Technische Universität Berlin

Institut für Softwaretechnik und Theoretische Informatik

Quality and Usability Lab

Bachelor Thesis

Predicting personality traits from touchscreen interactions

By

Ludwig Küster

Matriculation number: 361453

Course of study: B.Sc. Wirtschaftsingenieurwesen

Apes with exploratory tendencies may spend more time around and manipulating an experimental apparatus, which may enhance learning simply because these individuals spend more time with the task, rather than because they exhibit greater cognitive ability.

"Chimpanzee intellect: personality, performance and motivation with touchscreen tasks" (Altschul et al. 2017)

Picture: *Curious George*, by Margret and Hans Rey

Credits: *My thank goes to Dr.-Ing. Jan-Niklas Voigt-Antons for his supervision and advice. I thank Carola Trahms and Giorgio Colombo for sharing their code and data, which made this thesis possible.*

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilde und ausschließlich unter Verwendung der Aufgeführten Quellen und Hilfsmittel angefertigt habe.

_____ _____

Datum Unterschrift

Content

Introduction

A characteristic feature of our time is the shifting of life from analog to digital spheres. Daily tasks and joys are infused by data driven technologies. This leads to improved efficiency and access to unseen opportunities: We can move around our cities in shared-cars, rent the apartments of strangers on the other side of the globe or simply stay in touch with old school friends in an unprecedented ease. Hand in hand with these enrichments comes another novel phenomenon: Our daily movements are chiselled into an endless record, stored on servers without capacity limits. This data allows accurate behaviour predictions to be extracted by those who own it. A Donald Duck conceived today would jump into a Money Bin filled with digital footsteps.

Shocked by Trump's victory in 2016, the liberal world rubbed their eyes in disbelief. A few months later, the swiss newspaper *Das Magazin* offered a simple explanation: it is the data's fault. The editors, Grassegger and Krogerus, traced a psychometric method used by the Trump team to psychologist Michal Kosinskia, the man who created the method a few years before. Kosinskia applied classification methods to connect the publicly available "Likes" of Facebook-users to the five-factor personality model, the Big-5. The model captures most of the individual differences in human personality. By analysing the personalities of liberals and conservatives, Carney et al (2008) found that liberals are more open-minded, creative, and novelty seeking, whereas conservatives are more orderly, conventional and better organized. Building on this research, Kosinskia used logistic and linear regression to predict a person's sexuality and ethnicity solely from their Facebook Likes with around 90 percent accuracy. For the Big-5 factor of "Openness", the accuracy reached levels close to the test–retest accuracy of a standard personality test. Trump's teams had allegedly used these methods to make personality-profiles for millions of US citizens and influence them by targeted advertisement via Facebook. Kosinksia claims that Clinton and internet companies also use these methods and are "just not so stupid to talk about it" (Kosinksia, 2017). Further, the narrative of *Das Magazin* has been refuted by Trump's data-team themselves, Cambridge Analytica, admitting that they had overplayed their role in the election and that only conventional statistical methods were used (Golem,

2017). However, the notion that our data-driven era brings new levels of personally tailored content persists. Designers, journalists and developers will make use of these tools, predicting one's taste and creating adaptive content just the way we like it.

Different data sources have been used to infer the five-factor information. These methods require (a) an extensive use of the device and (b) comprehensive access to personal information. By contrast, touchscreen data can be easily obtained by applications. In this way, developers could inform adaptive content. It would also serve the needs of designers. Their representative 'personas' (model customers) can be more powerful if used to complement, not replace, a full range of quantitative and qualitative methods (Pruitt et al. 2003). To date, there is no publication of personality prediction from the touch modality.

The first chapter will present the concept of the Big-5 model and summarise related research from the fields of phone-use-based personality predictions, touchscreen-based emotion detection and user-authentication via touch. It concludes by a discussion of all touchscreen-based metrics that appeared in the literature. Chapter two will cover the feature and data aggregation, and describe the data of two previous studies on which this thesis is built. The third chapter will analyse the distribution of the Big-5 factors and reveal the different classifiers that were applied to the data. Results will be presented for the algorithms Automatic-Linear-Modelling, Support-Vector-Machine, Random-Forest, K-Nearest Neighbours, Naive Bayes and Logistic Regression. The paper will conclude with ideas for improving the classification accuracy.

Related work

This chapter will present the concept of the five-factor model ("Big 5"). Existing publications have not covered research on prediction of personality traits from touchscreen data. To gather best-practice approaches, three adjacent fields will therefore be explored. The touchscreen-metrics that are used in the screened data will be discussed.

Personality assessment via the Big-5 model

The five-factor model, commonly referred to as the Big-5, is one of the most influential models in psychology (Vinciarelli, 2013). It uses a lexical analysis to group a large body of adjectives that people use to describe each other or themselves, and condense them into a preferably small number of umbrella-terms. These terms are regarded the underlying dimensions which make up the model. Since the 1960s, researches have repeatedly arrived at five factors but it was not until 1993 that Lewis Goldberg coined the name, "Big 5". The adjectives are grouped on continuous bipolar scales which encompass the dimensions Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism, summarised in a mnemonic trick as OCEAN. An example for related adjectives on both ends of the scales can be found in Table 1 - OCEAN dimensions. The left side of the table lists adjectives which are associated with high levels, the right side shows low manifestations. For example, a person with high levels of Openness would be described as imaginative.

Table 1 - OCEAN dimensions

High	dimension	Low
imaginative, independent, interested	**Openness**	practical, conforming, routines
organized, careful, disciplined	**Conscientiousness**	disorganized, careless, impulsive
sociable, fun-loving, affectionate	**Extraversion**	retiring, somber, reserved
softhearted, trusting, helpful	**Agreeableness**	ruthless, suspicious, uncooperative
Anxious, insecure, self-pitying	**Neuroticism**	calm, secure, self-satisfied

As a hierarchical model of personality traits, it represents personality at the broadest level of abstraction, yet is believed to capture most of the individual differences in human personality (Gosling, 2003). Psychologists generally agree that the five-factor model has some predictive power for future behaviour, although the extent to which situations play into prediction remains controversial. A meta-analysis (Barrick, Mount, 1991) has shown correlations around of r=30 for conscientiousness and professional success, while Openness correlates with educational achievement. Conscientiousness is in systematic relation to life-expectancy, most likely due to an inclination towards health enhancing behaviour.

Mobile-phone based prediction of personality

Since the omnipresence of smartphones and their rich sensor data, a multitude of works have focused on using social-interaction (phone-usage) data, proximity data (via Bluetooth) and information about which apps were used on the phone. Valid personality predictions have even been carried out based on mere phone-logs (itemised bills).

Chittaranjan et al (2011) found that several aggregated features obtained from smartphone usage data can be indicators of the Big-5 traits. A software collected anonymized logs of calls, SMS Logs, Bluetooth scans (allows to compute the number of Bluetooth devices in proximity to the user) and app-usage. They first noted significant correlations between their data and personality-metrics, then they used a sequential backward feature selection algorithm and an SVM classifier (RBF) in a binary classification task. The median in each dimension was used to split the target date into two classes. Their classification reached a performance accuracy between 0.54 and 0.59.

Montjoye et al (2013) computed a novel set of indicators based on mobile-phone logs, made available by phone-companies. They divided every dimension of the Big-5 into three classes (low, average, high values). Their model can predict Big-5 dimensions with an accuracy of 0.61 on this three-class prediction problem. At the time, this was the most accurate prediction based on phone usage. For data, they used basic-phone

use (text and calls), the degree of activity in the use-behaviour (time till a text is answered), the user's locations (number of places from which calls have been made), regularity (frequency of usage) and diversity (number of interaction by number of contacts ratio).

Monsted et al. (2016) collected data from the smartphones of 730 individuals. In general, only extraversion and to some degree, neuroticism, could be predicted by smartphone usage patterns, whereas the remaining personality traits of the Big-5 had limited predictability. Again, a three-class prediction system was used which amounted to a predictability increase of 0.11 above random. The data was akin to the studies. Eventually, they investigated alternatives to the Big-5 scoring system in a factor analysis of the questions underlying the Big-5 inventory.

Predicting emotion from touch

In games, the chance to adapt difficulty or decide evolution of gameplay based on the current emotional state of the user is of great value. Extraversion and Neuroticism have the closest link to emotions: Whilst extraversion is a strong predictor of positive emotions, neuroticism is a strong predictor of negative emotions (Gomez, 2002). In person to person communication and acted scenarios, it could be shown that with touch-behaviour as the only communication modality, not only valence but also the type of an emotion can be decoded by the receiving person (Hertenstein et al. 2009). After this finding, a strand of research strived to build emotion prediction models from touch.

For example, Shah et al (2015) based their research on touchscreen behaviour, with a special focus on swiping and tapping. In an experiment, users are asked to complete a task. Previously, a minimum number of necessary steps for solving a task was defined. The deviation from this optimal path is measured as a feature, along with measuring the speed of touch-movements. In their experimental setting, they invoked the desired moods through presentation of videos prior to experiment.

Two levels of arousal and two levels of valence are used in Gao et al's 2012 study. In addition, four emotional states (excitement, relaxation, frustration, boredom) are measured. The results show that a differentiation between four emotional states is possible with 0.69-0.77 accuracy. Two levels of arousal and two levels of valence could be classified with 0.89 accuracy. Fruit Ninja (the test app game) was played twenty times with a target point to reach and each session lasting 30 seconds. The difficulty of the game increased throughout the course of the experiment. The finger stroke behaviour is tracked throughout the game. After each session, subjects do a self-assessment through a list of emotional words. The words were grouped according to the quadrants of the valence-arousal space. Gao et al find that the pressure feature differentiates frustration states from the other three states. Stroke speed and directness features distinguishes between different levels of arousal whilst stroke length features separates mainly boredom from the relaxed state.

Touch-screen based user authentication

In pursuit of research that infers features from touchscreen, the field of touchscreen-based user authentication provided a rich body of work. This field presents continuous user identification as a behavioural biometric and generally shows that touch-behaviour can be regarded as an invariant feature of human behaviour. As authentication techniques require a very precise classification, great attention is paid to an extensive feature set.

Frank et al's (2013) pioneering research asks whether a classifier can at all continuously authenticate users and based on their touch interactions. 30 behavioural features are extracted from touchscreen logs and data, collected from smartphone usage. They reach 0.02-0.03 accuracy for intra-sessional authentication. Frank points to the fact that distinctive emotional states might manifest in a user' touch. They distinguish between slower swipes and faster flicks (the fast scroll-movement known from reading on tablets). The complete swipe trajectory is stored. This allows for calculating the deviation from the straight line that hypothetically connects start and end of each swipe. If the users are not given a specific task, they tend to use distinctive screen areas for their strokes.

Zheng et al 2014) reveal that users display unique tapping patterns, with a different strength, rhythm and

angle of the finger movements. The touch-screen input is complemented by acceleration data (which is proportional to the tapping force applied) and angular acceleration. Here, touch size (a scaled value of approximate size of given pointer) is extracted from the screen. This touch-size point is related both to the general size of the finger (called baseline size) and the tapping pressure which differs throughout the usage.

Discussion of possible touch-screen based metrics

Both touch-based user-authentication and emotion prediction rely heavily on a set of meaningful input variables. A smart aggregation of touch-screen raw-data into higher-level metrics is the pivotal element which eventually divides high from low classification scores. This section aims to provide an overview of which features were used in aforementioned studies. The table allows a comparison between which kinds of raw data are recorded from user-interactions.

Table 2 - Feature overview compares the metrics of four papers that entail a comprehensive description of their feature sets. The list does not claim completeness about documenting which single feature was used in which single study, but it serves more as a general collection of metrics. New metrics were added to the collection in the order that they were found in the articles. All approaches compute basic numbers from the touchscreen logs. The finger-down, move and finger-up events are always recorded with (x, y) coordinates and a time-stamp. It is worth noting that there is no perfect tap, as there is always a small movement by the finger. A strike is therefore distinguished from a tap by a certain distance of up and down point. From this data, the number of touches and strokes as well as the length and speed of a stroke can be calculated.

