged with generally similar features, which means that due to no strong separation between clusters the algorithm is not able to separate clusters properly. Sessions generally not being grouped neatly into clusters is also not unexpected because of the variability in usage overall. Therefore, DBScan seemed difficult to use for this case because it did not allow to capture multiple types of usage at all.

Another approach, hierarchical clustering, allows to separate clusters by utilising a distance metric (e.g., euclidian) instead of a fixed count for clusters. This method allows to choose closely related sessions to be included in a cluster with usage that produced similar features vectors. However, in this case it is very sensitive to changing the configured distance and it is possible to end up with thousands of clusters or only a few. Additionally, for both of these methods (DBScan and hierarchical clustering) it is required to tune parameters individually for different features (e.g., on-screen time has to be configured differently than low-level events) which seemed counterintuitive when

capturing naturally occurring types of usage. Therefore, the best method for analysis of cluster differences which does not require differential parameter tuning between high-level and low-level features was chosen to be K-means.

The results confirm that clusters based only on high-level features may misrepresent the commonality between smartphone sessions. The additional information present when capturing low-level events is a useful tool to infer more about a session beyond how active a user was. Five different types of use that would not be transparent with previous techniques are captured and can be described. In addition, these clusters represent usage that applies to almost all users, with 63 out of 64 users in the dataset showing at least one session in each cluster.

3.5 Conclusions

While the field surrounding the capture of user behaviour (particularly of smartphones) has recently seen advances in terms of utilising the vast amount of data available, a large part of the literature still utilises high-level features such as screen-on time. While these are often easy to collect and evaluate they create a set of limitations (Section 3.1) which restrict possible findings. Some branches such as digital phenotyping [14, 108] or the labelled "psychoinformatics" [97] have started to recognise the value that considering the lower level interactions can add to the understanding of usage.

In this chapter, current methods of usage characterisation were evaluated. Isolated features, and with that commonly used features such as screen-on time, were found to have inherent limitations (L1 and L2) when used for behaviour representation. This supplies initial explorations into the issues posed in RQ1 and adds towards the evidence of C1.

Additionally, a new framework called the Behaviour-From-Usage-Stream model is proposed which is designed to transform streams of usage events. The model lays the foundation for how user usage data can be processed in a way to extract further

knowledge while being flexible enough to enable changes that might occur with future data capture or advancements in vectorisation methods. By choosing the parameters of the steps within the model properly the limitations identified in Section 3.1 can be addressed. Section 3.3.3 discusses the balance needed for compression methods and how using TF-IDF for session vectorisation allows the inclusion of multiple features at the same time and alleviates issues of single features identified in L1. This builds the main contribution for C2 and is the foundation for the following contributions in the next chapters.

The initial utility of the model was validated by applying it to the Tymer dataset in which it was able to create clusters of usage. By applying the model with exactly the same steps for bounding sessions and application, the only change was introduced was the vectorisation of variables. This revealed clusters that were not previously detectable using commonly applied features such as screen-on time. This shows that user-interface interactions can be used as valid features for decoding usage as questioned in RQ2. Furthermore, since there seems to be information hidden in those interactions that were not detectable before this shows that the transformation of low-level features compared to summative features may address the skew issues (L2) which are common in session data.

In the following chapters, the model is further tested with respect to its effectiveness, specifically in relation to stable user traits as posed in RQ3. This includes further results for its application with correlation tasks using independent variables surrounding mental health. Additionally, the model's robustness is tested with different kinds of bounding and vectorisation methods.

Chapter 4

Isolated Features for User Classification

In the previous chapter, a new model for decoding the encapsulated usage information from interaction event streams was proposed, alongside a demonstration of its application to explore clusters of usage. This chapter continues the validation process of the model which will focus on the relationships between independent variables and the vector space created post 'Bounding' and 'Vectorisation' steps.

Observing correlations between usage behaviour and latent, psychological states has been proposed as a means for a better understanding of changes in mental state (e.g., [4, 156]) as well as potentially problematic behaviour while using smartphones (e.g., [104, 96]). However, thus far research has predominantly focused on summative metrics, rather than direct interactions as discussed in Section 2.1.3. The more specific behaviour shown in individual application categories instead of usage across all applications has also been identified to be relevant. Therefore the efficacy of isolated features will be tested within and without categories, this is facilitated by the flexible BFUS model since only the vectorisation step has to be adapted while bounding and application will remain identical. Thus, in this chapter the focus will be on examining the potential utility of using isolated features (i.e. summative *and* UI events) to motivate whether single-faceted features provide notable utility (e.g. scrolling as seen in previous studies), or whether multi-modal models should be considered.

As discussed in Section 4.2, the participants of the Tymer dataset were independently profiled in multiple aspects of their personality and regular usage. The results of the MCQ and SAS will be used to validate the effectiveness of the BFUS in relation to independent variables that are representative of impulsivity and smartphone addiction risk respectively. L3.1 in Section 3.1 has discussed the potential issues with using isolated features to distinguish user behaviour. To validate those claims, this chapter acts as a power analysis of isolated features in varying scenarios as part of the BFUS. This will be continued in Chapter 5 where the isolated features will be compared to a multi-modal transformation of usage behaviour.

4.1 Refining Event Selection

The Tymer dataset used in Section 3.4 includes events that are user instigated (e.g., taps, scrolls) and system-instigated (e.g., notifications, battery state changes). As the focus of this Chapter is on exploring links with latent independent variables, the features considered are reduced to user-instigated events.

Not all events in the data represent actual interactions invoked by the user. For example, notifications or the battery state (while intrinsically linked with the usage of a user) are not initially triggered by the user. Instead, they represent an external influence which might be controlled by other individuals, the device, or other scheduled operations. In turn, these events are removed from the vector of considered event types to maintain integrity with actual inputs.

Also, unlock events only occur at most once per session (apart from extremely rare cases where the phone was relocked via software without turning off the screen). This contributes only very little additional information and therefore these are also dismissed.

The choice of TF-IDF as a vectorisation technique prompted the choice of including a 30-second cut-off for pauses. This addition addressed some of the issues of the TF-IDF

compression by including a time metric to the vectorisation. To align further with previous literature this was optimized to a 45-second cut-off. This aligns with the cut-off between tasks as proposed by Van Berkel et al. to be notable for user behaviour on their smartphone [152]. While this bump does not represent a substantial change, it was amended to better represent these psychological boundaries between user goals. Following this and upon reviewing the statistics of usage sessions as presented in Section 3.1, 198,981 of 301,024 (66%) sessions did not exceed the 45-second mark. Thus, to embed the psychological boundaries at 45 seconds, but to also retain some form of temporal dependencies for shorter sessions and as an indicator for potential momentary hesitation a shorter idle timeout is added at a 1-second interval.

The final events considered for vectorisation are “Tap”, “Long tap”, “Text input”, “Application switches”, “Scrolling”, “Short Idle (1 second)”, “Long Idle (45 seconds)”. This corresponds to an adjusted event type vector of $T = \{t_1, ..., t_7\}$.

4.2 Assigning Trait Labels to Sessions

As part of the Tymer study, in addition to capturing the user’s low-level interactions, the participants were asked to complete mandatory briefings before and after the collection period. These briefings included the SAS and MCQ. The exact questions of the surveys can be found in Appendix E and Appendix F. Table 4.1 shows the distribution of responses for the participants for the briefing, debriefing and overall. The SAS briefing (M=26.31, SD=8.46) and debriefing (M=25.41, SD=7.37) are similar where the overall score is M=25.86, SD=7.55. One of the participants did not complete the debriefing session surveys so the before and after are not perfectly balanced. The log-scaled briefing (M=-5.29, SD=1.53) and debriefing (M=-5.33, SD=1.5) also didn’t differ greatly. Resulting in a geometric mean of M=-5.18, SD=1.44.

These labels are strictly captured per user, however, to analyse the impact of user traits on a session level, labels per session are required. Capturing labels per session would

Survey		M	SD	Min	Max
SAS-SV	Briefing	26.31	8.46	11	53
	Debriefing	25.41	7.37	15	48
	Mean	25.86	7.55	14.5	50.5
MCQ	Briefing	-5.29	1.53	-8.74	-1.39
	Debriefing*	-5.33	1.5	-8.74	-2.65
	G. Mean*	-5.18	1.44	-8.74	-2.06

Table 4.1: The results of the 64 Tymer participants for the SAS-SV and the log transformed k-values of the MCQ. Values marked with * were missing a single reading.

not only be difficult, but from a psychological standpoint incorrect. Firstly, the logistics of having to show the user a screen of dozens of questions every time they pick up their phone is not realistic. Presenting the user with that many interruption will with high certainty influence how they answer the questions (e.g., out of frustration or boredom). Being this invasive in the real usage of a users smartphone would likely skew the data to be unusable. More importantly though, user traits should be stable in each user, this means that these surveys which aim to codify the nature of a user should not change between multiple sessions. At the very least the results should not change over a small period of time. So while, not every single session from a user encodes the signals that would make them smartphone addicted or impulsive, it is the closest approximation possible to detect whether or not a session may have those signals. Therefore, every user session is labelled with the label of its corresponding user.

This still introduces a potential issue going forward, as in the first instance sessions will be evaluated and not the users directly. While this does constitute a limitation on the results, given the large amount of sessions per user and the traits stability in each user, the overall picture of all sessions combined for each user should still reflect their usage. It also means that some level of uncertainty is introduced for every individual session, but being aware of this uncertainty allows the further investigation

Class	Users	Sessions	M	Mdn	SD	Min	Max
Low (L_I)	15	83189	-7.33	-6.91	0.97	-8.74	-6.35
Medium (M_I)	41	160585	-5.03	-5.12	0.63	-5.99	-4.14
High (H_I)	8	57250	-2.81	-2.92	0.74	-3.56	-1,39

Table 4.2: Distribution of 64 users and 301,024 sessions across the reward discounting classes L_I, M_I, and H_I and their respective log-transformed MCQ results.

in Section 6.1 to find potential ways in isolating session that show stronger signals of addiction or impulsivity later on.

This means that on the scale of all sessions, the assignment of user labels for each session should reflect each user's trait correctly. However, it means that interpreting the results for every individual session should be handled with care, as the session itself might have none of the underlying patterns that constitute problematic use, even if it was generated by a user that is labelled addicted or impulsive.

Experiments 2.1 and 2.2 summary

Aim: Test the efficacy of isolated features by probing them individually to check if they could viably be used to detect impulsivity (Experiment 2.1) or SA (Experiment 2.2) in users.

Input sessions: All available sessions after pre-processing (as in Chapter 3, Experiment 1), also each session receives a class label based on their user's impulsivity or SA.

Features: High-level: Screen-on time, application switches and time to unlock Low-level: Event counts and TF-IDF vectors created from UI events. Each are tested without and within application categories.

Output: Effect size of the features capability to separate the sessions based on their classes of impulsivity and SA.

4.3 Impulsivity

Section 2.1.4.3 introduced the concepts that link smartphone use and impulsivity. While correlations have been made (e.g., [128]), actually inferring impulsivity from direct screen interactions has not been explored. Therefore, this section will focus on this by applying the BFUS model with an aim to add evidence to RQ3.

Firstly, using the MCQ survey scores of the start and end of the Tymer study (see Section 4.2), a paired t-test shows that the samples collected before (M=-5.29, SD=1.53) and after (M=-5.33, SD=1.5) the collection period did not significantly change (p=.77). Also, the results were correlated in themselves (r=.73, n=64, p$<$.001). The MCQ reading of one user at the end of the study was missing, therefore because the samples did not significantly differ before and after, from here onwards only the results of the initial briefing are used (for the MCQ).

A Shapiro-Wilk test for normality shows that the scores are not normally distributed (p$<$.001, W=.4). Therefore, the approach of [140] is emulated to create three reward discounting classes with one adjustment of creating tertiles with same distance cut-offs instead of equal-sized bins. Given the smallest and highest scores (-8.74 and -1.39) equal distance cut-off points are defined at -6.3 and -3.84. This creates the classes H_I (high) ($-3.84 < k$) M_I (medium) ($-6.3 < k <= -3.84$) and L_I (low) ($k <= -6.3$). Table 4.2 shows the distribution of users for each group.

4.3.1 Cross-category Feature Results

From the results of Chapter 3, surrounding the use of screen-on time, a hypothesis can be made that impulsivity will show a stronger effect on overall user behaviour than just screen-on time by itself. This is tested by applying a Kruskal-Wallis test followed by a Dunn's posthoc tests, which will identify whether or not the distributions of a feature

[2] AUC cells are coloured based on their performance as discussed in Section 2.3.2.1.

Feature	AUC	H	C1	Mdn	SD	C2	Mdn	SD
Screen-on	.575	5286	H_I	10.9	320.3	M_I	25.4	322.4
time	.552	5286	H_I	10.9	320.3	L_I	21.0	285.7
	.519	5286	M_I	25.4	322.4	L_I	21.0	285.7
App	.540	1641	M_I	3.0	14.2	L_I	2.0	14.9
switches	.521	1641	H_I	1.0	39.3	M_I	3.0	14.2
	.512	1641	H_I	1.0	39.3	L_I	2.0	14.9
Time to	.545	1389	M_I	1.88	77.2	L_I	2.12	76.4
unlock	.538	1389	H_I	2.16	47.7	M_I	1.88	77.2
	.502[x]	1389	H_I	2.16	47.7	L_I	2.12	76.4
Event	.568	4496	H_I	13.0	878.7	M_I	34.0	729.3
count	.535	4496	H_I	13.0	878.7	L_I	25.0	724.5
	.532	4496	M_I	34.0	729.3	L_I	25.0	724.5

[x] p>=0.05

Table 4.3: Pairwise Dunn's tests of screen-on time, app switches, time to unlock and event count between discounting classes.[2] H_I (8 users), M_I (41 users) and L_I (15 users) classes shown as C1 and C2. Screen-on time is in seconds. p<0.001 unless indicated otherwise.

significantly differ between any of the three impulsivity groups. All resulting p-values are Bonferroni corrected. This will provide a more substantial basis for the limitations discussed in Section 3.1 and add to the evidence for RQ1. The focus of summative features will be screen-on time as this is the most common metric used in previous literature, however additional, summative metrics will also be drawn from application switches and the total count of events that were generated.

A pairwise test between the groups for each feature enables to target each feature in an isolated context. Of particular interest is the effect size of every individual test, as this gives an indication as to how strongly the isolated feature can separate the

samples. A pairwise test enables this with a low computational overhead for dozens to hundreds of features at a time to check the effect between two groups. While a regression would be the usual method for comparison of multiple features the focus in this case lies in testing whether it would be feasible to collect and evaluate single, isolated features instead of a classic feature ranking. Additionally, pairwise tests offer a helpful metric to compare the features across different models via a comparable effect. Extracting effect from a linear regression would usually be possible by evaluating the coefficients between each other. This gives an indication of the strength of each feature additionally to its significance. However, this only works for coefficients of the same model, for comparable effect sizes (i.e., effect sizes between models with different features) it requires retraining of the model for each feature individually and extracting the effect size of the entire model. This will be relevant because application categories are introduced as part of the vectorisation process and a comparison is drawn between features isolated within and outside of application categories.[3]

Table 4.3 shows how the features perform in each pairing with another group. All results show that the separation between groups is statistically significant with $p<0.001$. Users with high impulsivity have the lowest average screen-on time by some distance compared to users in L_I or M_I. This could be reflective of shorter bursts of usage in highly impulsive individuals. Screen-on time has its highest separability between H_I and M_I, where H_I-L_I and M_I-L_I have very similar results. However, these are very low on the AUC scale (<0.6) and therefore provide limited confidence of predictive power for this metric between groups. Similar results can be observed for the count of events, where high impulsivity users generate the least amount of interaction. Application switches and time to unlock, while fairly similarly distributed across all pairs, show that all pairs also have a very low effect size. Notably, in all instances does the highest AUC occur between M_I and another class rather than what may be expected at the 'extremes' between L_I and H_I.

[3] A predictive approach follows in Chapter 5 were these features are then used within a regression.

Count	**AUC**	**H**	**C1**	**Mdn**	**SD**	**C2**	**Mdn**	**SD**
Short idle	0.564	3776	H_I	10.0	713.7	M_I	20.0	535.1
App switch	0.559	2394	H_I	3.00	10.7	M_I	4.00	9.37
App switch	0.554	2394	H_I	3.00	10.7	L_I	4.00	9.64
TF-IDF	**AUC**	**H**	**C1**	**Mdn**	**SD**	**C2**	**Mdn**	**SD**
App switch	0.580	3851	H_I	0.40	0.21	M_I	0.36	0.17
App switch	0.559	3851	H_I	0.40	0.21	L_I	0.37	0.17
Single tap	0.553	1307	H_I	0.39	0.16	M_I	0.34	0.16

Table 4.4: Pairwise Dunn's tests of event count and TF-IDF weights between discounting classes showing the top 3 strongest effect sizes out of 21. H_I (8 users), M_I (41 users) and L_I (15 users) classes shown as C1 and C2. p<0.001 for all results.

Between all summative features, this shows that while these tests are statistically significant, screen-on time and event count produce a marginally stronger effect than app switches and time to unlock. However, there is little support to deem them effective in separating the different levels of impulsive use between groups because of the poor AUC scores (<0.6) for all of them.

It could be argued that the summative nature of these features obscures too much information about the underlying usage behaviour. Table 4.4 presents the same analysis of features for each interaction event type (defined in Section 4.1), as counts and TF-IDF scores. Alongside the combined result of all pairs overall, only the pairs of the three strongest effect sizes are shown. This enables a clearer comparison and a brief overview of the relative strength of each approach without the noisiness of every single result individually.

In comparison to summative features, there is not much difference in the results of isolated UI events. For event count features, events show an overall median AUC of .524 (SD=.019) and the TF-IDF single event features do not show many different results

Screen-on time								
Category	**AUC**	**H**	**C1**	**Mdn**	**SD**	**C2**	**Mdn**	**SD**
Sports	0.781	122	$H_I(2)$	2.00	10.8	$L_I(3)$	30.6	89.9
Finance	0.746	211	$H_I(3)$	4.69	29.5	$M_I(18)$	34.1	100.8
Sports	0.703	122	$H_I(2)$	2.00	10.8	$M_I(4)$	20.4	110.6
App switches								
Category	**AUC**	**H**	**C1**	**Mdn**	**SD**	**C2**	**Mdn**	**SD**
Sports	.654	43	$H_I(2)$	2.0	2.04	$L_I(3)$	1.0	1.16
Sports	.645	43	$H_I(2)$	2.0	2.04	$M_I(4)$	1.0	.76
Productivity	.620	623	$H_I(7)$	2.0	3.73	$M_I(39)$	1.0	1.55
Event count								
Category	**AUC**	**H**	**C1**	**Mdn**	**SD**	**C2**	**Mdn**	**SD**
Weather	.745	98	$M_I(5)$	31.0	76.7	$L_I(4)$	5.0	41.8
Tools	.706	3364	$M_I(41)$	1.0	128.7	$L_I(15)$	8.0	141.2
Sports	.664	35	$H_I(2)$	13.0	52.7	$L_I(3)$	49.0	218.4

Table 4.5: Pairwise Dunn's tests of top 3 of screen-on time, app switches and event count when taking app categories into account. H_I, M_I and L_I classes shown as C1 and C2, count of representing users in brackets. All p-values are Bonferroni corrected and $<$0.001.

with an average AUC of .523 (SD=.022). While TF-IDF's *App switches* (AUC=.580) do show the highest overall effect size it still is very close to any other result. This indicates that when any metrics are considered across a whole session, the nuances that make up behaviour might get squashed and that neither summative nor UI events as features are able to distinguish groups of impulsivity effectively. Once again, features compared to M_I actually show the strongest effect, instead of L_I-H_I. This motivates exploring more granular levels of usage, such as previous findings surrounding application categories.

Count									
Feat[a]	**Category**	**AUC**	**H**	**C1**	**Mdn**	**SD**	**C2**	**Mdn**	**SD**
ST**	Trivia[b]	0.790	7	$H_I(1)$	94.5	135.1	$M_I(4)$	13.0	28.6
TI**	Photography	0.763	8	$H_I(3)$	9.50	13.4	$M_I(6)$	3.00	4.02
ST*	Casual[b]	0.756	10	$H_I(3)$	6.00	3.13	$M_I(5)$	1.00	1.27
TF-IDF									
Feat[a]	**Category**	**AUC**	**H**	**C1**	**Mdn**	**SD**	**C2**	**Mdn**	**SD**
ST*	Trivia[b]	0.891	13	$H_I(1)$	0.52	0.04	$M_I(4)$	0.34	0.08
LI**	Weather	0.872	10	$H_I(1)$	0.14	0.01	$L_I(1)$	0.45	0.12
AS	Sports	0.766	85	$H_I(2)$	0.41	0.17	$L_I(3)$	0.19	0.11

[a] SC=Scrolling, AS=App Switch, ST=Single Tap, LT=Long Tap, TI=Text Input, SI=Short Idle, LI=Long Idle

[b] Game category

* $p<0.01$, ** $p<0.05$

Table 4.6: Pairwise Dunn's tests of top 3 of count and TF-IDF when taking app categories into account. H_I, M_I and L_I classes shown as C1 and C2, count of representing users in brackets. All p-values are Bonferroni corrected and $<$0.001, unless otherwise indicated.

4.3.2 Considering Usage within App Categories

The literature has shown that usage within particular application categories can add more insight into the structures behind behaviour and this motivates exploring whether examining interactions in specific categories may be better than entire sessions. In this section, the previous results are repeated with the addition of the application categories of where the specific features occurred. As the BFUS model is designed to be flexible nothing about the bounding or application stages have to change. Instead the only part that is updated is the vectorisation that will now take place with categories included. This way it is possible to provide a stable comparison between usage with and without categories.

The categories are determined from the Google Play Store based on application identifiers. Using the application switch event it is possible to identify the relevant application of each interaction. From this, we retrieve the category for each of those applications and combine them. All apps which were not available on the Google Play Store were placed in an *Other* category, resulting in a total of 45 categories (see Appendix C).

From this, given the set of all event types $T = \{e_1, ..., e_7\}$ and the set of all categories $C = \{c_1, ..., c_{45}\}$, a feature combination vector $f_{c_{315}d}$ can be constructed based on $E \times C$, containing information on each event type for each category. After constructing the final vector it was reduced from 315 to 278 features because 37 combinations of events and app categories which did not occur in the dataset were removed, this was done to not confuse completely empty vectors with those of low occurrence. This decision was made since unknown interactions may not necessarily be equal to deliberate non-interactions.

Given the 45 categories identified from the App Store, for every summative feature where previously 3 Dunn's comparison pairs existed, now there are $3 * 45 = 135$ and for UI event based features (7) there are $3 * 7 * 45 = 945$ individual tests. As mentioned previously, to address some of the issues introduced by this amount of multiple tests, the results in the following section contain Bonferroni corrected p-values.

