Forschungsinstitut
Cyber Defence
Universität der Bundeswehr München

Have you used this password before?

Investigating gaze behaviour to detect password reuse

Johannes Schütte
Bachelorarbeit

Abgabedatum:	31. März 2020
Betreuerin/Betreuer:	Yasmeen Abdrabou
	Ken Pfeuffer
Prüfer:	Prof. Dr. Florian Alt

Zusammenfassung

Heutzutage müssen Nutzer eine Vielzahl von online-Accounts verwalten. Viele dieser Accounts enthalten sensible Daten. Ein Schutz dieser Daten, die von persönlichen bis zu Bankdaten reichen, liegt im Interesse des Nutzers. Diesen Schutz versuchen Anbieter im Internet durch Authentifizierung anhand von Nutzernamen und Passwörtern zu gewährleisten. Bei einer steigenden Anzahl der genutzten Anbieter steigt somit auch die Anzahl dieser Logins. Da die meisten Menschen ihr Gedächtnis zur Verwaltung dieser Logins nutzen, verwenden viele der Einfachheit halber ein Passwort auf mehreren Seiten. Dies birgt jedoch die Gefahr, dass ein Passwort das in falsche Hände gerät, einem Angreifer Zugriff auf gleich mehrere Accounts mit gegebenenfalls sensiblen Daten gewährt. Aus diesem Grund gilt es die Wiederverwendung von Passwörtern zu verhindern.

Während ein Anbieter keine Möglichkeit besitzt festzustellen, ob ein Password bereits bei anderen Anbietern verwendet wird, könnte das Verhalten des Nutzers, zum Beispiel beim Tippen oder Ansehen der Webseite dieses anzeigen. Noch gibt es keine Möglichkeit dieses Verhalten zu nutzen, doch mit der Entwicklung moderner Eyetracker, die sogar in Smartphones eingebaut werden können, kann sich das bald ändern. Mit dem Auswerten des Nutzerverhaltens könnte interaktiv ein Verhalten erkannt,und direktes Feedback erlaubt werden. So könnte eine Webseite den Nutzer zum Beispiel auffordern ein neues Passwort zu vergeben, wenn das Verhalten auf Wiederverwendung hindeutet.

In dieser Arbeit wurde eine Studie durchgeführt um das Verhalten während Passworteingaben zu untersuchen. Durch das Anwenden von maschinellem Lernen auf in der Studie gesammelte Blickdaten wollten wir herausfinden, ob es möglich ist, Nutzer anhand von Blickdaten zu klassifizieren. Dabei unterschieden wir nach den Verhaltensmustern dass sich ein neues Passwort ausgedacht wurde, ein Bekanntes geändert, oder ein Existierendes wiederverwendet wurde. Darüber hinaus wollten wir das Verhalten von Nutzern während der Passworteingabe erforschen und besser verstehen.

Die Ergebnisse der Arbeit zeigen, dass die Wahl des Passworts einen Effekt auf das Nutzerverhalten bei der Eingabe hat. Leider gestaltete es sich als schwer, diese Unterschiede klar zu extrahieren und für maschinelles Lernen klassifizierbar zu machen. Dadurch war es uns nicht möglich ein allgemeines Modell des Nutzerverhaltens aufzustellen. Dennoch ermöglichte es die Analyse und Interpretation verschiedener Charakteristiken, einen Einblick in das Nutzerverhalten zu erlangen und eine Grundlage für weitere Forschung zu legen.

Abstract

Nowadays users have to manage multiple accounts in their daily digital life. These accounts contain a lot of sensitive data. On behalf of the user it is important to protect this data including personal and financial information. Services across the internet use logins by username and password. Increasing use of these services leads to an increasing number of logins to manage. To cope with the challenge of managing multiple accounts and remembering the password for them, users begin to reuse passwords. This is problematic because once such a reused password is cracked, a possible attacker would gain access to not only one but multiple accounts. Therefore, users are advised to refrain from using the same password for various accounts.

While there is no way an online service can confirm that the password being created has not been used on a different platform, some typing behaviours might indicate that. A feature that could extract this behaviour interactively built into a website, could alert the user to create a new and secure password. Up to this day there have not been any possibilities to determine this behaviour. But with the development of eye trackers, a tool to measure it becomes convenient. It is already built into devices like smartphones. To implement such a feature, the user behaviour first has to be investigated and understood.

In this work we conducted a study to observe gaze behaviour during password creation. We extracted gaze features from collected eye movements and used them to classify whether a user created a new password, changed an old one or reused an existing one. We applied machine learning to the features in order to classify users automatically. Also, we wanted to find out if and how the behaviour differs.

The results indicate that differences regarding user behaviour do exist. Nevertheless, the machine learning algorithms that we used were not able to classify behaviours based on these exact differences. This did not allow to implement a model to automatically classify how users chose their password due to various characteristics of gaze behaviour. Still, it allowed us to gain insight into different features of eye movements and introduce ideas as a foundation for future research.

Task

The aim of this project is to conduct a study to investigate typing behaviours that can tell us if the user is reusing a password. One of the challenges of this project is to find a way to save the typing behaviours (e.g. pauses, typing speeds, errors), without storing the actual password that was entered.The collected data should be analyzed thoroughly. The outcome will then be used to build a model that can predict whether the user is reusing an old password, making a slight addition to an old password, or creating a completely new one based on the user's input behaviour.

Tasks:

1. Literature review on gaze behaviour and password creation
2. Implementation of the platform for passwords collection
3. Running a user study to collect data in the wild
4. Running machine learning on the collected data
5. Analyze gaze data and machine learning output and report the results

Eigenständigkeitserklärung

Hiermit versichere ich, die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst, die Zitate ordnungsgemäß gekennzeichnet und keine anderen, als die im Literatur/Schriftverzeichnis angegebenen Quellen und Hilfsmittel benutzt zu haben.

Ferner habe ich vom Merkblatt über die Verwendung von studentischen Abschlussarbeiten Kenntnis genommen und räume das einfache Nutzungsrecht an meiner Bachelorarbeit der Universität der Bundeswehr München ein.

München, 2. April 2020 .....

Contents

1 Introduction

Nowadays logins are required everywhere across the internet. In a world that is constantly becoming more digital users face the challenge of managing multiple accounts for these logins. Accounts are not only used to protect user relevant resources, for example bank accounts, but also to identify and track users through user profiles for an online newspaper as an example. Due to the lack of alternatives, most accounts are accessed through a user name and a password. The more accounts a user owns, the more logins are required. According to a study from the year 2016 a user has to authenticate averagely 47 times per day [31]. Every authentication requires the belonging login data a user has to remember. Most people rely on their memory to manage accounts. A problem when using only this method is that passwords are often kept short and easy to remember them better. Also passwords get written down on paper or in a digital file if there are too many [12]. Finally passwords are reused across different accounts when it gets too hard to think of them and remember different login information.

Different strategies and tools already tried to deal with these problems. Password managers safe entries for different websites encrypted on a device. They can offer a secure way of managing passwords. Unfortunately they are either unknown or often seem too complicated for users. Also some people evaluate their current strategy (based on memory) as sufficient[1]. To force longer or allegedly stronger passwords, an increasingly number of websites and companies use so called 'password policies', which compel a certain length or the inclusion of special character-sets in a password. Unfortunately, the good intention may result in the opposing effect[29]. To cope with the new challenge of remembering these longer and more complicated passwords users might reuse already created ones across different accounts.

For IT-security means it is indispensable to avoid poor password choices and password reuse. So how is it feasible to indicate this security risk? A way of changing users behaviour and attitude towards passwords is to create awareness and educate the user. As interactive feedback during creation has an effect on people [51], an effective approach could be to do so while creating the password. Gaze analysis offers an ideal possibility to interact in this situation. It does not affect the person while typing a password since it happens in the background. In addition, gaze is a raw resource since it is a product of unconsciousness and not controlled. Therefore, gaze data is hard to manipulate and offers individualizing features. Gaze allows to access new properties like areas of interest, focus times or fixations on different environment elements. User behaviour can be researched on an additional level, going beyond self reporting, experience reports and questionnaires.

This thesis aims to find a way of classifying password creations by evaluating gaze data. A user study was conducted where behaviour of people during password creation was captured. It was captured by an eye tracker and resulting data was evaluated regarding different gaze metrics. By training a machine learning algorithm with these features, passwords should be categorized regarding the categories: entirely new ones, reused ones and reused ones that were changed by the user. Afterwards this model could be adapted to real logins and give advice to the users that reused a password instead of creating a new one interactively. In addition, advice like thinking of a different password or using tools like password managers could be given.

The thesis starts by presenting related work in the field of password management, creation and reuse. Also, applied gaze metrics are introduced. In the following concept-subsection the research questions and the resulting study are described. Different elements of the study are detailed and their usages are explained in the Design Consideration subsection. The implementation subsection gives an overview over software, starting with password and gaze collection and finishing with applying the machine learning algorithm. Afterwards, the study's setup is presented and the procedure is described. Finally, we present our findings and discuss them. Based on the outcome we give suggestions for future work and improvements in the last section.

2 Related Work

To our knowledge we are the first to investigate gaze behaviour during password creation processes. However, various research has already done regarding the different subtopics 'Password Creation', 'Password Management' and 'Password Reuse'. Furthermore, we give an overview of 'gaze metrics' and 'gaze as a biometric'. Finally, the machine learning is introduced and the algorithms we implemented are described regarding their functionality.

2.1 Password Usage and Analysis

Passwords are relevant not only to protect possessions like personal data or bank data, but also to identify and track user activities [12]. For example on Wikipedia no sensitive data has to be protected. Only the administration needs to know who changed something and when. This results in logins with creation of passwords, where not even a resource is protected. Still the user is left with the task to remember another password in his repertoire.

Even though there are alternatives that are much safer than a regular login by username and password for sensitive resources (hardware tokens[7], phone-based authentication[36]) and more convenient methods to identify users (OpenID [44], Biometrics[26]) these alternatives rarely get used since they are not as convenient or deployable [5]. As a direct consequence logins with passwords are used everywhere across the online and offline world.