The studies are different in the degree to which a user-interaction is predefined. Some have completely tailored tasks and compare the user's solution to some minimum-number-of-interaction benchmark, which was theoretically determined prior to the study. This approach seems promising in a laboratory setting with a test-app that offers only a limited set of essential functions. But the closer such scenario gets to real world conditions, the more ambiguous is a minimum benchmark. On a standard phone, there is a multitude of equally correct ways of e.g. searching a word in the web. Given multiple valid paths a minimum-number of steps is not so meaningful.

Task based apps like Pfelgetab or Zengh's Ninja-Turtle offer performance metrics as well as delay- and completion times which can be mapped on to the condition of the respective game.

Gao and Frank store the complete touch trajectory of swipes. This allows Frank et al. to identify distinctive areas on the screen, which are preferred by different users. They also calculate the speed at the start and end of a swipe. Therefore, the five fist and last touch-coordinate recording of a swipe are analysed. The complete trajectory also allows for measuring the deviation of a theoretical straight line from touch-start and touch-target, the stroke-directness-index.

Table 2 - Feature overview

	Gao	*Shah*	*Zengh*	*Frank*	*Pflege-Tab*
number of strokes / touches	x	x	x	x	x
stroke length	x	x	x	x	x
stroke speed	x	x	x	x	x
(average) delay: time lag between completion of one task and start of subsequent one.		x			x
time to complete a task		x			x
deviation from target point					x
touch duration					x
touch point size	x		x		
trajectory stored	x			x	
stroke Directness Index (DI)	x			x	
number of strikes/taps (for predefined task)		x			
angular acceleration			x		
pressure			x		
acceleration (proportional to tapping force applied)			x		
median velocity of last / first five points					
distinctive screen area for stroke				x	

Regarding emotion and personality, one other feature seems promising: The size of the finger-tip which is used for the touch interaction. Before the game starts the dimension of the fingertip is incorporated by

computing a baseline touch-area. Deviations from this area are then a good indicator for the touch-pressure exerted by the user. It is plausible to suppose that depending on a person's impulsivity, reactions to challenges which appear during the interaction can be captured by the pressure parameter. An impulsive person might press a non-reacting button with increasing firmness, while a controlled person might not vary pressure. This is sketched in *Figure 2 - Touch sequence.* The studies account for this phenomenon by receiving the touch-point-size from operational-system APIs, via the accelerometer or in the case of PflegeTab with touch-duration as a proxy.

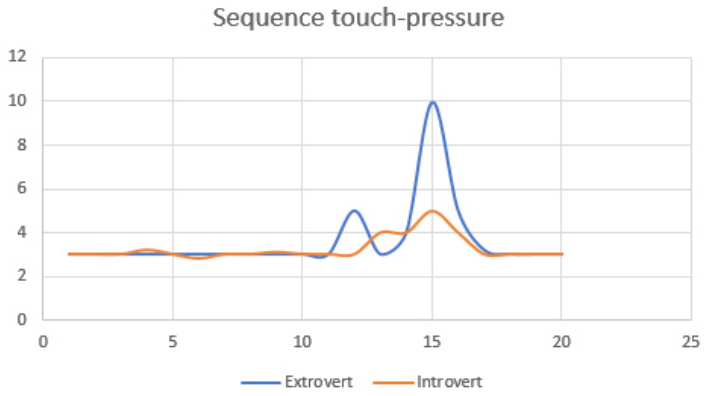


Figure 2 - Touch sequence

These assumed characteristic reactions might to some degree express personality traits. This could be used by running a sequence analysis on the feature. The touch-point-size is therefore regarded a quasi-continuous signal. The characteristic function-process of an extrovert vs. introvert can then be taught to a classification network and subsequently used as a prediction parameter. Even without a sophisticated sequence analysis, these differences could hypothetically manifest in a different variance for extroverts vs. introverts.

Methods

Description of the two underlying experiments

This work builds on the data of two previously conducted studies. The first set of interaction was recorded by Carola Trahms with 31 people in November 2016. She examined the quality-perception of subjects given certain transients, compared to the normal state. The second study was conducted by Giorgio Colombo under supervision of Carola Trahms, and the first 44 subjects from July 2017 are included here. The latter study strives at distinguishing left- from right-handed people via characteristic touch-artefacts. The two studies retrieved the Big-5 dimensions as part of thorough experiment routines and always included segments of data from normal-usage scenarios, which they used as comparison groups.

It is the aim of my thesis to build a predictive model which maps normal-usage interaction-data on to the Big-5 dimensions. The two studies contain this data. It was however produced in experimental settings which followed other purposes. The conditions under which the data can be used for my aim shall be examined here.

Study 1: Quality-perception of subjects given certain transients

In her study of quality perception, Carola Trahms used different disturbances in the experimental sessions. Every participant played 12 sessions of each around two minutes length. Two of the PflegeTab games, Quiz and Spell, were played six session each. Within these six sessions three different conditions (tiny icons, normal, freezing) appeared twice, in randomized order.

The *Quiz* game displays a question to the subject, both in written and audio form. A multiple-choice task then asks the player to tap on one of four answers displayed in buttons on the right side of the touch-screen. The experiments were both done on Apple-Tablets *iPad Air 2*.

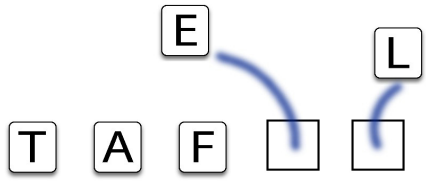


Figure 3 - Spell game

The *Spell* game asks the user to spell a word which is displayed on the bottom side of the screen. The letters are randomly scattered around the screen. The player must tap on the sought letter and drag it down to its position denoted by a placeholder.

While the Quizz setting captures taps and allows for performance metrics, it does not include swipe movements. It was therefore decided that the Spell scenario should be focussed on. This also has greater degrees of freedom, as the letters are randomly distributed.

The first experiment was filtered for 'Spell' and 'Normal' conditions, which leaves two sessions per participants with sessions lasting for around two minutes each.

Study 2: Prediction of Handedness from touchscreen interaction

In this second study Giorgio Colombo aims to identify right- and left-handed people by their touchscreen-imprint. For this encompassing experiment, 75 participants were invited to play the Spell game under 12 different conditions. The data of the first 44 participant was available by the time this thesis was written. Each of the conditions lasted for about one minute. Half of the participants were right handed, the other half left-handed. Every subject was instructed to play six conditions with the right hand and six conditions with the left, meaning that everyone did the experiment once with their main hand and once with the other. In addition, it was controlled for the position the tablet was used in: Either laying on the table, standing in a tilted table-stand or being held in the other hand. Again, these three conditions were combined with normal-size and tiny icons.

In line with the requirements of the question of this thesis, people should use their main hand and be shown icons of normal size. In Study 1, conducted in November 2016, it is not known in which position the tablet was held, while Study 2 is particularly controlled for this parameter. This point certainly deserves special attention during statistical analysis.

The combined sample

Merging the two studies, my data analysis will comprise of records from the *Spell*-game, where people drag-and-drop normal-sized icons with their dominant hand. Study 1 contributes two, two-minute chunks for each of the 31 persons. Study 2, entailing 44 subjects, brings three, one-minute sessions with specific tablet-positions to the record. It should be noted that in Study 1 subjects had to sit for 24 minutes pure time for the tablet test only, whereas in Study 2 they played for 12 minutes total. The data-snippets are taken from a randomised order design. It can clearly also distort the data, whether a chunk was recorded at the beginning or end of an experiment. Eventually the data might not only depend on the position of the recording within the experiment, but also on the absolute time a subject had to sit in the laboratory. Finally, the possibility of experiment specific other factors distorting a session (e.g. the examiner) could also skew the data.

It is left to the data-analysis whether a) the tablet position and b) the origin of the data (Study 1 or Study 2) has a significant influence on the feature form. Only then it can be decided whether and under which limitations a merge of the two studies is permitted.

Aggregation of raw-data into features

The PflegeTab application stores a detailed record of the user-interaction on the server. The information is distributed in different tables. Figure 6 - Spell raw datagive an insight into the data-format obtained for the specific game and shows the simultaneous record of touch-interactions. These raw-data tables from the server must be combined in a meaningful way. I received the data from both studies pre-processed through a script by Carola Trahm in the statistical language R.

In this intermediate state, the raw-data is combined into the *data.interactions.combined* table. One minute of play still has around 100 rows. Every row denotes either one game event (like ‘button loaded’ or ‘correct button pressed’), or it displays one touch event. Touch-up events are counted as swipes if their coordinates are far enough from the preceding touch-down event.

They already contain information such as speed and length of a swipe. Over all, the table contains 68 variables which broadly express the following information:

Game-event-type	User_id	Timestamp	Condition	x-y coordinates
size and place of buttons	Correctness of user-action		Touch-duration	Touch-accuracy
Swipe-length	Swipe-speed	Unique session number		Task number

This information needs to be further combined into a meaningful set of features. For Study 1, the information was already aggregated by Carola Trahms: The table *data.features.aggregated* lists 114 variables and 372 observations for 31 participants. 372 can be broken down into 12 conditions times 31 participants. These features are with a few exemptions the same ones as in the final combined data-set. A detailed description can be found in the 'The feature set' section.

Study 2: The Script

The data of Study 2 including the order-of-conditions and demographic information was kindly shared by Giorgio Colombo. The interaction data came in the *data.interactions.combined* format as described before. The R-Script Carola kindly provided exploits of all the possible and meaningful ways of aggregating the raw-data into features. A major portion of work was to learn R from scratch so that I could understand and adapt the 700-lines script which makes use of the shortcuts and particularities of this language. Therefore, Andrie De Vries' *R for Dummies* and *Reinhold* Hatzinger's *R - Einführung durch angewandte Statistik* were insightful.

The R script combines the intermediate data-format into an aggregated feature list. The script had to be adapted according to the changed conditions and timings in Giorgio's study. This entailed, for example, a new extraction script for the duration of the sessions, embedding the external demographic and course-of-experiment information into the code, and fixing irregularities in the raw data.

Irregularities in the raw-data

The data set from the server contained all sessions run on the device. During the experiments however, some sessions were run either as trial-session, aborted prior to the 60 seconds time target, or were shorter because the app crashed. This led to matching problems in the feature-aggregation section of the code, which is designed for 12 different sessions per subject.

At this point, the randomized order of the conditions cannot be matched with the data from the server anymore. One wrong person who strayed into the experiment had to be removed and a few sessions merged. For more details see Appendix "Outlier removal".

NEO-FFI scoring

The subjects were given paper versions of the German NEO-FFI questionnaire. It comprises of 60 statements and the degree of agreement or disagreement must be rated by the subject on Likert scales. The 5-point scale values of each item are to be weighed with 0=strongly disagree to 4=strongly agree, partially their ordering is reversed. Each of the 5 OCEAN dimensions have 12 statements associated. A neutral answer in all items would result in a score of 24 in each dimension.

The ratings were digitalized into a table and reversed where necessary by an R-Script. Subsequently the values for each dimension are summed up, divided by the number of given answers in this dimension and multiplied by 12 (Costa, 2008).

Merging the two data sets

For the combination of the two datasets, some few columns that existed only in the Study 1 set had to be dropped while the others had to be renamed. Once the data was merged, it became obvious that a group of count-variables had to be standardized. A group of features counts game events per session. This makes sense given the structure of the feature-aggregation script, whose logic builds on the session numbers. However, the variance of the session duration in both data sets would skew the data following a simple per-session count (Figure 4). The count per-session variables were therefore divided by the session lengths to obtain event / second features. On the other hand, the category of *mean per session of 'count' per task / interaction* variables do not have the same problem. As they aggregate the counts per a container (task/interaction) and the mean is taken over several containers, it is independent from the total session duration.