The results in Table 4.5 show an improvement over the previous tests which did not include category information. For in-category screen-on time (AUC: M=.577, SD=.057) the *sports* category shows a higher effect size between H_I and L_I than any of the previous tests. This constitutes an increase in effect size of $>$20% compared to the highest effect size observed for screen-on time with all categories included. Similar, if not slightly worse, results can be observed for event count (AUC: M=.570, SD=.053) even for its highest AUC=.745.

Short bursts of interactions being normal for smartphone sessions [111] translates to only a few application switches per session for most sessions. This is even more no-

ticeable once those events are further split between multiple categories. Application switches did not show similarly strong separability between groups overall (AUC: M=.548, SD=.036) and also did not peak as high with its highest AUC being .654 between H_I and L_I in the *sports* category.

In comparison, Table 4.6 shows the top 3 statistically significant feature pairs of interface interaction event count and TF-IDF. Just like summative features, the top features for these have large improvements in effect size. For example, when counting individual events the users with high impulsivity tapped more frequently in *trivia games* than those with medium levels. Comparatively, the top feature for TF-IDF has the highest effect sizes so far, firstly also for taps in *trivia games* (AUC=.891) between H_I and M_I but also for *long idle* events in *weather* applications between H_I and L_I. However, overall the effect sizes for both of these transformation methods do not show the same large improvements. Count has an average effect size of AUC=.566 (SD=.051) and TF-IDF reports AUC=.579 (SD=.062).

In all of these category-specific tests an observation about user representation can be made. These effect sizes are all backed by a low sample size since not every user will have used applications from each category. Moreover, once the events get more specific (resulting in larger vectors), less frequent events such as application switches or long taps are only backed by a handful of users each. For example, only 9 users collectively used sports apps (2, 4 and 3 users respectively for H_I, M_I and L_I). While similar results can be observed for the remaining pairs some categories were used by almost all users.

This shows that there might be behaviours encoded in smaller, more specific groups of usage which causes an interplay with high effect size for lower user representation and on the other hand low effect size with higher user representation. Considering this trade-off between effect size and user representation means that a process which evaluates user behaviour should be designed to balance the specificity and summativeness of its features.

4.3.3 Discussion

To recapitulate, the overarching goal of this section has been to determine the efficacy of different summarisations of usage using the the BFUS in correlating with a user's impulsivity score. In this, previous methods of usage capture, such as summative features like screen-on time are evaluated along with features representing user interactions from the BFUS.

For the previous discussions surrounding (isolated) summative features (e.g., screen-on time), their results match expectation. All tests (except time to unlock H_I-L_I) being statistically significant at $p<0.001$ matches with previous findings in the literature that these measures are correlating. However, between screen-on time, app switches, time to unlock and event count no features effect size exceeded an acceptable range for a strong effect. While the average of the features between classes does show differences of sometimes more than double (e.g. screen-on time H_I=10.9 and M_I=25.4 seconds) the high standard deviation in each feature causes too much variability which affects the effect size. This already offers some insight towards RQ1 as it seems like summative features likely do not encode enough information to separate user impulsivity by itself.

The outcome of repeating the tests with low-level features identified in Section 4.1 is two-fold. It addresses whether low-level features in isolated form offer any improvements over summative features as could be expected following Section 3.4 and secondly it will contribute to verifying L3.1. In the test results, it is clear that isolated low-level features (count or TF-IDF) do not improve on the summative features in any significant way. This confirms the previous assumptions that isolating single features and stretching their range across entire sessions just compresses too much information to capture the markers of different levels of impulsivity in a user.

The previous tests are repeated by emulating previous literature where the correlation between user traits and smartphone use often existed more prominently when isolating specific application categories [104, 135, 26]. When inspecting behaviour bound

to application categories, both summative features and UI events show much better effect sizes but smaller representation of users per feature compared to general use. Summative features see their best result for screen-on time at an AUC of .781, while UI events that are TF-IDF transformed go as far as .891. These are decent to great results according to the AUC scale, but they come with the drawback of having very low representation among users. This also shows that there is a difference between the samples of sessions as inferred from their user labels if they are broken down with sufficient detail. While there will still be some uncertainty in the ground truth of the labels, this shows that the sessions produced by the user can be separated based on their assigned label.

However, not every user has used (or will ever use) every application category or will take every possible action (e.g. long taps) in every category. Some exceptions exist where almost all users were represented by specific category combinations, for example almost all users switched to productivity application at some point or another where an AUC of .620 is still better than any of the non-category based results. As a conclusion to category features, the average AUC for all category results are very similar to those of isolated features ($<.6$). This indicates that while these features can be a better estimator, they do not generalise to all users.

Across all results of pairwise tests it was possible to observe a phenomenon which may be considered unexpected. In many of these cases the effect size between the two groups that may be expected to show the furthest spread (L_I and H_I) actually did not show the strongest effect. Often the effect between one of the groups and the M_I class were the strongest. This further indicates that separation on just isolated features alone might be problematic because the traces in actual usage do not correspond linearly with the level of impulsivity.

In order to add depth to the investigations on usage behaviour against latent independent variables, a second variable, smartphone addiction is explored.

Class	Users	Sessions	M	Mdn	SD	Min	Max
Not addicted (N_{SA})	51	221670	23.11	23	5.00	14.5	32
Addicted (A_{SA})	13	79354	38.61	25.5	5.69	32.5	50.5

Table 4.7: Distribution of 64 users and 301,024 sessions across addicted and non-addicted users according to [72] and their respective SAS results.

4.4 Smartphone Addiction

Similar to impulsivity, SA has been connected to isolated features such as time spent in applications [124] or application changes in fragmented use [25]. To validate our findings against another potential user trait that has an effect on user behaviour we repeat the analysis steps of the previous section. It will follow the same general structure of Section 4.3 where the effectiveness of isolated features is evaluated to separate the user groups. The difference for SA is that users are only divided into two instead of three groups. These groups are A_{SA} (addicted) and N_{SA} (non-addicted).

At the start and end of the Tymer study, participants were asked to complete the SAS survey (see Section 4.2). The results of the SAS collected before (M=26.31, SD=8.46) and after (M=25.41, SD=7.37) the data collection period did not significantly differ (p=.14). Additionally, the samples were highly correlated with each other (r=.82, n=64, $p<.001$). In this instance, since all data were available, the results for each user were combined and the mean of their answers was used to label them addicted or not addicted. The cut-off points for SA based on a user's SAS score are already well defined by [72], and these were used as the basis for partitioning users. This decision and its inherent, potential uncertainty is discussed in more detail in Section 6.1 and Section 5.6. As Table 4.7 shows 51 user belong to N_{SA} (M=23.11, SD=5) and 13 users were identified as addicted in A_{SA} (M=25.5, SD=5.69).

			Addicted		Non addicted	
Feature	**AUC**	**U**	**Mdn**	**SD**	**Mdn**	**SD**
Screen-on time	.505	8686911416	20.0	324.1	20.4	276.7
App switches	.536	9513955856	2.0	14.6	2.0	34.1
Time to unlock	.520	1833312400	2.01	78.3	1.85	51.8
Event count[x]	.501	8814163664	28.0	725.6	26.0	844.9

[x] p>=0.05

Table 4.8: MWU tests of screen-on time, app switches, time to unlock and event count between A_{SA} (13 users) and N_{SA} (51 users). Screen-on time is in seconds. p<0.001 unless indicated otherwise.

4.4.1 Cross-category Feature Results

In this section, the analysis of isolated and high-level features is expanded to groups separated by SA (addicted: A_{SA}, not-addicted: N_{SA}), complementing the analysis undertaken for impulsivity in Section 4.3.1. Since the label is binary, a Mann-Whitney U test is more applicable for SA instead of a Kruskal-Wallis and Dunn's test, but as they are functionally identical the same metrics are reported.

Table 4.8 shows that the average high-level features did not deviate a lot between the A_{SA} and N_{SA} groups. The low effect size for all of these statistically significant summative features mirrors the results of impulsivity (AUC<0.6). The high standard deviation of all features compared to their averages likely means that their data is too irregular to separate them.

Given the assumption about the loss of nuance in session-compressed features, this should also be the case for event count and TF-IDF transformations of UI events. Table 4.9 shows that while significant, the effect size measured by the AUC (M=.531, SD=.021) is low for all individual event types. In comparison to high-level features, most features show that medians are also very similar between the classes. The only

Count			Addicted		Non addicted	
Feature	**AUC**	**U**	**Mdn**	**SD**	**Mdn**	**SD**
Text input	.573	492978166	10.0	144.5	22.0	171.5
Long idle	.546	197616102	2.0	18.1	2.0	15.5
Long tap	.535	1170076	1.0	2.21	1.0	1.14
Scrolling	.529	1169988738	5.0	436.7	6.0	659.3
Single tap	.524	1657905034	3.0	27.4	3.0	27.5
App switch	.509	3170474236	4.0	9.66	4.0	9.91
Short idle	.506	8335055119	16.0	538.5	16.0	617.0
TF-IDF			Addicted		Non addicted	
Feature	**AUC**	**U**	**Mdn**	**SD**	**Mdn**	**SD**
Short idle	.560	7293223043	.66	.26	.83	.26
Scrolling	.548	1111583024	.42	.18	.46	.18
Single tap	.545	1742656121	.36	.17	.33	.16
Long tap*	.532	1028618	.36	.18	.38	.18
Text input	.525	567149570	.61	.18	.64	.19
App switch	.521	3406799519	.38	.18	.36	.17
Long idle	.519	186302310	.46	.23	.42	.24

* $p<0.01$

Table 4.9: MWU tests of event count and TF-IDF weights between A_{SA} (13 users) and N_{SA} (51 users). $p<0.001$ unless indicated otherwise.

exception to this is text input events where non-addicted users produced over double (Mdn=22, SD=171.5) the amount of text input events than users in A_{SA} (Mdn=10, SD=144.5). The results for TF-IDF scores show equally low effect sizes to counts (M=.536, SD=.014) individually and overall. The only noticeable change in medians between the classes is short idles where the value for non-addicted users is almost 20% higher. Notably, the highest AUC for either method was achieved by the feature that

Screen-on time			Addicted			Non addicted		
Category	**AUC**	**U**	**N**	**Mdn**	**SD**	**N**	**Mdn**	**SD**
Education	.692	13399	4	15.7	147.8	16	70.8	129.2
Sports	.689	38236	2	23.5	84.8	7	3.71	102.0
Tools	.688	98492634	13	1.6	118.7	51	8.86	176.6
App switches			Addicted			Non addicted		
Category	**AUC**	**U**	**N**	**Mdn**	**SD**	**N**	**Mdn**	**SD**
Sports	.756	7610	2	1.0	.71	7	2.0	2.05
Simulation[a]	.731	13485	1	2.0	2.84	3	1.0	.42
Productivity	.619	7666702	12	1.0	2.04	49	2.0	3.23
Event count			Addicted			Non addicted		
Category	**AUC**	**U**	**N**	**Mdn**	**SD**	**N**	**Mdn**	**SD**
Simulation[a]	.799	15364	1	42.0	80.6	3	5.0	25.1
Trivia[a**]	.686	94	4	345.0	303.2	5	826.0	2928.9
Tools	.661	18961633	13	2.0	116.3	51	11.0	212.4

[a] Game category

[*] $p<0.01$, ** $p<0.05$

Table 4.10: Results of an MWU test for the top ten features (count of each event type in an app category) with the highest effect sizes. N refers to the count of users in each group. $p<0.001$ for all unless indicated differently

showed the largest gap between medians, exceeding those of high-level features.

These findings indicate that considering some UI events may provide slightly stronger predictive power in comparison to overall screen-on time, count of application switches, and the overall count of UI events. In general, effect sizes are low for every isolated, non-category feature, which is similar to those findings of impulsivity in Section 4.3.1. The next focus is on these features when considered within specific categories of apps as behaviour is expected to retain more of its nuance.

Count				Addicted			Non addicted		
Feature	**Category**	**AUC**	**U**	**N**	**Mdn**	**SD**	**N**	**Mdn**	**SD**
Scrolling*	Trivia[a]	.867	1	3	1.5	.50	3	19.0	20.6
Single tap*	Trivia[a]	.809	10	4	3.0	4.5	3	41.0	187.7
Long idle	Simulation[b]	.794	8866	1	10.0	25.9	3	3.0	4.04
TF-IDF				Addicted			Non addicted		
Feature	**Category**	**AUC**	**U**	**N**	**Mdn**	**SD**	**N**	**Mdn**	**SD**
Scrolling*	Trivia[a]	.877	0	3	.17	.02	3	.39	.08
Text input	Education	.873	1155	2	.28	.12	9	.54	.13
Single tap*	Trivia[a]	.832	7	4	.23	.08	3	.42	.10

[a] Game category

* $p<0.01$

Table 4.11: Results of MWU tests of the top three features (count of each event and TF-IDF score) with the highest effect sizes. N refers to the count of users in each group. $p<0.001$ for all unless indicated differently.

4.4.2 Considering Usage within App Categories

Table 4.10 shows how the results change when considering time spent in specific categories. Some pairs of high-level features and categories show significant boosts in effect size. App switches of the *sports* category and event count in *simulation* applications both have AUCs $>.75$ which is approaching good results in terms of separability. While none of the screen-on time effect sizes exceed 0.7 for individual tests, it shows the highest average effect size (M=.59, SD=.567) compared to app switches (M=.571, SD=.063) and event count (M=.582, SD=.07). It should also be noted that just as before with impulsivity, these category-feature pairs under-represent the total sample since only 9 and 4 users respectively generated events in those categories.

Moving from summative features to UI-based low-level features, the first observation is an increase in overall effect size (M=.587, SD=.073) for application-enhanced counts

compared to general counts (the best feature was text inputs with AUC=.573). The best features outperform summative features where, for example, scrolling in *trivia game* apps, with an AUC close to 0.9, shows a very strong effect size. When employing TF-IDF vectorisation the effect size between all tests increases to the best overall score slightly above screen-on time (M=.591, SD=.077). Additionally, individual tests also show some of the highest effect sizes. Both, scrolling and single taps, in *trivia games* reappear in the top three for TF-IDF compared to counting. *Education* applications were the strongest feature for screen-on time and now reappear in the top three of TF-IDF features when focusing on text input events.

This builds upon the evidence for low-level features encoding information more effectively, especially when transformed through TF-IDF. This also further motivates addressing the issues of low user representation through a combination of those improved features.

4.4.3 Discussion

The results of isolated features separating the classes A_{SA} and N_{SA} show similar results to those observed for impulsivity. Even though mostly statistically significant, in terms of effect size summative and UI event features perform poorly by themselves. None of those effect sizes exceeds an AUC of 0.6 which in practical terms means that the distributions are almost indistinguishable when trying to separate them using those metrics. Overall, the results of isolated features are unlikely to be suitable for the goal of identifying SA.

When features are enhanced through application category data the overall effect sizes improve. However, the gains overall are usually minimal (e.g. screen-on time gains only a small percentage in terms of improvement) whereas when focusing on the very top selection of events there are some more distinct differences. This suggests that when all data points apart from a single category-feature combination are dismissed,

better results can be achieved for those sessions that remained because there is more specific usage data encoded in them. The caveat is that as only sessions are used where those combinations actually occurred, many sessions are lost for evaluation. This is also reflected by two factors: The test statistic (U value) is very low for some of the results. While low values are positive because they represent the proportion of potential false guesses when viewed in context to the U values of isolated features it shows that there were a lot fewer sessions to evaluate from overall.

Secondly, the results present interesting distinctions of notable features while considering categories. However, the samples are under-representing the actual usage behaviour of all users where, for the most part, less than a dozen of users are contributing to any single feature-category pair. This typically low 'N' creates similar limitations to impulsivity and suggests considering a balance of summative and low-level features or multi-modal models and motivates their exploration.

Shifting the focus to the results of low-level UI events, the differing results between the top features of count and TF-IDF can most likely be accounted to the scaling that takes place during the TF-IDF vectorisation. For isolated features, TF-IDF's ability to distinguish nuances in usage might be compromised by the very low vocabulary of only 7 features (i.e. event types). Once categories are introduced the TF-IDF scaling becomes more prevalent and improves the results of the pure count. Apart from the higher effect sizes for the top few features, overall they still perform similarly.

4.5 Conclusion

This chapter explored multiple aspects of isolated, summative and UI event features when applied via the BFUS model to evaluate user traits: impulsivity (measured by the MCQ delay discounting task) and smartphone addiction (measured by the SAS). In particular, this highlighted the previously discovered limitations of usage behaviour (L3.1 and L3.1) since isolated, summative features (but also isolated features overall)

were not effective in distinguishing user groups. This adds evidence for C1.

Additionally, in a novel approach, it adapted an NLP technique in the form of TF-IDF to show how it can improve the effect of identifying user traits through smartphone behaviour. This extends previous literature which used low-level events but without the structured nature of bounded sessions or TF-IDF transformations [104], affirming C2.

In practical terms, it would not be possible to actually separate users that way. This is further highlighted by instances of the impulsivity multi-class problem where cut-off points on the response variable are placed out of order, i.e. for binning purposes the expectation would be that the response variable shows a relationship in order so that as it increases or decreases it corresponds to the classes such that low→medium→high (or high→medium→low) impulsivity is present. Instead, there were instances of non-linear relationships between response variables and classes such that as the response variable increased the classes actually went through the transition of high→low→medium impulsivity (see *Event count* in Table 4.3).

Introducing application categories boosted the results so that the top pairs of features and categories performed significantly better than any non-category feature. The strongest feature peaked at an AUC of .877 which could be considered as strong. However, the splitting of sessions into many individual grouping causes the support for every test to be much lower than any of the previous tests. This means that not all users are represented by that pair which lessens confidence in their practical efficacy.

At this point, another important statistical observation has to be made. While all tests were Bonferroni corrected, conducting this amount of statistical tests at once is arguably sub-optimal to warrant any individual interpretation of results in detail. In fact, all comparisons in this chapter should be considered in unison with all other results in their respective group and not by themselves. Any individual test, especially if already close to the significance threshold of .05, needs to be carefully evaluated if to be used in a psychological profile for impulsivity or smartphone addiction. Similarly, as dis-

cussed in Section 4.2, impulsivity and addiction labels were applied to every session from the label that was created for each user. This means that these results represent the usage of user trait classes overall, but potentially not on an individual session level.

This motivates an approach which considers all features at the same time to include all session data. The higher performance of the TF-IDF results already considers the co-existence of events among others, and this motivates that the interplay of multiple features may produce strong results when applied to a trained model which then leads to C3. This can be achieved via a logistic regression and forms the focus of the next chapter.

Chapter 5

Extracting User Traits Embedded in Complex Behaviour

Following the approach of Chapter 4 gives an opportunity to explore the utility of multi-modal models to address the limitations of isolated feature models. For this, likewise to Chapter 4, independent models will be built for SA and impulsivity. In addition to checking the viability of any single feature to discern samples of classes as posed in RQ1, this chapter aims to harness the improved magnitude of effect while still maintaining the support of all users, which can be tested by training multi-modal logistic regression models capable of predicting a user's class (such as impulsivity or SA) in accordance with RQ3. It is expected that using high-level features will exhibit some form of mischaracterisation while identifying usage behaviours because of compression and aggregation. This means models built from low-level features should show comparable to better accuracy at predicting the user labels from their behaviour in sessions.

In the following sections, regression models are built from the category features introduced in Chapter 4. This methodology was constructed assuming that being able to utilise all combinations will be more accurate (as per the improved peaks of AUC) but it will also take into account all available interaction data instead of focusing on one specific pair.

This also avoids the issue of potentially misrepresenting a characteristic of this dataset

as generally applicable, which could be possible considering the number of multiple tests where some features could show correlations by chance even with a standard p-value threshold of .05. Table 4.6 shows multiple instances of the p-value being close to this threshold. However, given the precision of the other p-values ($p<0.001$ for many) that are reported, the chance of all of them showing false significance is low.

The regression models predict the probability of a user's session falling into any of the user trait classes. This has multiple benefits compared to the approach of single features. With a regression, all sessions from all users can be utilised, just like in the isolated feature comparison cases of the Kruskal-Wallis or Mann-Whitney analysis. Additionally, all of the possible combinations of UI events and categories can be utilised at the same time, which as displayed in Table 4.5 and Table 4.6 comes with a drastic improvement in effect size when inspecting their isolated cases.

5.1 Regression Preparation

Multiple multi-modal regression models are going to train on and test the users' trait classes based on a vector of input features (as extracted from the second vectorisation step of BFUS). These regression models are going to output a probability which corresponds to one of the classes for every single a session of a user. A standard 10-fold cross-validation at a 90/10 train-test split is used for all session data. Since the features are transformed using PCA to remove issues during the regression, the influence of specific features will not be discussed in detail as they are obscured and do not hold relevant information to user behaviour any longer. Chapter 6 will explore and discuss the benefits and drawbacks of an adjacent methodology which allows to maintain non-transformed session features and offers potential psychological interpretations beyond a predictive application.

To utilise all of the information present in the feature vectors while reducing the amount of noise, the features are compressed to a more tractable set by employing principal

component analysis (PCA). The noise from so many features is hard to avoid if the category data for every single event is to be retained. While it would be possible to only inspect one category at a time, this would then not include the data of each category for each event in case multiple different categories were used between sessions (this is also even more important for the application with TF-IDF which requires the entire corpus to be present). Configured so that a variance of at least 99% is required to explain each component, this transforms the raw feature vector to a vector with fewer total components. While PCA obscures the individual influences of the input features, it enables the same method of variable transformation for every input vector (screen-on time, summative features, event count and TF-IDF) without having to conduct an individual feature extraction. By observing the receiver-operator-characteristic (ROC) of these classes, the AUC for each class and their average can be extracted. Similar to the Kruskal-Wallis and Mann-Whitney tests, the AUC of each class will give an indication of how discernible any session in the individual class is compared to all other classes. This also established a comparable metric between those tests as an addition to the standard classification metrics such as precision, recall and accuracy.

Because there are fewer addicted (13) than non-addicted (51) users, they also produced vastly different amounts of sessions (79,354 and 221,670 respectively). To account for this skew of volume in the input the regression is trained with balanced training weights. The balance weight is calculated as follows: Let *N* be the sample size of training sessions ($N = 301024 \times 0.9 = 270,922$) and *F* the number of features (98 after PCA). When *K* is the count of available classes, let $y_i \in 1, ..., K$ be the label of the target variable for observation i and c_{yi} the count of observations in the current sample of *N*.

$$\frac{N}{F \times c_{yi}} = \frac{270,922}{98 \times c_{yi}} \tag{5.1}$$

Given a well-shuffled set of input features, every iteration of the following 10-fold cross-validation should contain enough data for each value of y_x to be stable.

The logistic regression calculates the probabilities for each class for every session in

the test data. To then get a result for each of the users, the probabilities of all their sessions are combined by taking the geometric mean of probabilities for each class. The class with the highest of those means is subsequently chosen as the final class for each user.