Various studies reveal that users tend to have different accounts for different use cases. A smaller study with 26 people conducted by Notoatmodo [34] found that on average users have around 12,9 accounts with almost 8,1 passwords to protect these accounts. Dhamija and Perrig [6] hosted a password study with a similar amount of participants (30) to reflect on how many unique passwords users use. They only found 1-7 unique passwords among 10-50 recorded password instances. The results of another study by Gaw and Felten [12] including 49 participants match with Notoatmodos results: an average of 12,5 accounts was found to be used by individuals. An average of 8,5 accounts were recorded by two studies: the first one, a survey study by Riley [45] including 315 college students and the second one by Hayashi and Hong [19], where 20 users wrote a diary about logins. Probably the biggest study with 544.960 participants was conducted by Florencio and Herley [10]. They found 6,5 unique passwords for an average of 25 online accounts over a period of 3 months. With increased use of online services the number of accounts per person will most likely grow. That challenges users to create increasing numbers of passwords. Even though the amount of accounts and passwords varies strongly, it is noticeable that the number of passwords never matches the number of accounts. The reason for this is a reuse of passwords across different websites. This reuse is a risk since attackers might get access to multiple logins with capturing just one of the passwords. The phenomenon of reuse is also researched very well.

2.2 Password Creation and Password Policies

According to experts strong passwords are randomly created and include different sets of characters. The longer they are, the harder they are to crack. Also, user's safety requires having different unique passwords for each account [22]. However, such passwords are also hard to remember for the human brain, since it is not a combination of meaningful words, numbers or names as in our regular language. Especially with the described rising number of accounts, a random combination of characters for every single one is a very tough challenge.

When people create passwords they often follow a certain pattern. According to Ur et al. [52] 88% of users in their study either reused one password or few passwords across multiple logins, had a personal algorithm to find a new combination or built new creations around reused elements. These elements include names, dates or things in users range during creation. A study

done by Inglesant et al. [21] found that users tend to create passwords using elements found on their working desk. In a diary study by Duggan, Johnson and Grawmeyer [8] password-characteristics were recorded by 22 users for 7 days. More than 50% of all passwords were made up from a 'word or name or an abbreviation of a meaningful phrase (the initial letters of a sentence, e.g. the first letters of the sentence 'Who has sent me new mail?' would be 'Whsmnm?'). In 2010, Shay et al. [48] had almost 500 users fill out a survey about new passwords. 78% of the respondents claimed to create their password based on just a word or a name. To create new passwords, users do choose randomly but based on a personal preference, which results in shorter passwords based on personal elements in order to be able to remember them. Once they have a satisfactory selection they reuse it to memorize it better. The passwords users make up based on simple elements like names or dates oftentimes are easily guessable. In addition, they are mostly short, since they are easier to remember.

One way to avoid these weak and short passwords is to apply password policies on the creation process. Popular policies like 'minimum length of 8' which is also recommended by the 'National Institute of Standards and Technology' (NIST)[35] or 'include at least x special characters' got quite popular and aim to achieve longer and more complex creations resulting in harder-to-guess passwords. Unfortunately the effect of password policies is not always the planned one. Since many users already have their set of passwords memorized and do not see a need for a change but still have to abide by the new rules, the existing passwords fall under one of those categories:

1 A password fits the password policy. Often no need for a change is seen. Even though a user could take the chance to change the old password and make up a new one there seems no actual need for it from the user's perspective. This leads to a reuse, a behaviour documented by Komanduri et al. [29] in a large scale study.

2 The user makes up a new password according to the password policy. The new password is longer and contains defined character classes (like special characters, numbers or big letters). Ideally it is constructed in an unobvious way which means it does not follow an obvious pattern (like 'initials+birthday'). This is the goal admins want to achieve with their new regulation. Unfortunately, applying password policies does not always result in this change.

3 A user does not want to create a new password to align to the new policy but a reused one is too weak to comply with the conditions. As a conclusion he modifies a reused password until it matches the needs. Either incrementally characters are added (e.g. include an exclamation mark at the end if the policy requires special characters) or passwords are modified until they meet the conditions (e.g. '12345' becomes '12345678' if the policy requires 8 characters). This behaviour was documented by Komanduri et al. [29], too. They also found the modifying behaviour typical for more complex password policies, where more conditions were applied and users challenged to make up passwords with many features. Inglesant and Sasse [21] claim that users keep " 'good' passwords (that are memorable and conform to the policy) as a 'resource' " for multiple new ones. Unfortunately Zhang et al. [56] found modifications to be predictable since they often are easily guessable (for example replace 's' with $ or add '.' at the end if special characters are required).

4 This option is possible in addition to 2 and 3: The user adapts to the new requirements but in order to cope with the challenge of remembering a new or changed password he starts to write it down. Besides Inglesant and Sasse also Stanton et al. [50] found evidence for this behaviour. Today, on the contrary to earlier recommendations, some experts discuss about writing a password down in a special way as a memory aid. This

might be a possible way to cope with the challenge of remembering increasingly longer, more complex and a higher amount of passwords (for example one can create a paper with an individual hint and not the actual password).

For the efficient use of password regulations a happy medium has to be found between forcing more complex passwords and making these still memorable for a user. Shay et al. [47] investigated policies with a simulation tool, surmising stricter rules to lead to better passwords but also to the noting of them, while weaker policies leave systems vulnerable. This simulation matched results of a study by Proctor et al. [41], where stricter policies lead to hard-crackable passwords but also to hard memorable ones, which is called 'memorability/security trade off'.

2.3 Password Management and Reuse

As we introduced in subsection 2.1, users face the challenge of managing an increasing number of passwords. As this number rises with age and the ongoing process of digitalization, more and more passwords have to be managed. Tools like password managers are unpopular and find little use, resulting in three coping strategies for users: sharing, writing down and reusing passwords. In Shay et al.'s study [48], 28% admitted sharing their old or current passwords with other people. Meanwhile, over 80% conceded they used a set of passwords for multiple accounts. A study by Zviran and Haga [57] from 1999 found 35% of 997 participants wrote down their password on paper. Recent research done by Shay [48] showed that only 13% of all users said they wrote down passwords, which could indicate a trend away from writing passwords down, towards reusing them. As noted by Grawmeyer [18], in a diary study including 22 participants, 86 out of 175 recorded passwords were reused across different websites . Florencio and Herley's large scale study [10] discovered an increase of reuse with more accounts, meaning that a higher number of accounts did not correlate with a higher number of passwords. Concretely speaking, 50% of all users had just 3 or less password families (where a family could also be a resource for numerous but similar passwords like 'password1' and 'password2'). In addition the strongest passwords were used on few websites, while weaker passwords were widespread across remaining websites. An explanation for this behaviour is provided by Wash et al. [53] who conducted a study with 134 participants and presumed that reused passwords are frequently entered passwords. Users seem to choose an easy to memorize password which often is weak but offers the comfort of not having to create a new one. If then a login is required, there is no need to search through many passwords.

As a coping strategy for managing a high number of passwords, reuse is a comfortable solution. It does not require any physical premise like tokens and even allows a user to choose a (earlier described) 'strong' password, because the effort of learning it decreases the more it is used on multiple websites. Regrettably, behaviour like this results in a huge security issue. With just one revealed instance of a reused password, an attacker gets access to, in the worst case, all accounts using this password, resulting in a 'domino effect of password reuse' [23].

Especially for security purposes, the avoidance of password reuse is important. The use of gaze data can allow us to detect and nudge the user in case of password reuse. Since eye trackers are ubiquitous nowadays, with major enterprises like Apple buying SMI (Senso Motoric Instruments), or Facebook owning 'The Eye Tribe', eye trackers offer a prevalent possibility.

The review of literature shows that password reuse is a problem and leaves systems vulnerable. Often, created passwords are too weak because they are also easy to memorize. Even though the number of logins increases in a world that is becoming more digital, no solution to the problem of weak, reused passwords has been found yet.

2.4 Eye Tracking as a Tool

In the field of eye tracking (as far as we were able to find out) there has not been any research related to the process of password creation. Nevertheless, the analysis of gaze has been used in human computer interaction research for a long time and useful features have been explored. An application of eye tracking enables a deeper insight to human behaviour with metrics, only in this field available.

Gaze Metrics

Today many different metrics can be found. The most important ones include fixations, saccades, scanpaths and the pupil diameter.

Fixations occur when the gaze maintains focus on a specific location. This means the eye lays stable on a position for a minimum duration of at least one hundred milliseconds [25]. During a fixation content can be processed. The duration of fixations can show an increased interest in elements or indicate complex content which has to be decoded [14]. In the year of 2004 a study by Poole et al. [39] found out that the number of fixations per area of interest indicated a higher focus and more importance to a user. An area of interest is a declared subsection of an interface. The developers of the interface define the area of interest. Poole et al. also set the number of fixations per area of interest in relation to the typed characters in order to adjust different text lengths. In Justs and Carpenters work from 1975 [27] they stated that a longer duration of a fixation indicated either problems extracting information or a higher level of interest in a certain element compared to others.

Saccades are quick eye movements between fixations. The focus is shifted from one point to another. During a saccade no content can be decoded. Still they can be used as a metric to evaluate an interface. Goldberg and Kotval [15] discovered that more saccades indicated more searching for content. Goldberg et al. [16] and Sibert et al. [49] found that cues have an impact on saccades. Meaningful cues lead to a higher distance of saccades while regressive saccades are an indicator for less meaningful cues.

Scanpahts are saccade-fixate-saccade sequences [38]. They represent the process of a focus shift. The direction of a scanpath can be used to indicate the efficiency of arranged elements in a user interface, for example an ideal scanpath for search page would be a straight line to the desired target with short fixations. Goldberg and Kotval [15] found a correlation between the efficiency of an interface and the duration and length of a scanpath.

The pupil diameter is a metric to measure the cognitive effort (stress) that a user is put under. Larger pupils may indicate a higher cognitive workload [32], [37], but since the pupil size is also influenced by other factors like light levels and tiredness, this metric may be falsified [17].

Gaze as a Biometric

The idea to use gaze as a measurement for physical characteristics is more than one hundred years old [25]. Eye movements are very intuitive and hard to manipulate because they are unconscious. They are individual to a user and can be helpful to distinguish between them. Also, they reveal a lot of information about areas of interest and user experience. Modern eye trackers are unobtrusive and not perceived spurious at all [33]. The use of these modern eye trackers allows the measurement of natural and unbiased user behaviour, from which different features can be extracted. Another advantage is that eye tracking leaves no further restrictions

regarding usability. Gaze data can be collected in the background and processed, which makes it convenient for users[33]. The access to eye movements can be permanent and collectable.

Today, gaze can even be used as a biometric [55]. That means it provides a way to identify a user based on his physical and biological properties. A good biometric should be 'unique, universal, and permanent over time, easy to measure, cheap in costs, and have high user acceptance' [4]. Eye movements are unique, universal and can be collected over a permanent period of time. In addition with the dissemination of eye trackers, gaze becomes easy to measure and cheaper. Also with the usage of eye trackers in daily devices like smartphones, gaze is more easily collectable and the collection accepted (sometimes not even consciously) by the user [28].