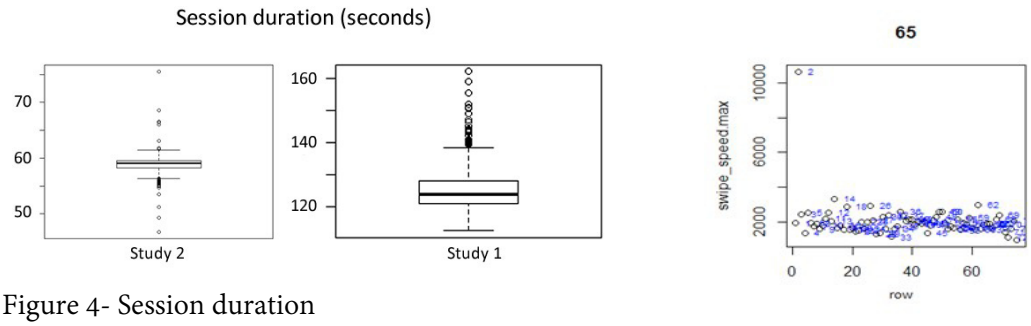


Figure 4- Session duration

Outlier removal

The combined data set was visually examined for extreme distortion. There were a few outliers, especially related to the speed and time metrics, which seem prone to flaws (Figure 5). Three sessions from three different subjects contained an implausible number of distortions in many different variables and were removed. Any person in the data set has at least two sessions associated. Where one session was removed, at least one other remains and the subject is maintained in the data set. Sessions of subjects (VP.Nr) 4 and 31 (from Study 1), as well as of VP-Nr. 124 from Study 2 were removed (more detail: Appendix "Outlier removal").

The feature-set

After separately aggregating the two studies into features, merging, cleaning and outlier removal, the data set contains 110 variables. This big number contains the statistical parameters for a smaller number of core features. Carola's script exploits all metrics that can be computed from the raw-data at hand. They can be summarised under these headlines:

Performance metrics

Counts and means of "Button … [touched, missed, correctly touched] aggregated both per task and per interaction	Count of tasks, interactions, touches,

For each of the general features, a set of statistical variables was computed:

General features		*Statistical variables*	
touch duration	X	mean	median
swipe length and speed		sd	range
time between touches		kurtosis	skew
difference touchpoint to button		mad	se
touch accuracy		min	max

Demographic and control variables

The variable 'control' denotes the position the tablet is held in: Either laying on the table, in a stand, or held in the other hand. The 'source' variable distinguishes the data origin between Study 1 and Study 2. Age and gender are included, as they are known to have a significant influence on the shape of the Big-5 dimensions. In addition, the handedness variable is included as differences in hand geometry are known to influence touch-characteristics.

Handedness	Age	Gender	Condition	Source

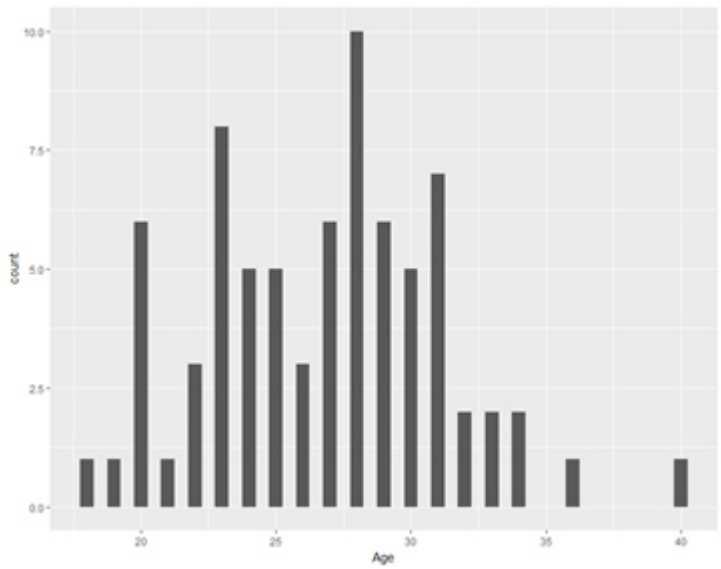
Results

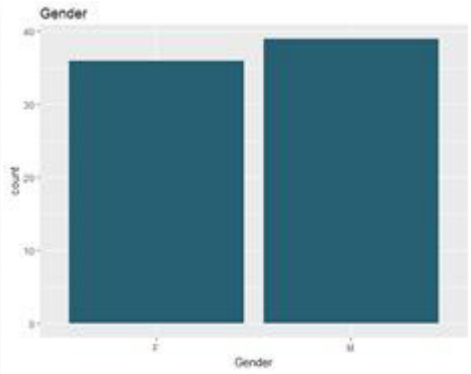
Distribution of personality types in the sample

The sample compared to cross population data

The final data set comprises Big-5 personality data for 75 subjects. The NEO-FFI has 60 questions with 12 items for each of the 5 personality dimensions – and a maximum of 4 points per item. This amounts to a theoretical maximum of 48 points for each dimension. Answering all questions with indifference (ticking the middle bullet of the 5-items Likert scale) would make a score of 24 points.

In reality, medium and average values are quite different. The sores are closely related to external factors, most prominently age and gender. Women score higher than men in all 5 scales of the NEO-FFI. In addition, all five scales correlate significantly with age. Older people tend to be less neurotic, less open and less extroverted while having higher values for agreeableness and conscientiousness.





Gender and age in the sample:
The distribution of gender is about equal. Age wise, the bulk and median of subjects is between 21 and 29 years old.

Figure 5 - Age and Gender

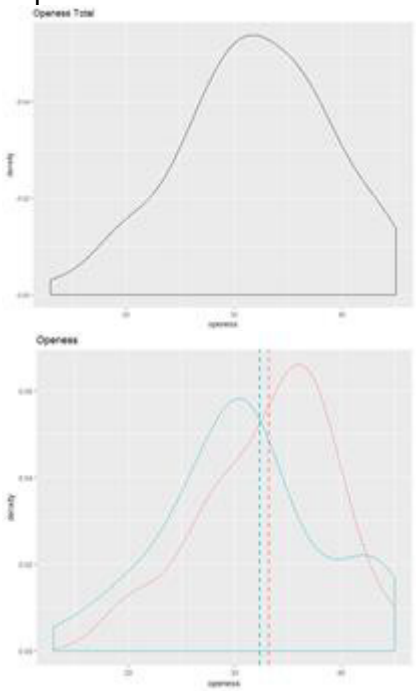
An empirical benchmark for comparison is offered by the NEO-FFI manual. Their data set on the German population covers 50 individual studies with 11.724 cases altogether. In the age group 21-29, which is predominant among the 75 subjects, the following mean values are found:

	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
➔ Population means					
Men	19,8	28,2	32,3	28,9	29,9
Women	23,2	29,1	33,1	30,9	30,6
➔ Sample means					
Men	16.2	28.6	30.7	30.6	33.3
Women	20.5	26.3	33.1	31.5	32.4

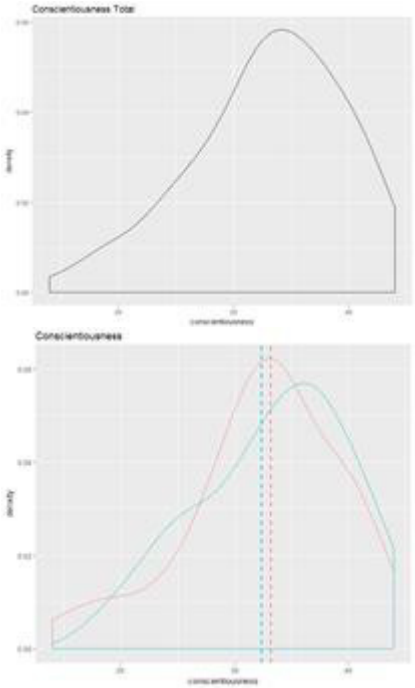
For a test of the distribution of the sample used in this thesis, the Anderson-Darling normality test was used. It shows that on an alpha-level of 5% the Hypothesis that the data is normally-distributed could not be rejected for 4 out of 5 dimensions. The test signalled a non-normal distribution for agreeableness.

	O	C	E	A	N
p-value	0.7688	0.1068	0.1335	0.03548	0.1514

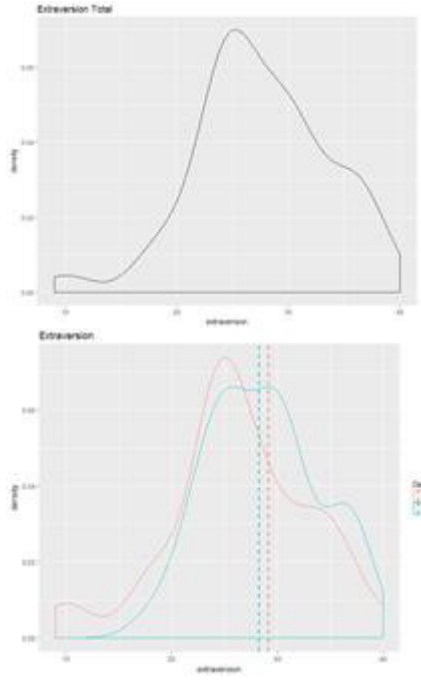
Openess



Conscientiousness

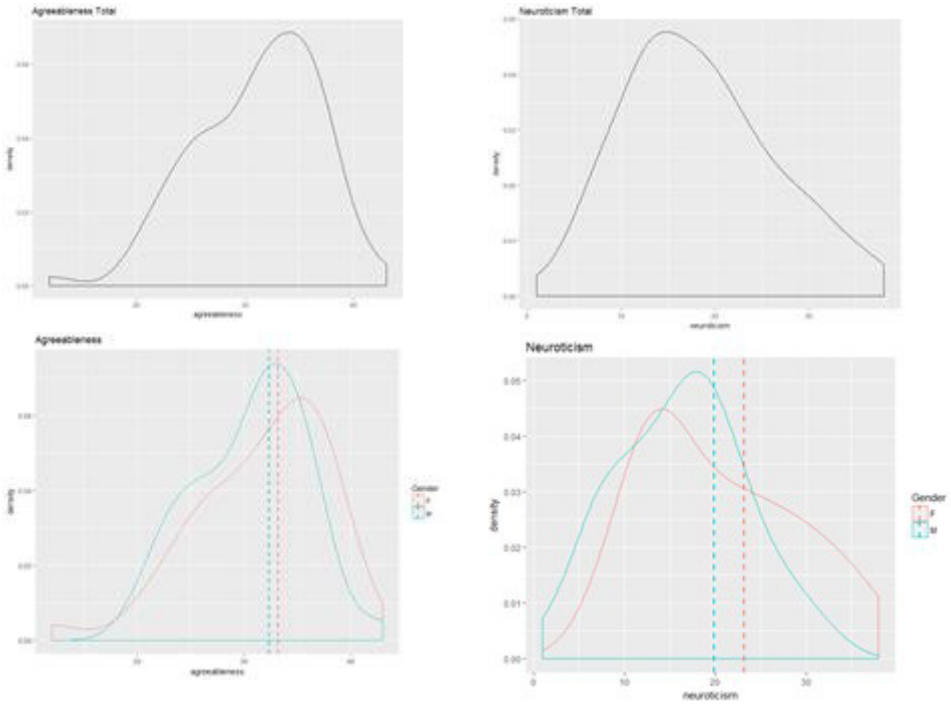


Extraversion



Agreeableness

Neuroticism



The distribution of the Big-5 dimensions in the sample. *The respective upper plot is the density over the total sample. The lower plot displays the densities depending on the gender. Two dotted lines displays the empirical medium for the German population in the 21-29 age group, for men and women. It is an empirical truth that women display higher levels in all dimensions. Consequently, the one dotted line further on the right is always the female-population mean, the other one, the mean of the male-population.*

Figure 6 - OCEAN distribution

The table above summarises the distribution in the five dimensions as density curves. On population level, female means are generally higher for all dimensions. In the sample, this is true only for N, O, A and reversed for E and C. The sample means are generally in line with population means. The highest deviation of 3.4 points is observed with conscientiousness, which could be explained by the high share of students in the sample. Educational success and high levels of conscientiousness are correlated. A look at the plots shows how agreeableness is obviously skewed. This explains the result of the Anderson-Darling test, which called agreeableness not normally-distributed.