Let W be a matrix of coefficients where each row W_k represents one class k. Then the regression predicts the class probabilities $P(y_i = k|X_i)$ as[1]:

$$\hat{p}_k(X_i) = \frac{exp(X_i W_k + W_{0,k})}{\sum_{l=0}^{K-1} exp(X_i W_l + W_{0,l})} \tag{5.2}$$

Let C be the constant of regularization strength and $r(W)$ be a regularisation term, then the optimization is calculated as follows:

$$\min W - C \sum_{i=1}^{n} \sum_{k=0}^{K-1} [y_i = k] log(\hat{p}_k(X_i)) + r(W) \tag{5.3}$$

The Iverson bracket $[y_i = k]$ evaluates to 0 or 1 depending on the result being false or true, respectively.

To improve model fit the regression is configured with an L2 penalty regularisation term which calculates the square root of the sum of the squared vector values. Given m is the number of features then this penalty is defined as follows:

$$\frac{1}{2}\|W\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{K} W_{i,j}^2 \tag{5.4}$$

To solve, the regression uses a limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) method as proposed by Byrd et al. [20].

[1] As implemented by scikit-learn (version 1.3.1), see `https://scikit-learn.org/1.3/modules/linear_model.html`

	Predicted			
Actual	$\mathbf{L_I}$	$\mathbf{M_I}$	$\mathbf{H_I}$	**Avg**
$\mathbf{L_I}$	8	0	7	
$\mathbf{M_I}$	9	11	21	
$\mathbf{H_I}$	0	0	8	
Precision	.471	1.00	.583	.685
Recall	.533	.268	1.00	.601
Accuracy	.750	.531	.563	.615
AUC	.572	.588	.617	.601

(a) Screen-on time

	Predicted			
Actual	$\mathbf{L_I}$	$\mathbf{M_I}$	$\mathbf{H_I}$	**Avg**
$\mathbf{L_I}$	10	0	5	
$\mathbf{M_I}$	13	15	13	
$\mathbf{H_I}$	1	0	7	
Precision	.417	1.00	.280	.566
Recall	.667	.366	.875	.636
Accuracy	.703	.594	.703	.667
AUC	.580	.582	.630	.607

(b) Screen-on time, app switches and event count

	Predicted			
Actual	$\mathbf{L_I}$	$\mathbf{M_I}$	$\mathbf{H_I}$	**Avg**
$\mathbf{L_I}$	14	0	1	
$\mathbf{M_I}$	16	21	4	
$\mathbf{H_I}$	2	1	5	
Precision	.438	.955	.500	.631
Recall	.933	.512	.625	.690
Accuracy	.703	.672	.875	.750
AUC	.571	.577	.592	.586

(c) Event count

	Predicted			
Actual	$\mathbf{L_I}$	$\mathbf{M_I}$	$\mathbf{H_I}$	**Avg**
$\mathbf{L_I}$	9	6	0	
$\mathbf{M_I}$	1	40	0	
$\mathbf{H_I}$	2	0	6	
Precision	.750	.870	1.00	.873
Recall	.600	.976	.750	.775
Accuracy	.859	.891	.969	.906
AUC	.679	.677	.727	.702

(d) TF-IDF

Table 5.1: Confusion matrices for classifying the delay discounting class of each user when utilising features which are PCA transformed category-specific vectors originating from summative features or user UI events of each session.

Experiments 3.1 and 3.2 summary

Aim: Multi-modal models which combine isolated features into a single regression task are used to classify impulsivity (Experiment 3.1) and SA (Experiment 3.2). This combination addresses the user representation problem while including the category specific features which showed a high effect size.

Input sessions: All available sessions with class labels after pre-processing (as in Chapter 4, Experiment 2.1 and 2.2).

Features: High-level: Per-category screen-on time and a combination of summative features; Low-level: Per-category event counts and TF-IDF vectors created from UI events.

Output: Probability of a session showing signs of impulsivity or SA.

5.2 Impulsivity is Encoded in Complex Behaviour

This section explores the results of user impulsivity classification based on a range of input features such as screen-on time or other summative features but also low-level UI features. All results for the various kinds of regression models and their resulting classification of the 64 participants are displayed in Table 5.1. In this, an *inverse classification* will be used to describe the following kind of misclassification for a user: A user that belongs to class L_I is classified to be part of class H_I or vice versa, implying an inversion of a user's original predisposition to impulsivity. This special case is defined because it demonstrates a particular inability of a model to classify the data even on the extreme ends of the classification spectrum.

For in-category screen-on time, 27 of 64 users are correctly classified. Per class, this method correctly identifies 8 of 15 users in L_I, 11 of 41 users in M_I and 10 of 8 users in H_I. The model is most accurate at detecting low impulsivity users at a 75% accuracy, however, the precision (47.1%) and recall (53.3%) are still fairly low. M_I has

no false positives at all and misses 30 out of 41 true positives. Overall, the classes (L_I AUC=.572, M_I AUC=.588) and H_I AUC=.617) perform similarly with an effect size for the entire model AUC=.601, SD=.004. The strength of this effect size is also reflected in the macro average accuracy of 60.1%.

When utilising a larger vector compared to just screen-on time by expanding it with application switches and general event count per category, results improve slightly to 32 of 64 correct classifications. Other than that the classification is similar to screen-on time by itself. The accuracy is 5% better (66.7%) but the precision is worse (56.6%). Even though more users are correctly identified, the effect size (AUC=.607, SD=.003) is slightly lower than that of screen-on time by itself. These regressions show that using just a user's high-level features is not enough to accurately classify them into their correct group (<=50% correct classification). This motivates exploring the space of low-level multi-modal features.

Instead of counting event totals per category, it is also possible to take every event as an individual feature. This is the same method as in the previous section. When using these features in the regression another bump in classification accuracy occurs. 40 out of 64 users are identified correctly. This represents an increase of almost 10% overall accuracy to 75%. In comparison to the accuracy bump provided by adding more kinds of summative features, the precision and recall are closer to, or even higher than what can be observed for screen-on time alone. This method caused a first false positive for M_I but also halved the number of inverse classifications.

The results for TF-IDF based features are higher for all classes individually and also combined. 55 of 64 users were classified correctly. H_I being the most distinguishable class (AUC=.727) where 6 of 8 users were classified correctly while two were missed as inverse classifications. M_I (AUC=.677) and L_I (AUC=.679) then follow as the second and third highest effect sizes of all previous models. Respectively that is 40 of 41 (89.1% accuracy) and 9 of 15 (89.1% accuracy) users correctly identified in those classes. Additionally, no more users apart from the previously mentioned two

from H_I were inversely classified. The average performance of the model is AUC=.702, SD=.002 is 10% better than previously and the overall accuracy of 95% improves the results achieved by event count vectorisation by 15%. For this application of user classification, low-level features have improved the accuracy of the models significantly.

5.2.1 Classification Result Discussion

The previous sections displayed the model's capability to decode smartphone behaviour at different granularities as a proxy to observing user traits. Section 2.3.2.1 discussed how an effect size of 0.7 falls into the range of acceptable results. Only the TF-IDF models were able to match that performance while summative features hovered around .6 AUC and count-based models didn't even reach this level. However, this is still above the averages observed for any of the isolated tests and also they are more consistent with standard deviations being in the 1% range.

These effect sizes only apply to classifying a single session by itself, not the overall disposition of a user. The effect sizes in these models apply to detecting the disposition of each separate session, not a user's total disposition to impulsivity as reflected in the profile of all sessions they produce. By combining the class probabilities of all sessions a much more accurate result for a user can be found than by checking each session individually.

Summative features, both for screen-on time and a combination of summative features to build a larger vector, do not perform well even after this transformation. At most half of the users get classified correctly, with a relatively high amount (10%) of inverse classifications. Perhaps counter-intuitively, using the UI event count vector produces better accuracy results even though the models have worse effect sizes. While it seems that models that are worse at classifying individual sessions should also show worse results when combining those results, this could be the result of UI event models classifying a smaller amount of sessions correctly, but the probabilities for correct predictions being

higher resulting in a stronger influence when combining.

Finally, TF-IDF models predict the discounting classes of 55 from 64 users correctly. Also, only 2 users ended up being completely misclassified (*low* as *high*). Compared to the other methods this only improves the accuracy of complete misclassification by 1, however classifications for M_I are drastically improved. This constitutes an overall classification accuracy of 90%. Given the variability of all the input factors and the general noise and complexity within human behaviour, this can be considered a very good result for a classification of this kind. The jump in accuracy from the previous count model could be a result of the data being well distributed as an effect of the TF-IDF transformation. The reduction of variability from abstracting away pure screen-on time and rather focusing on the internal characteristics of the usage session might cause a regression model to learn more intrinsic features about usage. From this, it can also be deducted that the vectorisation method for BFUS likely needs careful selection, since results can differ starkly based on the desired result (per-session or per-user) and processing in the application stage (pairwise tests or regression). It seems that the vectorisation method plays a part in how well users can be classified but especially on the extreme ends TF-IDF did not improve detection significantly, it is mostly the medium cases of impulsivity that were captured more accurately.

In conclusion, TF-IDF weighted UI events as features are an improved way of distinguishing between the delay discounting classes L_I, M_I, and H_I. This method enables detecting a user's impulsivity class based on their delay discounting by only processing their user-app interactions without requiring further interaction with a user, such as needing them to fill out a survey. Given enough data points in form of sessions, it captures the nuances of their usage for each session and will calculate a result based on their overall behaviour. It is able to do so over other methods such as utilising summative features or simpler vectorisation methods such as counting UI events. However, while it did perform slightly better for the extremas most of the improvements are seen for more accurately detecting whether a user is of the M_I class.

This directly offers some answers to the research questions in Chapter 1. RQ1 is now likely as high-level summative were less or even in-effective. Also, the results of applying the BFUS with TF-IDF features gives a partial answer to RQ2 as it shows that deducting user traits from usage is possible in some scenarios. To add further support, analysis of multi-modal models for SA are investigated next.

	Predicted							
	Screen-on		Summative		Count		TF-IDF	
Actual	$\mathbf{A_{SA}}$	$\mathbf{N_{SA}}$	$\mathbf{A_{SA}}$	$\mathbf{N_{SA}}$	$\mathbf{A_{SA}}$	$\mathbf{N_{SA}}$	$\mathbf{A_{SA}}$	$\mathbf{N_{SA}}$
$\mathbf{A_{SA}}$	6	7	10	3	6	7	9	4
$\mathbf{N_{SA}}$	2	49	10	41	3	48	4	47
Precision	.750		.500		.667		.692	
Recall	.462		.769		.462		.692	
Accuracy	.859		.797		.844		.875	
AUC	.603		.607		.588		.724	

Table 5.2: Confusion matrices for classifying the SA class of each user when utilising features which are PCA transformed category-specific vectors originating from summative features or user UI events of each session.

5.3 Smartphone Addiction is Encoded in Complex Behaviour

In this section the feature vectors described in Section 5.1 will be reused to classify users in relation to their proneness to smartphone addiction. The process is very similar apart from addiction being binary and not a multi-class problem (like the impulsivity classes in Section 5.2), therefore a differentiation of an *inverse* classification does not exist (as every misclassification is inverse from the desired outcome).

Table 5.2 shows, with an accuracy of .859, that using only screen-on time as the re-

gressor predicts 55 of 64 users correctly. While there were only 2 user falsely classified as addicted, 7 addicted users were missed. The per-session model performance is not very high (AUC=.603).

Summative features (screen-on time, application count and total event count) perform worse than screen-on time by itself, only classifying 51 users correctly. A total of 10 users were falsely classified as addicted, however only 3 addicted users were missed (AUC=.607). Similar to the combined high-level feature models for impulsivity, the addiction models also struggled to learn the session specific markers and therefore encourage the investigation of models using low-level features.

Moving to counts of low-level features as the regressor shows similar performance to screen-on time with a 54 users being classified correctly. This method also correctly identifies 6 users as addicted and misses 7, but interprets one more user in N_{SA} as addicted. Similar to the regression for impulsivity features in Table 5.1, event count seems to be the hardest to learn for the models (AUC=.588).

Finally, TF-IDF features do perform best overall with 56 correctly identified users. 9 users are correctly identified as addicted while minimising wrong classifications (4 in either case). This means that while it is not strictly the best for every metric it does maximise the total predictive power. This is also reflected by the model fit (AUC=.724) which is the strongest performing model across all regression models in this chapter (including the impulsivity models).

5.3.1 Classification Result Discussion

Smartphone addiction shows some parallels to impulsivity in terms of possible classification. The models achieve adequate results given the extreme variability of the input vectors. The power of the results is comparable to those of impulsivity while having much less variance. The accuracies of the SA models are all within 10% of each other compared to up to 30% for impulsivity.

The models trained with TF-IDF features again show the best performance, at least on a session-by-session basis. They also have the highest accuracy overall by a small amount (with one more correct prediction) but do not have the highest precision and recall. TF-IDF improved on screen-on time by only a single additional correct classification. There is multiple possibilities for why they are settled so close to each other, while the AUC is improved by over 10% for TF-IDF models, there might be an issue of sample size arising. While every user has generated a sizeable amount of sessions, it is possible that adding even more session could show more improved results for TF-IDF features. This is especially the case because of how user labels are attached to each session as discussed in Section 4.2, it is possible that for some of the users this difference does not crystallise out from their regular usage enough in a binary classfication problem. Alternatively, it is possible that screen-on time is actually a fairly decent predictor for SA, however, it should be noted that in this case when it came to predictions of specifically the true positive case A_{SA} it performed worse to TF-IDF by a more significant degree (6 and 9 correct classification respectively).

This means that which one of these models is the best in that situation may depend on the desired outcome. Two potential alternate real world scenarios could be constructed:

- The correct detection of addicted users is vital. The system has to be sure that once a user is flagged they are actually showing these negative tendencies. In this case maximising the precision would be preferred to overall accuracy, because missing users on the edge of detection has less consequence than including wrong classifications.
- For screening purposes as many of the addicted users are sought to be included. This enables reducing the set of potential participants with issues that have to be addressed and can be followed up by additional steps. Here a high recall would be ideal because less addicted users would be missed in the selection.

While these are not the only potential scenarios it demonstrates how considering al-

	Impulsivity		
Addiction	**High (H_I)**	**Medium (M_I)**	**Low (L_I)**
Addicted (A_{SA})	4 (31%)	7 (54%)	2 (15%)
Not addicted (N_{SA})	11 (21%)	34 (67%)	6 (12%)

(a) Total overlap of all users.

	Impulsivity		
Addiction	**High (H_I)**	**Medium (M_I)**	**Low (L_I)**
Addicted (A_{SA})	3 (30%)	4 (40%)	2 (30%)
Not addicted (N_{SA})	6 (15%)	30 (75%)	4 (10%)

(b) Overlap of correctly classified users by TF-IDF based regression models.

Table 5.3: The count (and percentage) of overlapping users between the addiction and impulsivity classes.

ternative models could be beneficial. Given the better model fit of TF-IDF and how close the actual number values are though, in this case it is likely that when user count is extrapolated further the results would favour TF-IDF overall.

Another aspect of these models is that misclassifications are (as discussed in the introduction to this section) always 'hard'. Generally misclassifications are never desired, but given certain scenarios it can be less of an issue than in others. For example, in a multi-class problem the difference between classifying a user as low instead of medium might not be as impactful as if there was only low and high. Which in a binary problem (such as SA in this case) is always going to be the case.

In combination this has shown that SA and impulsivity individually can be detected in users from just their usage behaviour, following is a discussion about how these two classes (which might be perceived as similar) should be valid to allow an extrapolation to other cases.

5.4 Generalisation of User Trait Extraction

The results in this chapter show that the BFUS model can be configured in such a way to enable user trait identification. Those findings were validated in two separate instances by identifying SA measured via the SAS and impulsivity measured via the MCQ in user groups. This gives confidence in the model functioning not just specifically with SA or impulsivity but rather is generally applicable.

A caveat to generalisation may be the perceived closeness of smartphone addiction and impulsivity in terms of what they cause in the behaviour of a user. The argument could be made that the issues between SA and impulsivity are linked as both topics describe some impact on the human psyche. One could come to the conclusion that a large user overlap exists between the addicted and high impulsivity groups. However, there is no correlation between the SAS and MCQ results of each user (r=.196, p=.121). This also extends to the specific pair of the groups for addicted (A_{SA}) and highly impulsive (H_I) users (r=.266, p=.379). Additionally, Table 5.3 shows that the overlap between impulsivity groups is similarly distributed across addicted and non-addicted users. This means that the SAS and MCQ are not biased in such a way that they capture similar groups of users by default.

This distribution does not significantly change if only the users that were correctly classified using the TF-IDF models are considered. When focusing on the high impulsivity group the gap does get larger for correct predictions of addicted compared to non addicted users. This is likely the case because impulsivity and addiction alike have been identified to have an influence on a user's behavioural patterns and the likelihood that *some* matching usage behaviour exists is relatively high. Apart from those assumptions the connection between smartphone addiction and impulsivity and its effect on the prior classification tasks seems to be small as addicted and non addicted users alike can be identified with similar accuracy in relation to their impulsivity groups. In conclusion, as part of this analysis impulsivity and smartphone addiction can be considered separate variables as there is no substantial overlap between their groups.

5.5 Implications and Considerations in predicting User Traits

Due to the sensitive nature of the topic there are implications of hard misclassifications if represented as strict binary classification labels (addicted when not addicted or vice versa). The classification that was used relies on combining all session probabilities to create a single probability of how impulsive or addicted a user is. The magnitude of per-session probability is expected to be relevant when combining it into a total probability. For example, instead it would be possible to count how many sessions were assigned a label and pick the highest count but that would actually remove valuable information of how impactful any single session was in context of all sessions.

Conceptually, when extending this to applying the final class label to a user it creates a similar problem. Once the total probability is retrieved, applying a class label actually loses information. Considering that issues such as problematic smartphone use does not share the exact same symptoms between each user, quantifying it as a simple yes or no question could be problematic. For a classification task (especially to evaluate model performance) it makes sense, however in a real-world setting it might be more practical to evaluate users based on the sliding scale that is presented when inspecting the identified probability of addiction.

As a consequence the findings could be reported in terms of potentially correlating factors and probabilistic risk, rather than treating it as a strict classification problem. Anything added beyond that should be treated merely as suggestions as to how this knowledge could be used in practice.

In Figure 5.1 this is demonstrated by showing the results of the TF-IDF model and how users are actually distributed when considering their combined session probabilities. The separation of users is not very clear, while there are some users who have stronger tendencies on either side of the spectrum the general distribution is a trend rather than a clear separation. Given this image, drawing a hard line at exactly the mid point and

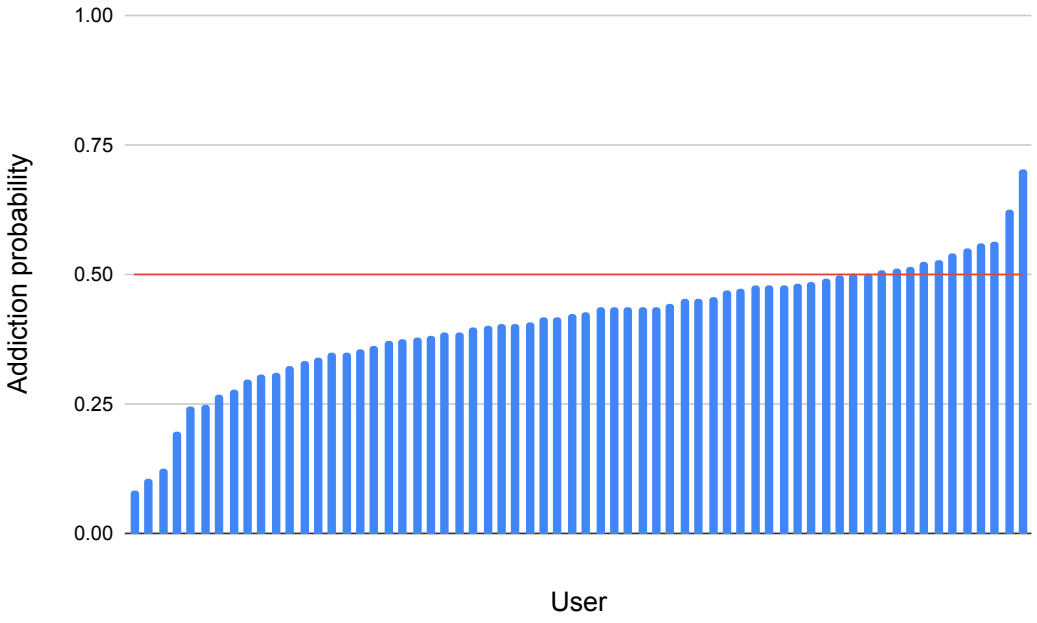


Figure 5.1: A sorted view of all 64 users in the Tymer dataset and their combined addiction probabilities based on a TF-IDF vectorisation. The red line represents a default cut-off value of 0.5 between an addicted and non-addicted class.

declare everyone above addicted seems questionable.

This shows the importance of considering the impact on users from these classifications and that any conclusion stemming from it should be carefully reviewed. However, depending on the accuracy of the results it is still likely that such a prediction is a solid indicator, even if used with such a cut-off method. A deeper exploration of this problem surrounding classifications and specifically uncertainty during the prediction process is featured in Chapter 6.

5.6 Stability of Trained Models and Survey Data

While the surveys to infer user traits used in this thesis (SAS and MCQ) are statistically validated, compared to the objective data which is collected from user interactions it

may be possible that they lack in some aspects of actual representation. The models presented in this thesis do not follow specific questions or rules that were previously defined, instead they learn from the user inputs and then relate users that show similar usage patterns. This could mean that these models will find relationships which are closer than those modelled from static questions, which would have an implications on the conclusions of prediction results in this thesis.

When false classifications are encountered, they are wrong from the perspective that the original label could not accurately be recreated from the supplied input data. The model identifies the input features of an unseen case to be similar to what it has learned and therefore relates it to the probability of it being of a certain class. The cause of this could be of two reasons: The first being that the learned features do not describe the variance of the independent variable closely enough. In this it can be assumed that the model does not have a good fit of understanding the relationships between features.

For the second point it has to be understood that in this analysis the SAS and MCQ were used as a proxy for SA and impulsivity. This means that the model is learning features which inform of the results of those surveys since they are markers which can be used for those traits. However, that is what they are, an approximation of our understanding for what constitutes SA or impulsivity.

The model has learnt the relationship of behaviour in the form of actual usage features at a low-level of interactions. It then groups users based on this usage instead of considering (potentially arbitrary) relationships of questions given to the user. It is possible that the models actually have learned habits and patterns of problematic user behaviour that cannot be detected by an in-person survey. Thus it may be that it has developed a more in-depth understanding of those conditions than can be described by survey questions.

So, while the analysis in this chapter showed that the response values of the SAS and MCQ can be predicted with high accuracy, this poses the discussion whether a predictive model like this is even better at decoding the markers of the underlying con-

ditions (SA and impulsivity) than those original surveys. Validation of such a concept would present a challenge since it would require 'perfect' labels (e.g., addicted or not addicted) for the training set of users and it is unclear how that could be produced because surveys are usually what is relied on to produce them.