2.5 Machine Learning

The notion of machine learning was introduced by Samuel Artur in 1959 [46] and is about computers using statistics and patterns to decide, rather than instructions. In the Cambridge dictionary, machine learning is defined as 'the process of computers changing the way they carry out tasks by learning from new data, without a human being needing to give instructions in the form of a program' [40].

The big difference to a classic software lays in the structure of the machine learning algorithm. While usually a code is written to complete a task in a static way, machine learning figures out its own way by 'learning' from examples and adapting them to a wider context [2]. Machine learning is very powerful when working on big data sets and data sets that are too big for humans to dive through [54]. A computer can process more and other types of data, for example numbers, way better than a person. It connects this advantage with the human strength of learning and adapting to different, changing conditions. Even though an approach to analyse the data, for example by visualizing it, is a human-friendly approach, it leaves a person with the impossible challenge of connecting lots of confusing features to different patterns. This is where machine learning comes to use.

To classify data two algorithms were used. One was the 'Support Vector Machine' (SVM) and the other one the 'Random Forest Classifier'. Both classifiers were supervised machine learning algorithms. That means that they use labeled input data to learn how to classify it. During the training phase the algorithms knew what the belonging label or outcome class (in our study the password choice: new, change, reuse) was and tried to find similarities in the belonging features (in our study durations, ratios or fixation metrics) for every class. They build a decision function based on these findings that helped them to distinguish between the different classes. After the training phase the algorithms should be able to apply this decision function to unknown unlabeled data, which means they classified a class just based on features they received.

The 'Random Forest Classifier' creates decision trees where all features are a node and their values a branch. The trees are built randomly. When a user-defined number of trees was grown (a 'forest'), all get the same input and their results are compared. The most appearing output is chosen as a result. The decision function is a majority vote of all tress. [30].

A 'Support Vector Machine' uses an algebraic approach to classify data. In this algorithm, each data item is represented as a point in a n-dimensional space, with n being the number of features [42]. For example a data point with two features would be represented in a two-dimensional space as a point with x- and y- coordinates. To classify groups the algorithm creates a hyperplane with a maximised margin between labeled groups. In a two-dimensional space it would be a line between different groups. The hyperplane is then used as the decision function for new data. Depending on the position to the hyperplane the unlabeled data point is categorized. In a two-dimensional space a unlabeled point would be classified depending on which side of the line it is localized [3].

3 Concept

The concept of this thesis included a study to research gaze behaviour of people while creating passwords, the evaluation of collected gaze data and a machine learning setup classifying the features. Our goal was to be able to tell from gaze characteristics how a user-password was chosen.

In order to be able to differentiate between gaze behaviours we set up 3 main research questions:

Question 1: Have you used this password before? Is it possible to tell from gaze analysis whether a user reuses a password, slightly changes an old one or creates a completely new one?

Question 2: How can gaze relate to password creation? In other words: What does gaze behaviour look like during the process? Are there features that indicate a specific behaviour?

Question 3: Which features better reflect password reuse from gaze data? The collected gaze data offered possibilities to analyse different features and metrics. The goal was to find out if some indicated a specific behaviour better than others.

In order to design the study optimally a few hypothesises were set up in advance and later verified and refined by results from pilot testing.

As seen in Figure 1 the structure of the study was divided into different stages. After the consideration of different factors for our study we started implementing the interface and the gaze collection. When the setup was completed the study was conducted with volunteers. They clicked through the interfaces we implemented and made entries, while simultaneously, their eye movements were measured. We included two interfaces, a known mail-interface and a popular newspaper. Users had to create passwords on both interfaces and answer a questionnaire at the end. After finishing the study, gaze data was analyzed to find areas of interest and good features to classify users' password choices. When the extraction and processing of this was done, a machine learning algorithm was applied to learn to distinguish between different password choices based on these features. The trained algorithm then could work as a classification tool for other people that created a password.

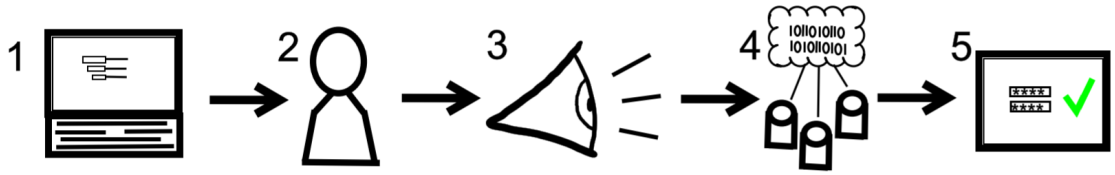

Figure 1: **Procedure:** Prepare two interfaces in the study (1), let participants do the study (2), collect eye movements(3), analyse gaze data (4), classify a persons behaviour by a machine learning algorithm (5)

4 Methodology

In order to be able to extract as much information from the users as possible, we carried out a focus group to understand users better and to be able to identify the metrics and scenarios. Then we continued by designing the study based on the considerations.

4.1 Focus Group

The meeting of the focus group was held in the community room of the lab of the usability and privacy department. A number of 9 people participated, all of them were researchers in the field of usability and privacy. Four participants were male and five were female. Short questions were prepared and individual answers noted on post-its, so they could be presented on a board. When all post-its were collected, the results were attached to the board and grouped. The grouping resulted in hypothesises or further questions. The asked questions and their results were:

How does behaviour differ between creating a new, changed and reused password?
In order to distinguish between different behaviours we first had to figure out how these behaviours may look like. The focus group was asked to answer this question not only based on their own experience, but also by what they estimated a general behaviour might be. The answers agreed on different hypothesises: The time to create a new password would be the longest because users had to figure out a new combination. In addition, they would have to find the corresponding keys on the keyboard without being able to rely on their cognitive memory, which remembers an often typed combination. To reuse a password would take the least effort and least amount of time, since the password was directly available and could be typed immediately. In between these behaviours the change would be a mix of both behaviours, tending more to one of the sides, depending on the change that was made.

Which Scenario is good for such a study?
The focus group worked out different criteria to choose a fitting scenario. First of all the interface should be well known to many participants, so they could identify with the use of it. Second, it should be accessed as often as possible with a login process since we wanted a behaviour that is as natural as possible for our login. Finally, it should be a serious website with relevant resources to protect. The reason for that is the fact that multiple people in the focus group reported that their behaviour varied between the sensibility of the website, for example a stronger rated password was used for a bank account compared to a less relevant website like a newspaper. In order to investigate these reports, we decided to involve a second interface, categorized by the focus group as 'non-sensitive' to be able to find out about the difference between those as well. It was also discussed to include even more than one interface to find distinctions between types of interfaces, for example social media websites, seldom visited websites or online shops. Eventually, we abandoned this idea to avoid creating too much of a cognitive workload for participants, a behaviour that was observed by an attendee of the focus group in one of her previous studies.

What should the interface look like?
Regarding the design and features of the study, the focus group proposed a re-entry of the password for different reasons. With the re-entry of a password we would have the possibility to compare between two immediate typing events. The duration per field might be different regarding the new or changed password, because for re-entering a password there would be no time needed to think about a new combination. Also, the re-entry field could serve as a

comparison field, where regular user typing behaviour could be collected and compared to the first entry field.

Another feature the focus group suggested was a re-entering in the end. With a repeated entry it might be possible to distinguish between users who might have just typed random entries for the sake of time reduction and because they wanted to finish the study fast. Users who made up a 'short-time-password' would not successfully complete this re-entering test, because they had to memorize their choice not only in the short-term but at least in the mid-term memory.

Where should the study take place?
We knew in advance that our study should not be too time consuming for a single participant. With this frame as a foundation the focus group suggested not to hold the study in a lab and invite people separately, but to ask them spontaneously in their daily life. The place that crystallized in the focus group was the canteen of the university. This way, we could reach more people more easily, since they did not have to visit a lab or make an appointment for the study. Also participants were accessed in a natural environment where they did not have to prepare for a study. Finally, we could access a complete spectrum of university personnel as the canteen was a place all people (not only soldiers nor only academic staff) visited.

Other findings
There were further findings in the focus group but they were not considered in our study because of time limitations. Also we did not want to focus on too many aspects at once. Still, some suggestions could be the foundation for future work. For example one person proposed to hold the study under different circumstances regarding privacy. It may be interesting to know if users behaved different if they knew people were observing them or if they felt private.

4.2 Design Considerations

Based on the results of the focus group we designed the study. Other than that, further thoughts and hypothesises we proposed were put into consideration when designing the study as well.

4.2.1 Interfaces Used

Since users should be able to relate to the interface, we wanted to choose a well known and important website. Different possibilities were considered such as a google account, as most people nowadays own one and access it quite often. But because there was no use for a regular login since many people stay logged in, we abandoned this interface. At the end two options were left. One was 'Übersicht, Verwaltung und Abfrage des Studierendenbereich' (UEVAS), an online managing service by the military part of the Bundeswehr University. Since soldiers, the biggest part of the students at the university where this study was conducted, are obliged to check UEVAS at least two times a day, the interface and login is well known to them. The big disadvantage was that really only soldiers were able to use the service and the only ones to know the interface. Because we wanted to include a wider range of people we decided against UEVAS. Our final choice was the mail-interface of the university, shown in Figure 2.