The median as the split for classification

It was previously mentioned that no publication reported a successful prediction of personality from touch-interaction. This thesis will therefore take a binary-classification approach and split the metric dimensions into two groups. It remains debated as to which splitting point should be chosen. Even if the normally-distribution assumption was not rejected, it does not seem wise to split the data along the mean. The density plots show that not only the agreeableness dimension is skewed to one end of the distribution. As a result, the mean is shifted in the same direction and an unequal number of subjects would be found in the two groups. On one hand, some classifiers require classes with equal size. On the other, equal classes promise

generally higher prospects for classification accuracy. The data is therefore split along the median of the sample.

Feature elimination and covariates

Feature elimination

Many of the 110 explanatory variables in the final feature set had high inter-correlations. A smaller number of features would however increase the efficiency of the classification-algorithms and make the data-set easier to handle. For this purpose the caret package in R provides a *findCorrelation()* function. It removes variables with a correlation higher than an adjustable cut-off. Here, 50 features remain with absolute correlations less than 80%.

Covariates

Merging firstly the two data sets and secondly three different ways of holding the tablet, assumes a non-significant influence of these factors. Then again, an influence of the condition (the way of holding the table) or source (Study 1 or Study 2) could be subtracted out of the data when handling these two variables as covariates. An Anova-test was therefore calculated for each of the 50 variables of the correlation-reduced feature set: First for the source, then for the condition.

Condition - The condition variable includes information both about the hand which was used and the position the tablet was hold in. As this thesis examines normal usage scenarios only session where people interact with their naturally dominant hand are included. The levels {1,3,5} on the x-axis of Figure 7 mean a use of the right hand, {7, 9, 11} of the left. The respective first entry of the triple says: The tablet was laying on the table; Second: The tablet was held in a stand on the table; Third: The tablet was held in the other hand. A significant dependence on the condition-variable first means that left- and right-handedness has a significant influence on the features. A look at *Figure 7 - Condition left vs. right.* reveals that differences are mainly *between* the two handedness groups, not within.

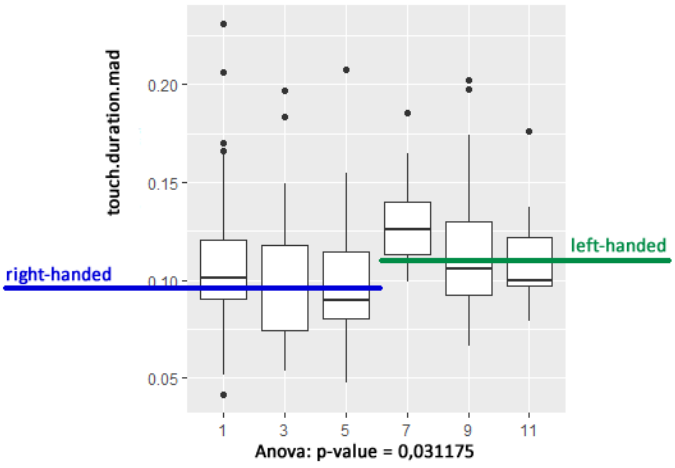


Figure 7 - Condition left vs. right

Therefore, conditional dependencies must be considered. Given that people had a dominant right hand, the condition (whether the tablet was held in the hand, put on the table or rested in a stand) had little effect. Only two variables were significantly different, dependent on the condition: (time_between_touches.mean; p-value: 0,0468) and (difference.touch_buttonCenter_x.kurtosis; p-value: 0,0058). Given a left-hand dominance, the conditions did not matter at all.

Source – Whether a session came from the older or newer study made a greater difference than the condition. Of the 50 variables in the correlation-reduced data set, 11 were in significant relation to the nominal source variable (*Figure "Correlation to the source variable"*).

"difference.touch_buttonCenter_x.se" p=0,0300; "touch.duration.se" p=0,0019;
"swipe_speed.kurtosis"p=0,0283;"touch.duration.skew" p=0,0319;
"difference.touch_buttonCenter_x.max" p=0,0284; "swipe_length.max" p=0,0003;
"difference.touch_buttonCenter_y.min" p=0,0443; touch.duration.min p=0,0276;
"touchAccuracy_x.mean" p=0,0375; "time_between_touches.mean" p=0,0001;
"Count.ButtonTouchedPerSession" p=0,0001; (11/50 variables)

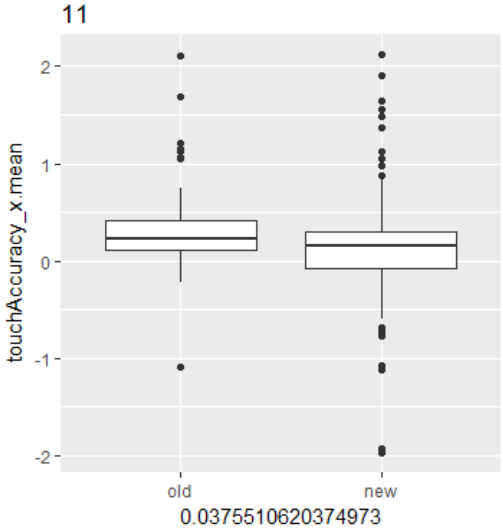


Figure 8 - Correlations with 'source'

Given the number of relations, the source variable should be considered for a covariates correction. This was postponed due to time-constraints.

Handling of age, gender, handedness - In a normal touch scenario, externally validated information such as gender, handedness or age cannot be assumed to be known. It therefore skews the prediction accuracy of the classifier to consider them at all: Both included as predictors or used to perform a pre-classification covariates correction. Considering that the aim of the classification is to make prediction purely based on the touchscreen-log or application data, these external variables should be entirely left out. However, is worth noting that acknowledged studies from the field of phone-metrics-based predictions include gender as a predictor (Montjoye et al. 2013). As gender and age are known to have a great influence on the Big-5, it is hardly surprising that this boost their results. However, reflecting Montjoye's use of phone-logs as a source, it does not seem necessarily plausible to have this external information, given their data collection scenario.

Classification

R, the caret package and additional resources

The statistical language R offers a multitude of extensions and several hundreds of classification and feature-preparation algorithms. It is not always obvious to which standards these libraries adhere. This need for validation and user friendliness is addressed by the caret package. This "is a set of functions that attempt to streamline the process for creating predictive models." (Kuhn, 2017). It offers functions for data splitting, pre-processing, feature selection, model tuning using resampling and variable importance estimation, all of which were used here. It is especially commendable for the ease at which different models can be tried on the data and later compared in the same framework.

The way into machine learning with R was paved by the 7-part lecture "Introduction to Data Science with R" by David Langer. He recommends picking R over python for especially for caret, as it makes various activities such as k-fold cross validation and parallel processing easy to handle. A resource for the practical implementation of caret routines were excerpts of Jason Brownlee's book *Machine learning mastery with R*.

ALM

Automatic linear modelling in SPSS is routed in machine-learning and part of the predictive analytics package. It calculates a linear regression function for the prediction of a metric variable. In a first step, the data is automatically prepared and transformed. The best-subset method is used for feature-selection with R^2 as optimization criterion. The model performance is then increased with bagging. This resulted in the following R^2 values:

Extra: 0.325 Neuro: 0.259 Open: 0.283 Agree: 0.149 Con: 0.224

For evaluation purposes, the model was applied to the feature-set and produced prediction values. These were plotted against the real observations and complemented by a correlation analysis (Pearson). The correlation coefficients are:

Extra: 0.58 Neuro: 0.52 Open: 0.53 Agree: 0.39 Con: 0.46

Extraversion scores the highest in both measures and the ordering of the other dimensions is equal.

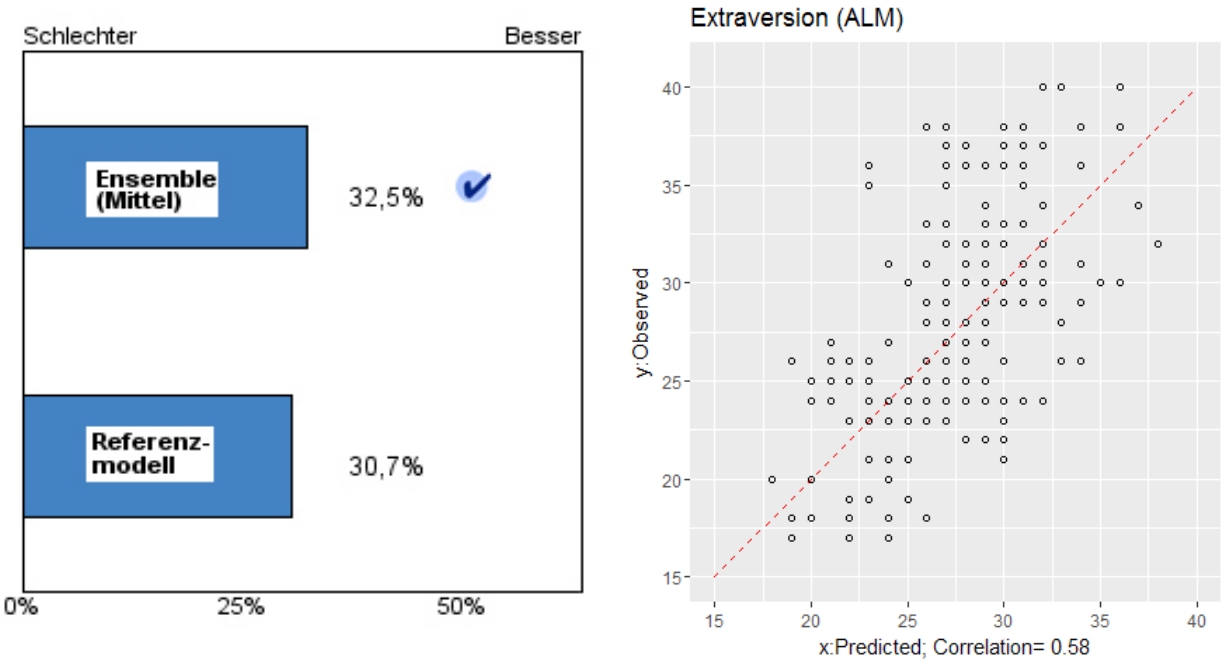


Figure 9 - ALM extraversion

On the left side of *Figure 9 - ALM* extraversion the model-output of SPSS ALM states the accuracy measure s R^2 exemplary for extraversion, as it has the best performance. The image on the right plots the observed extraversion values against the one predicted by the model. The dashed line marks what would be a perfect prediction.

Train control: k-fold cross validation

The training and testing was set up along with a validation set. 90 percent of the data was used in a 10-fold cross validation and repeated 10 times. The remaining 10 percent were kept for validation purposes. The data partitioning is done while maintaining an equal distribution of the target classes in both sets. Overall this amounts to 100 runs for a single version of an algorithm. Caret requires to define the training-testing configuration in a single "traincontrol" object. This setting is then used in all the different algorithms. In addition, the random-number generator is set to the same seed before every run. The same seed with the same traincontrol-object ensures comparable results between different classifiers. An inbuild pre-processing functions allows to centre and scale the data. Caret provides a uniform interface for many algorithms, which all have different knobs for tuning accuracy. The framework allows to set an integer

'tuning' parameter, which is the number of parameters which will be automatically optimized in each algorithm.

The random forest for example uses different total number of trees. All different combinations of parameters are computed and only the best performing version of every algorithm is kept for further use or evaluation. The computation is run with the doSnow tool. It allows for parallel processing and reduced computation time by 70%.