5.7 Conclusion

In this chapter logistic regression models were used to predict user trait classes from their transformed usage event stream. Between impulsivity and SA the results overall are positive, where classification accuracy is as high as 90%. It further validates previous literature which has found markers of usage to be correlated with user traits such as addiction (e.g., [96]) by adding method which is able to effectively separate groups of users based on low-level events. It also achieves this by only monitoring a user's usage stream instead of the use of surveys (e.g., [21]). This compliments the isolated feature analysis of Chapter 4, it shows that by transforming low-level features they can be user to infer user behaviour as posed in RQ2 and also provides answers to RQ3 in that it is possible to predict user traits based on those features which produces (at least partly) improved results to high-level features. Therefore it presents significant evidence which supports C2 and also supports the application of the BFUS model of C3.

Perhaps surprisingly, the models for either user trait show similar performance. In Section 5.4 the analysis of an overlap between the trait groups shows that there is no significant similarities especially between the "true positive" groups. This is interesting considering SA was classified based on previously defined cut-off points outside of the scope of this thesis while MCQ cut-off points were derived from their distribution. This might inform of a general level of user trait information that is encoded in usage behaviour.

The application with multiple features also reveals an observation but also potential

limitation in the form of feature selection. For impulsivity models, low-level feature vectors like event count and TF-IDF performed much better than the other representative summative high-level features. For SA, while the training effect (AUC) was better for TF-IDF, there was a less clear separation between summative and low-level features. Generally, selection of features seems to be a very important consideration for model performance.

At a more general level, it is important to highlight that as well as SA and impulsivity detection being used for positive and supportive purposes, such as to help diagnosis or manage a condition, this work highlights that automatically inferring a user's traits from simple smartphone behaviour is possible. This reaffirms the importance of clarity and ownership of one's low level data and the inferences that are drawn from it. It also highlights the significant trust that is placed in third-party organisations aligned to individual data, even if the data is anonymous but bound to a specific device. These issues surrounding the current data governance principles (or lack thereof) of automated, potentially intelligent, systems which handle and process human characteristics are the subject of ongoing development and debate in the wider literature [132, 16, 48].

A psychological interpretation of results is complicated since TF-IDF encodes the events as weights and PCA then obfuscates the incoming features completely. After this the regression model can show that some features have lower or higher β values but they do not represent any real-world explainable features. The results show that the presented methodology creates an effective way of capturing and representing usage, so this poses the question of a possible extension which either does not obscure the input parameters or is able to evaluate them.

The previous chapters culminate and are compounded here, the investigation within this thesis moved through the multiple stages of smartphone behaviour analysis, from its history and limitations to a new model with novel applications for detection purposes. In this, the aim of this thesis, the detection and potential utilisation of user traits from just smartphone interactions is considered successful. From here, the next chapter

offers a collection of extensions to the process which addresses some of the limitations or edge cases remaining.

Chapter 6

Extensions to Behaviour Modelling

The previous chapters have discussed the efficacy of methods in behaviour research (e.g., a focus on high-level or isolated features) and how those can be improved with methods such as proper bounding or consideration of low-level events and applied in a context such as user trait detection. In this chapter, recommendations for user trait model parameters are explored which extend upon the previous additions towards RQ3. Those suggestions offer an in-depth analysis of how user behaviour can be evaluated instead of just predicted. This expands the set of potential tools for the evaluation of usage as presented in Chapter 4 and contributes to C5.

6.1 The User-Session Relationship

In classification tasks the relationship between class (e.g., A_{SA}or N_{SA}) and label should be one-to-one. This means when estimating the class of a user, such as that in Chapter 5, the input features should describe the label of the user themselves directly and not be an amalgamation of their session behaviour. By considering each session as an individual classification problem, the scope was extended to a one-to-many relationship (many sessions for one user). The outcomes were then combined to a single result based on their average. Implementing such a strategy meant that every session can be predicted individually but also introduced a certain amount of uncertainty for each label. This leaves room for a potential concentration of data points with stronger indic-

ators of the user's class traits and therefore improve predictions on a per-user basis.

This becomes relevant given the construction of the learning set for the models presented in Section 4.2. Since this is a novel application of low-level UI events for user trait detection there is no available data on how they are influenced by user behaviour. Concretely, for the SA models (but similarly the impulsivity models) this means that sessions are labelled based on the user's addiction label as a binary value. This assumes that every session of an addicted user will be distinct from that of non-addicted users. In reality, this is likely not the case; only a subset of sessions will exhibit distinct problematic characteristics. Additionally, the dataset contains sessions that are short and have limited UI events, where correlating characteristics with addiction may be prevalent. This also could be a source for sessions which are similar between addicted and non-addicted users.

Based on this, we may expect that a lot of sessions exist that influence the classification by training the model with conflicting information. Currently, the labelling of sessions assumes an inherent base stability in the input data. If every data point (e.g., session) in the one-to-many relationship is labelled correctly then the naive solution of including all data points should be the best representation of the original data. Instead, if the labels of the input data cannot be verified accurately, then there might be cases where incorrectly tagged data can introduce issues during the training of the models. Excluding these sessions from the evaluation of the users' trait classes may improve the accuracy of the models.

To examine this effect and the impact on the modelling, sessions that are close to the boundary of classification of the current models can be excluded, as they as they are likely to provide little to no valuable information for the actual task of user classification. The intention of this is to isolate a type of 'uncertain' usage session which is common in both addicted and non-addicted users and then remove it.

To account for these uncertain cases of interaction, we continue by training a binary classifier, but then evaluate and adjust the classification of the training data to account

for uncertainty in the model. Interpreting ranges in logistic probabilities for uncertain cases has been discussed in the literature before [69]. In these cases, a low confidence area around the classification threshold is created which, besides the labels (in our case, *addicted* and *not addicted*), creates a third label, *uncertain* as similarly explored by Li et al. [78] and performed by Johannes Landsheer for detection of cognitive impairment (CI) [74]. In this, Landsheer was able to isolate 27.5% (N=1379) of CI cases as uncertain, which allowed him to boost his accuracy for results of positive and negative classifications.

6.1.1 Examining the Effects of Evaluation Thresholds

The regression model calculates the probability that a given session is contained in a positive (e.g., is addicted) class. A classification threshold is a cut-off point in $[0, 1]$ which is defined to separate the two classes. By default, a value of 0.5 is used, however it can be adjusted to account for cases of imbalanced classification. A more suitable threshold can be calculated via the maximum Youden index [166] to optimise the break-even point between the false positive and true positive rate.

$$sensitivity = \frac{\text{true positives}}{\text{true positive} + \text{false negatives}} \tag{6.1}$$

$$specificity = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \tag{6.2}$$

$$J = sensitivity + specificity - 1 \tag{6.3}$$

To demonstrate the effects of such an approach, the smartphone addiction TF-IDF models of Section 5.3 can be re-used, as they represent a fairly classic binary yet imbalanced classification problem. With this model, the Youden index (J) is calculated as .567, which is very close to the default threshold. This can be explained through the model being trained based on a balanced input (see Section 5.1), which means that the input data is weighted to be equally as likely to occur and therefore adjusts the output probabilities. Doing so is fine for prediction tasks, but because it does not accurately

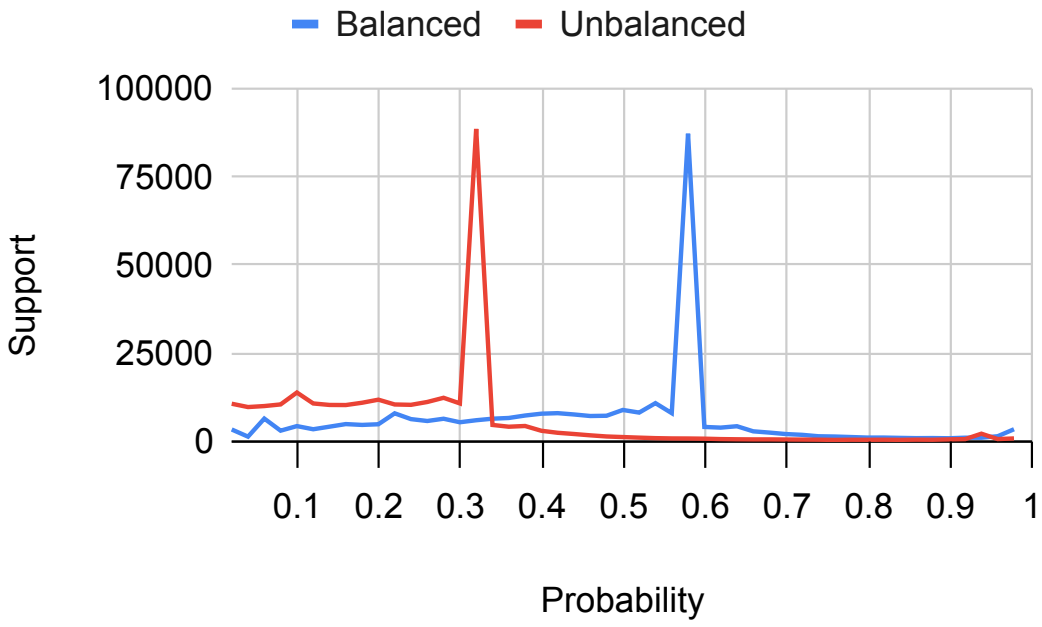


Figure 6.1: The amount of session support for all probabilities when a model is trained with balanced and unbalanced input weights. A probability closer to 0 corresponds to N_{SA} and 1 to A_{SA}.

represent the underlying distributions of occurrence it shifts the Youden index away from its original position.

To identify the Youden index as it occurs in the original distribution of classes, the model needs to be adjusted so that the inputs are no longer balanced for class imbalance. The issues this introduces can be later addressed by shifting the Youden index into a more ideal position than the base cut-off. Figure 6.1 shows a histogram of the probabilities of sessions between a balanced and unbalanced dataset. The major difference is the cut-off point which can be observed as the point where the most uncertain sessions settle, compared to a model with balanced features it moved away from the centre to ~ 0.31 for the unbalanced set. This occurs because of the larger volume of sessions with a low probability of being in a certain class, the model does not correct for the skew anymore and learns the distribution which will be a high concentration of low probabilities.

The resulting Youden index (M=.31, SD=0.01) for the unbalanced dataset can be used as the threshold of probabilities when classifying users (as $A_{SA}$and N_{SA}) which means this method should be close to equal to balancing the input data before. In this case, the prediction results (9 true positives, 4 false positives, 4 false negatives and 47 true negatives) are exactly the same as the balanced model and even the AUCs are within 1% (compare Table 5.2). This way the same accuracy of a pre-balanced model is achieved while maintaining a model which is trained using a distribution of sessions of how they actually occurred. This shows that either method is suitable for a normal classification task, however for further computation with uncertainty along the probabilities, an unskewed distribution is preferred.

Experiment 4.1 summary

Aim: Users may not exhibit signs of addiction in every usage session. Removing the sessions which only have a small impact on the classification may improve its accuracy.

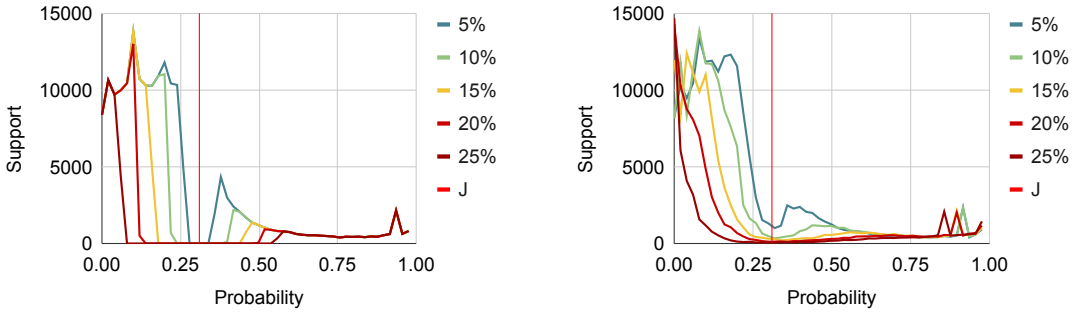
Input sessions: All available sessions with class labels after pre-processing (as in Chapter 4, Experiment 2.1 and 2.2).

Features: PCA transformed TF-IDF vectors created from UI events (as in Chapter 5, Experiment 3.2).

Output: Probability of addiction for each session (as in Chapter 5, Experiment 3.2).

6.1.2 Classification of the Uncertain

For the low-confidence (or 'uncertainty') range, a balance between removing sessions and probability should be considered. Choosing it too high might remove too many sessions and reduce the overall confidence in the model being generalisable, but leaving it too small causes an ineffective amount of sessions to be removed. Figure 6.2



(a) Probabilities before retraining

(b) Probabilities after retraining

Figure 6.2: The amount of session support for all probabilities when an unbalanced model is trained but sessions in the uncertainty range of 5/10/15/20/25% expanded from the Youden index (J) are removed.

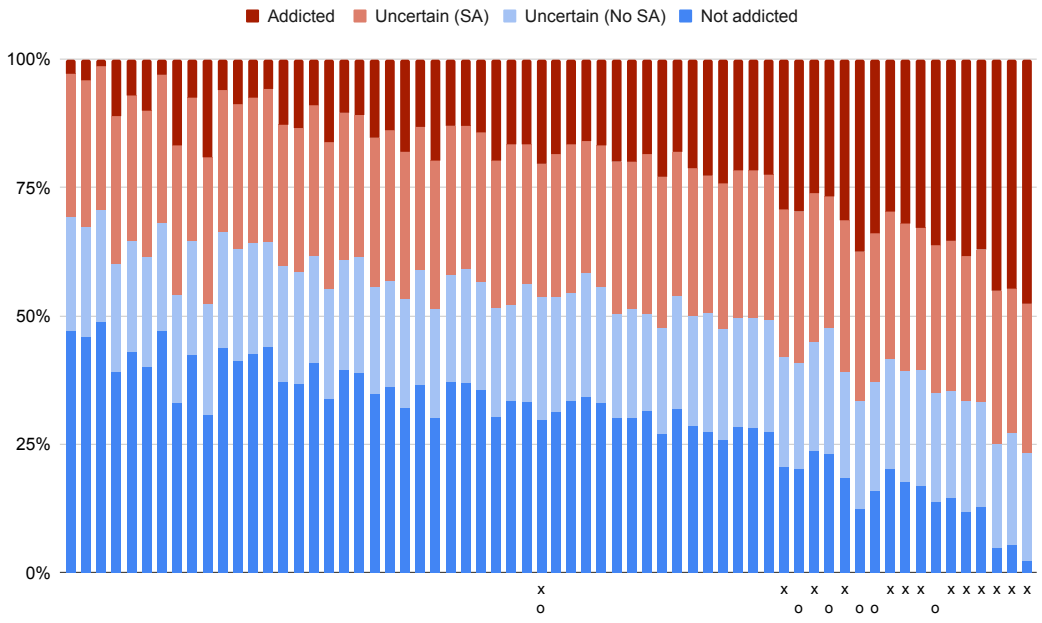


Figure 6.3: The distribution of sessions that are considered addicted, non-addicted or uncertain for every user. Uncertain is split into sessions that would have been classified as addicted or not addicted at the probability threshold. The users are sorted by their mean addiction probability of all their sessions. Symbols along the x-axis: 'x' is a user that is addicted according to the SAS. 'o' shows a false classification (depending on 'x')

	Predicted									
	5%		10%		15%		20%		25%	
Actual	$\mathbf{A_{SA}}$	$\mathbf{N_{SA}}$	$\mathbf{A_{SA}}$	$\mathbf{N_{SA}}$	$\mathbf{A_{SA}}$	$\mathbf{N_{SA}}$	$\mathbf{A_{SA}}$	$\mathbf{N_{SA}}$	$\mathbf{A_{SA}}$	$\mathbf{N_{SA}}$
$\mathbf{A_{SA}}$	11	2	12	1	12	1	12	1	12	1
$\mathbf{N_{SA}}$	10	41	7	44	5	46	5	46	8	43
Sessions	176,385		141,265		107,978		78,363		46,457	
Precision	.524		.632		.706		.706		.600	
Recall	.846		.932		.923		.923		.923	
Accuracy	.813		.875		.906		.906		.859	
AUC	.794		.826		.865		.909		.950	

Table 6.1: Confusion matrices for detecting user smartphone addiction when evaluated with an uncertainty range surrounding a probability threshold. Reference for 0%: 301,024 sessions (TP=9, FP=4, FN=4, TN=47), Precision & Recall=.692, Accuracy=.875, AUC=.724

shows how different cut-off points affect the number of sessions that get removed surrounding the probability threshold. Those predictions are often incorrect because of the low probabilities assigned by the regression model, therefore the overall classification accuracy of the model can be raised.

Figure 6.3 visualises how different thresholds affect the sessions removed by the uncertainty range affect the number of considered sessions. Every user produced sessions of each category but either side of the scale shows that users predominantly create sessions with their respective labels. Additionally, it shows how users that were falsely classified (marked with 'o') fit in with their neighbours based on the volume of sessions of the opposite label they produced.

In Table 6.1 the impact on classification is shown. As the adjusted models remove the peak around the Youden index, even 5% removes almost half of all classification sessions (∼300k to ∼175k). Given the large volume of sessions this still leaves a lot

of sessions per user to be evaluated (M, Mdn, Std) while removing a lot of sessions that do not contribute a lot of distinguishable information. However, in the case of 5%, the overall prediction quality is slightly worse with 52/64 correct predictions, this is likely a result of removing previously available information which helped classification while still retaining too much noise. Once 10% are removed the results show slight improvements with 56 correct classifications. This is better than the previous models where no thresholding took place and evidences some support that addictive behaviour may not be exhibited in all sessions a user performs.

There is a relationship between the threshold increasing and more and more sessions being removed (roughly 30-40k per step). The results for 15% and 20% are identical with an accuracy of 58/64 correct classifications. Simultaneously, the AUC for the remaining session is increasing steadily as sessions which are less relevant are removed. The accuracy for remaining sessions increases, even if it does not directly translate to improved results for users. One aspect to consider when applying such a technique is that the removal of an increasingly high volume of data points might cause issues with the general applicability of the regression models, risking fitting it too close to the specific data.

For example, at 25% (with support of only 46,457 sessions) the accuracy reduces back to 55/64 correct classifications. At this point, so many sessions are removed that some users only have <10 sessions to be tested for each iteration of the cross-validation. The loss of trainable data at a threshold this high removes not only confidence in the generalisability of the method as a whole but also removes too much data from the model to make good guesses on a user's session.

6.1.3 Validity and Discussion

By applying an uncertainty range the gap required to create a more definitive addicted or non-addicted session is widened. In Figure 5.1 it was possible to observe how the

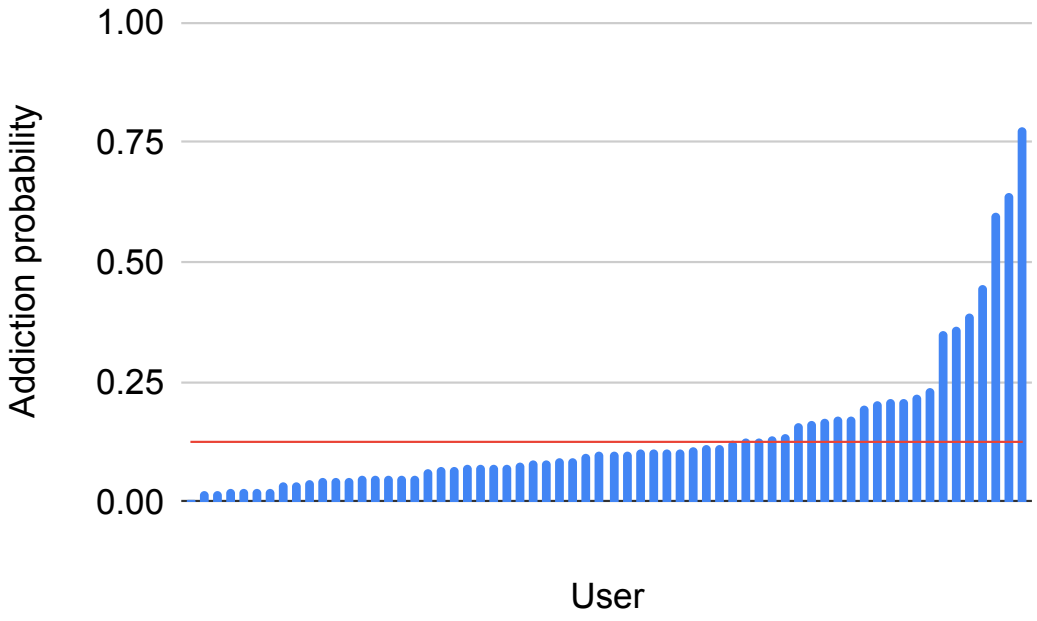


Figure 6.4: Distribution of probabilities for a regression model with unbalanced input data and after removing sessions in a 15% uncertainty range around a 0.31 threshold represented by the red line.

probabilities of all users aligned. This shows that drawing a hard line at a threshold of 0.5 is probably not a practice that is applicable in the real-world. Figure 6.4 shows a distribution of all users after applying a 15% uncertainty range and retraining the model with those sessions. While there is a clearer separation between the peaks of addicted and non-addicted users the threshold is still in an area where it should not be applied automatically (note that for demonstration purposes the probability threshold is adjusted as it changes again after retraining).

This shows how the one-to-many relationship between users and usage sessions can be utilised and also adds another layer to the discussion of classification strategies in Section 5.5. Improvements to accuracy are possible, however, depending on the available data and methods used they can introduce issues with pruning too much data. Choosing to use an uncertainty range for classification in a user trait identification task

can therefore only make sense if tuned appropriately. Additionally, while the best-performing uncertainty range in this section is 20%, it may be not the exact threshold that will be applicable for other feature sets or users. It shows that removing some of the less impactful data can be useful to improve classification accuracy and sets an example of how to choose such a range instead of a simple threshold without uncertainty considerations.

Another point that needs to be addressed is that this actively interferes and adjusts the outcomes of a regression rather than improving the understanding of input parameters to the model. This would be a bigger issue if the input data, specifically the user labels, are completely accurate. However, the way that user labels are applied to each session (as discussed in Section 4.2 means that the ground truth on a per-session level is not guaranteed. Thus the main motivation to this approach stems from the fact that it is not possible to retrieve the per-session labels but the analysis is based on a per-session basis. Following this, it is difficult to tune the models further for more accurate readings on the labels themselves because on a session level it may be unknown whether they are more or less reflective of a user trait if they are not already firmly predictive of one or the other. This also means that this method is not recommended in a situation where per-session labels exist, because then the goal should instead be to identify the markers for 'neutral' sessions directly. In this case, where the label assignment causes the uncertainty of which features may be influential, thresholding represents a compromise since it does necessitate discarding information in the form of some sessions but may help to strengthen a predictive model which may otherwise not be possible.