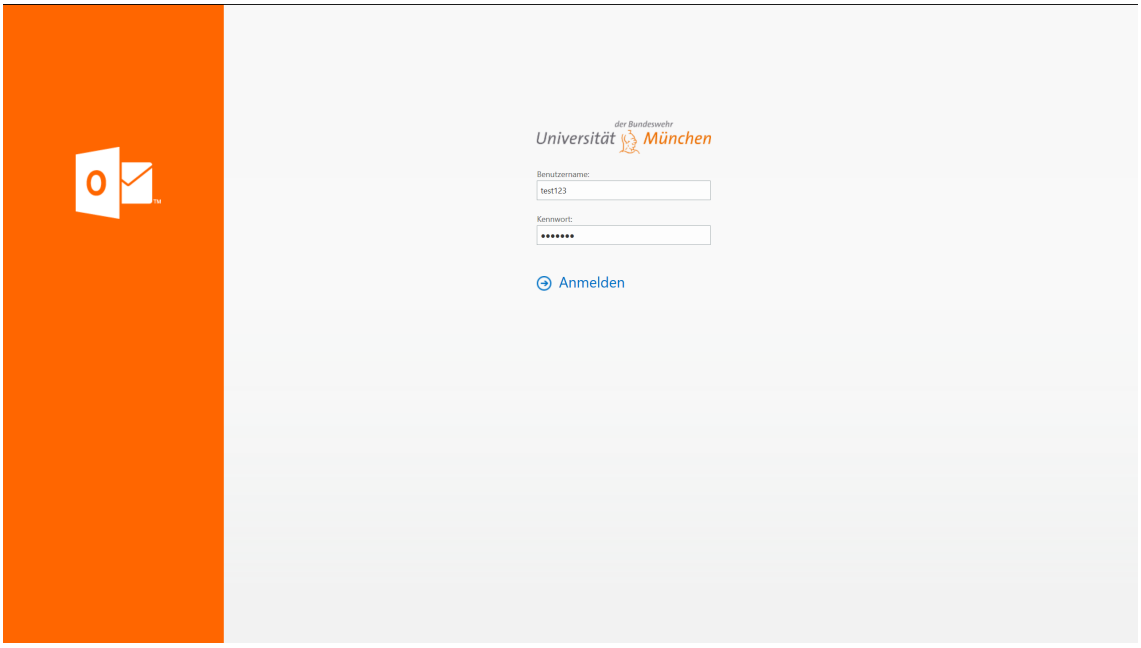

Figure 2: Original Mail-Interface

It gave us the possibility to include not only soldiers but also academic staff and university employees in our study. Since the study was aimed to be held completely on the campus, this would cover all people at the university. The mail-interface is also a website that is visited very often. Students are required to check it as frequently as the previous option (UEVAS) and receive important information regarding academic- and campus-news in this way. So we had an interface users knew well and used regularly. For our study, we decided to mimic the interface design as realistically as possible so users would get the feeling of an actual login situation.

Since we also wanted to find answers whether the chosen scenario had an impact on password creation we decided to introduce another interface. The difference compared to the first one should be that users would rate the website less sensitive in order to behave differently. As earlier researches could not clearly find evidence for or against a varying behaviour regarding password creation on different sensitivity-rated websites, we wanted to cover this aspect, too. As we already covered a website that involved sensitive data, we had to find an interface that seemed 'unimportant' to a user. Our first idea was to mimic a public Wi-Fi like an airport or train Wi-Fi, but these usually just required an e-mail address. Another idea involved a newspaper website. We chose the second highest circulated newspaper in Germany, the 'Sueddeutsche Zeitung' [24] shown in Figure 3 so users were not confronted with a random interface but a situation they could relate to. The tasks were equal to the first ones, but now users also had to think of a new username. We did not apply a password policy to the second interface to support the feeling of a less sensitive website.

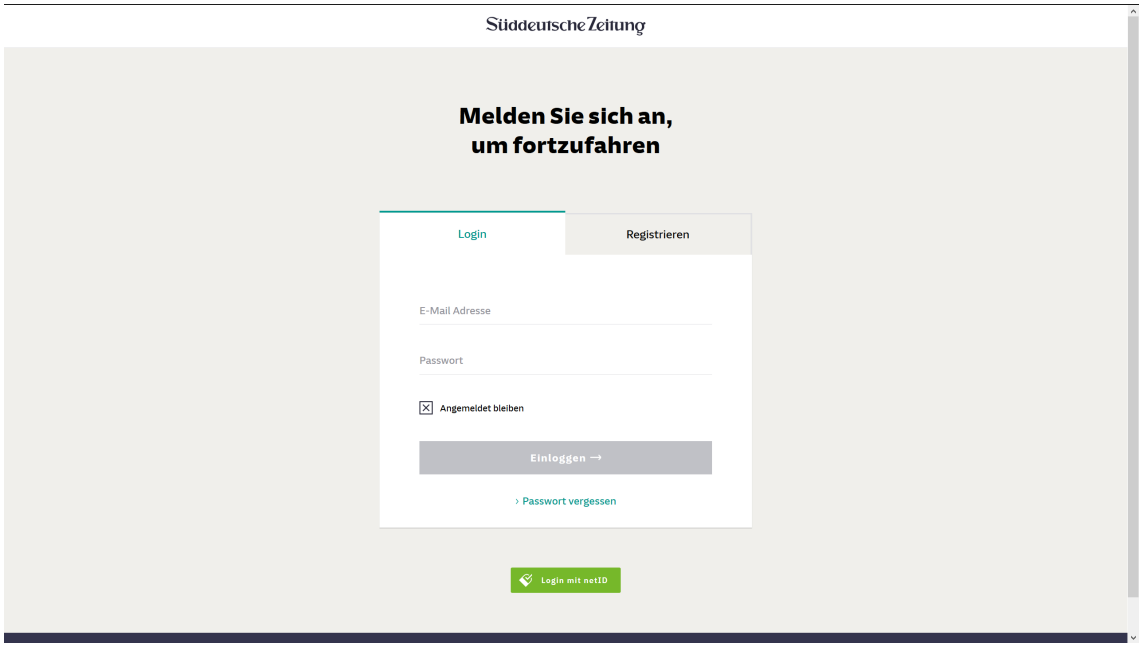

Figure 3: Original Newspaper-Interface

As we wanted to achieve the most natural behaviour possible we also did not want to reveal the true nature of our study (security research related to password creation) and decided to explain users that the goal of our study was to investigate the usability of interfaces at first. When the study was completed we explained the real nature and aim of our study.

4.2.2 Gaze Behaviour

While completing the study we outlined three main areas where a person could look during password creation. The computer screen is the first one. A user has to interact with the creation interface and look for entry fields but might also check during the creation what he or she entered. Area number two is the keyboard. If typing has to be approved, the gaze will most likely switch from the screen to the keyboard and check the pressed keys. This might lead to saccades between these areas or short fixations on keys or the screen. Finally the last possible area that could attract gaze was the 'outside screen area'. This area covered the rest of the environment. People might look here to consciously shift their focus (for example when they think about a new password).

We supposed that gaze behaviour would differ between reuse, slight change and new creation. If an old password is used, people would know in advance what they will be typing. The time on the screen might be longer in relation to the time spend on the keyboard. Depending on the complexity some might not even look at the keyboard at all. In contrast, this area might get attention when a new password is worked out. Gaze may wander around between the three areas while thinking. Also the focus would switch more often between the screen and the keyboard, or might even stay on the keyboard, because users need to verify their input. In total, the time spent on the interface will be longer. Finally, the change of an old password might result in a mixture of both already mentioned behaviour types with a propensity to one side depending on the amount of change. If for example only the last character was changed, it would tend more to the reuse behaviour, while on the opposite the change of a big part would lead to a behaviour more on the side of new password creation, for example if a name in a password had to be changed. Although the user already has a 'password-resource' in mind he or she might need some time to think about changes he or she wants to apply. The focused areas could involve all three with a frequent change between screen and keyboard like in the

second setup described. Depending on the grade of change the behaviour might be closer to a reuse (for example when just a number is changed) or to a new creation (for example when users think about new words in a pattern).

As a main gaze metric fixations were chosen because they offered multiple further metrics (2.4). The total number of fixations could indicate how often a user had to shift his or her focus while creating the password. A high number, meaning lots of switches between areas of interest like different keys or entry fields could indicate insecurity related to the written input, for example while creating a new password. Meanwhile, a low number could indicate a safe and calculated behaviour as when reusing a known password. The area of fixation could reveal the main focus. A new creation of a password might result in fixations in the area of the keyboard since users might look for characters to choose. A higher number of fixations in the area of the screen therefore may be caused by an intentional behaviour, typical for a reuse. Finally the duration of fixations could be another indicator for a special behaviour. Longer durations may indicate either missing need to shift attention (for example away from the entry fields if a user reuses a password), or difficulty extracting information, for example when looking for a screen (while creating a new password).

In addition to the fixation-values different durations were estimated to allude user behaviour as well. The total duration for each interface, the duration for every entry field while it was active and the initial duration. The total duration covers the time between the first sight of the interface after the user clicked the 'continue'-button on the page before and the successful submitting of all entries. The total duration was assumed to be longer, if a new password was created as the time for typing a new combination may be higher and extra time where a person thinks about the new password may be necessary. The initial duration or 'thinking-time' covers the time between the last keystroke event of the 'username'-field and the first keystroke event in the 'password'-field. We estimated a higher thinking time if a person thought about a new password, compared to a password-reusing person who already would knew what to type. The process of thinking about a new password would take longer and it may also be delayed for a changed password, since a change requires time to be made up. The last duration we covered was the duration for each entry field. Also related to the password choice a more time for the password-entry field was estimated when people thought about a new password. Meanwhile, the other entries could be useful as a comparison value. The duration covered the active time of a field, meaning that it had the focus, after it was selected and clicked with the mouse.

Based on the calculated fixations and durations we could extract different ratios in order to be able to compare users. As described, for a new password a higher number of fixations and fixation switches with longer fixation durations was estimated together with more time spent on each interface and a higher thinking time. But not only the password choice may influence the durations. The length of the entries may also leave a trace, because a user would need more time to type a longer entry with more characters or maybe deleted a whole entry to write it again. To keep track of this factor, the number of fixations was divided by the time. The result is the ratio of fixations per second. A higher ratio could mean users had to extract more information (like finding keys on the keyboard).

From the number of keystrokes, also the average time between key presses was accessible. This ratio could also indicate how long a participant needed compared to others. A higher value, which means users took longer on the interface, could be caused by more time to think about a new password. Another ratio that was covered was the rate of screen fixations to keyboard fixations. We estimated this rate to be more on the side of the screen when a user was sure what he was typing, meaning he reused a password. If a new password was created, the user may look more to the keyboard to identify keys he may use. The ratio was calculated by dividing the amount of fixations for the screen and the keyboard.

The base for all hypotheses were own experience and other people's report about password creation. Together with the focus group they influenced the study's design, especially the features covered by and calculated from the gaze data software.

4.2.3 Password Generation

We presumed that the created password itself would also have an influence on user's behaviour. For example, a complex password with different character sets may take longer in creation than one made from letters. To be able to include this in our analysis we wanted to log this feature. Since plain text passwords should not be saved because of data protection, we decided to log password characteristics. These characteristics included different character sets (lowercase, uppercase, digits and special characters) as these could give feedback about the complexity of the entry. More character sets could result in a stronger password, but also in a longer creation process.

4.2.4 Questionnaires

We decided to include two questionnaires in our study. The first one was a demographic questionnaire. Firstly, it was needed to group users and get an overview of different distributions regarding gender, age or experience. Secondly, we also may find correlations between these traits and a password choice, like Notoatmodjo [34], who found that women tend to reuse a password more often compared to men.

Because we did not want to reveal the nature of our study during the entering of password, we decided to place the questionnaire completely at the end and not to ask the questions relating to the first password right after the first interface.