Classifiers

In a Support Vector Machine (SVM) data points with different classification are tried to be separated by an optimal boundary. The vectors themselves are the lines between the separation boundary and the closest points. Two hyperparameters must be found for this model: the misclassification cost C (the number of points that can deviate from the separation) and the sharpness sigma of the Gaussian basis functions (Brownlee, 2017). This can be imagined like a linear line separating two groups of data points in a two-dimensional space. With higher dimensions however, a linear line or hyperplane is unlikely to be separate points well. This thesis uses a radial-basis-function (RBF) instead. Therefore, the mentioned distance measure between separation boundary and points is measured in higher dimensions, in a non-Euclidian space. The two hyperparameters C and sigma can either be left to be automatically tuned or handed in as a defined search grid. The later approach generally yields better results. The left side of *Figure 10 – tuning* displays how every Sigma value is paired with every C values from the grid, and run 100 times. Only the best performance combination is kept as a result. The following search values were used:

Sigma {0.01, 0.02, 0.025, 0.03, 0.04, 0.05, 0.06, 0.07,0.08, 0.09, 0.1, 0.25, 0.5, 0.75, 0.9 }

C {0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2,5}

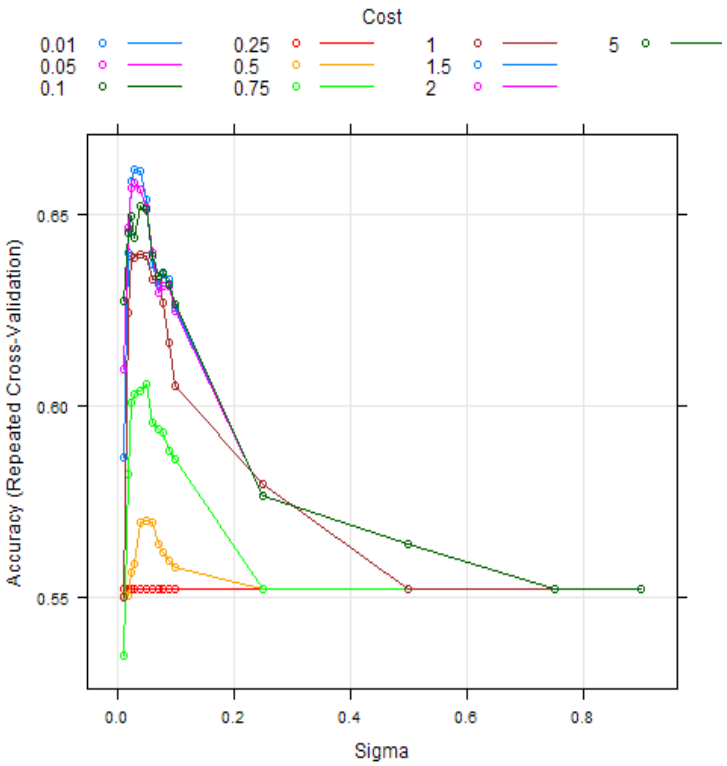


Figure 10 – tuning. Left: SVM. Right: RF

The right side of *Figure 10 displays the tuning of the predictor-number in the Random-Forest (RF)* algorithm. It works well for exploratory modelling and had an implicit feature-selection mechanism, which arises from the structure of the classifier (Langer, 2014). The k-nearest neighbour (KNN) algorithm does not require the input variables to follow certain distribution. In this aspect, it is different to many linear models, which require an equal variance distribution and some normally distributed data. These requirements are overcome by a generalized-linear-model as also offered in SPSS. A popular classifier for binary target data is the logistic regression from caret's "GLM". Caret offers a uniform *train* function for easily running and comparing different algorithms. After setting up the greater framework the Naive Bayes (NB) was added to the classifier set, to increase the diversity of non-linear algorithms (Brownlee, 2017).

In all cases "accuracy" was used as a target parameter for optimization. This requires the two binary target classes to be of the same size, which is the case with the median split. For each algorithm, the best model was eventually applied to the validation set. The general best-practice to gauge a classifier by its

performance on a validation set only works, if the validation-set is a representative excerpt of the total data sample. With 191 data samples and 90% for the classifiers, only around 20 samples remain in the validation set. The assumption of representation must be dropped then, which renders a validation set useless due to its small size.

The validation set results are, if not representative for the overall accuracy, yet interesting applied to caret's model-comparison output. The 10x10 runs of the best models of each classifier are used to compute a probability distribution of the algorithm's accuracy. This distribution can be displayed as boxplots, alongside with confidence intervals including information about whether the results are significantly different from random chance. The accuracy distributions for Neuroticism are plotted in Figure 11 - Accuracy: Boxplots and Confidence.

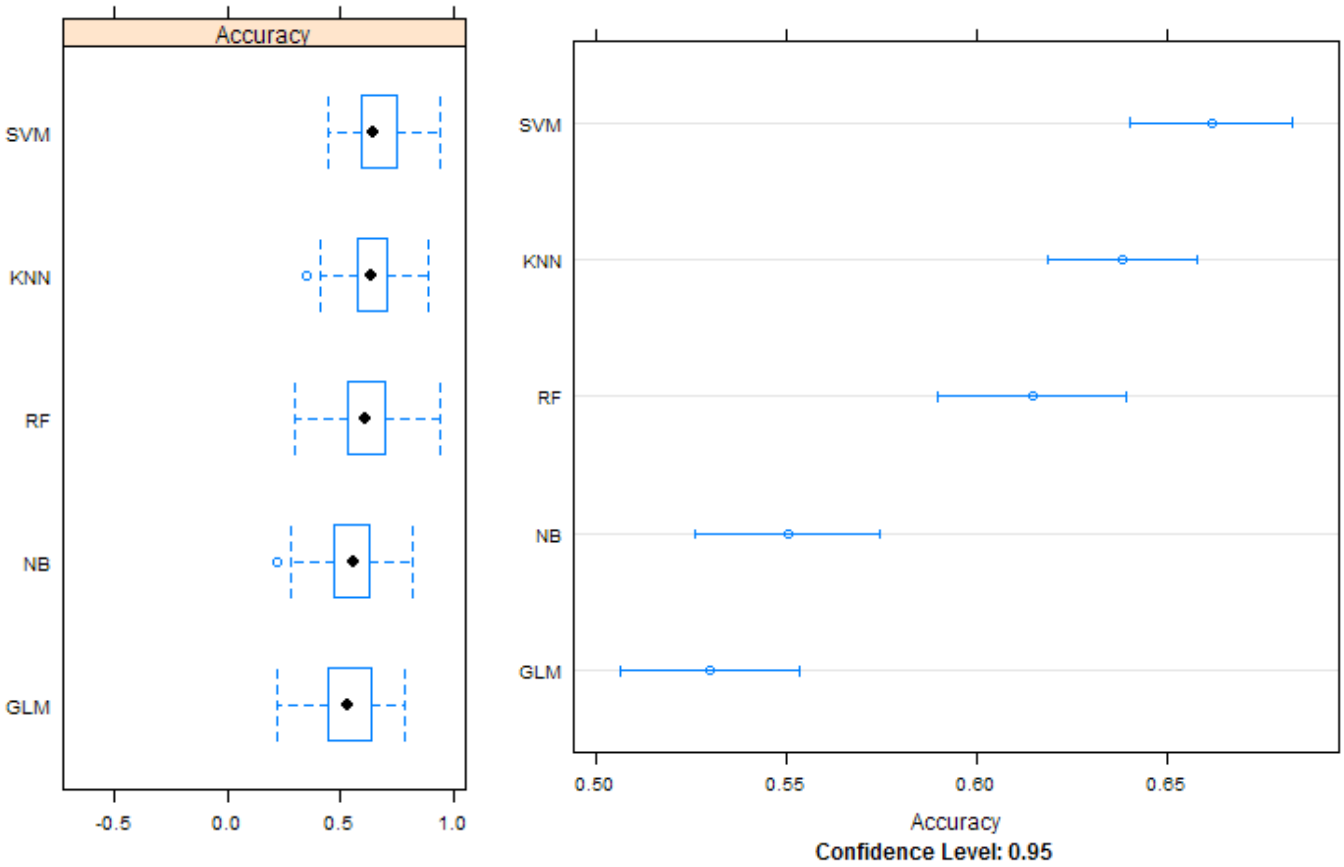


Figure 11 - Accuracy: Boxplots and Confidence

The following mean-accuracy values can be seen in the plot: SVM (66,1%), KNN (63,8%). RF (61,4%), NB (55%) and logistic regression from the generalized-linear-model GLM (53%). The confidence-interval plot clearly reveals that all values are significantly higher than the random chance baseline of a binary classification problem of 0,5. The validation set values (in the same order as accuracies) for these cases are 42,1%, 47,4%, 58%; 31,6% and 42,1%. These values fall into the range of the boxplots on the left side of *Figure 11 - Accuracy: Boxplots and Confidence*.

A list of similar result plots for all five personality dimensions can be found in the Appendix "Classifier rankig, boxplots and confidence", alongside with significance values and confidence intervals of the mean accuracies (Appendix: "Final results classification"). Significance was manually computed via t-test based on the probability distribution of the accuracies minus the random chance of 0,5.

A recursive feature elimination routine was applied to SVM, RF and GLM and did not yield improved results. All classifications are based on the complete feature set of 50 variables. Caret's *rfe* function was used for this purpose, which is supposed to process feature elimination aligned to a specific algorithm. Variable-baskets of size {1,3,5,10,15,20,25,30,35,40,50} were computed (see appendix "RFE plot"). It remains an open question if higher accuracies could be reached with an optimized feature set. One explanation for non-increasing accuracies is that certain algorithms already have an implicitly inbuild feature selection, arising from the logic of the classification.

Throughout the research all algorithms were run four times for all dimensions. The last two runs were computed with the same configuration, only the seed of the random-number generator was set to a different starting point. Mean-accuracy values of these two runs are listed in *Table 3 final accuracies*. The green value denotes the best algorithm for every dimension in the first run, the yellow cells are the best-cases for the second run. Accuracies of both runs are very similar, thus stable. For O, A and N the SVM shows the pest performance. This is in line with this classifier being especially suitable for problems with a small sample size compared to the number of explanatory variables (Brownlee, 2017). For extraversion, the logistic-regression (GLM) scores the highest, which is surprising giving the high dimensionality of the task.

For conscientiousness, the random-forest (RF) performs the best.

Table 3 - Final accuracies

		Open	Con	Extra	Agree	Neuro
SVM	Run 1	0.6470	0.5688	0.5907	0.6279	0.6512
	Run 2	0.6312	0.5863	0.5907	0.6083	0.6618
KNN	Run 1	0.6111	0.5384	0.5402	0.5839	0.5665
	Run 2	0.6093	0.5646	0.5403	0.5781	0.6383
RF	Run 1	0.6470	0.5834	0.6084	0.6144	0.6221
	Run 2	0.6301	0.5942	0.6085	0.6035	0.6146
NB	Run 1	0.5555	0.5504	0.5648	0.5361	0.5362
	Run 2	0.5604	0.5707	0.5649	0.5200	0.5502
GLM	Run 1	0.5835	0.5075	0.6119	0.4844	0.5002
	Run 2	0.5598	0.5055	0.6120	0.4648	0.5298

Table 4 shows the best three predictors for the respective best performing classifier. It can be noted that metrics derived from touch-duration, touch-accuracy, swipe-speed and time-between-touches constitute this best-of list. Inferences about variable importance still must be taken with a grain of salt as the difference in the ranking is partly marginal.

Tabelle 4 - Top 3 predictors

Openess (SVM)	touch.duration.se	swipe_speed.max	touch.duration.min
Conscientiousness (RF)	touch.duration.min	touchAccuracy_x.median	time_between_touches.median
Extraversion (GLM)	time_between_touches.median	swipe_speed.mad	touchAccuracy.max
Agreeableness (SVM)	touchAccuracy_y.median	touch.duration.skew	touchAccuracy_y.mad
Neuroticism (SVM)	time_between_touches.skew	touchAccuracy.max	touch.duration.se

Using the best classifiers for every dimension this results in the final prediction accuracies:

Discussion

Related results

A literature research did not reveal publications which report predicting personality from touch-interaction. A benchmark for the classification accuracy can nevertheless be derived from related fields. Montjoye reached a mean of 0.61 accuracy with mobile-phone logs, in a three-class prediction setting. Similar values were reached by Monsted. Chittaranjan reached between 0.54 and 0.59 with two classes in 2011. Gao predicts four emotional states with between 0.69 and 0.77 percent, and arousal and valence with 0.89 accuracy.