The results suggest that the classification of SA can be reduced down to the occurrence (or not) of a small number of specific smartphone sessions to a user. This raises questions, along with the contributions of Chapters 4 and 5, about the psychological interpretation of the influences of individual features on trait predictions.

6.2 Balancing Model Interpretability and Accuracy

The choices made in the implementation of the BFUS model to examine linkages between smartphone behaviour and user traits in Chapter 5 were focused on maximising classification accuracy. In particular, training the regression models with raw feature vectors causes issues with highly correlated features (such as singular matrices during training). By applying PCA, those collinearity issues are resolved automatically while maintaining almost the total variance (99%) in the data. This is a common strategy for regression models but comes at the cost of losing their interpretability, as they now represent a compressed version of the original input features which can no longer be accessed after a PCA transformation. However, the implementation can be adjusted to allow for the extraction and examination of the effects of specific behaviours in specific app types that can bring useful conclusions in their own right, albeit at the potential expense of maximising model accuracy.

6.2.1 Regression Coefficients

To circumvent the feature obfuscation which is introduced by using PCA an alternate method to resolve collinearity issues during the regression is needed. Collinearity occurs for features which are highly correlated, so removing features is an option. However, it adds the additional task of having to decide which features to remove or keep. Since this is a manual step it may introduce issues for creating regression models because of reasons such as:

- It can become very laborious for a large number of features
- Singularity can also arise from inter-dependencies of multiple variables that are linearly linked
- Ultimately it is an arbitrary selection of which feature to keep or discard

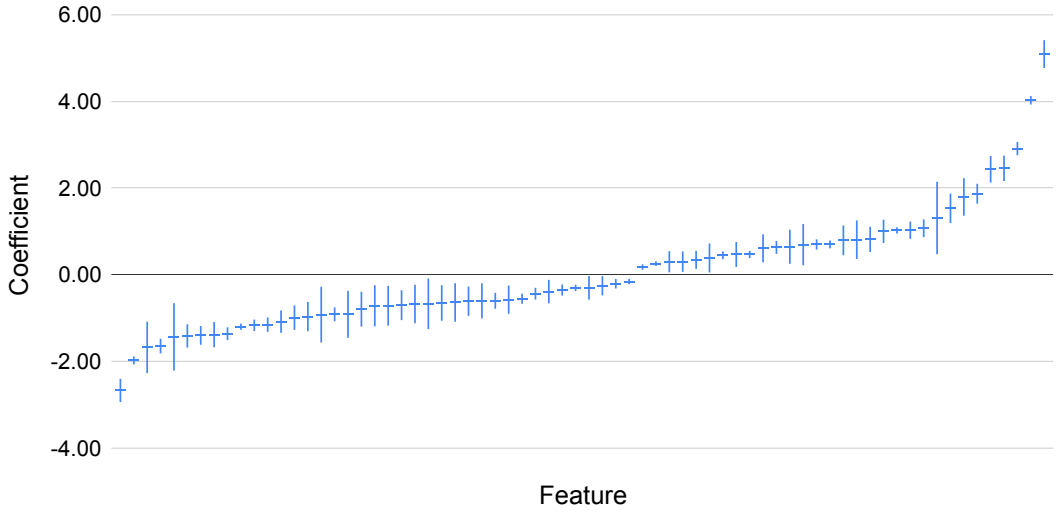
One way around this issue could be hierarchical regression, however, this is difficult because of the high amount of features. Since any session may include multiple categories, they need to be included in their entirety alongside all features in the dataset. This means that building many stepwise models for the amount of sessions in the Tymer dataset is a very computationally expensive process. Thus, for a faster selection this section opts to remove collinaearity issues removing highly correlated features.

Assuming a successful removal of correlated features, they can be used as input features for the regression models with the same process as Chapter 5. In the case of the Tymer dataset, with input features being category-specific TF-IDF transformed UI events, removal of features showing a correlation of $>= 0.9$ proved to effectively resolve all collinearity issues. This section will go on to show how the BFUS can be used to identify the individual influences of features which make up the whole of a user trait classification.

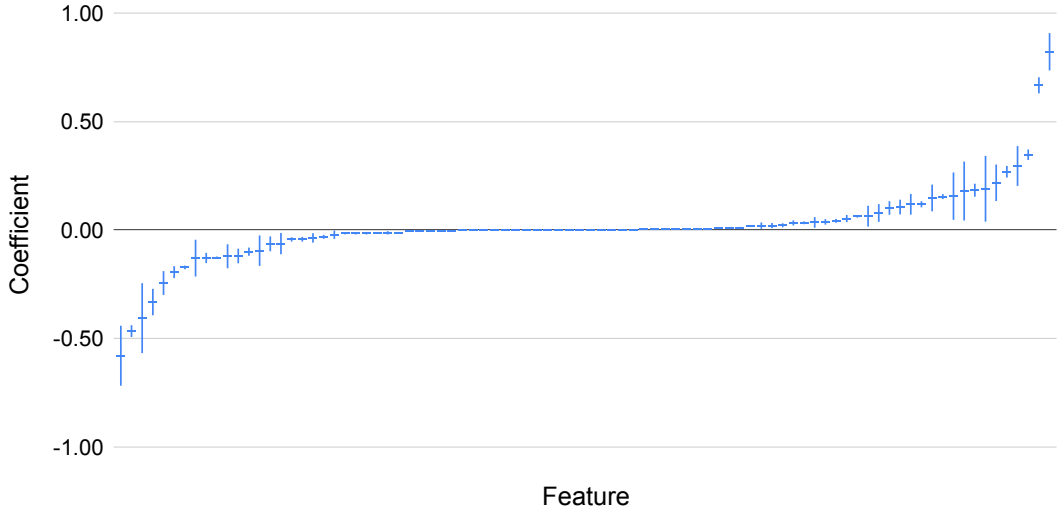
Once the logistic regression models are adjusted in this way, feature importance can be extracted by utilising the models' coefficients. This shows how important specific regressors were in the context of the current model, even if they might not be directly comparable to a general measure or between different models. This gives insight into the influence of features on the decision function of the model made based on the behaviour of a user.

The regression coefficients are a measure of how much any given regressor contributes to the outcome of the final prediction. A higher, positive value corresponds to a positive outcome (*addicted* in this case) and a lower, negative value corresponds to a negative outcome (*not addicted*). Features which are close to zero show that they are either too similar or hold too little variance information to be useful for predictive tasks.

Figure 6.5a and Figure 6.5b visualise all coefficients in the model with statistical significance ($p<.05$) after Bonferroni correction. The full range of coefficient tables is available in Appendix D. TF-IDF has a fairly gradual, linear distribution of coefficients with a few features on either end of the curve having a particularly strong influence.



(a) Using TF-IDF



(b) Using event count

Figure 6.5: Shape comparison of feature coefficients and a 5% confidence interval for all statistically significant features in the logistic regression. View Appendix D for the full list of coefficients.

This shows that any prediction with these models is an accumulation of many reasonably weighted features. In contrast, the event count is distributed such that only a few events have a strong influence on the result.

Model	Mean	Median	Std	CV	Min	Max
TF-IDF	-.014	-.304	1.384	.488	-2.852	5.093
Count	.012	0	.178	.301	-.579	.821

Table 6.2: Descriptive statistics of coefficients for TF-IDF and event count models. CV=Coefficient of variation.

This can also be seen in Table 6.2, even though the means between coefficients for both models are similar, the deviation from the mean is much higher for the TF-IDF model (M=-.014, SD=1.384, CV=.488) than for the event count model (M=.012, SD=.178, CV=.301). Because of the unrelated ranges between model coefficients, this is true even when the normalised coefficient of variation (CV) is considered instead of just the pure standard deviation. This shows that TF-IDF has a better distribution of feature influences and the model learns from more factors of information instead of only really considering a small count of available features.

This variation could be the reason for the TF-IDF models outperforming count models in classification tasks as more features meaningfully contribute to the total outcome. However, given the transformation by TF-IDF, these coefficients do not necessarily hint towards the frequency of the events occurring in a session, but rather towards the relevance of the event. It makes a psychological interpretation difficult because TF-IDF's relevance calculation obfuscates the raw information (i.e., counts) used to calculate it. In turn, this makes interpreting the potential influence of the abundance or scarcity of events not possible in comparison to using raw event counts instead. This motivates the exploration of feature influences in an event count model.

Feature	coef	[0.05	0.95]
App switch	.154	0.143	.165
Text input	.003	.002	.004
Short idle	0	0	0
Scrolling	0	0	0
Tap	N/A		
Long idle	-.042	-.053	-.032

(a) Coefficients for count model

Feature	coef	[0.05	0.95]
App switch	1.027	.827	1.227
Text input	.698	.579	.817
Short idle	.468	.384	.553
Scrolling	-.209	-.317	-.100
Tap	-.602	-.786	-.419
Long idle	-1.401	-1.628	-1.185

(b) Coefficients for TF-IDF model

Table 6.3: Coefficients for the statistically significant features (events) in applications of the *Social* category.

Experiment 4.2 summary

Aim: Depending on the use case, the psychological interpretability of the input features may be desirable. In the previous regression models, the input features were obfuscated using PCA to resolve collinearity issues. Instead, their correlation can be used to manually remove features.

Input sessions: All available sessions with class labels after pre-processing (as in Chapter 4, Experiment 2.1 and 2.2).

Features: TF-IDF vectors created from UI events, where all features with a correlation of $>= 0.9$ were removed.

Output: Probability of addiction for each session (as in Chapter 5, Experiment 3.2) and an interpretable regression coefficient for each feature.

6.2.2 Evaluating SA Regression Coefficients

While event count models attain less accuracy than a model built using TF-IDF scores as features, they provide an opportunity to drill down into the more specific causes

of behaviour changes. To give an example of how this usage could be interpreted, Table 6.3 shows the coefficients of events for the *Social* category. This category was selected as it is frequently referred to in the literature as a key category for its impact on user behaviour and its potentially problematic aspects with respect to SA [161, 114, 109].

Application switches in this category are a factor associated with SA, this suggests that users that frequently switch from and to this category are more prone to SA. Text input, short idle and scrolling are (or close to) non-contributing factors. Taps were not a significant coefficient. Long idle is a slightly negative feature, meaning that longer pauses in these applications are a slight indicator for more considered and less SA-prone session behaviour. This aligns with previous findings that show that forcing pauses (effectively locking device usage) before using applications which are linked with overuse (such as social network applications) can induce regularisation of smartphone use [59, 60].

TF-IDF models show some parallels in the results, where it shows that as app switches become the dominant feature, it increases the probability of SA behaviour. Additionally, the same is true for the opposite effect with long idles. This shows that for these two events, their total frequency but also their relevance in a session have an impact. On top of these parallels, TF-IDF shows some more variation in the rest of the coefficients. For example, tapping and scrolling are slightly negative factors. This could be due to the fact that these are very normal actions performed by every user, so when they are the focus of a social application session it's potentially not correlated with addiction. On the other hand, text inputs (could be things such as writing a post or replying to a comment) and short idles seem to be evaluated as factors that could be interpreted as problematic. This could be due to them being the result of more engaged behaviour that requires a user to interact more closely with their device than just scrolling.

As evident, the lines of reasoning for the interpretation of TF-IDF models are much more difficult compared to those with event count. Being able to connect a higher count of an event with a higher chance of problematic behaviour is easier to understand than

the relative score of a transformation. However, in the previous chapters adding this kind of additional level has been helpful in differentiating the highly nuanced behaviour between users.

In the context of the BFUS model, this shows that choosing the parameters for all three of its steps carefully is vital to generate desired results. For example, in a classification task, maximising probability is desirable even at the cost of obfuscating the input variables. Whereas for more real-world explainable parameters, especially when it comes to trying and resolving issues of potentially problematic behaviour, it might be more appropriate depending on the application to choose a slightly worse fitting model that provides an easier means to examine the impact of event counts.

6.3 Conclusion

This chapter showed that considerations of uncertainty and interpretability should affect the creation of behaviour models - using Smartphone Addiction models as a demonstration to provide design considerations going forward. Firstly, in a novel approach addressing a problem caused by the sessionisation of the event stream through the BFUS model, the user-session relationship is explored. User traits may not be reflected by each session a user produces but is assumed to via the application of the user labels. This causes similarity of types of sessions between addicted and non-addicted users and creates uncertainty which can be leveraged to produce more precise results. Every user produces many sessions which may or may not be indicative of a user trait and there is no absolute certainty in how strongly a user's behaviour is reflected by their usage. It is possible to construct a set of training data (i.e., user sessions) that isolates more pronounced markers of a user's behaviour by removing those sessions that fall into a range of 'uncertainty'. The results show that by focusing on sessions with more certainty, probability-based prediction tasks can be improved. However, this should be approached with caution due to the effects this may have on data quantity and overfit-

ting. This is a novel approach to tackling uncertainty in smartphone behaviour and in this example which uses user traits also adds evidence for C3.

Furthermore, the challenge of building models using interpretable features is discussed. While Chapter 5 showcases the predictive power of usage transformed into TF-IDF score features, there are limited opportunities to interpret and contextualise the impact of each raw event on the classification produced by the model. Such interpretations are helpful to identify possible cause-effect relations between smartphone usage and a user's traits. Alternative strategies using event counts present a design consideration in balancing model accuracy with improved interpretability through regression coefficients.

Chapter 7

Conclusions

The overall efforts in this thesis provide contributions to the advancement of the field of behavioural research surrounding digital devices. In this chapter, these contributions are summarised and key results are highlighted, as well as a discussion surrounding its wider impact in the literature. This includes an assessment of the techniques that are used throughout the thesis. This paves the way for potential future work surrounding the future of behaviour research with digital devices, in which key directions are discussed following the thesis summary.

7.1 Thesis Summary

7.1.1 Assessment of Smartphone Behaviour Research

In the literature, a wide variety of methods have been used to record and extract the data surrounding smartphone use. This includes information specific to the device that can be used to infer habits based on usage patterns (e.g., derived from application usage, battery status, SMS and phone calls) and other sensors that collect external information (e.g., GPS location or signal strength). One area of application of understanding this behaviour is exploring links between behaviour and user states and traits. Furthermore, the use of surveys for mental states (e.g., mood or boredom) has enabled the collection of user personalities and traits (e.g., smartphone addiction or impulsivity). These

metrics have been used to infer various aspects of the behaviour of users such as rule mining or problematic use patterns.

Within these research areas, some conventions bring progression into an otherwise unstructured landscape. Recent work has been moving away from strict survey data and towards large-scale data processing because of the advanced capabilities of on-device capture but also progress in the understanding of aligning user states with their usage. However, aspects such as feature selection still remain to be identified and it motivated the exploration of alternative approaches such as low-level UI events as the basis for contextualising on-device usage. It is likely that they will evolve even further in the future as device capabilities change, instead of there being a permanent solution.

7.1.2 Evaluation of Isolated and High-Level Features

Parts of this thesis were motivated by questions surrounding commonly used metrics in the literature (specifically RQ1). The assessment found isolated (oftentimes summative) features to be common metrics used in the literature (Chapter 2) for usage behaviour evaluation. However, with the rise of modern data-driven approaches and their capability to observe finer-grained device usage events, the effectiveness of these common metrics can provide more detail on device use compared to model usage.

For this, the thesis utilises the Tymer dataset which fulfills the various requirements for low-level feature capture. The Tymer dataset includes data collected over 8 weeks, with a reasonable amount of users (64) which generated a sizeable amount of session (>300,000). While this is comparable to or even more data than other studies in this research area, there are also others that collect data for hundreds or thousands of users, and collect data over much longer periods of time. On an individual level, these larger studies often collect data less frequently and with smaller individual datapoints, or data in a more compressed form. However, this still offers more diverse data over a longer period of time. The Tymer dataset is the largest (public) dataset that captures

data of this low-level nature (combined with user labels) and should largely be representative of user behaviour, but it is limited in size and diversity when it comes to the consideration of universal applications.

The analysis in Chapters 3 and 4 has shown that isolated features (of any level) do not suffice to accurately capture the complexity of user behaviour compared to a combination of features in a multi-modal model. To do so Chapter 4 focuses heavily on the features in isolation and their comparative effect sizes using multi-class pairwise tests. While other methods of analysis such as regression models would have been possible, the focus on these pairwise tests enables a direct comparison not just between the features within one model but lend itself to comparison when other vectorisations are introduced (such as added application categories). As such the outputs of Chapter 4 are purely meant to highlight the inconsistency of how frequently isolated features and are highlighted through a standardised effect size. Additionally, Chapter 5 highlighted the utility of low-level UI events in comparison to summative features. For two examples of groups of independent variables (e.g., smartphone addiction risk, impulsivity groups), TF-IDF transformed low-level features provided stronger predictive power compared to high-level summative features.

As discussed previously, low-level features as represented by UI events here should offer a better representation of the nuances in behaviour in the form of a users usage. This is because high-level features compress and misrepresent the actual usage. Reutilising TF-IDF in a non-NLP context enables a novel kind of 'importance' extraction for events within each session and its fixed vector output is ideal for many forms of follow up analysis. However, this method intrinsically also represents a form of compression. By using TF-IDF the direct temporal relationship between events within a single session is lost since it computes the relevancy of events for each session across all usage of users. While it is possible to encode timings and hesitation by introducing 'pause' pseudo-events, this is otherwise only circumventable by utilising n-grams which are computationally very expensive. It would also be possible to only focus on

transitions within each session through methods such as markov chains, but this loses the relationship to the rest of the 'corpus'. It appears that for to analyse usage of user behaviour there is balance for feature compression which moves between usability, computability and (mis-)characterisation.

The findings of the analysis applied with isolated and high-level features leads to the following contribution:

C1 Identification of issues with current common methods for representing user behavior, which tend to focus on single, isolated features and high-level characteristics.

7.1.3 The Behaviour-From-Usage-Stream Model

The limitations identified as part of C1 motivated the design of a generalisable model for usage processing as a basis for comparison between usage features. Chapter 3 introduced the BFUS model which acts as a framework for user behaviour modelling. The model assumes that users generate a stream of interactions with their device which can be used to explore their behaviour. It provides multiple steps to aid the construction of a user evaluation model while maintaining indifference towards individual parameters such as feature selection or possible applications. The abstraction is necessary to enable future compatibility with developments in device capabilities, data capture methods and advancements in usage behaviour research.

C2 The proposal of the Behaviour-From-Usage-Stream model represents a formal framework to process and evaluate user behaviour data.

7.1.4 Trait Prediction from Behaviour Stream

Within this thesis, the BFUS model was designed to utilise any factor of usage behaviour and relate it to independent variables. This enabled research questions that stable

user traits reflect on a user's behaviour and can be detected as such to be investigated. This is posed by RQ3 and tested in Chapters 4 and 5. Isolated features were not sufficiently capable to separate groups of different levels of addiction or impulsivity. However, a combination of features resulted in strong performance in predicting user trait groups for SA and impulsivity.

In this impulsivity and smartphone addiction were used as separate traits to demonstrate the models ability to adapt to more than one variable. The traits were determined using previously established surveys (the MCQ and SAS-SV respectively), but there are other methods that have been developed for them. The focus of this thesis was not to obtain a definition of how impulsivity and smartphone addiction influence the users behaviour but rather that user traits in general will show influences in the users usage habits. This was successful in that low-level features were able to predict classes of users based on the markers identified by these surveys. Furthermore, both traits are expected to encourage negative behaviours and could be construed to be too similar to be considered individual traits. While psychologically there may be overlaps in how this affects a person, the limited overlap between the highly affected users of SA and impulsivity gives confidence in the generalisability of applying the BFUS framework for applying to different user traits. This leads to the following thesis contribution:

C3 A case study of UI event based user behaviour capture being powerful enough to distinguish users based on psychological traits such as addiction or impulsivity.

7.1.5 Uncertainty and Interpretability Recommendations

Predictive tasks surrounding smartphone behaviour offers multiple candidates for viable evaluation methods. Chapter 6 expands this knowledge by supplying analysis-backed suggestions for topics such as feature selection, psychological interpretation, and a novel concept for decisions on the uncertain aspects of the user-session relationship. It presented an in-depth exploration into suggestions for the evaluation tools

which apply to predictive tasks and can be used in conjunction with the BFUS model.

It introduces the concept of leveraging the one-to-many relationship between users and sessions to deal with the uncertainty introduced by user trait labels. This uncertainty is originally created by assigning the users labels (SA and impulsivity) to each session they created individually. It is unlikely that every session they generated actually displayed their traits in a detectable manner. Therefore there is a form of base inconsistency in the labels when the sessions are evaluated. This also prevents really drilling down into how the behaviour caused by these traits affects session usage. After all it is not really possible to determine how the behaviour is correlated to the session interactions. The uncertainty threshold may contribute a method to reduce the sessions with low impact to the overall evaluation. It should be noted, that instead of identifying the root behaviours, this just removes cases which are hard to identify, in a usual scenario with correct labels this should be avoided. Furthermore, this also applies for the discussion in feasibility and trade-offs of utilising these predictive models of smartphone usage for psychological interpretations.

Extracting the psychological relationships when low-level features are transformed via TF-IDF is difficult, and relying on counts or screen-on time may be reasonable options when making inferences for audiences that need more explainable influences. However, as discussed this does introduce issues with compressing actual behaviour. When using transformed low-level features for prediction tasks and still desiring to understand the influences of the individual features better, it is possible to remove features to avoid collinearity issues while building the regression models. Therefore, feature correlations are used to build models that retain their input features (instead of utilising PCA) for collinearity issues. While there are other standard ways to accomplish this, such as a stepwise linear regression, such an approach would be very computationally expensive for the high amount of features and sessions in this situation. By addressing this issue, non-PCA transformed features enable a comparison between the coefficients of different features. This gives information to how user traits (such as

SA or impulsivity) have changed the usage in sessions. It should be noted that in the case of TF-IDF a direct comparison is still difficult because it obscures the meaning of "more" or "less". It can only inform of the relevancy of the feature and that in a given session it is impactful for a trait. In contrast, using the count of events gives a more natural, explainable result but comes at the cost of a worse performing prediction model.

7.2 Future Work

This thesis builds on the current body of literature surrounding smartphone behaviour research and addresses key limitations. The contributions outlined in 7.1 provide a basis for further exploration of understanding usage behaviour and the BFUS framework can be applied in various different contexts. This section explores how these concepts could be applied or extended in future work by shifting the context of their application.

7.2.1 Extension to Other Digital Devices and Multi-Device Use

Capture and processing of usage data are often concentrated on smartphones or other wearable devices because, for a large proportion of users, these are devices which they interface with constantly throughout the day. Simultaneous use of, or switching between multiple devices is a growing trend [106]. Using multiple devices enables ways of multitasking and context switching which are not possible to be achieved by a single device and likely creates interaction effects which are not well understood yet. This motivates the exploration of the effects stemming from multi-device usage.