4.3 Machine Learning

The eye tracker we used provided 3 to 5 data points per second. This results in a big amount of data. Also, gaze data and extracted features had to be set in context to the questionnaire results. As a consequence, machine learning was regarded useful to analyse the raw and unexplored data. An algorithm could be trained to find patterns in previously confusing data, too complicated and unclear for humans to interpret. This is the advantage of machine learning: without a strict limitation to the procedure of the code, the computer can still use its capabilities of fast calculations. Unexplored data like we received was legit input for a machine learning algorithm.

The condition the algorithm had to meet was the ability to classify multiple classes (new, change or reuse). Also we had many different features for every user, that had to be processed well. After some research we decided to use two classification machine learning algorithms: 'Support Vector Machines' and the 'Random Forest Classifier'.

More concretely it would be possible to tell if a user created a new password, changed an old one or reused an existing one, just by analysing his or her gaze behaviour.

5 Implementation

For our study three main parts had to be implemented. The guided user interface (GUI), leading the participant through the study, the concurrent collection of gaze data and the evaluation software including processing data and machine learning.

5.1 Guided User Interface

We implemented the GUI in python. Python is a freely available all-purpose programming language. With different packages that provided additional capabilities the creation of a graphical interface is possible. The package we used was 'Tkinter'[11], the most popular package for user interfaces in python [9]. With a canvas as foundation, different elements, so called 'widgets' can be placed to create a site with various functions. Every page of our study consisted of a canvas and different widgets. The widgets we used included:

Buttons Buttons provided a simple way to interact with the user. They could be connected with functions from our code while also helping the user to guide through the study. Other than that, they could be labeled to make them more informative. The sliding through sites was implemented by calling a function when the user clicked a button. This function closed the current page and showed the next one. Also the collection of gaze data started and ended with click on the button.

Check-Boxes These boxes could be labeled and visually showed a tick, when they were selected. The selected value ('clicked' or 'not clicked') was retrievable and was used to implement queries. This feature was needed while implementing the questionnaires. Check-boxes with different labels were placed on the canvas and all answers entered through ticks could be accessed and saved. By comparing different values of a question, we could guarantee that only one answer was chosen.

Entry-Boxes These boxes allowed user inputs by typing. Entries could be saved to a variable and thus were treatable by functions from the code and storable. This widget was used for open questions in the questionnaires and for the mail interfaces. When participants entered their username and password they did it in these Entry-Boxes. With a provided functionality it was possible to show '*'-symbols while filling the password fields.

Image Especially for the interfaces we used the image widget to build the design as close to the real website as possible. For example, the button widget was connected with an image to mime the submit-button of the mail-interface.

Messages Message-widgets allowed to place text elements on the canvas. They were used every time users received instructions or a description was needed.

Widgets were placed on the canvas with coordinates as well as height- and width-parameters.

5.2 Gaze-Collection

With a python binding for the software development kit (SDK) of the eye tracker we used, the collected data should have been accessible for logging. Unfortunately, during the implementation we found some features of this SDK to be available only with a special licence. Since the affordance would have taken too much time, the use of the python binding was no option. As a solution the gaze collection was shifted to a c# program where the access to collected gaze data was possible without an extra licence. As the whole GUI was already written in python we refrained from completely switching the study to c#. The solution to connect both was to send the accessed gaze data from c# to the python software via UDP. Gaze data was

formatted and sent through a predefined socket on the localhost address (127.0.0.1). In the python code a listener on this socket was implemented. While the user clicked through the study, this python-implemented UDP-listener started to log the data when the user accessed the mail or newspaper-interface and stopped it as soon as he or she left it. x- and y- Position and a timestamp were received 4 to 6 times per second. They were replenished with additional information from the study and saved to a CSV file. The information we added included the participants ID, the active entry field and the interface's name (mail or news).

5.3 Machine Learning

The algorithms were implemented in python with the SciKit-Machine Learning Tool[43]. It provided a 'Support Vector Machine' and a 'Random Forrest Classifier', modifiable in different features. Since there is no predefined best solution to classify data, the best configuration for our problem had to be found by experimenting.

For the SVM, we modified the kernel and the sample decision function, which was changed to 'one-versus-rest' (ovr). The kernel type modifies the hyperplane by which data points are separated and classified. While the default is a linear kernel type (a straight line in a two-dimensional space) it can also be changed to a polynomial kernel (a wave-like line in two-dimensional space) or a gaussian kernel type (multiple clusters in a two-dimensional space). The different kernels give different results and the best fitting one had to be found through experimenting.

The 'sample decision function' defines the way the decision function is applied to a data point that needs to be classified. Either it can be 'one-vs-rest', where a point is compared to all other data points simultaneously, or 'one-vs-one', where a point is compared to every single other data point separately.

The 'Random Forestclassifier' was implemented in the SciKit-Tool as well. It needed no further modifications except for the number of trees.

To run the algorithms, the input had to be prepared. The input of supervised machine learning algorithms is twofold: One contains the features and one contains the belonging label groups.

To evaluate the algorithms, we decided to use a confusion matrix and the classification report provided by SciKit. A confusion matrix, shown in table 1, displays which class features belonged to and how the algorithm classified it. The classification report was based on this confusion matrix and calculated different performance values.

		Prediction outcome		
		new	**change**	**reuse**
Actual value	**new**	1	2	3
	change	4	5	6
	reuse	7	8	9

Table 1: Example for a confusion matrix: in field 1 values that were 'new' were classified as 'new', in field 2 values were 'new' but were classified as 'change' and in field 3 values were 'new' but classified as 'reused'

6 Evaluation

This chapter discusses the evaluation process starting with the study design. It continues by introducing the setup of the study and the apparatus used. We give an overview over the participants and finally introduce the procedure they had to pass through.

6.1 Study Design

The study was designed as a between-subjects experiment with one independent variable: the interface, which included the mail- and then the newspaper-page. All participants did both conditions in the same order: first mail-, then newspaper-interface.

As for the dependent variables, we had 6 variables:

1. The total duration spent on an interface

2. The number of fixations on the screen, on the keyboard and overall

3. The duration between the first sight of the interface and the first keystroke

4. The time spent on every entry field

5. The number of keystrokes per field and overall

6. The time between keystrokes

6.2 Setup

The study took place for a total duration of two weeks. Participants were not recruited through email or postings. Instead we decided to recruit them in the university canteen. The idea for this setup originated in the focus group 4.1. Since the canteen was most visited during lunchtime we looked for participants during that time. The setup was placed where many people came across, close to the only entrance and exit. As a motivation we offered treats. People were asked to support this study by offering up a little bit of their time to take part in it. When asked, the purpose of this study that we communicated with the potential participants was usability research regarding interfaces to cover up the actual purpose. The eye tracker allegedly was used to elicit areas of interest on websites in order to find out more about user-computer interaction. We did not tell participants about the true research question of password behaviour, rather, we wanted them to behave as naturally as possible. behaviour should not be manipulated as it could lead to more complex passwords or the absence of password reuse which would not be the case when not under observation.

Figure 4: The setup in the university canteen

A corner close to the main exit was chosen as the location of this study. The setup itself was shielded by multiple canvases so participants did not feel observed during the study. People were welcomed in front of the testing setup and upon the participant's agreement, were introduced to the study's instructions. Volunteers were required to fill out a consent form first. With their signature they agreed to the recording of characteristics of all entries they made. Also, they agreed to the recording of gaze data. The consent form expressly emphasized that all data was going to be treated anonymously. If participants had questions about the results of the study, they were provided with the contact details of the hosts. After the consent was completed, users were accompanied to the recording setup as seen in Figure 4. When the participant arrived, the application was already launched and showed the first page, welcoming the participant. When a participant sat down, he or she first calibrated the eye tracker and was then instructed to start the study. If the participant had no more questions after the calibration, the supervising personnel left, to give the user privacy. If questions came up in the course, the study participants always had the opportunity to ask for help. At the end of the study, the last screen instructed the participant to contact the supervisor again and to leave the work space as the study was finished. If no problems occurred, we thanked the participant for their support and compensated their time with treats. Furthermore, if they were interested, we explained the true nature of the study and what the research goal was.

6.3 Apparatus

The setup involved three parts: the computer, the eye tracker and a connected mouse. As a computer we used a Lenovo yoga 900s 12ISK laptop with a 12,5" screen. The eye tracker was placed between screen and keyboard without any limitations to the user regarding the use of both, screen and keyboard (F.4). To the user the running device only showed inconspicuous dark red lights in the middle and on the sides. With the attached mouse the user was able to navigate through the study and could leave out the laptop's touchpad.

6.4 Participants

A total number of 52 participants was recruited over the course of two weeks. Ages ranged from 17 to 54 with a median of 22. The mean age was 25,27 years with a deviation of 6,76 years. 10 participants were female and 42 male. We also examined the expertise level regarding computer science and security knowledge in the IT sector. In general, users rated themselves as non-experts with 92% in a range from 1 to 3 (Fig. 5). The distribution of professions is shown in Figure 6. 30 participants were students and 10 worked as academic staff. The remaining 12 were either soldiers or had another, undefined job at the university.

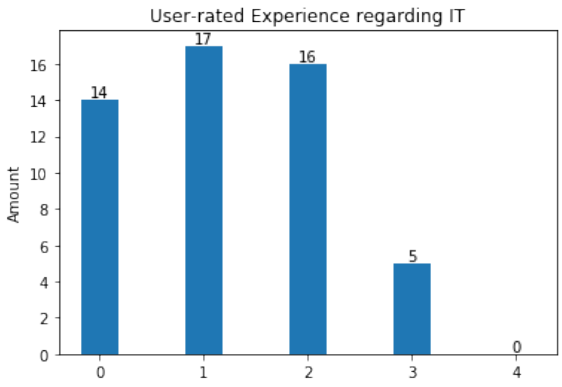

Figure 5: Self-reported User-Experience

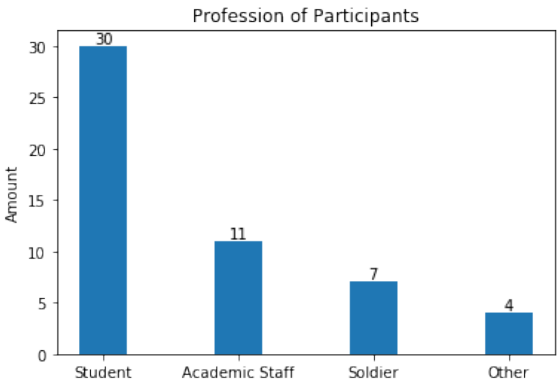

Figure 6: Profession of Participants

Every participant stated, that they would check their mail at least once a day. As Figure 7 shows, the maximum was 20 times and the median 4 times a day. In order to do so, people logged in between 0 times (when they used an email program) and 10 times, with a median of 3.