It should also be noted that with Speech (Polzehl 2015), Social-Network and text-analysis approaches (Kosinskia, 2012) other data-sources are powerful for personality perditions.

Room for manoeuvre

This thesis is a first exploration of the relation between touch and personality. It has shown that stable and meaningful connections exist. First classification results are significantly higher than random chance and the range of 59-66% accuracy is in line with research from related fields. Most certainly however, these results can be substantially improved.

Covariates. It was previously mentioned that a covariates-correction for the variable distinguishing the

data-origin between Study1 and Study2 would be an enriching step prior to classification. The reason why results from the two studies are very slightly different remains open to debate.

More samples. Making not 80% but 90% of the samples available to the training-testing phase (instead of putting them into the validation set) substantially increased the results. With more samples, accuracy will most likely further increase. On one hand, this could be achieved with more subjects. On the other, it is also worth a thought to aggregate the features into shorter time-containers. If for the 75 subjects, features were aggregated into 30 second containers instead of averaging over several minutes, particularities could be caught in greater detail.

Clustering. An interesting idea during the topic-presentation was the possibility to cluster the personality dimensions into types. This approach is also taken by (Polzehl, 2015) and promises better results.

More features. Extending the scope of the raw-data. While touch-duration is thought to be a proxy for touch-pressure, this could be supplemented by adding accelerometer or touch-point-size data to the data records. A quasi-continuous signal would also allow for new classification techniques, such as sequence detection in combination with artefacts. Another possibility would be a more detailed outline of the touch-trajectories.

References

Carney, D. R., Jost, J. T., Gosling, S. D. and Potter, J. (2008), The Secret Lives of Liberals and Conservatives: Personality Profiles, Interaction Styles, and the Things They Leave Behind. Political Psychology

BARRICK, M. R. and MOUNT, M. K. (1991), THE BIG FIVE PERSONALITY DIMENSIONS AND JOB PERFORMANCE: A META-ANALYSIS. Personnel Psychology

Brownlee, Jason; Machine learning mastery with R; machinelearningmastery.com; 2017

Chittaranjan •. J. B. •. D. Gatica-Perez, "Mining large-scale smartphone data for personality studies," 2011.

Costa und McCrae, NEO-Fünf-Faktoren-Inventar nach , 2. Neu normierte Auflage, 2008. Borkenau, Peter; Ostendorf, Fritz.

Frank, Biedert, Ma, Martinovic, Song; "Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication"; IEEE Transactions on information forensics and security; 2013.

Gao, Bianchi-Berthouze,Meng, "What does touch tell us about emotions in touchscreen-based gameplay?" Transactions on Computer-Human Interaction , 2012

Golem (newspaper); Beuth, Patrick; „Die Luftpumpen von Cambridge Analytica"; golem.de; 2017

Gomez, A., Gomez, R.: Personality traits of the behavioural approach and inhibition systems: Associations with processing of emotional stimuli. Pers. Indiv. Differ. 32(8), 1299–1316 (2002)]

Gosling SD, Rentfrow PJ, Swann W (2003) A very brief measure of the big-five personality domains. J Res Pers 37:504–528

Hertenstein, Holmes , McCullough , Keltner; "The Communication of Emotion via Touch"; 2009; American Psychological Association.

Kosinskia Michal, Quote in a presentation at UdK Berlin, 2017

Kuhn, Max; "Introduction to the caret package"; lead author of caret; online; 2017

Kuhn, Max; "Predictive Modelling with R and the caret Package"; 2013

Langer, David; Introduction to Data Science with R; daveondata.com; 2014

Monsted, Mollgaard, Mathiesen "Phone-based Metric as a Predictor for Basic Personality Traits" University of Copenhagen, 2016.*

Montjoye, "Predicting Personality Using Novel Mobile Phone-Based Metrics," 2013.

Polzehl, Tim; Personality in Speech, "Assessment and Automatic Classification" 2015; T-Labs Series in Telecommunication Services

Pruitt, John, and Jonathan Grudin. "Personas: practice and theory." Proceedings of the 2003 conference on Designing for user experiences. ACM, 2003.

Shah, J. Narasimha Teja and Samit Bhattacharya. "Towards affective touch interaction: predicting mobile user emotion from finger strokes." Journal of Interaction Science (2015)

Vinciarelli , Alessandro Member; IEEE, and Gelareh Mohammadi, "A Survey On Personality Computing", IEEE Transactions On Affective Computing, 2013.

Zheng ,Bai, Huang, Wang; "You are How You Touch: User Verification on Smartphones via Tapping Behaviors"; IEEE 22nd International Conference on Network Protocols; 2014

Appendix

Final results classification

Agreeableness

	median	pvalue	Sig	conf.start	conf.end	testset
SVM	0.6083	0.0000	TRUE	0.5881	0.6285	0.5263
KNN	0.5781	0.0000	TRUE	0.5537	0.6025	0.3684
RF	0.6035	0.0000	TRUE	0.5812	0.6258	0.6316
NB	0.5200	0.0799	NA	0.4976	0.5425	0.3684
GLM	0.4648	0.0018	TRUE	0.4431	0.4866	0.6316

Conscientiousness

	median	pvalue	Sig	conf.start	conf.end	testset
SVM	0.5863	0.0000	TRUE	0.5644	0.6082	0.6842
KNN	0.5646	0.0000	TRUE	0.5436	0.5855	0.6842
RF	0.5942	0.0000	TRUE	0.5756	0.6128	0.6316
NB	0.5707	0.0000	TRUE	0.5527	0.5887	0.5263
GLM	0.5055	0.6287	NA	0.4831	0.5278	0.6316

Extraversion

	median	pvalue	Sig	conf.start	conf.end	testset
SVM	0.5907	0.0000	TRUE	0.5707	0.6107	0.7895
KNN	0.5403	0.0003	TRUE	0.5189	0.5617	0.6842
RF	0.6085	0.0000	TRUE	0.5884	0.6286	0.7895
NB	0.5649	0.0000	TRUE	0.5450	0.5847	0.7895
GLM	0.6120	0.0000	TRUE	0.5904	0.6335	0.8947

Openess

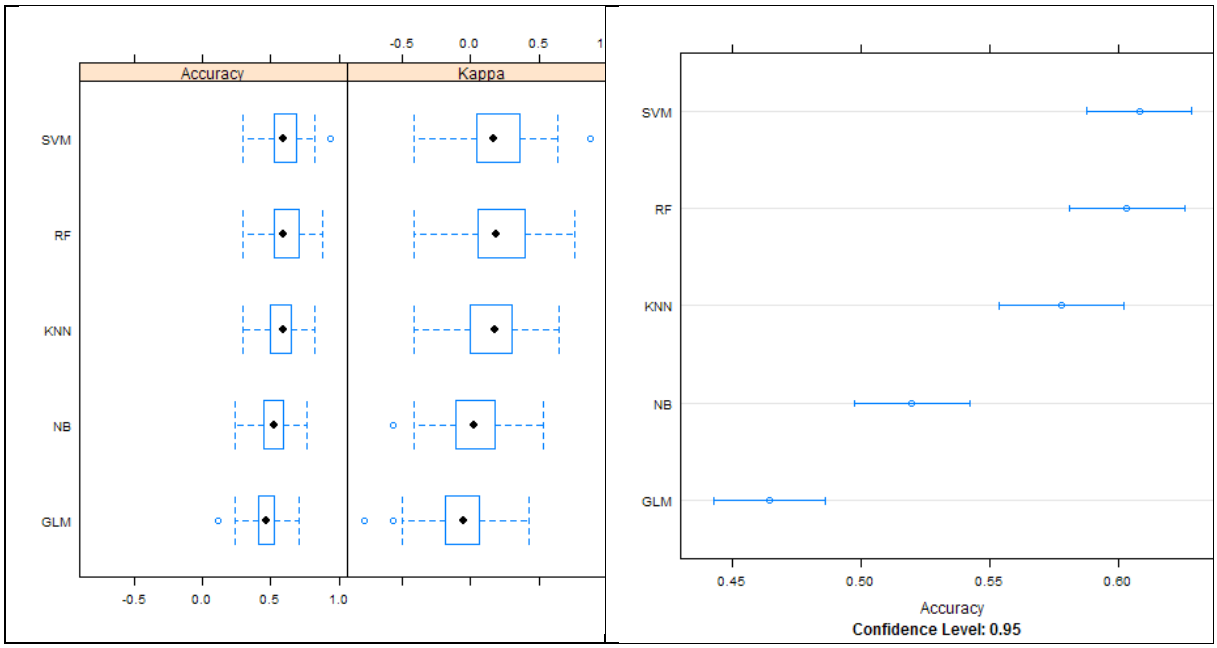
	median	pvalue	Sig	conf.start	conf.end	testset
SVM	0.6312	0.0000	TRUE	0.6137	0.6486	0.6842
KNN	0.6093	0.0000	TRUE	0.5896	0.6290	0.7895
RF	0.6301	0.0000	TRUE	0.6095	0.6507	0.6842
NB	0.5604	0.0000	TRUE	0.5358	0.5850	0.5263
GLM	0.5598	0.0000	TRUE	0.5393	0.5803	0.5263

Neuroticism

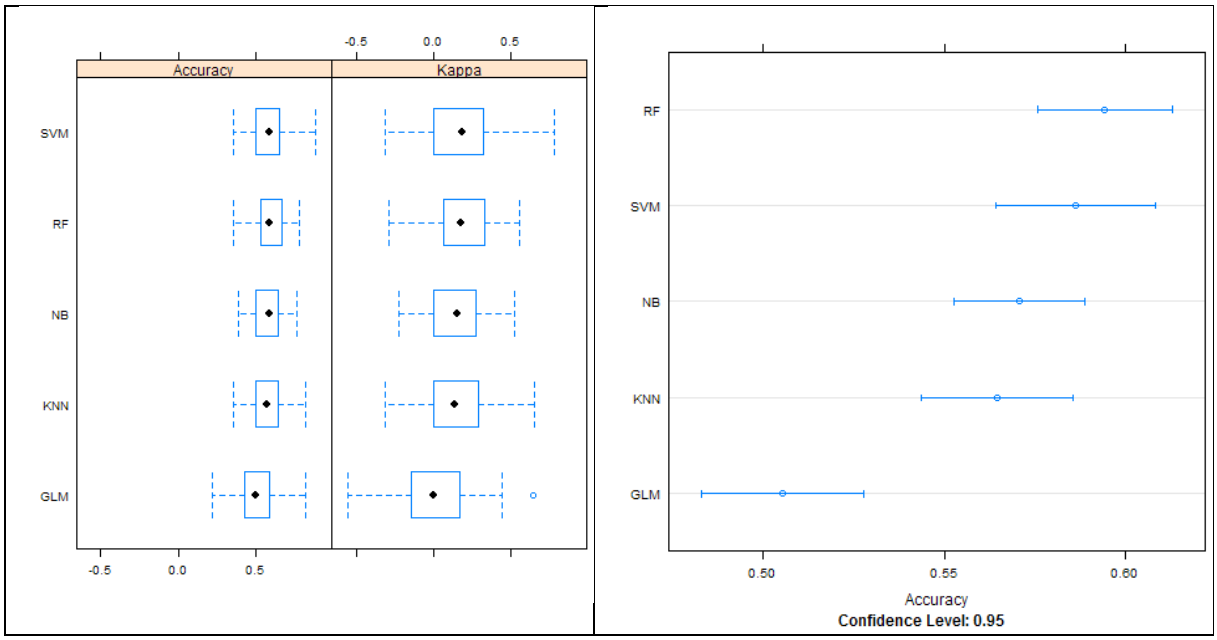
	median	pvalue	Sig	conf.start	conf.end	testset
SVM	0.6618	0.0000	TRUE	0.6405	0.6831	0.4211
KNN	0.6383	0.0000	TRUE	0.6186	0.6579	0.4737
RF	0.6146	0.0000	TRUE	0.5898	0.6395	0.5789
NB	0.5502	0.0001	TRUE	0.5259	0.5745	0.3158
GLM	0.5298	0.0138	TRUE	0.5062	0.5534	0.4211