These devices often also bridge personal and business use and may not be restricted to certain times of the day. In many ways, they are often ubiquitous to every aspect of life. This is reinforced by other devices such as laptops or personal computers which

represent a different category of digital devices altogether. Datasets for these devices are not as frequent or have enough detail to be considered useful for the transformation in this thesis. However, the general idea of collecting many low-level data points is not completely foreign as seen in the example of the Behacom dataset [127]. This is an opportunity for the research community to design studies generating suitable datasets of single and multi-device use with low-level interaction events.

Parallels between the usage of smartphones and desktop computers can be established. High-level features such as screen-on time, application switches or categories are very similar. Low-level features are also very comparable, all primary and most supplementary actions are well-defined in both cases and oftentimes directly applicable through some sort of semantic mapping. For example taps and clicks, scrolling with a finger or mouse wheel, and inputting text using the onscreen or an external keyboard.

The model framework as outlined in Chapter 3, enables flexibility in areas such as vectorisation and application such that it could theoretically be applied to other technologies, given the relative closeness of platforms in terms of interaction possibilities. There might be different parameters and features that need to be chosen as these devices to have different characteristics than smartphones such as portability, screen sizes or computational power that add additional layers of complexity for evaluating behaviour. Use cases such as multiple applications being used at the same time do not directly translate to the capabilities of most smartphones and therefore would require special consideration.

7.2.2 Robustness and Configuration of BFUS

When the BFUS model is introduced in Chapter 3 it is described with a three-step process (Bounding, Vectorisation and Application). In this, it formalises a process of user behaviour analysis that previously was left entirely to the researcher themselves.

Over the course of this thesis, its individual steps have been utilised and validated,

specifically with respect to Bounding where screen-to-screen sessions were created, Vectorisation in the form of utilising high-level summative features or UI-based features and Application with internal consistency and external variables in the form of SA and impulsivity. Choosing the method to use in each one of these steps was based on the surrounding literature and the assumptions made of the outcome. Screen event-based breakpoints for sessions, low-level event TF-IDF transformation and logistic regressions provided the best results for the Tymer data.

However, given the extreme variance in user behaviour data not only between-subject but also within-subject, these methods might not translate to every possible use-case; on top of there potentially being better methods being discovered in the future. The model provides a flexible framework for future research with alternative goals. Therefore there is a wide breadth of options to consider when choosing how to construct such a model. The aim of this section is to encourage the careful selection of those options and add to the scope of considerations.

7.2.2.1 Considerations of Alternative Model Parameters

As discussed in Section 3.3.1 **Bounding** can be approached in multiple ways. Various methods including screen event-based separation (e.g. screen on to off), cognitive timeouts (e.g. 45 seconds of idle) or time windows (e.g. day of the week or hour of the day) have been proposed before. In this thesis, screen event separation is used for bounding but timeouts are included to encompass multiple facets of the current state of research surrounding cognitive boundaries. Alternative methods of slicing the event stream (e.g., application launch boundaries) may be useful in other scenarios.

Vectorisation establishes the features that are available to relate to the outcome (e.g. independent variable). Commonly this has been simply screen-on time or similar high-level features. As part of this thesis, the use of transform low-level events in the form of TF-IDF has been explored.

NLP offers a lot of parallels with its various methods to transform the high variance in a string of strictly defined tokens. A different approach to TF-IDF is Word2Vec which instead of extracting the relevance of a token in a document, encodes the information of how different tokens relate to each other. This means that instead of the importance of the word itself, the semantic similarity to each other word is computed. In terms of natural words, this could be that the word “man” is the combination of matrices for “human” and “male” whereas “woman” combines “human” and “female”. In the space of NLP, this can be very useful information as it introduces semantic connections that were previously known to humans but not computers. This information could potentially be useful to relate to user interactions in sessions rather than words and sentences. The relationship of actions taken based on content or time of day could resolve new findings in the user behaviour space. Additionally, it encodes this information in a fixed-size vector matrix, which provides a flexible means for further processing.

Similarly, transition matrices are used to represent transitions between different states of a system. They are used to model the probability of transitioning from one state to another. They can also be used to predict the future states of a system and to determine the most likely paths between states. In the context of smartphone UI events, they can be difficult to tune because of the overwhelming amount of some events over others (e.g. scrolling). Markov chains can be used to visualise (and therefore contextualise) transition matrices easily which could be useful in scenarios where the state model is of interest.

TF-IDF as a bag-of-words method loses these relationships between words completely. Instead, it is possible to construct N-grams or apply sliding windows for usage within sessions to retain some of this information. However, this introduces an exponential growth of features to consider which might increase compute times beyond acceptable levels. It also increases the count of features significantly which means overfitting issues can arise depending on the data which is used to fit regression models.

For the analysis in this thesis, application categories were extracted from the app store

as-is. The reason for this was to retain accurate input data without adding a layer of potentially biased clustering of applications for the sake of the model. But Al-Subaihin et al. identified that categories from the app stores can be misleading because of a lack of granularity and specificity [5]. By transforming the input vectors via alternative measures to create categories for applications (e.g., clustered description text) it may be possible to increase the model's accuracy and interpretability.

Alternative approaches for **Application** include internal disambiguation methods, for example, one of the various other clustering algorithms. For example, hierarchical clustering as an unsupervised machine learning technique can be used to group sessions into clusters based on similarity (similar to the approach of K-Means). This approach allows for the creation of a hierarchical structure of clusters, which can be used to highlight how usage is connected. It can also highlight decision values which separate usage clusters from each other and could be used for further psychological interpretations of usage behaviour.

7.2.3 Practical Applications

In this thesis, the feasibility of detecting user traits using the BFUS model was demonstrated. This motivates constructing potential scenarios in which this technology could be applied with a real-world effect. Simultaneously, given that such systems will handle sensitive personal data it raises the need for consideration of potential ethical and privacy issues.

With the focus on problematic smartphone use, scenarios in a clinical setting in which such a system is extended to a platform used by professionals to evaluate patients' psychological profiles could be a research direction to explore. This would enable them to explore the efficacy of personal device usage data alongside a patient's development without any need for more invasive methods. For example, not only would it remove the need for repeated surveying of the user to track how treatment impacts them over

time.

Since on-device capabilities have evolved to the point where this kind of computation might also be feasible right 'in-the-moment'. With traits of a user being tracked as they use their device in everyday life. This could provide invaluable insight and feedback to users about their own habits and usage patterns that they might not be aware of.

Contextualising smartphone usage from UI interactions could also be explored for intervention-based applications (e.g., [60]) and could also prompt a user to reduce problematic patterns or inform them of how their behaviour reflects those issues.

7.2.3.1 Preventing Impulsive Use

The detection of sporadic or impulsive usage may also be of interest for user traits that are not specifically bound to smartphone usage. Certain aspects of user behaviour have previously been linked with issues in cyber security (e.g. [100, 2]). A user's behaviour can identify them, which can be useful in cases where it is desirable to know if the user of a device is its actual owner solely based on usage traits. This can be extended to a security feature which is called continuous authentication. In this, the behaviour of the current user is monitored constantly and checked against a learned pattern of the owner of the device [70, 86]. If the behaviour does not match the previously learned behaviour the user gets effectively locked out of using the device. While current approaches are not secure enough to replace traditional measures such as traditional biometrics or PIN completely, it is a promising additional security layer that is difficult to fake. The link between cyber security and impulsive characteristics in user behaviour and adding such warnings could be beneficial to protect against the breach of critical systems.

7.2.3.2 Suggestion and Recommendation Systems

One of the key areas of smartphone usage concerns itself with directly utilising certain information to generate a benefit for the user. The smartphone and its rich application

ecosystem enable many different tasks to be accomplished and users tend to individualise their phones to fit processes in their own lives. With dozens of applications installed on a single device, it can be difficult to organise the limited screen space. Previously, by observing the usual context and usage data it is possible to observe common switching patterns [149] and use those patterns to predict which apps users will likely seek to use next [11, 56]. This can be as simple as a drawer of frequently used applications that a user is most likely to return to but using approaches such as rule mining (Section 2.1.2) can enable more powerful, dynamic options. Instead of just displaying frequent applications this approach can show less frequent or more niche applications based on previous application sequences, the location and time or other context information.

Another aspect is maximising the longevity of a constantly draining battery. With strains on different components of the device (based on the applications that are being used), managing the available resources can be a valuable tool. Throttling or optimising the drain of processing and memory modules can be an effective strategy to minimise battery drain for specific tasks [133]. Combined with next-app prediction it has been shown that pre-loading expected application launches in the background can beneficially impact battery drain [164]. Also more generally applied, modern low-power modes already implement some of these ideas by restricting resources that reduce responsiveness to an extent but do not affect the user experience. For example, these can include limiting notification frequency, reducing network speeds or reducing background processes while the display is sleeping.

This has been possible through learning and adapting to usage in the form of time spent on the device and application switch patterns. It may be possible to also utilise the captured signals of low-level events to direct a users next move. This is could be useful in isolation (e.g., for better next app predictions), but also with additions of user labels such as SA to drift them away from other problematic or addiction-seeking usage patterns.

The evolution of the smartphone and future capabilities in sensing and computing

means that continued development surrounding these personal devices is likely. To further observe and understand how people use (and want to use) their smartphones will be relevant to drive this development.

7.3 Final Remarks

In this thesis, a new model for decomposing usage from behaviour in the form of an event stream was presented. Utilising this model in combination with various features showed their correlating strength with personal user traits. Usage at the low-level in the form of UI events inherited the strongest signals of those traits when compared with summative time-based features. This has shown the efficacy of utilising low-level interactions such as UI events as the basis for contextualising behaviours in smartphone usage.

Bibliography

[1] Assessing the Fit of the Model. In *Applied Logistic Regression*, chapter 5, pages 153–225. John Wiley & Sons, Ltd, 2013.

[2] Zahra Aivazpour and Zahra Ap. Impulsivity and Risky Cybersecurity Behaviors: A Replication. July 2019.

[3] Zahra Aivazpour and V. Srinivasan (Chino) Rao. Information Disclosure and Privacy Paradox: The Role of Impulsivity. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 51(1):14–36, January 2020.

[4] Yeslam Al-Saggaf, Rachel MacCulloch, and Karl Wiener. Trait Boredom Is a Predictor of Phubbing Frequency. *Journal of Technology in Behavioral Science*, 4(3):245–252, September 2019.

[5] A. A. Al-Subaihin, F. Sarro, S. Black, L. Capra, M. Harman, Y. Jia, and Y. Zhang. Clustering Mobile Apps Based on Mined Textual Features. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '16, pages 1–10, New York, NY, USA, September 2016. Association for Computing Machinery.

[6] Mohammed A. Alqarni, Sajjad Hussain Chauhdary, Maryam Naseer Malik, Muhammad Ehatisham-ul-Haq, and Muhammad Awais Azam. Identifying smartphone users based on how they interact with their phones. *Human-centric Computing and Information Sciences*, 10(1):7, February 2020.

[7] Ionut Andone, Konrad Błaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. How age and gender affect smartphone usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pages 9–12, New York, NY, USA, September 2016. Association for Computing Machinery.

[8] Ionut Andone, Konrad Blaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. Menthal: Quantifying smartphone usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pages 559–564, New York, NY, USA, September 2016. Association for Computing Machinery.

[9] Johan Andrén and Peter Funk. A Case-Based Approach Using Behavioural Biometrics to Determine a User's Stress Level. In *ICCBR Workshops*, pages 9–17, 2005.

[10] Sung-Man Bae. The relationship between the type of smartphone use and smartphone dependence of Korean adolescents_ National survey study. *Children and Youth Services Review*, page 5, 2017.

[11] Ricardo Baeza-Yates, Di Jiang, Fabrizio Silvestri, and Beverly Harrison. Predicting The Next App That You Are Going To Use. In *Proc. WSDM '15*, pages 285–294. ACM, February 2015.

[12] Nikola Banovic, Christina Brant, Jennifer Mankoff, and Anind Dey. ProactiveTasks: The short of mobile device use sessions. In *Proc. MobileHCI '14*, pages 243–252. ACM, September 2014.

[13] Nathaniel Barr, Gordon Pennycook, Jennifer A. Stolz, and Jonathan A. Fugelsang. The brain in your pocket: Evidence that Smartphones are used to supplant thinking. *Computers in Human Behavior*, 48(C):473–480, July 2015.

[14] Harald Baumeister and Christian Montag. Digital Phenotyping and Mobile Sensing in Psychoinformatics—A Rapidly Evolving Interdisciplinary Research Endeavor. In Christian Montag and Harald Baumeister, editors, *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*, Studies in Neuroscience, Psychology and Behavioral Economics, pages 1–9. Springer International Publishing, Cham, 2023.

[15] Marta Beranuy Fargues, Andrés Chamarro Lusar, Carla Graner Jordania, and Xavier Carbonell Sánchez. [Validation of two brief scales for Internet addiction and mobile phone problem use]. *Psicothema*, 21(3):480–485, August 2009.

[16] Bettina Berendt. AI for the Common Good?! Pitfalls, challenges, and ethics pen-testing. *Paladyn, Journal of Behavioral Robotics*, 10(1):44–65, January 2019.

[17] Nicole Blabst and Sarah Diefenbach. Whatsapp and wellbeing: A study on whatsapp usage, communication quality and stress. In *Proc. HCI '17*, pages 1–6. BCS Learning & Development Ltd., July 2017.

[18] Matthias Böhmer, Brent Hecht, Johannes Schöning, Antonio Krüger, and Gernot Bauer. Falling asleep with Angry Birds, Facebook and Kindle: A large scale study on mobile application usage. In *Proc. MobileHCI '11*, pages 47–56. ACM, August 2011.

[19] Michael Borenstein, Larry V Hedges, Julian P T Higgins, and Hannah R Rothstein. Converting among effect sizes. In *Introduction to Meta-Analysis*, pages 45–49. Wiley,, Chichester :, 2009.

[20] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, September 1995.

[21] Anne-Linda Camerini, Tiziano Gerosa, and Laura Marciano. Predicting problematic smartphone use over time in adolescence: A latent class regression analysis of online and offline activities. *New Media & Society*, page 1461444820948809, August 2020.

[22] Daniel Chen and Roel Vertegaal. Using mental load for managing interruptions in physiologically attentive user interfaces. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, pages 1513–1516, New York, NY, USA, April 2004. Association for Computing Machinery.

[23] Matteo Ciman and Katarzyna Wac. Individuals' Stress Assessment Using Human-Smartphone Interaction Analysis. *IEEE Transactions on Affective Computing*, 9(1):51–65, January 2018.

[24] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, Hillsdale, N.J, 2nd ed edition, 1988.

[25] Tao Deng, Shaheen Kanthawala, Jingbo Meng, Wei Peng, Anastasia Kononova, Qi Hao, Qinhao Zhang, and Prabu David. Measuring smartphone usage and task switching with log tracking and self-reports. *Mobile Media & Communication*, 7(1):3–23, January 2019.

[26] Xiang Ding, Jing Xu, Guanling Chen, and Chenren Xu. Beyond Smartphone Overuse: Identifying Addictive Mobile Apps. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 2821–2828, New York, NY, USA, May 2016. Association for Computing Machinery.

[27] Mark R. Dixon, Janice Marley, and Eric A. Jacobs. Delay Discounting by Pathological Gamblers. *Journal of Applied Behavior Analysis*, 36(4):449–458, 2003.

[28] Trinh Minh Tri Do, Jan Blom, and Daniel Gatica-Perez. Smartphone usage in the wild: A large-scale analysis of applications and context. In *Proc. ICMI '11*, pages 353–360. ACM, November 2011.

[29] Trinh Minh Tri Do and Daniel Gatica-Perez. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing*, 12:79–91, June 2014.

[30] Jon D. Elhai, Jason C. Levine, Robert D. Dvorak, and Brian J. Hall. Fear of missing out, need for touch, anxiety and depression are related to problematic smartphone use. *Computers in Human Behavior*, 63:509–516, October 2016.

[31] Jon D. Elhai, Jason C. Levine, Robert D. Dvorak, and Brian J. Hall. Non-social features of smartphone use are most related to depression, anxiety and problematic smartphone use. *Computers in Human Behavior*, 69:75–82, April 2017.

[32] Jon D. Elhai, Haibo Yang, Dmitri Rozgonjuk, and Christian Montag. Using machine learning to model problematic smartphone use severity: The significant role of fear of missing out. *Addictive Behaviors*, 103:106261, April 2020.

[33] Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, and Deborah Estrin. Diversity in smartphone usage. In *Proc. MobiSys '10*, pages 179–194. ACM, June 2010.

[34] Denzil Ferreira, Jorge Goncalves, Vassilis Kostakos, Louise Barkhuus, and Anind K. Dey. Contextual experience sampling of mobile application micro-usage. In *Proc. MobileHCI '14*, pages 91–100. ACM, September 2014.

[35] Björn Friedrichs, Liam D. Turner, and Stuart M. Allen. Discovering Types of Smartphone Usage Sessions from User-App Interactions. In *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops)*, pages 459–464, March 2021.

[36] Björn Friedrichs, Liam D. Turner, and Stuart M. Allen. Utilising the co-occurrence of user interface interactions as a risk indicator for smartphone addiction. *Pervasive and Mobile Computing*, page 101677, August 2022.

[37] Surjya Ghosh, Kaustubh Hiware, Niloy Ganguly, Bivas Mitra, and Pradipta De. Emotion detection from touch interactions during text entry on smartphones. *International Journal of Human-Computer Studies*, 130:47–57, October 2019.

[38] Andrej Gisbrecht and Barbara Hammer. Data visualization by nonlinear dimensionality reduction. *WIREs Data Mining and Knowledge Discovery*, 5(2):51–73, 2015.

[39] Şahin Gökçearslan, Filiz Kuşkaya Mumcu, Tülin Haşlaman, and Yasemin Demiraslan Çevik. Modelling smartphone addiction: The role of smartphone usage, self-regulation, general self-efficacy and cyberloafing in university students. *Computers in Human Behavior*, 63:639–649, October 2016.

[40] Sian Gooding, Yevgeni Berzak, Tony Mak, and Matt Sharifi. Predicting Text Readability from Scrolling Interactions. *arXiv:2105.06354 [cs]*, November 2021.

[41] Charles Gouin-Vallerand and Neila Mezghani. An analysis of the transitions between mobile application usages based on markov chains. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 373–378, New York, NY, USA, September 2014. Association for Computing Machinery.

[42] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, April 1982.

[43] Marian Harbach, Alexander De Luca, and Serge Egelman. The Anatomy of Smartphone Unlocking: A Field Study of Android Lock Screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*,

CHI '16, pages 4806–4817, New York, NY, USA, May 2016. Association for Computing Machinery.

[44] Jeffrey Heer and Ed H. Chi. Separating the swarm: Categorization methods for user sessions on the web. In *Proc. CHI '02*, pages 243–250. ACM, April 2002.

[45] Daniel Hintze, Rainhard D. Findling, Muhammad Muaaz, Sebastian Scholz, and René Mayrhofer. Diversity in locked and unlocked mobile device usage. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 379–384, New York, NY, USA, September 2014. Association for Computing Machinery.

[46] Daniel Hintze, Rainhard D. Findling, Sebastian Scholz, and René Mayrhofer. Mobile Device Usage Characteristics: The Effect of Context and Form Factor on Locked and Unlocked Usage. In *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia*, pages 105–114, Kaohsiung Taiwan, December 2014. ACM.

[47] Ke Huang, Chunhui Zhang, Xiaoxiao Ma, and Guanling Chen. Predicting mobile application usage using contextual information. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 1059–1065, New York, NY, USA, September 2012. Association for Computing Machinery.

[48] Zohra Ihsan and Adrian Furnham. The new technologies in personality assessment: A review. *Consulting Psychology Journal: Practice and Research*, 70(2):147–166, 2018.

[49] Chakajkla Jesdabodi and Walid Maalej. Understanding usage states on mobile devices. In *Proc. UbiComp '15*, pages 1221–1225. ACM, September 2015.

[50] Zhao Xia Jin, Tom Plocher, and Liana Kiff. Touch Screen User Interfaces for Older Adults: Button Size and Spacing. volume 4554 of *LNCS*, pages 933–941. Springer, 2007.

[51] You Jin Jeong, Bongwon Suh, and Gahgene Gweon. Is smartphone addiction different from Internet addiction? comparison of addiction-risk factors among adolescents. *Behaviour & Information Technology*, 39(5):578–593, May 2020.

[52] Simon L. Jones, Denzil Ferreira, Simo Hosio, Jorge Goncalves, and Vassilis Kostakos. Revisitation analysis of smartphone app use. In *Proc. UbiComp '15*, pages 1197–1208. ACM, September 2015.

[53] Hyunjin Kang and Wonsun Shin. Do Smartphone Power Users Protect Mobile Privacy Better than Nonpower Users? Exploring Power Usage as a Factor in Mobile Privacy Protection and Disclosure. *Cyberpsychology, Behavior, and Social Networking*, 19(3):179–185, January 2016.

[54] Joon-Myung Kang, Sin-seok Seo, and James Won-Ki Hong. Usage pattern analysis of smartphones. In *Proc. APNOMS'11*, pages 1–8, September 2011.

[55] Juuso Karikoski and Tapio Soikkeli. Contextual usage patterns in smartphone communication services. *Personal and Ubiquitous Computing*, 17(3):491–502, March 2013.

[56] Katerina Katsarou, Geunhye Yu, and Felix Beierle. WhatsNextApp: LSTM-Based Next-App Prediction With App Usage Sequences. *IEEE Access*, 10:18233–18247, 2022.

[57] David J. Ketchen and Christopher L. Shook. The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*, 17(6):441–458, 1996.

[58] Hyun-Jun Kim and Young Sang Choi. Exploring emotional preference for smartphone applications. In *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, pages 245–249, January 2012.

[59] Jaejeung Kim, Hayoung Jung, Minsam Ko, and Uichin Lee. GoalKeeper: Exploring Interaction Lockout Mechanisms for Regulating Smartphone Use. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):16:1–16:29, March 2019.

[60] Jaejeung Kim, Joonyoung Park, Hyunsoo Lee, Minsam Ko, and Uichin Lee. LocknType: Lockout Task Intervention for Discouraging Smartphone App Use. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12. Association for Computing Machinery, New York, NY, USA, May 2019.

[61] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. Understanding eye movements on mobile devices for better presentation of search results. *Journal of the Association for Information Science and Technology*, 67(11):2607–2619, 2016.

[62] Kris Kirby, Nancy Petry, and Warren Bickel. Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *Journal of experimental psychology. General*, 128:78–87, April 1999.

[63] Kris N. Kirby. Bidding on the future: Evidence against normative discounting of delayed rewards. *Journal of Experimental Psychology: General*, 126(1):54–70, 1997.

[64] Kris N. Kirby and Nancy M. Petry. Heroin and cocaine abusers have higher discount rates for delayed rewards than alcoholics or non-drug-using controls. *Addiction*, 99(4):461–471, 2004.