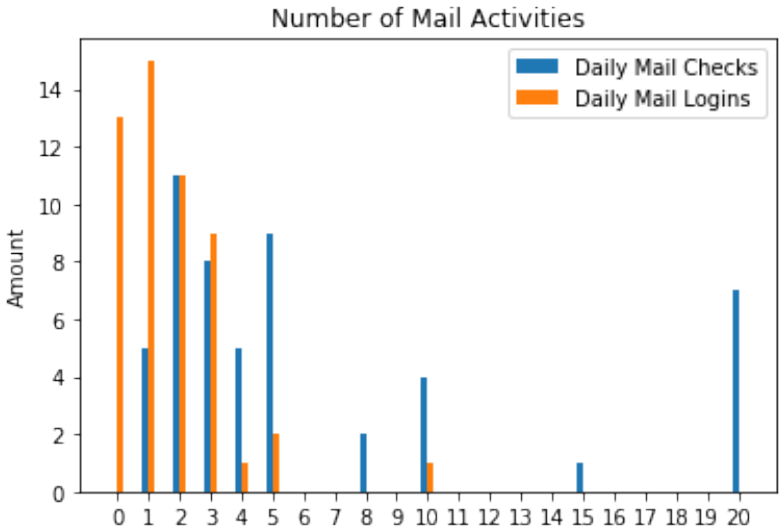

Figure 7: Daily Mail-Checks and Mail-Logins of Participants

6.5 Procedure

Every time the study software was launched, it provided a session ID which was used during the study to anonymously connect different data entries like demographic results and gaze data. No names were recorded by the software and users were not linked to the assigned ID in order to keep individuals anonymous. The first page that participants saw was a starting page, shown in Figure 8. This page welcomed the participant and gave two options: either to calibrate the eye tracker or to start with the study immediately. When choosing to calibrate the eye tracker they were directed to another page showing the keyboard shortcut Tobii provided. As the calibration was a feature included in the software, a new window popped up and guided the user through the calibration. After finishing it, it also closed automatically and showed the calibration site again, where users could go back to the main menu.

Figure 8: Starting Page that was shown when the participant took a seat

When the button to start the study on the menu was clicked, a questionnaire shown in Figure 9 opened up. Fields could be clicked with the mouse and selected entry fields could be filled by keyboard entries. The questionnaire offered both, checkboxes and open entry fields.

The questionnaire covered demographic questions to identify our target group. To separate the demographic questions from the rest of the study, the questionnaire was put at the beginning of the study before the participant was redirected to the first (mail-) interface. Questions involved age, gender, profession and experience in the field of IT. Finally, users were asked to estimate how often they checked their mails per day and how often they logged in to do so. We asked this question to find out if the number of authentications had an influence on password choices. Also we had an indicator if and how relevant the interface was to participants.

After completing this demographic questionnaire, users had to click the 'submit' button. If an answer was invalid (for example someone entered a word in the age entry field), an error message was shown. When all answers were submitted correctly, results were automatically logged and the demographic questionnaire was closed.

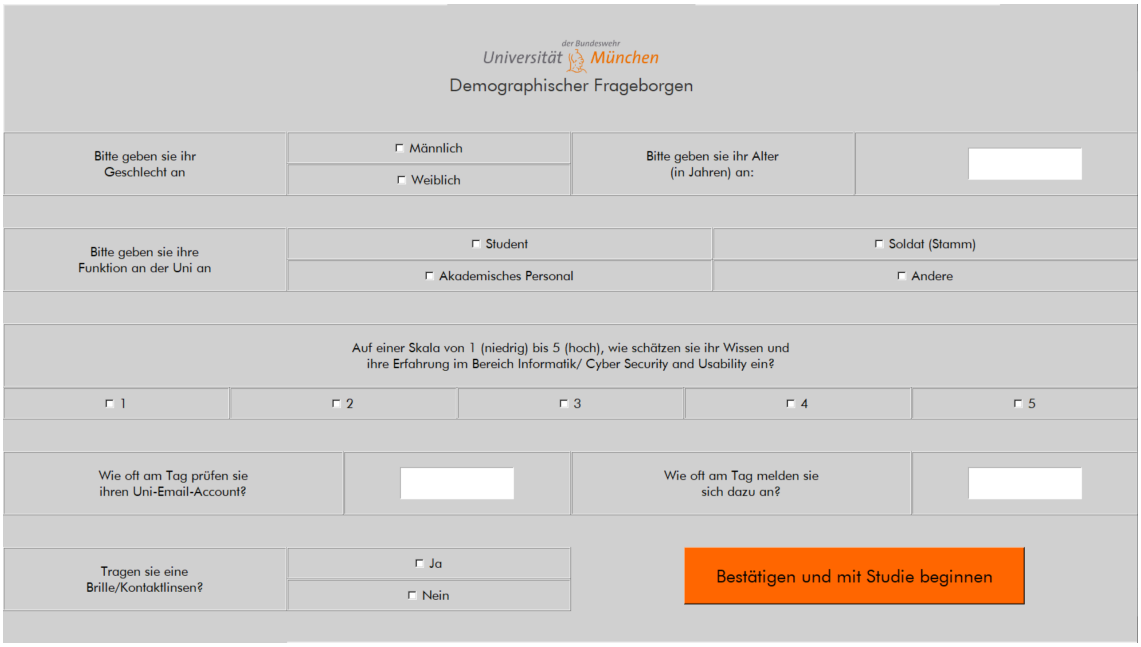

Figure 9: The demographic questionnaire

Next, users were instructed with their first task. As seen in Figure 10, on a plain page users read that the following interface would be the known university-mail-interface (4.2.1). They were advised to behave like they would on the regular web page. We also told them that they would have to create a new password and should treat this process like they would do the same on the real interface. To create a memorable password, users were told that they would have to re-enter the password at the end of the study. At this point we stated once again, that no passwords would be recorded. The instructions could be read but no interaction was possible except for a clickable 'continue'-button. When this button was clicked, not only the mail-interface was shown, but simultaneously the collection of gaze data started.

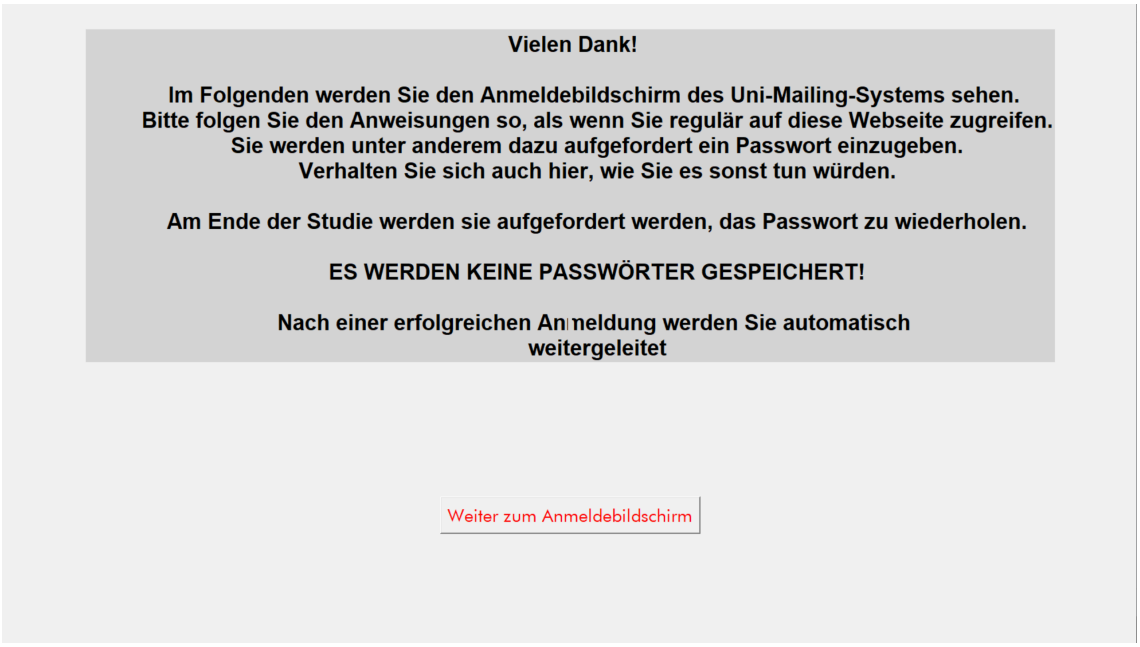

Figure 10: The instruction screen, shown before the mail-interface

The mail-interface's look, seen in Figure 11, was very close to the real interface (compare Fig. 2). A difference from the original website was an added red written description of the task ('create a new password and confirm it') and the re-entry field. On the interface we made, users were advised to create a new password since we did not just want to investigate regular password typing behaviour but the creation of passwords. Even though the advice on the interface clearly requested a NEW password, due to our foreknowledge we predicted that some users would reuse or change an old password, so we did not specifically ask for this behaviour. This hypothesis was supported by Shay's study, where only 30% of all users created a completely new one when asked, while the rest reused an old password or changed an old one [48].

Totally, we utilized three entry fields on the interface.The first entry field required the username. Since this name is deployed by the university, users had to make an entry that was already known to them. If users chose a random username it was not regarded as a mistake and thus possible as well. Then the actual password-entry followed. Because of findings and results from earlier studies we wanted to make sure users avoided entering nonsense passwords randomly. A conclusion was the application of a password policy and a re-entering field. The policy requested at least 8 characters, which is popular across the internet and recommended by NIST [35]. In addition, it prevented users from entering passwords that are too short like 'a' or '1234' just to complete the study fast. For this reason we also introduced the re-entering field. If a password had to be retyped, users could not enter a random sequence since they are required to know which characters they used in the previous entering field.

While the username was shown in alphanumeric characters, password-entry and re-entry just showed '*' (which was similar to the original website). The three entry fields had to be clicked with the mouse and could then be filled out with the keyboard. In this whole process we recorded the duration a user spent on the different entry fields (username, password-entry, password re-entry) and the number of entered characters.