Classifier ranking, Boxplots and Confidence

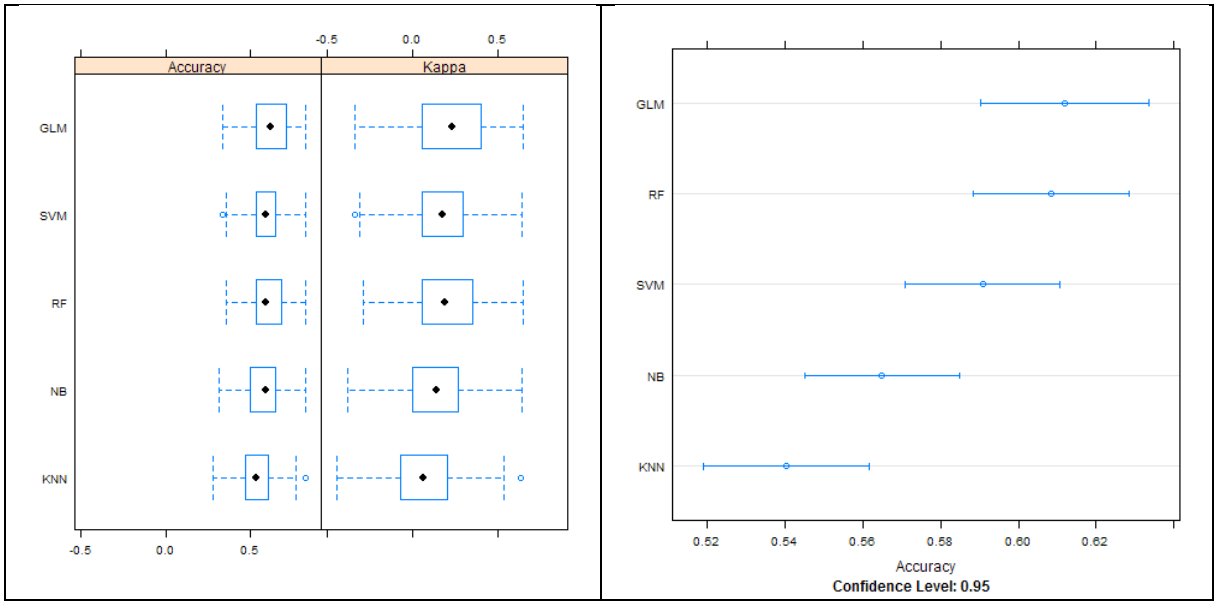
Agreeableness



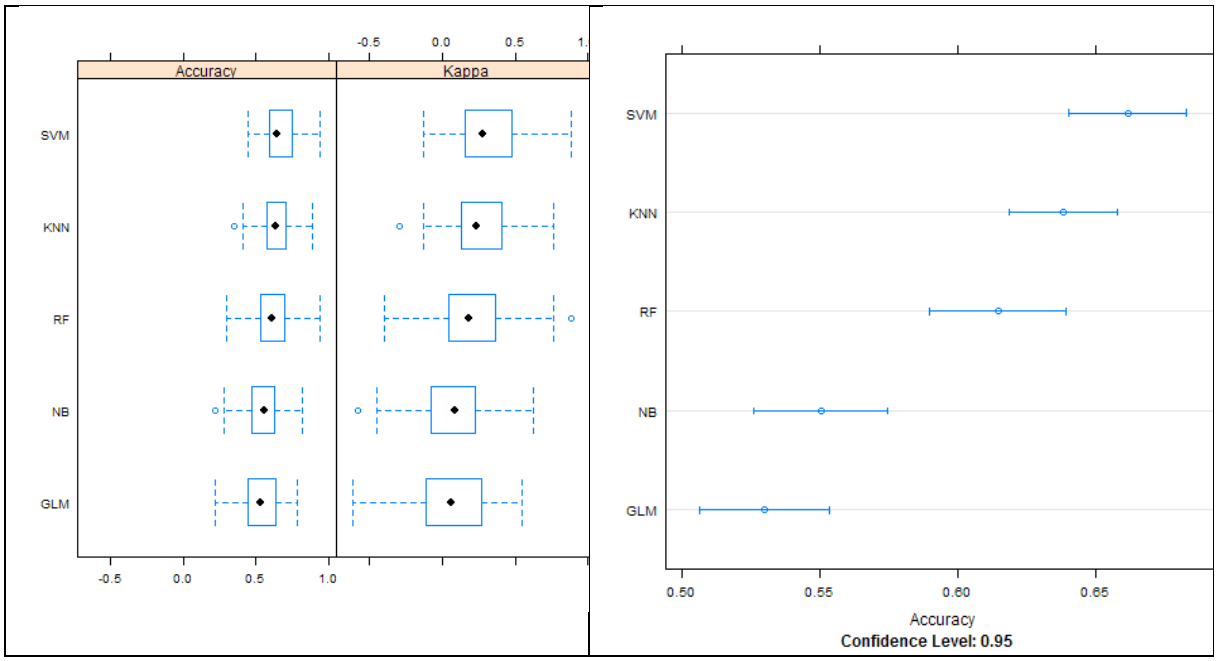
Conscientiousness



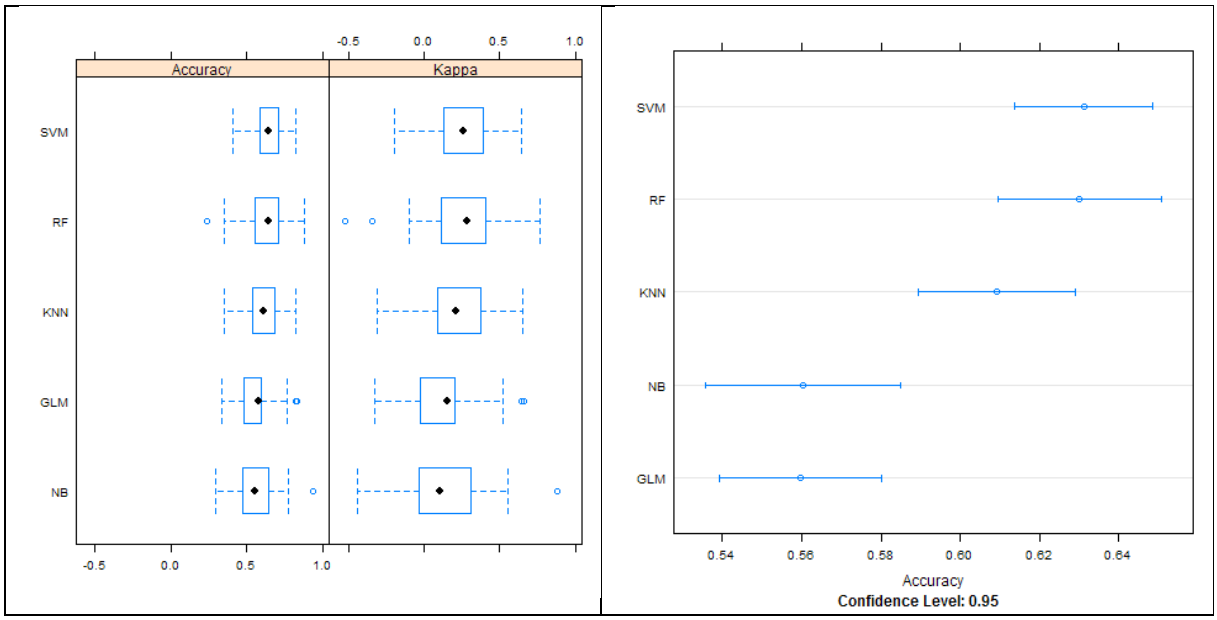
Extraversion



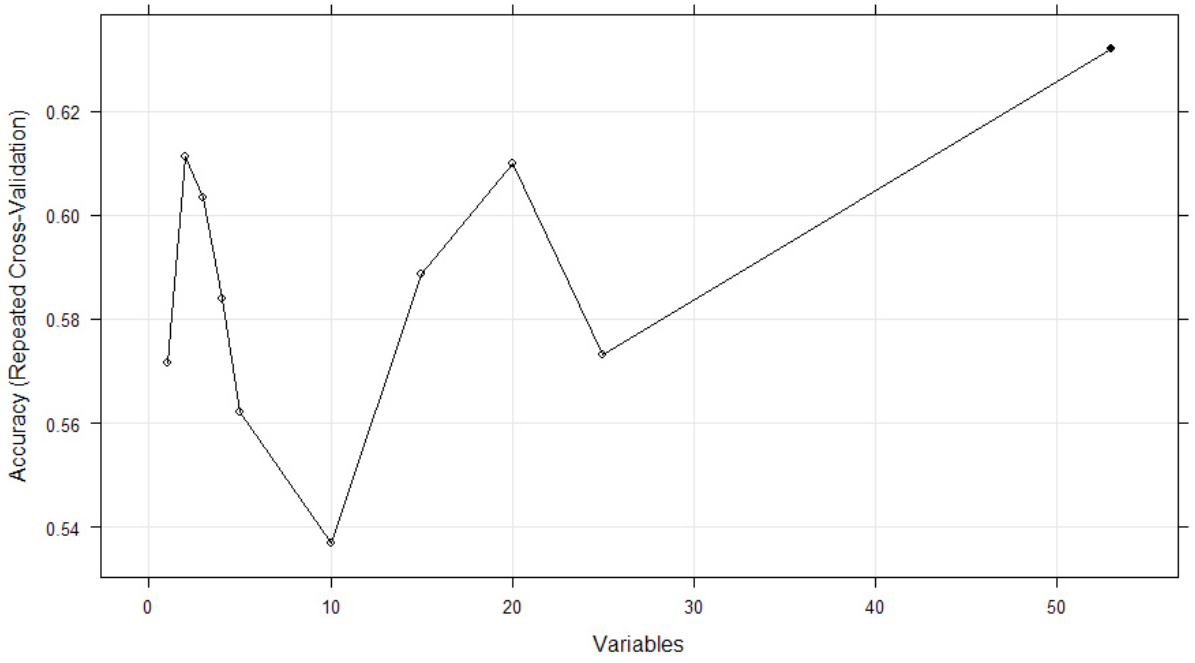
Neuroticism


Accuracy
Kappa
SVM
KNN
RF
NB
GLM
-0.5
0.0
0.5
1.0
0.50
0.55
0.60
0.65
Accuracy
Confidence Level: 0.95

Openness



RFE plot



Outlier removal

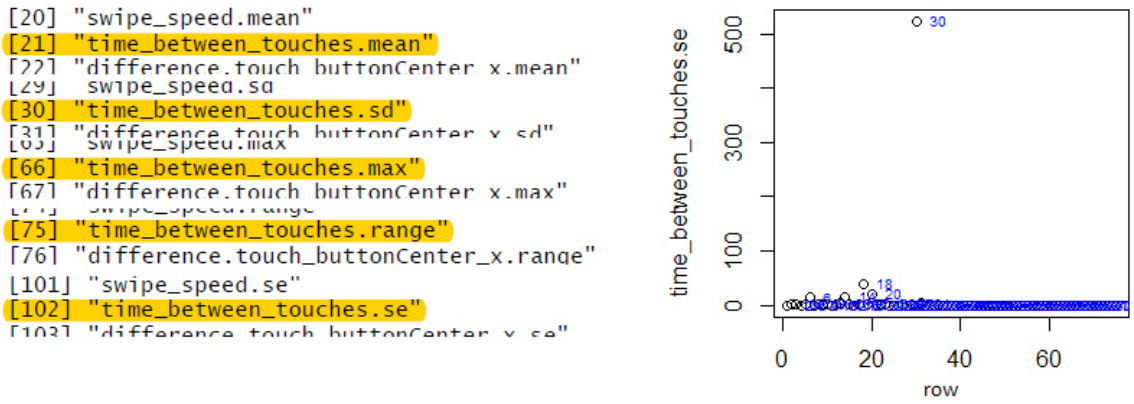
In the section VP [611.655] the irregularity that more than 12 session per person exist, occurred with 4 subjects (user_id: 216, 234, 231, 220). After consultation with Giorgio it was said that 216 (VP-Nr. 637) must be left out, because it was a wrong person who strayed into the experiment. In general, the App sometimes crashed after a time less than one minute so the subjects were made to play the remaining time in a new session. For 22, 231, 234 sessions were merged together into one. This process was reconfirmed by the Conditions table.