[65] Erez Kita and Gil Luria. Differences between males and females in the prediction of smartphone use while driving: Mindfulness and income. *Accident Analysis & Prevention*, 140:105514, June 2020.

[66] A. Kołakowska. A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *2013 6th International Conference on Human System Interactions (HSI)*, pages 548–555, June 2013.

[67] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, April 2013.

[68] Vassilis Kostakos, Denzil Ferreira, Jorge Goncalves, and Simo Hosio. Modelling smartphone usage: A markov state transition model. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 486–497, New York, NY, USA, September 2016. Association for Computing Machinery.

[69] Damjan Krstajic, Ljubomir Buturovic, Simon Thomas, and David E. Leahy. Binary classification models with "Uncertain" predictions. *arXiv:1711.09677 [stat]*, page 1, December 2017.

[70] Rajesh Kumar, Vir V. Phoha, and Abdul Serwadda. Continuous authentication of smartphone users by fusing typing, swiping, and phone movement patterns. In *Proc. BTAS'16*, pages 1–8, September 2016.

[71] Min Kwon, Dai-Jin Kim, Hyun Cho, and Soo Yang. The Smartphone Addiction Scale: Development and Validation of a Short Version for Adolescents. *PLOS ONE*, 8(12):e83558, December 2013.

[72] Min Kwon, Joon-Yeop Lee, Wang-Youn Won, Jae-Woo Park, Jung-Ah Min, Changtae Hahn, Xinyu Gu, Ji-Hye Choi, and Dai-Jin Kim. Development and Validation of a Smartphone Addiction Scale (SAS). *PLOS ONE*, 8(2):e56936, February 2013.

[73] R. Lambiotte and M. Kosinski. Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12):1934–1939, 2014.

[74] Johannes A. Landsheer. Impact of the Prevalence of Cognitive Impairment on the Accuracy of the Montreal Cognitive Assessment: The Advantage of Using two MoCA Thresholds to Identify Error-prone Test Scores. *Alzheimer Disease and Associated Disorders*, 34(3):248, July 2020.

[75] Juha Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Bornet, T.-M.-T Do, Olivier Dousse, Julien Eberle, and Markus Miettinen. The mobile data challenge: Big data for mobile computing research. Nokia Research Center. January 2012.

[76] MyungSuk Lee, MuMoungCho Han, and JuGeon Pak. Analysis of Behavioral Characteristics of Smartphone Addiction Using Data Mining. *Applied Sciences*, 8(7):1191, July 2018.

[77] Tong Li, Tong Xia, Huandong Wang, Zhen Tu, Sasu Tarkoma, Zhu Han, and Pan Hui. Smartphone App Usage Analysis: Datasets, Methods, and Applications. *IEEE Communications Surveys & Tutorials*, 24(2):937–966, 2022.

[78] Yuefeng Li, Libiao Zhang, Yue Xu, Yiyu Yao, Raymond Yiu Keung Lau, and Yutong Wu. Enhancing Binary Classification by Modeling Uncertain Boundary in Three-Way Decisions. *IEEE Transactions on Knowledge and Data Engineering*, 29(7):1438–1451, July 2017.

[79] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. MoodScope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '13, pages 389–402. ACM, June 2013.

[80] Soo Ling Lim, Peter J. Bentley, Natalie Kanakam, Fuyuki Ishikawa, and Shinichi Honiden. Investigating Country Differences in Mobile App User Behavior and Challenges for Software Engineering. *IEEE Transactions on Software Engineering*, 41(1):40–64, January 2015.

[81] Chun-Hao Liu, Sheng-Hsuan Lin, Yuan-Chien Pan, and Yu-Hsuan Lin. Smartphone gaming and frequent use pattern associated with smartphone addiction. *Medicine*, 95(28):e4068, July 2016.

[82] Chung-Chu Liu. Mobile phone user types by Q methodology: An exploratory research. *International Journal of Mobile Communications*, 6(1):16–31, December 2008.

[83] Alex Lloyd, Ryan McKay, Todd K. Hartman, Benjamin T. Vincent, Jamie Murphy, Jilly Gibson-Miller, Liat Levita, Kate Bennett, Orla McBride, Anton P. Martinez, Thomas V. A. Stocks, Frédérique Vallières, Philip Hyland, Thanos Karatzias, Sarah Butter, Mark Shevlin, Richard P. Bentall, and Liam Mason. Delay discounting and under-valuing of recent information predict poorer adherence to social distancing measures during the COVID-19 pandemic. *Scientific Reports*, 11(1):19237, September 2021.

[84] Karina Loid, Karin Täht, and Dmitri Rozgonjuk. Do pop-up notifications regarding smartphone use decrease screen time, phone checking behavior, and self-reported problematic smartphone use? Evidence from a two-month experimental study. *Computers in Human Behavior*, 102:22–30, January 2020.

[85] Haiping Ma, Huanhuan Cao, Qiang Yang, Enhong Chen, and Jilei Tian. A habit mining approach for discovering similar mobile users. In *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web*, pages 231–240, April 2012.

[86] Upal Mahbub, Jukka Komulainen, Denzil Ferreira, and Rama Chellappa. Continuous Authentication of Smartphones Based on Application Usage. *IEEE*

Transactions on Biometrics, Behavior, and Identity Science, 1(3):165–180, July 2019.

[87] Eric Malmi and Ingmar Weber. You Are What Apps You Use: Demographic Prediction Based on User's Apps, February 2016.

[88] Alexander Markowetz, Konrad Błaszkiewicz, Christian Montag, Christina Switala, and Thomas E. Schlaepfer. Psycho-Informatics: Big Data shaping modern psychometrics. *Medical Hypotheses*, 82(4):405–411, April 2014.

[89] Aleksandar Matic, Martin Pielot, and Nuria Oliver. Boredom-computer interaction: Boredom proneness and the use of smartphone. In *Proc. UbiComp '15*, UbiComp '15, pages 837–841. ACM, September 2015.

[90] Abhinav Mehrotra, Fani Tsapeli, Robert Hendley, and Mirco Musolesi. MyTraces: Investigating Correlation and Causation between Users' Emotional States and Mobile Phone Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):83:1–83:21, September 2017.

[91] Francisco Melo. Area under the ROC Curve. In Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, editors, *Encyclopedia of Systems Biology*, pages 38–39. Springer, New York, NY, 2013.

[92] John V. Monaco, Ned Bakelman, Sung-Hyuk Cha, and Charles C. Tappert. Developing a Keystroke Biometric System for Continual Authentication of Computer Users. In *2012 European Intelligence and Security Informatics Conference*, pages 210–216, August 2012.

[93] Christian Montag. The Neuroscience of Smartphone/Social Media Usage and the Growing Need to Include Methods from 'Psychoinformatics'. In Fred D. Davis, René Riedl, Jan vom Brocke, Pierre-Majorique Léger, and Adriane B. Randolph, editors, *Information Systems and Neuroscience*, Lecture Notes in Information Systems and Organisation, pages 275–283, Cham, 2019. Springer International Publishing.

[94] Christian Montag, Harald Baumeister, Christopher Kannen, Rayna Sariyska, Eva-Maria Meßner, and Matthias Brand. Concept, Possibilities and Pilot-Testing of a New Smartphone Application for the Social and Life Sciences to

Study Human Behavior Including Validation Data from Personality Psychology. *J*, 2(2):102–115, June 2019.

[95] Christian Montag, Konrad Błaszkiewicz, Bernd Lachmann, Ionut Andone, Rayna Sariyska, Boris Trendafilov, Martin Reuter, and Alexander Markowetz. Correlating personality and actual phone usage: Evidence from psychoinformatics. *Journal of Individual Differences*, 35(3):158–165, 2014.

[96] Christian Montag, Konrad Błaszkiewicz, Bernd Lachmann, Rayna Sariyska, Ionut Andone, Boris Trendafilov, and Alexander Markowetz. Recorded Behavior as a Valuable Resource for Diagnostics in Mobile Phone Addiction: Evidence from Psychoinformatics. *Behavioral Sciences*, 5(4):434–442, December 2015.

[97] Christian Montag, Éilish Duke, and Alexander Markowetz. Toward Psychoinformatics: Computer Science Meets Psychology. *Computational and Mathematical Methods in Medicine*, 2016:e2983685, June 2016.

[98] Diana Moreira and Fernando Barbosa. Delay discounting in impulsive behavior: A systematic review. *European Psychologist*, 24(4):312–321, 2019.

[99] Aske Mottelson and Kasper Hornbæk. An affect detection technique using mobile commodity sensors in the wild. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 781–792, New York, NY, USA, September 2016. Association for Computing Machinery.

[100] Ahmed A. Moustafa, Abubakar Bello, and Alana Maurushat. The Role of User Behaviour in Improving Cyber Security Management. *Frontiers in Psychology*, 12, 2021.

[101] Abhishek Mukherji, Vijay Srinivasan, and Evan Welbourne. Adding intelligence to your mobile device via on-device sequential pattern mining. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 1005–1014, New York, NY, USA, September 2014. Association for Computing Machinery.

[102] Sarfraz Nawaz and Cecilia Mascolo. Mining users' significant driving routes with low-power sensors. In *Proceedings of the 12th ACM Conference on Em-*

bedded Network Sensor Systems, SenSys '14, pages 236–250, New York, NY, USA, November 2014. Association for Computing Machinery.

[103] Beryl Noë, Liam D. Turner, David E. J. Linden, Stuart M. Allen, Gregory R. Maio, and Roger M. Whitaker. Timing rather than user traits mediates mood sampling on smartphones. *BMC Research Notes*, 10(1):481, September 2017.

[104] Beryl Noë, Liam D. Turner, David E. J. Linden, Stuart M. Allen, Bjorn Winkens, and Roger M. Whitaker. Identifying Indicators of Smartphone Addiction Through User-App Interaction. *Computers in Human Behavior*, 99:56–65, 2019.

[105] Beryl Noë, Liam D. Turner, and Roger M. Whitaker. Smartphone interaction and survey data as predictors of snapchat usage. In *Proc. UbiComp/ISWC '19 Adjunct*, pages 438–445. ACM, September 2019.

[106] Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K. Dey, and Hideyuki Tokuda. Reducing users' perceived mental effort due to interruptive notifications in multi-device mobile environments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 475–486, New York, NY, USA, September 2015. Association for Computing Machinery.

[107] Adam J. Oliner, Anand P. Iyer, Ion Stoica, Eemil Lagerspetz, and Sasu Tarkoma. Carat: Collaborative energy diagnosis for mobile devices. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, SenSys '13, pages 1–14, New York, NY, USA, November 2013. Association for Computing Machinery.

[108] Jukka-Pekka Onnela. Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology*, 46(1):45–54, January 2021.

[109] Alvaro Ortigosa, Rosa M. Carro, and José Ignacio Quiroga. Predicting user personality by mining social interactions in Facebook. *Journal of Computer and System Sciences*, 80(1):57–71, February 2014.

[110] Antti Oulasvirta, Tye Rattenbury, Lingyi Ma, and Eeva Raita. Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing*, 16(1):105–114, January 2012.

[111] Antti Oulasvirta, Sakari Tamminen, Virpi Roto, and Jaana Kuorelahti. Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile HCI. In *Proc. CHI '05*, pages 919–928. ACM, April 2005.

[112] Panagiotis Papapetrou and George Roussos. Social context discovery from temporal app use patterns. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 397–402, New York, NY, USA, September 2014. Association for Computing Machinery.

[113] Abhinav Parate, Matthias Böhmer, David Chu, Deepak Ganesan, and Benjamin M. Marlin. Practical prediction and prefetch for faster access to applications on mobile phones. In *Proc. UbiComp '13*, pages 275–284. ACM, September 2013.

[114] Jihwan Park, Jo-Eun Jeong, and Mi Jung Rho. Predictors of Habitual and Addictive Smartphone Behavior in Problematic Smartphone Use. *Psychiatry Investigation*, 18(2):118–125, February 2021.

[115] Joonyoung Park, Jin Yong Sim, Jaejeung Kim, Mun Yong Yi, and Uichin Lee. Interaction Restraint: Enforcing Adaptive Cognitive Tasks to Restrain Problematic User Interaction. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, pages 1–6, New York, NY, USA, April 2018. Association for Computing Machinery.

[116] Jim H. Patton, Matthew S. Stanford, and Ernest S. Barratt. Factor structure of the barratt impulsiveness scale. *Journal of Clinical Psychology*, 51(6):768–774, 1995.

[117] Ella Peltonen, Eemil Lagerspetz, Jonatan Hamberg, Abhinav Mehrotra, Mirco Musolesi, Petteri Nurmi, and Sasu Tarkoma. The hidden image of mobile apps: Geographic, demographic, and cultural factors in mobile usage. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '18, pages 1–12, New York, NY, USA, September 2018. Association for Computing Machinery.

[118] Xixian Peng. Investigating user switching intention for mobile instant messaging application: Taking WeChat as an example. *Computers in Human Behavior*, page 11, 2016.

[119] Charlie Pinder, Jo Vermeulen, Adhi Wicaksono, Russell Beale, and Robert J. Hendley. If this, then habit: Exploring context-aware implementation intentions on smartphones. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '16, pages 690–697, New York, NY, USA, September 2016. Association for Computing Machinery.

[120] Erika Pivetta, Lydia Harkin, Joel Billieux, Eiman Kanjo, and Daria J. Kuss. Problematic smartphone use: An empirically validated model. *Computers in Human Behavior*, 100:105–117, November 2019.

[121] A. Rahmati, C. Shepard, C. Tossell, L. Zhong, and P. Kortum. Practical Context Awareness: Measuring and Utilizing the Context Dependency of Mobile Usage. *IEEE Transactions on Mobile Computing*, 14(09):1932–1946, August 2012.

[122] Ahmad Rahmati and Lin Zhong. A Longitudinal Study of Non-Voice Mobile Phone Usage by Teens from an Underserved Urban Community. *arXiv:1012.2832 [cs]*, abs/1012.2832, December 2010.

[123] Timothy Regan, Bethany Harris, Matthew Van Loon, Namrata Nanavaty, Jordan Schueler, Solangia Engler, and Sherecce A. Fields. Does mindfulness reduce the effects of risk factors for problematic smartphone use? Comparing frequency of use versus self-reported addiction. *Addictive Behaviors*, 108:106435, September 2020.

[124] J.A. Roberts, L.H.P. Yaya, and C. Manolis. The invisible addiction: Cell-phone activities and addiction among male and female college students. *Journal of Behavioral Addictions*, 3(4):254–265, 2014.

[125] Mohammad Salehan and Arash Negahban. Social networking on smartphones: When mobile phones become addictive. *Computers in Human Behavior*, 29(6):2632–2639, November 2013.

[126] Pedro Miguel Sánchez Sánchez, Jose María Jorquera Valero, Alberto Huertas Celdrán, Gérôme Bovet, Manuel Gil Pérez, and Gregorio Martínez Pérez. A Survey on Device Behavior Fingerprinting: Data Sources, Techniques, Application Scenarios, and Datasets. *arXiv:2008.03343 [cs]*, page 1, August 2020.

[127] Pedro M. Sánchez Sánchez, José M. Jorquera Valero, Mattia Zago, Alberto Huertas Celdrán, Lorenzo Fernández Maimó, Eduardo López Bernal, Sergio López Bernal, Javier Martínez Valverde, Pantaleone Nespoli, Javier Pastor Galindo, Ángel L. Perales Gómez, Manuel Gil Pérez, and Gregorio Martínez Pérez. BEHACOM - a dataset modelling users' behaviour in computers. *Data in Brief*, 31:105767, August 2020.

[128] Tim Schulz van Endert. Addictive use of digital devices in young children: Associations with delay discounting, self-control and academic performance. *PLOS ONE*, 16:e0253058, June 2021.

[129] Clayton Shepard, Ahmad Rahmati, Chad Tossell, Lin Zhong, and Phillip Kortum. LiveLab: Measuring wireless networks and smartphone users in the field. *ACM SIGMETRICS Performance Evaluation Review*, 38(3):15–20, January 2011.

[130] Choonsung Shin and Anind K. Dey. Automatically detecting problematic use of smartphones. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 335–344, New York, NY, USA, September 2013. Association for Computing Machinery.

[131] Choonsung Shin, Jin-Hyuk Hong, and Anind K. Dey. Understanding and prediction of mobile application usage for smart phones. In *Proc. UbiComp '12*, pages 173–182. ACM, September 2012.

[132] Ben Shneiderman. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4):26:1–26:31, October 2020.

[133] Alex Shye, Benjamin Scholbrock, and Gokhan Memik. Into the wild: Studying real user activity patterns to guide power optimizations for mobile architectures. In *Proc. MICRO '09*, pages 168–178, December 2009.

[134] Chris Smith-Clarke and Licia Capra. Beyond the Baseline: Establishing the Value in Mobile Phone Based Poverty Estimates. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 425–434, Republic and Canton of Geneva, CHE, April 2016. International World Wide Web Conferences Steering Committee.

[135] Sei Yon Sohn, Philippa Rees, Bethany Wildridge, Nicola J. Kalk, and Ben Carter. Prevalence of problematic smartphone usage and associated mental health outcomes amongst children and young people: A systematic review, meta-analysis and GRADE of the evidence. *BMC Psychiatry*, 19(1):356, November 2019.

[136] Tapio Soikkeli, Juuso Karikoski, and Heikki Hammainen. Diversity and End User Context in Smartphone Usage Sessions. In *Proc. NGMAST 2011*, pages 7–12, September 2011.

[137] Vijay Srinivasan, Saeed Moghaddam, Abhishek Mukherji, Kiran K. Rachuri, Chenren Xu, and Emmanuel Munguia Tapia. MobileMiner: Mining your frequent patterns on your phone. In *Proc. UbiComp '14*, pages 389–400. ACM, September 2014.

[138] SrinivasanVijay, KoehlerChristian, and JinHongxia. RuleSelector: Selecting Conditional Action Rules from User Behavior Paterns. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, March 2018.

[139] Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D. Gosling, Gabriella M. Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, Heinrich Hussmann, Bernd Bischl, and Markus Bühner. Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30):17680–17687, July 2020.

[140] C. Henrico Stam, Frederik M. van der Veen, and Ingmar H. A. Franken. Individual differences in time estimation are associated with delay discounting and alcohol use. *Current Psychology*, July 2020.

[141] TalkingData. TalkingData Mobile User Demographics. https://kaggle.com/competitions/talkingdata-mobile-user-demographics, May 2017.

[142] Chang Tan, Qi Shuai Liu, Enhong Chen, and Hui Xiong. Prediction for Mobile Application Usage Patterns. Nokia MDC Workshop '12, page 1, 2012.

[143] Subrata Tikadar, Sharath Kazipeta, Chandrakanth Ganji, and Samit Bhattacharya. A Minimalist Approach for Identifying Affective States for Mobile

Interaction Design. In Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler, editors, *Human-Computer Interaction - INTERACT 2017*, Lecture Notes in Computer Science, pages 3–12, Cham, 2017. Springer International Publishing.

[144] John Torous, Mathew V Kiang, Jeanette Lorme, and Jukka-Pekka Onnela. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Mental Health*, 3(2):e16, May 2016.

[145] Chad Tossell, Philip Kortum, Clayton Shepard, Ahmad Rahmati, and Lin Zhong. Exploring Smartphone Addiction: Insights from Long-Term Telemetric Behavioral Measures. 9(2):7, 2015.

[146] Huawei Tu, Xiangshi Ren, Feng Tian, and Feng Wang. Evaluation of Flick and Ring Scrolling on Touch-Based Smartphones. *International Journal of Human–Computer Interaction*, 30(8):643–653, August 2014.

[147] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. Push or Delay? Decomposing Smartphone Notification Response Behaviour. In Albert Ali Salah, Ben J.A. Kröse, and Diane J. Cook, editors, *Human Behavior Understanding*, LNCS, pages 69–83. Springer International Publishing, 2015.

[148] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. The influence of concurrent mobile notifications on individual responses. *International Journal of Human-Computer Studies*, 132:70–80, December 2019.

[149] Liam D. Turner, Roger M. Whitaker, Stuart M. Allen, David E. J. Linden, Kun Tu, Jian Li, and Don Towsley. Evidence to support common application switching behaviour on smartphones. *Royal Society Open Science*, 6(3):190018, 2019.

[150] Niels van Berkel, Simon D'Alfonso, Rio Kurnia Susanto, Denzil Ferreira, and Vassilis Kostakos. AWARE-Light: A smartphone tool for experience sampling and digital phenotyping. *Personal and Ubiquitous Computing*, November 2022.

[151] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys*, 50(6):93:1–93:40, December 2017.

[152] Niels van Berkel, Chu Luo, Theodoros Anagnostopoulos, Denzil Ferreira, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. A Systematic Assessment of Smartphone Usage Gaps. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4711–4721, San Jose California USA, May 2016. ACM.

[153] Steven Van Canneyt, Marc Bron, Andy Haines, and Mounia Lalmas. Describing Patterns and Disruptions in Large Scale Mobile App Usage Data. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1579–1584, Republic and Canton of Geneva, CHE, April 2017. International World Wide Web Conferences Steering Committee.

[154] Jan Van den Bulck. Adolescent Use of Mobile Phones for Calling and for Sending Text Messages After Lights Out: Results from a Prospective Cohort Study with a One-Year Follow-Up. *Sleep*, 30(9):1220–1223, September 2007.

[155] Tim Schulz van Endert and Peter N. C. Mohr. Likes and impulsivity: Investigating the relationship between actual smartphone use and delay discounting. *PLOS ONE*, 15(11):e0241383, November 2020.

[156] Aku Visuri, Zhanna Sarsenbayeva, Jorge Goncalves, Evangelos Karapanos, and Simon Jones. Impact of mood changes on application selection. In *Proc. UbiComp '16*, pages 535–540. ACM, September 2016.

[157] Aku Visuri, Zhanna Sarsenbayeva, Niels van Berkel, Jorge Goncalves, Reza Rawassizadeh, Vassilis Kostakos, and Denzil Ferreira. Quantifying Sources and Types of Smartwatch Usage Sessions. In *Proc. CHI '17*, pages 3569–3581. ACM, May 2017.

[158] Liang Wang and Xin Geng, editors. *Behavioral Biometrics for Human Identification: Intelligent Applications*. IGI Global, 2010.

[159] Hongyi Wen, Michael Sobolev, Rachel Vitale, James Kizer, J. P. Pollak, Frederick Muench, and Deborah Estrin. mPulse Mobile Sensing Model for Passive Detection of Impulsive Behavior: Exploratory Prediction Study. *JMIR Mental Health*, 8(1):e25019, January 2021.

[160] Stephen P. Whiteside and Donald R. Lynam. The Five Factor Model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences*, 30(4):669–689, March 2001.

[161] Lea-Christin Wickord and Claudia Michaela Quaiser-Pohl. Does the Type of Smartphone Usage Behavior Influence Problematic Smartphone Use and the Related Stress Perception? *Behavioral Sciences (Basel, Switzerland)*, 12(4):99, April 2022.