All entries were monitored as we applied the 'minimum-8-characters' policy users had to fulfil. Also, users could submit their entries only if password entry and re-entry matched. When the 'submit'-button was clicked and all conditions met, the user was redirected to an instruction screen. In addition, the gaze collection stopped and all data saved to a CSV-file

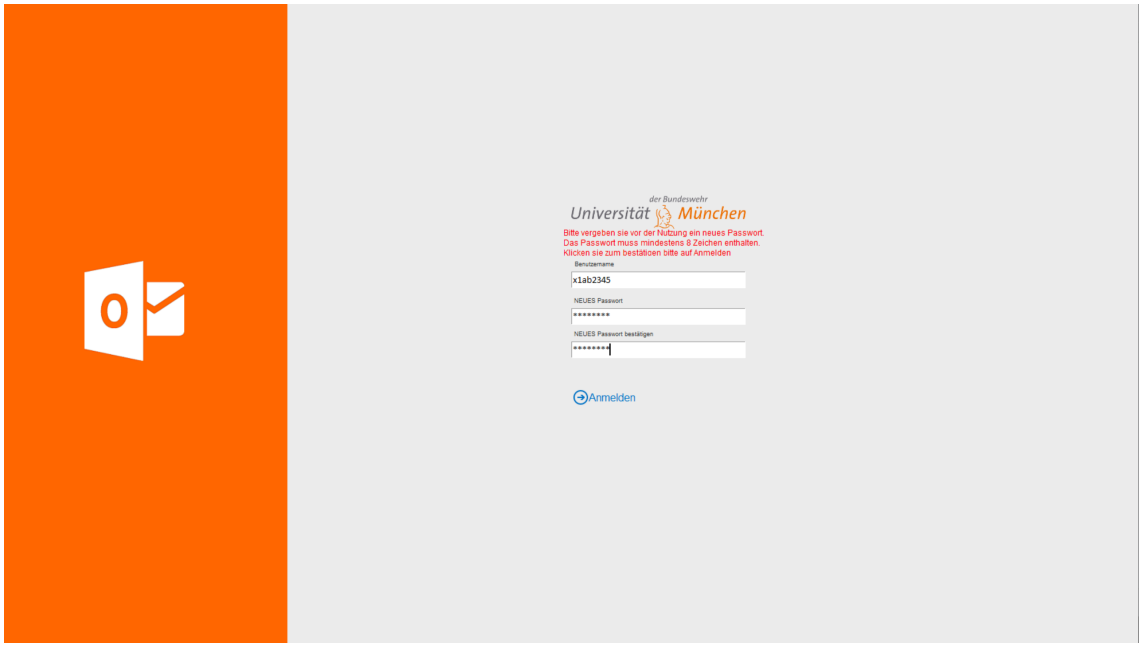

Figure 11: Implemented Mail-Interface

Next was another instructions page, shown in Figure 12. The instructions on the screen advised users for the second interface. They were told that on the upcoming newspaper-interface they had to create a password again. In addition, they now had to come up with a username as well, compared to the mail-interface where usernames were provided by the university. Once more, they were instructed to behave like they normally would do on such a website and keep in mind that they would have to re-enter the password at the end of the study. A click on the 'agree'-button opened the newspaper-interface and started the second gaze collection.

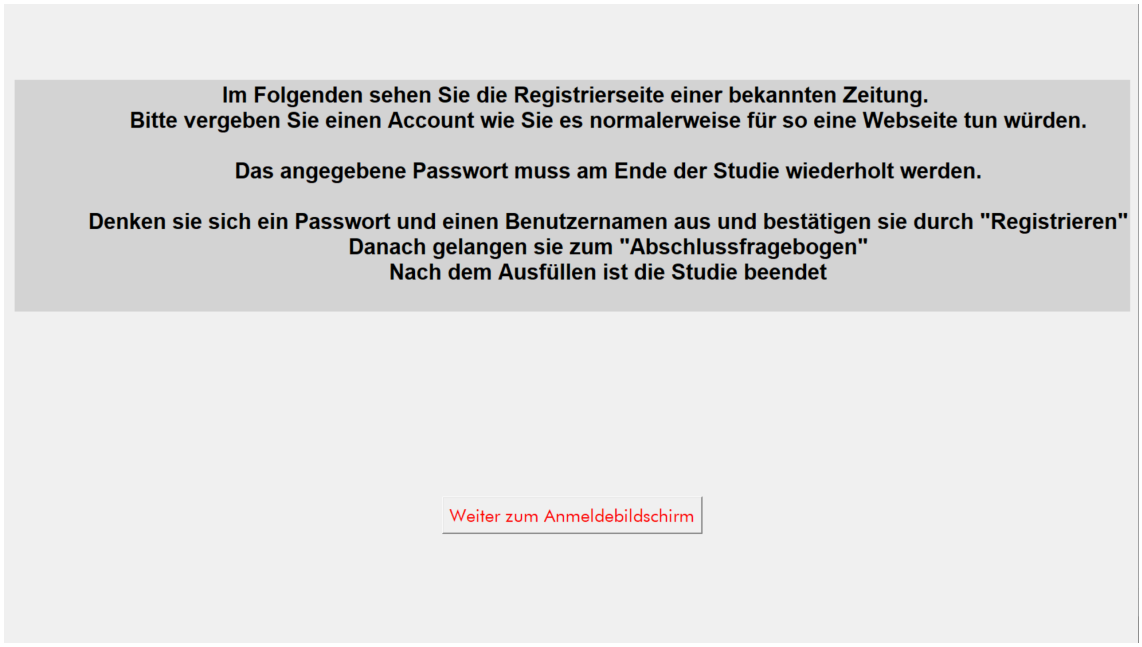

Figure 12: The second instruction screen.
It was shown between the mail- and the newspaper-interface

The newspaper-interface, shown in Figure 13 was designed to be as similar as possible to the

original 'Sueddeutsche Zeitung' website (Fig. 3) as well. Now a username entry followed by a password entry and the confirmation entry was required. We did not apply a password policy to simulate a website with less sensitive data to protect. Fields had to be chosen with the mouse and filled out by the keyboard another time. To mime the original website, a check-button was introduced where users could confirm the terms of service with a tick in a checkbox. These terms of service did not exist and the tick in the checkbox was optional. No error message was shown when the box was not selected. This means the only necessary condition participants had to fulfill to be directed to the next page was that the password and its repetition matched. The password entries were examined regarding similarity, but not regarding length since no password policy was applied this time. When all entries were successfully committed by clicking the "submit" button, participants were automatically transferred to the final site, a questionnaire about their password choices.

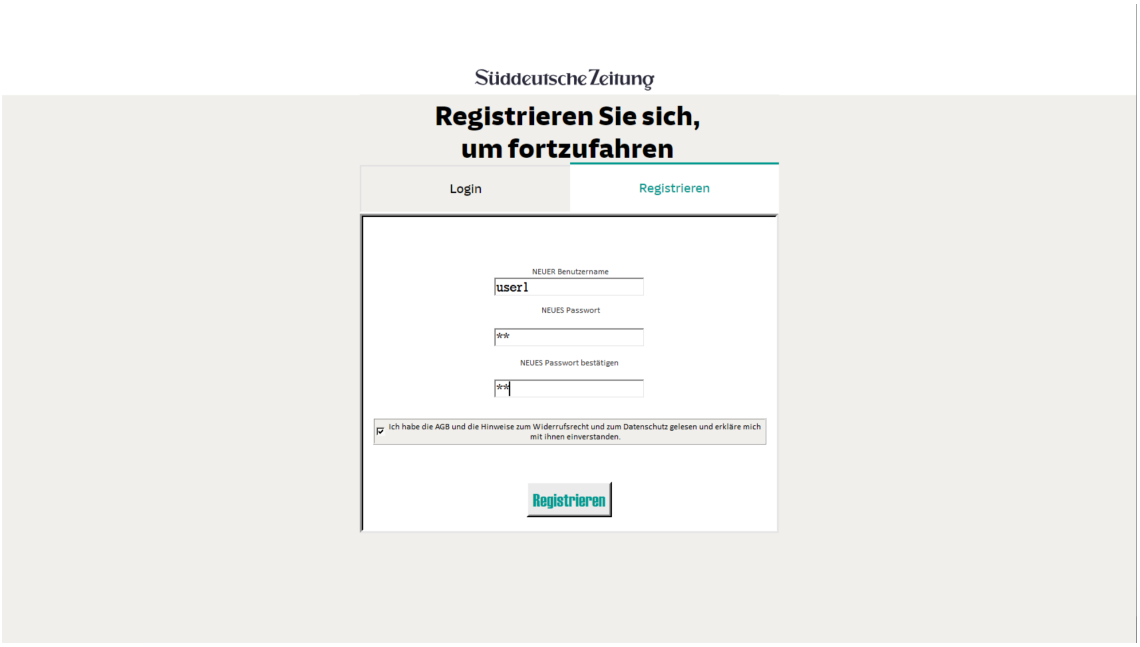

Figure 13: Implemented Newspaper-Interface

As a last step, a final questionnaire had to be filled out. As shown in Figure 14 it included checkbox questions and Likert-scales as well as entry fields which included free text entering. Only one answer per question was choosable. It covered the same questions about the first and second interface to compare them easily. For the first answer users had to specify how they chose the password (reuse an old one, change an existing one or creating a completely new one). Then the characteristics of a password were inquired. We offered predefined answer-possibilities ('consisting of a word/words', 'consisting of meaningful letters and/or digits', 'with a personal algorithm' or 'randomly') and an open answer possibility where users could specify their characteristics. As announced earlier, the passwords had to be retyped. This way, we were able to get an insight of the memorability of the entry. On a Likert-scale from 1 (low) to 5 (high) we additionally asked how well the password could be remembered. With this question we added a subjective and personal user-estimation of the memorability. At the end two questions referring to the difference between the two passwords created were asked ('Did your choice between the two passwords differ?', 'Why or why not?'). The second question was an open one. These questions may help to identify if users chose different passwords for various scenarios and the reason for that. At the end a 'submit'-button checked if not too many or too few fields were clicked or entered and if so, closed the final questionnaire.