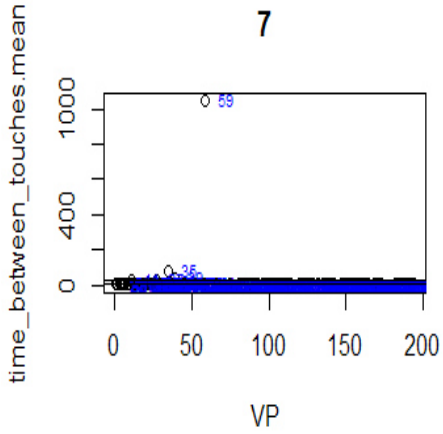
- 216 completely removed
- 220: merge (360+361) (363+364)
- 234: merge (537, 538)
- 231: merge (495,496)

Now: 39 subjects with exactly 12 sessions each. Yet, the session 360,363,537,495 deserve special attention and will yield wrong session-durations.

Distortion profile subject 4 (from the old data set). All swipe_speed parameters for subject 4 are removed, as such deviations seems non-plausible and this metric is vulnerable to flaws.

Distortion profile of subject 32 (from old data set). All time_between_touches parameters for subject 32 are removed, as such deviations seems non-plausible and this metric is vulnerable to flaws.





Raw data snapshots

id	1	2	(...)	22	23	24	25	26
timestamp	14.11.2015 14:52	14.11.2015 14:52	(...)	14.11.2015 14:53	14.11.2015 14:53	14.11.2015 14:53	14.11.2015 14:53	14.11.2015 14:55
app	9	9	(...)	9	9	9	9	9
event	0	2	(...)	0	2	4	4	0
fly_on	FALSCH		(...)	FALSCH				WAHR
percentage	0		(...)	0				3
sequencing	WAHR		(...)	WAHR				FALSCH
is_clickable	FALSCH		(...)	FALSCH				FALSCH
is_target	FALSCH		(...)	FALSCH				FALSCH
volume	2		(...)	2				6
size	8		(...)	8				1
letter_id		78	(...)		77	77	77	
user_id	15	15	(...)	15	15	15	15	16
word_id	97	97	(...)	113	113	113	113	115
device_name	PflegeTab1	PflegeTab1	(...)	PflegeTab1	PflegeTab1	PflegeTab1	PflegeTab1	PflegeTab1
installation_UUID	E170EFF5-D1D8-4D2A	E170EFF5-D1D8-4D:	(...)	E170EFF5-D1D8-4I	E170EFF5-D1D8-4I	E170EFF5-D1D8-4D	E170EFF5-D1D8-4	E170EFF5-D1D8-4D
placed_on_wrong_target_letter								
level_p								
level								

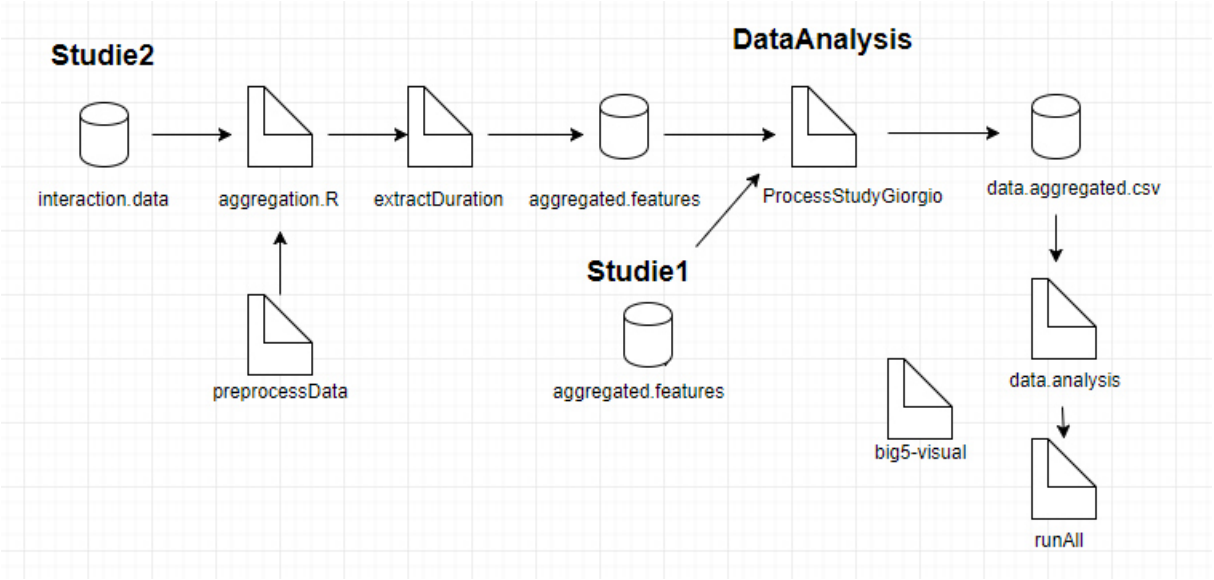
Figure 1 - Spell raw data

id	timestamp	app	event	x_location	y_location	user_id	device_name	installation_UUID	touch_intervall
4	28.10.2015 09:34	7	1	975	408	3	PflegeTab1	B27AB146-2F5C-4712-9349-00E6DA652EFE	1.55181E+15
6	28.10.2015 09:34	7	1	975	408	3	PflegeTab1	B27AB146-2F5C-4712-9349-00E6DA652EFE	1.55181E+15
7	28.10.2015 09:34	7	0	975	408	3	PflegeTab1	B27AB146-2F5C-4712-9349-00E6DA652EFE	1.55181E+15
8	28.10.2015 09:34	7	1	975	408	3	PflegeTab1	B27AB146-2F5C-4712-9349-00E6DA652EFE	1.55181E+15
11	28.10.2015 09:34	7	1	975	408	3	PflegeTab1	B27AB146-2F5C-4712-9349-00E6DA652EFE	1.55181E+15
12	28.10.2015 09:34	7	0	975	408	3	PflegeTab1	B27AB146-2F5C-4712-9349-00E6DA652EFE	
13	28.10.2015 09:34	7	1	975	408	3	PflegeTab1	B27AB146-2F5C-4712-9349-00E6DA652EFE	1.55181E+15
14	28.10.2015 09:34	7	0	975	408	3	PflegeTab1	B27AB146-2F5C-4712-9349-00E6DA652EFE	
15	28.10.2015 09:34	7	1	975	408	3	PflegeTab1	B27AB146-2F5C-4712-9349-00E6DA652EFE	1.55181E+15
16	14.11.2015 14:51	0	0	4195	1285	15	PflegeTab1	E170EFF5-D1D8-4D2A-BDC1-FF1473327AE1	
17	14.11.2015 14:51	0	1	473	479	15	PflegeTab1	E170EFF5-D1D8-4D2A-BDC1-FF1473327AE1	5.74403E+15
18	14.11.2015 14:51	0	0	4605	1195	15	PflegeTab1	E170EFF5-D1D8-4D2A-BDC1-FF1473327AE1	
19	14.11.2015 14:51	0	1	4605	1195	15	PflegeTab1	E170EFF5-D1D8-4D2A-BDC1-FF1473327AE1	9.99079E+15
20	14.11.2015 14:52	0	0	441	145	15	PflegeTab1	E170EFF5-D1D8-4D2A-BDC1-FF1473327AE1	
21	14.11.2015 14:52	0	1	441	145	15	PflegeTab1	E170EFF5-D1D8-4D2A-BDC1-FF1473327AE1	2.92412E+15
22	14.11.2015 14:52	0	0	334	153	15	PflegeTab1	E170EFF5-D1D8-4D2A-BDC1-FF1473327AE1	
23	14.11.2015 14:52	0	1	334	153	15	PflegeTab1	E170EFF5-D1D8-4D2A-BDC1-FF1473327AE1	3.31566E+15
24	14.11.2015 14:52	0	0	2345	1555	15	PflegeTab1	E170EFF5-D1D8-4D2A-BDC1-FF1473327AE1	
25	14.11.2015 14:52	0	1	2345	1555	15	PflegeTab1	E170EFF5-D1D8-4D2A-BDC1-FF1473327AE1	3.55666E+15

Figure 2 - Touch raw data

Description of attached code and data

Three folders can be found, with various code-files that are connected like this:



Study1

-- AggregateFeatures.R --
Load aggreagated data Study1
Filter out Spell game and 'normal' condition
Prepare set for later merge with Study2:
- create 'Handedness'
- manually code 'condition' variable following the standard of
Study2, described in the "Condition explanation table"
disabled code-parts contain:
- a routine to spot correlations and output correlation,
p-value, and lm-coefficient in one table for all detected
variables.
Output: AG.csv

Study2

-- aggregation.R --
The code is taken from Carolas aggregation and adapted to work on this
new data set. It aggregates InteractionDataStudie01.Rda into higher level features.
In the outsourced code-file
"preprocessData.R" is deactivated in big parts.
UserID=216 is excluded as it is
"the guy who showed up for the wrong study" (and did not complete it).
Note that this file run only, if there are exactly 12 sessions for every

subject. Otherwise the big loop in the end crashes.
line 54-83 provides tools for checking the number of sessions per VP,
and if necessary removing them.
output: 1) FeaturesStudie_cleaned.Rda
2) AggregatedFeaturesPerSession_cleaned.Rda
Note: The last loop in the code takes around 30 min to run. However, it works and
I simply had lunch during that time instead of fixing it.
But for the protocol, in my opinion the problem is:
For each VP the code runs through all 500 sessions. This is unnecessary as
every VP owns and needs only a handful of sessions, and not all.
The code calculates 497 values for every VP, which are not needed and
later discarted.

-- extract_duration.R --
for some reason the initial duration-computation does not work on the Study2 data.
this code a) calculated the sessions durations,
b) manually fixes some sessions which were 'split in two': because the App
sometimes crashed, Giorgio let the subject play the remaining time in what is
counted a new sessions. These few cases are manually merge into one session.

Data Analysis

-- ProcessStudieGiorgio.R --
load the two aggregated features set (studie1 and studie2)
divide the 'count' variables by the session length.
unify columns names, columns and variable-types between the
two data sets.
output: data_aggregated.csv --> contain both data sets.

-- DataAnalysis.R --
throw out conditions that are not 'normal' scenario.
dimensionality reduction with 0.8
interpolation of missing values in Big5 with means.
several handy loops for inspecting the data.

-- runAll.R --
the classification.
i found it very cumbersome in R to write functions which take

columns of data-frames as agruments. It is therefore much easier
to user the "search" toolbar on the upper side of the code-window,
type "extra" and in the replace window e.g. "neuro".
Once the code is written for one dimension it can be copy-and-pasted
this way within few seconds.
the example works with extraversion.
- code sets up table which neatly reports confidence intervals,
significance, etc for all classifiers.
- target class is made.
set.seed operators make sure results are compareable cross classifiers.

-- big5 visual --
number of plots for the dimensions.