[162] Henry H. Wilmer and Jason M. Chein. Mobile technology habits: patterns of association among device usage, intertemporal preference, impulse control, and reward sensitivity. *Psychonomic Bulletin & Review*, 23(5):1607–1614, October 2016.

[163] Annelot Wismans, Srebrenka Letina, Karl Wennberg, Roy Thurik, Rui Baptista, Andrew Burke, Marcus Dejardin, Frank Janssen, Enrico Santarelli, Olivier Torrès, and Ingmar Franken. The role of impulsivity and delay discounting in student compliance with COVID-19 protective measures. *Personality and Individual Differences*, 179:110925, September 2021.

[164] Tingxin Yan, David Chu, Deepak Ganesan, Aman Kansal, and Jie Liu. Fast app launching for mobile devices using predictive user context. In *Proc. MobiSys '12*, pages 113–126. ACM, June 2012.

[165] Haibo Yang, Zihao Wang, and Jon D. Elhai. The relationship between adolescent stress and problematic smartphone use: The serial mediating effects of anxiety and frequency of smartphone use. *Current Psychology*, June 2022.

[166] W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[167] Jiadi Yu, Haofu Han, Hongzi Zhu, Yingying Chen, Jie Yang, Yanmin Zhu, Guangtao Xue, and Minglu Li. Sensing Human-Screen Interaction for Energy-Efficient Frame Rate Adaptation on Smartphones. *IEEE Transactions on Mobile Computing*, 14(8):1698–1711, August 2015.

[168] Lin Liu Eric Yu. From Requirements to Architectural Design –Using Goals and Scenarios. *Proc. Workshop From Software Requirements to Architectures*, 1, May 2001.

[169] Sha Zhao, Julian Ramos, Jianrong Tao, Ziwen Jiang, Shijian Li, Zhaohui Wu, Gang Pan, and Anind K. Dey. Discovering different kinds of smartphone users through their application usage behaviors. In *Proc. UbiComp '16*, pages 498–509. ACM, September 2016.

[170] Sha Zhao, Feng Xu, Zhiling Luo, Shijian Li, and Gang Pan. Demographic Attributes Prediction Through App Usage Behaviors on Smartphones. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18, pages 870–877, New York, NY, USA, October 2018. Association for Computing Machinery.

[171] Sha Zhao, Feng Xu, Yizhi Xu, Xiaojuan Ma, Zhiling Luo, Shijian Li, Anind Dey, and Gang Pan. Investigating smartphone user differences in their application usage behaviors: An empirical study. *CCF Transactions on Pervasive Computing and Interaction*, 1(2):140–161, August 2019.

[172] Jonathan J. H. Zhu, Hexin Chen, Tai-Quan Peng, Xiao Fan Liu, and Haixing Dai. How to measure sessions of mobile phone use? Quantification, evaluation, and applications. *Mobile Media & Communication*, 6(2):215–232, May 2018.

[173] Sijie Zhuo, Lucas Sherlock, Gillian Dobbie, Yun Sing Koh, Giovanni Russello, and Danielle Lottridge. Real-time Smartphone Activity Classification Using Inertial Sensors—Recognition of Scrolling, Typing, and Watching Videos While Sitting or Walking. *Sensors*, 20(3):655, January 2020.

Appendices

A Tymer Demographics

Age	**M**	**SD**
Years	25.4	5.87
Gender	**N**	**%**
Male	34	53.13
Female	30	46.88
Employment	**N**	**%**
Student	38	59.38
Student & employed	13	20.31
Employed	12	18.75
Unemployed	1	1.56
Education	**N**	**%**
High school, no diploma	1	1.56
High school diploma or equiv.	5	7.81
Trade/technical/vocational training	1	1.56
Undergrad education, no degree	14	21.88
Bachelor's degree	19	29.69
Master's degree	21	32.81
Doctorate	2	3.13
No answer	1	1.56

Table A1: Demographics of the Tymer dataset. Table taken from [104].

B K-means elbow

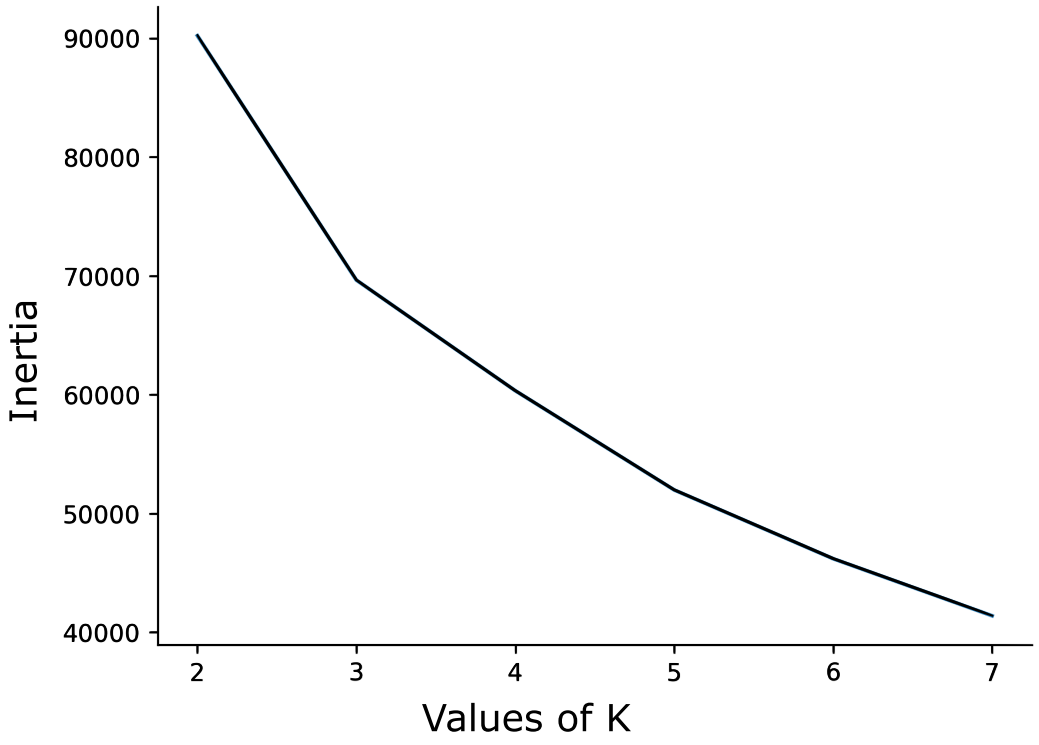


Figure A1: The elbow method of choosing K for K-Means. The number of clusters can be chosen by observing the inertia (sum of squared error per cluster) elbow. This elbow was created using clustered values of TF-IDF transformed user interaction events within the Tymer dataset.

C Application Categories

Category	Apps	Sessions	Category	Apps	Sessions
Books	29	1,097	Game sports	11	296
Business	25	720	Game strategy	10	833
Comics	2	5	Game trivia	12	48
Communication	57	103,658	Game word	12	170
Dating	2	47	Health and fitness	51	4,620
Education	31	574	House and home	2	39
Entertainment	57	1,187	Launcher	23	130,438
Finance	29	1,212	Libraries and demo	1	1
Food and drink	7	17	Lifestyle	69	19,037
Game action	15	77	Maps and navigation	29	1,220
Game adventure	8	6,823	Personalization	10	1,299
Game arcade	38	1,033	Photography	64	12,427
Game board	2	63	Productivity	102	17,234
Game card	5	19	Shopping	31	44,237
Game casual	24	1,656	Sports	16	528
Game educational	5	5	Tools	130	38,486
Game music	1	5	Travel and local	58	6,803
Game puzzle	42	320	Video players	33	4,975
Game racing	8	9	Weather	14	1,051
Game simulation	12	340	Web browser	5	4,818
None used	n/a	81,451	Other	442	175,441

Table A2: A list of all fetched application categories from the Google Play Store for the Tymer study. Including how many applications were captured for each category and in how many sessions (out of 301,024) the category appeared.

D Full Coefficient Tables

Category	Feature	p	Coef	[.05	.95]
Launcher	App switch	<0.001	-2.852135	-2.951985	-2.752285
Health and fitness	Short idle	<0.001	-2.671645	-2.939358	-2.403931
None	Text input	<0.001	-1.978947	-2.070161	-1.887734
Weather	Short idle	<0.001	-1.678748	-2.272526	-1.084969
Communication	Long idle	<0.001	-1.647544	-1.815843	-1.479245
Personalization	Short idle	0.001	-1.435098	-2.212702	-0.657494
Web browser	Short idle	<0.001	-1.413694	-1.682605	-1.144783
Social	Long idle	<0.001	-1.401252	-1.617531	-1.184974
Tools	Tap	<0.001	-1.381696	-1.671575	-1.091817
Productivity	Short idle	<0.001	-1.36148	-1.509132	-1.213828
None	Tap	<0.001	-1.200717	-1.27141	-1.130023
Lifestyle	Short idle	<0.001	-1.168593	-1.300682	-1.036505
Travel and local	Short idle	<0.001	-1.152798	-1.318996	-0.986599
Tools	Text input	<0.001	-1.083707	-1.342573	-0.82484
Video players	Short idle	<0.001	-0.991284	-1.273419	-0.709149
Tools	Long idle	<0.001	-0.965766	-1.303246	-0.628286
Game strategy	Short idle	0.029	-0.922575	-1.565069	-0.280082
Communication	App switch	<0.001	-0.915468	-1.076192	-0.754744
Video players	Scrolling	0.003	-0.914673	-1.45676	-0.372586
Maps and navigation	Short idle	0.001	-0.795469	-1.196938	-0.394
Entertainment	Short idle	0.021	-0.716507	-1.190453	-0.24256
Video players	Long idle	0.005	-0.715964	-1.175242	-0.256687
Productivity	Long idle	<0.001	-0.703567	-1.048754	-0.35838
Sports	Short idle	0.017	-0.676605	-1.120425	-0.232785
Video players	App switch	0.042	-0.670972	-1.254391	-0.087553
Music and audio	Long idle	0.006	-0.651788	-1.061487	-0.242089
News and magazines	Long idle	0.008	-0.640459	-1.085302	-0.195616
Lifestyle	Long idle	0.001	-0.611132	-0.948748	-0.273515
Productivity	Tap	0.005	-0.603145	-1.007512	-0.198777
Social	Tap	<0.001	-0.602403	-0.785716	-0.419091
Productivity	Text input	0.001	-0.575255	-0.900907	-0.249604
Photography	Short idle	<0.001	-0.554558	-0.671584	-0.437533

Communication	Tap	<0.001	-0.438293	-0.573774	-0.302812
Photography	Long idle	0.011	-0.389208	-0.659241	-0.119175
News and magazines	Short idle	<0.001	-0.353949	-0.482596	-0.225303
None	App switch	<0.001	-0.30404	-0.378253	-0.229828
News and magazines	Tap	0.039	-0.303973	-0.575135	-0.03281
Game adventure	Long idle	0.048	-0.256363	-0.479046	-0.03368
Social	Scrolling	<0.001	-0.20872	-0.317118	-0.100323
Communication	Text input	<0.001	-0.158144	-0.220712	-0.095575
Communication	Short idle	<0.001	0.177391	0.115099	0.239682
None	Short idle	<0.001	0.256269	0.202861	0.309677
Music and audio	Scrolling	0.039	0.298947	0.054167	0.543726
Productivity	Scrolling	0.014	0.301837	0.065896	0.537778
Tools	Scrolling	0.001	0.342677	0.135093	0.55026
Lifestyle	Tap	0.027	0.385634	0.051834	0.719434
None	Scrolling	<0.001	0.447581	0.359525	0.535638
None	Long tap	0.002	0.466842	0.179694	0.75399
Social	Short idle	<0.001	0.468252	0.383618	0.552885
Books and reference	Short idle	0.001	0.609201	0.286925	0.931477
Game adventure	Short idle	<0.001	0.630984	0.482078	0.77989
Travel and local	Scrolling	0.005	0.646952	0.252508	1.041397
Lifestyle	Text input	0.006	0.694262	0.217876	1.170648
Social	Text input	<0.001	0.697928	0.579194	0.816663
Communication	Scrolling	<0.001	0.699164	0.608512	0.789816
News and magazines	App switch	<0.001	0.792517	0.448465	1.13657
Web browser	Text input	0.001	0.808868	0.363443	1.254293
Lifestyle	Scrolling	<0.001	0.816276	0.526001	1.106552
Photography	App switch	<0.001	1.000054	0.73123	1.268879
Launcher	Short idle	<0.001	1.022458	0.951079	1.093837
Social	App switch	<0.001	1.026912	0.827227	1.226597
Music and audio	Tap	<0.001	1.072903	0.868518	1.277287
Game casual	App switch	0.004	1.30914	0.472058	2.146223
Game casual	Short idle	<0.001	1.533177	1.194198	1.872157
Game casual	Long idle	<0.001	1.795045	1.362459	2.22763
News and magazines	Scrolling	<0.001	1.869467	1.639643	2.099292
Game adventure	App switch	<0.001	2.433656	2.127719	2.739593

Productivity	App switch	<0.001	2.456935	2.164005	2.749866
Launcher	Scrolling	<0.001	2.912109	2.758879	3.065339
Tools	Short idle	<0.001	4.026339	3.928404	4.124275
Lifestyle	App switch	<0.001	5.093477	4.771957	5.414997

Table A3: All statistically significant coefficients for TF-IDF.

Category	**Feature**	**p**	**Coef**	**[.05**	**.95]**
Personalization	App switch	<0.001	-0.579053	-0.71728	-0.440826
Tools	App switch	<0.001	-0.465619	-0.492801	-0.438438
Weather	Short idle	<0.001	-0.406019	-0.567119	-0.244918
News and magazines	Long idle	<0.001	-0.331893	-0.392763	-0.271023
Video players	App switch	<0.001	-0.244471	-0.299519	-0.189424
Tools	Long idle	<0.001	-0.19405	-0.221682	-0.166418
Launcher	App switch	<0.001	-0.172302	-0.180678	-0.163926
Education	Tap	0.01	-0.129564	-0.214334	-0.044793
Launcher	Long idle	<0.001	-0.128969	-0.152479	-0.105459
None	Tap	<0.001	-0.12749	-0.132916	-0.122063
Game strategy	Short idle	<0.001	-0.120857	-0.176054	-0.06566
Productivity	Long idle	<0.001	-0.119671	-0.1534	-0.085942
Productivity	Tap	<0.001	-0.099738	-0.118549	-0.080928
Entertainment	Tap	0.019	-0.09506	-0.165237	-0.024884
Music and audio	Long idle	0.001	-0.063736	-0.097913	-0.02956
Shopping	Tap	0.029	-0.062576	-0.111723	-0.013428
Tools	Tap	<0.001	-0.043201	-0.052964	-0.033438
Social	Long idle	<0.001	-0.042443	-0.053233	-0.031653
Lifestyle	Long idle	0.001	-0.03643	-0.058229	-0.01463
Communication	App switch	<0.001	-0.031257	-0.040238	-0.022276
Lifestyle	Tap	0.032	-0.021648	-0.040915	-0.002381
None	Long idle	<0.001	-0.015664	-0.020103	-0.011225
Video players	Scrolling	<0.001	-0.015222	-0.020549	-0.009895
Communication	Long idle	<0.001	-0.014741	-0.020594	-0.008888
News and magazines	Tap	<0.001	-0.013668	-0.018175	-0.00916
Game puzzle	Short idle	0.002	-0.013008	-0.020592	-0.005423
Communication	Tap	<0.001	-0.012208	-0.015555	-0.008861
Travel and local	Short idle	<0.001	-0.004756	-0.005733	-0.003779

Web browser	Scrolling	<0.001	-0.00465	-0.006474	-0.002825
Shopping	Scrolling	0.024	-0.002876	-0.004954	-0.000799
Sports	Short idle	0.024	-0.002679	-0.004505	-0.000852
Productivity	Short idle	<0.001	-0.002312	-0.002821	-0.001804
Maps and navigation	Short idle	0.032	-0.001853	-0.00322	-0.000487
Health and fitness	Short idle	0.022	-0.001654	-0.002752	-0.000555
Entertainment	Short idle	0.001	-0.001551	-0.002375	-0.000727
Web browser	Short idle	<0.001	-0.001409	-0.001764	-0.001054
Books and reference	Short idle	0.019	-0.0014	-0.002402	-0.000398
Lifestyle	Short idle	<0.001	-0.001046	-0.00136	-0.000732
Photography	Short idle	0.001	-0.001046	-0.001581	-0.000511
Video players	Short idle	0.004	-0.000694	-0.00113	-0.000258
Social	Scrolling	<0.001	-0.000293	-0.000409	-0.000177
Communication	Short idle	<0.001	-0.0002	-0.000266	-0.000134
Social	Short idle	<0.001	-0.000093	-0.00014	-0.000047
None	Short idle	<0.001	0.000265	0.000152	0.000378
News and magazines	Short idle	<0.001	0.000296	0.000194	0.000398
Communication	Scrolling	<0.001	0.000317	0.000269	0.000365
Productivity	Scrolling	0.047	0.000352	0.000016	0.000688
Books and reference	Scrolling	0.047	0.001198	0.000163	0.002233
None	Text input	<0.001	0.001579	0.001079	0.002079
News and magazines	Scrolling	<0.001	0.0026	0.00177	0.003429
Tools	Scrolling	<0.001	0.002709	0.001963	0.003456
Social	Text input	<0.001	0.00296	0.002138	0.003781
Launcher	Short idle	<0.001	0.003018	0.002343	0.003693
Productivity	Text input	<0.001	0.003222	0.00197	0.004475
None	Scrolling	<0.001	0.003699	0.00297	0.004428
Music and audio	Scrolling	0.001	0.004752	0.002113	0.007392
Tools	Short idle	<0.001	0.007442	0.006983	0.007901
Photography	Scrolling	0.002	0.008993	0.003819	0.014166
Lifestyle	Scrolling	<0.001	0.011081	0.007666	0.014495
None	App switch	<0.001	0.019199	0.015539	0.022858
Shopping	Text input	0.024	0.020104	0.005519	0.034688
Travel and local	Tap	0.002	0.020444	0.008524	0.032363
Lifestyle	Text input	<0.001	0.021797	0.012687	0.030906

Travel and local	Long idle	<0.001	0.032581	0.020852	0.04431
Tools	Text input	<0.001	0.03316	0.026834	0.039486
Video players	Tap	0.009	0.035068	0.010563	0.059573
Game casual	Long idle	<0.001	0.037317	0.024722	0.049911
Web browser	Text input	<0.001	0.04221	0.03345	0.050971
Web browser	Tap	<0.001	0.052714	0.036198	0.069231
Launcher	Scrolling	<0.001	0.062883	0.057055	0.06871
Maps and navigation	Tap	0.032	0.06371	0.015872	0.111548
Music and audio	App switch	0.001	0.078418	0.037717	0.119119
Education	Text input	<0.001	0.101673	0.070412	0.132935
Game sports	Long idle	<0.001	0.105913	0.072087	0.139739
Game puzzle	Long idle	<0.001	0.118627	0.070878	0.166376
Launcher	Tap	<0.001	0.118687	0.104569	0.132805
Entertainment	Text input	<0.001	0.14762	0.085966	0.209275
Social	App switch	<0.001	0.153962	0.142574	0.165349
Books and reference	App switch	0.019	0.156014	0.046999	0.265029
Sports	App switch	0.028	0.179583	0.043892	0.315275
News and magazines	App switch	<0.001	0.18348	0.153661	0.213299
None	Long tap	0.014	0.190074	0.038633	0.341515
Game arcade	App switch	<0.001	0.217473	0.133394	0.301552
Game adventure	App switch	<0.001	0.268609	0.241875	0.295343
Finance	App switch	<0.001	0.294922	0.203135	0.38671
Productivity	App switch	<0.001	0.346977	0.322771	0.371184
Lifestyle	App switch	<0.001	0.666628	0.629541	0.703714
Game casual	App switch	<0.001	0.821195	0.735187	0.907202

Table A4: All statistically significant coefficients for event counts.

E Smartphone Addiction Scale

#	Items
1	Missing planned work due to smartphone use.
2	Having a hard time concentrating in class, while doing assignments, or while working due to smartphone use.
3	Feeling pain in the wrists or at the back of the neck while using a smartphone.
4	Won't be able to stand not having a smartphone.
5	Feeling impatient and fretful when I am not holding my smartphone.
6	Having my smartphone in my mind even when I am not using it.
7	I will never give up using my smartphone even when my daily life is already greatly affected by it.
8	Constantly checking my smartphone so as not to miss conversations between other people on Twitter or Facebook.
9	Using my smartphone longer than I had intended.
10	The people around me tell me that I use my smartphone too much.

Table A5: The final questions of the SAS-SV which are asked on a 6-point Likert scale. The possible answers were 'Strongly disagree', 'Disagree', 'Weakly disagree', 'Weakly agree', 'Agree' and 'Strongly agree' [71]

F Monetary Choice Questionnaire

#	Items	k
1	Would you prefer $54 today, or $55 in 117 days?	.00016
2	Would you prefer $55 today, or $75 in 61 days?	.006
3	Would you prefer $19 today, or $25 in 53 days?	.006
4	Would you prefer $31 today, or $85 in 7 days?	.25
5	Would you prefer $14 today, or $25 in 19 days?	.041
6	Would you prefer $47 today, or $50 in 160 days?	.0004
7	Would you prefer $15 today, or $35 in 13 days?	.10
8	Would you prefer $25 today, or $60 in 14 days?	.10
9	Would you prefer $78 today, or $80 in 162 days?	.00016
10	Would you prefer $40 today, or $55 in 62 days?	.006
11	Would you prefer $11 today, or $30 in 7 days?	.25
12	Would you prefer $67 today, or $75 in 119 days?	.001
13	Would you prefer $34 today, or $35 in 186 days?	.00016
14	Would you prefer $27 today, or $50 in 21 days?	.041
15	Would you prefer $69 today, or $85 in 91 days?	.0025
16	Would you prefer $49 today, or $60 in 89 days?	.0025
17	Would you prefer $80 today, or $85 in 157 days?	.0004
18	Would you prefer $24 today, or $35 in 29 days?	.016
19	Would you prefer $33 today, or $80 in 14 days?	.1
20	Would you prefer $28 today, or $30 in 179 days?	.0004
21	Would you prefer $34 today, or $50 in 30 days?	.016
22	Would you prefer $25 today, or $30 in 80 days?	.0025
23	Would you prefer $41 today, or $75 in 20 days?	.041
24	Would you prefer $54 today, or $60 in 111 days?	.001
25	Would you prefer $54 today, or $80 in 30 days?	.016
26	Would you prefer $22 today, or $25 in 136 days?	.001
27	Would you prefer $20 today, or $55 in 7 days?	.25

Table A6: The questions of the monetary choice questionnaire. Participants have the option to either choose the 'smaller reward today' or the 'larger reward in the specified number of days'. k represents the estimated rate of discounting. [62]