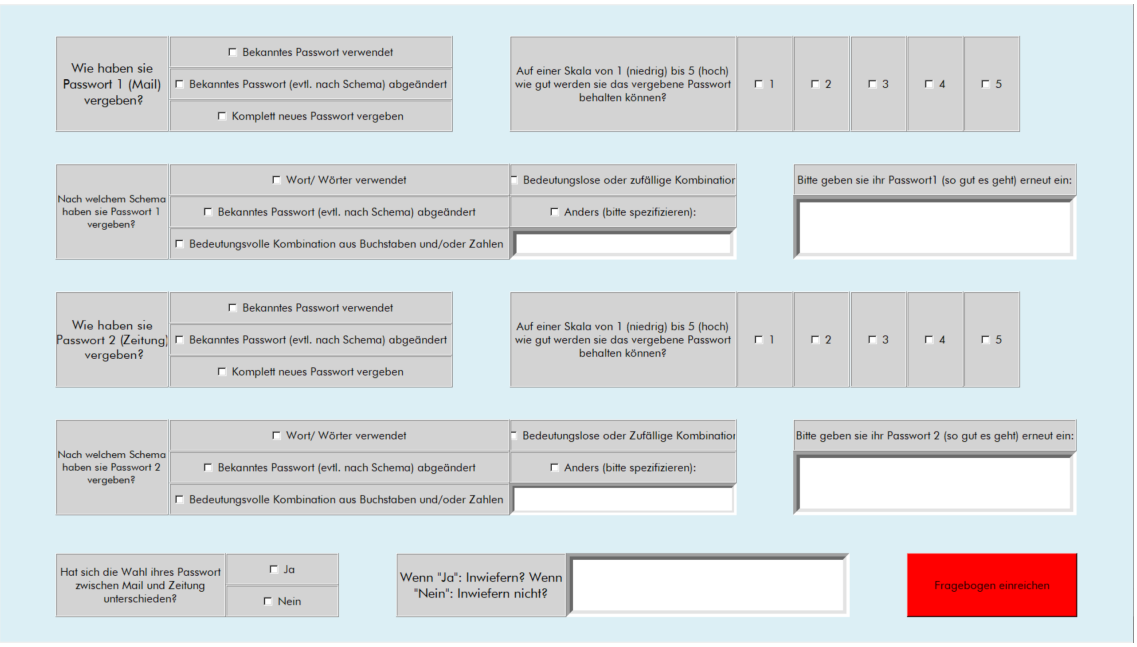

Figure 14: The final questionnaire

After submitting, participants were transmitted to the final screen shown in Figure 15. The final screen showed a message thanking the participant and advising them to contact the conducting personnel. A clickable button closed the whole study GUI and ended the session for the participant.

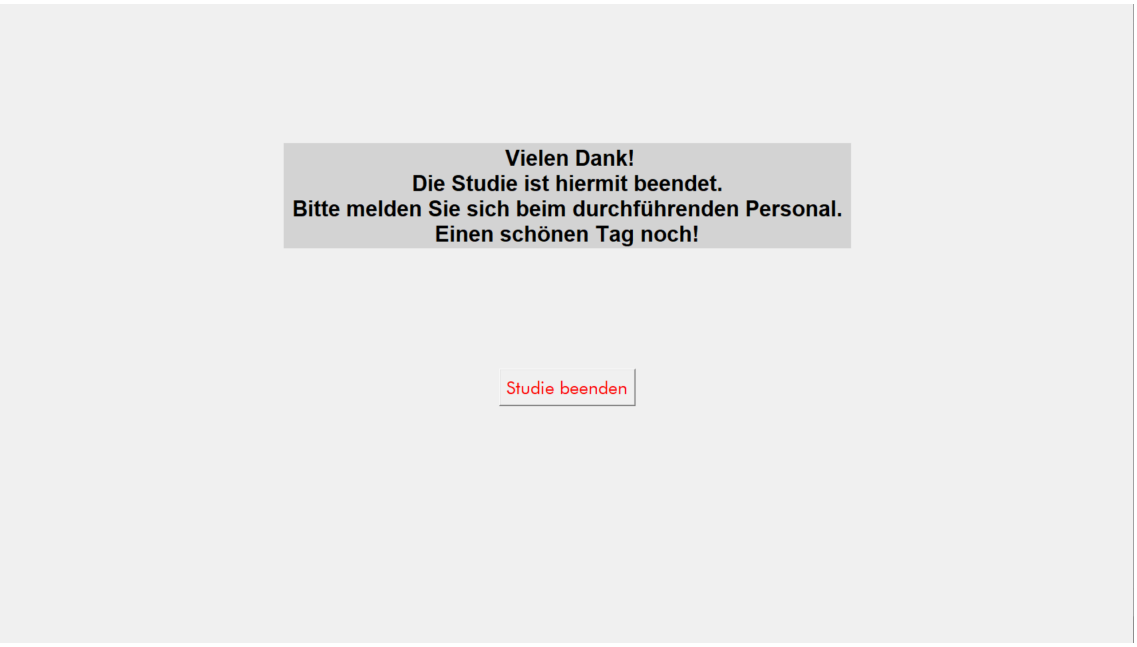

Figure 15: The final screen

7 Results

The results of the study do not only relate to gaze and the machine learning. They also allow an insight to users behaviour regarding password management and creation.

7.1 Qualitative Analysis

The qualitative analysis evaluates the questionnaires and the gaze data of each participant. It presents how passwords were chosen and how memorable they were. We compare passwords for the mail- and newspaper-interface and between password choices (new, change and reuse). Since we estimated a correlation between a password and the user behaviour we wanted to cover as many details of a password as possible without actually storing the password. Even though we did not record passwords we could get a lot of characteristics based on answers to questionnaires and logged password features.

7.1.1 Questionnaire

The questionnaires were evaluated with python. The results were presented as charts under the use of the matplotlib library [20] and also saved to CSV files to further process it. The different possible answers were compared and if applicable, password 1 (mail) and 2 (newspaper) were compared as well. Answers to the open questions were evaluated separately by hand.

Passwords Choice

The way people chose their password is shown in Figure 16. For both, the mail (first) and the newspaper (second) interface, people reused, changed and created a new password. Most people (25) chose a new password for the mail-interface, while 16 reused an already existing one and 11 changed a known one. For the second password on the newspaper-interface more people decided to reuse an existing password (18), rather than create a new one (23), while the number of people who changed their password stayed the same. From all participants 52% chose a reused password for at least one interface.

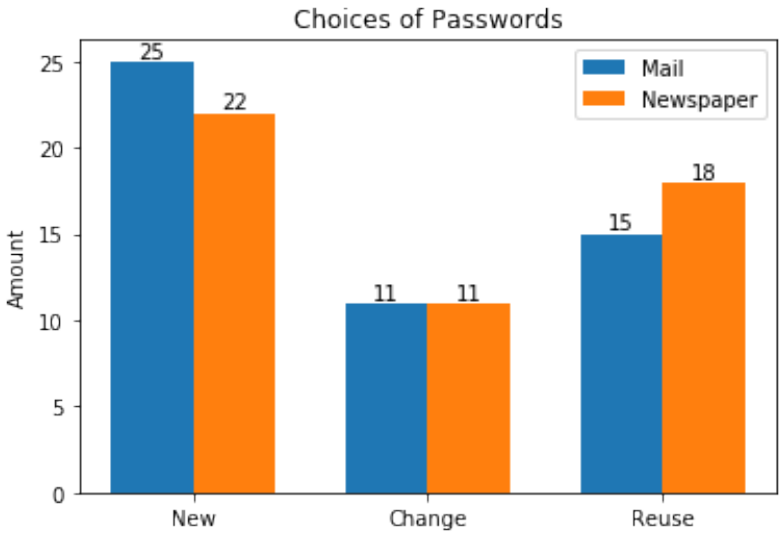

Figure 16: Comparison of password choices between the mail-password(password 1) and the newspaper-password(password 2)

Password Scheme

To find out if there was a correlation between the nature of a password and the password choice we asked participants on which scheme they based their password. The nature of used passwords is shown in Figure 17 and 18. It shows that most people (23=45,1% for the first, 21=41,18% for the second password) chose a meaningful combination of letters as their password. For both passwords an equal number of 14 people (27,45%) based their choice on an actual word or multiple words, but while on the mail-Interface 11 of these people chose a new password, only 7 people did on the newspaper-interface. Meanwhile the amount of people who reused a password based on one or multiple words rose from the mail-interface with 2 people to 6 on the newspaper-interface. Only 4 people for the first and 2 for the second interface chose a random combination of different characters. A creation scheme that was not covered by our questionnaire was chosen by 4 people on the mail-interface. The answers to the open question in the questionnaire were left blank in 2 cases and referred to what we called a 'personal scheme' in the other 2 cases ('password based on the postal code of my home town' and 'simple password for a study'). The answer to the same question for the newspaper-interface was chosen by 6 people. Their answers showed that 3 used the exact same password like on the first interface, while two described a personal scheme ('Name of a brand' and 'postal code of my home town') and one declared to have chosen a simple password for a study.

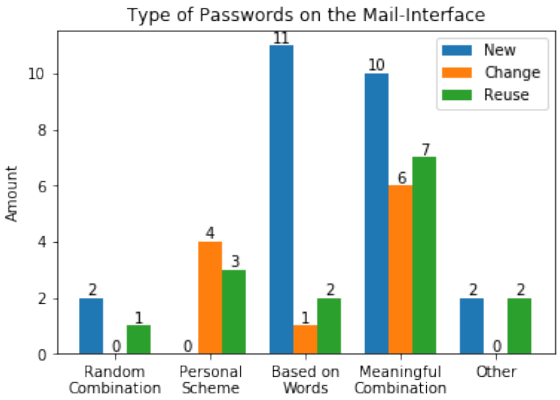

Figure 17: Password-creation schemes for the first password (Mail)

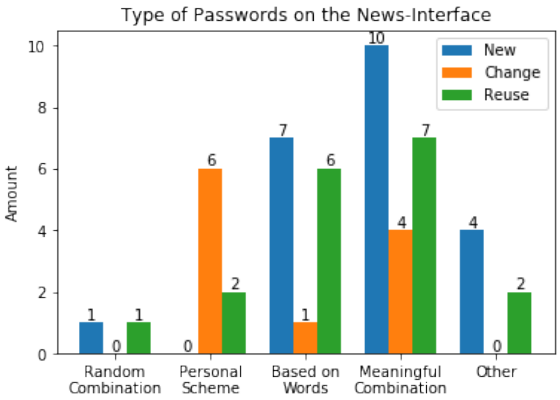

Figure 18: Password-creation schemes for the second password (News)

Comparing the three password choice-possibilities it is conspicuous that users who changed a password used a password based on a personal scheme or a meaningful combination of characters. Most of the participants who created a new password based it on one or multiple words and a meaningful combination (21=84% for the first and 17=77,27% for the second password). People who reused a password were found in all five categories with significantly more reusing a password based on a meaningful combination on the mail-interface (7=46,66%). On the second interface password-reusing participants again based their passwords on all five schemes but this time with nearly as much people reusing a password based on one or multiple words (6=33,33%) as based on a meaningful combination of characters (7=38,88%).

User-rated Memorability of Passwords

In order to find out whether people chose memorable passwords they would remember for a longer time, we first included a 5 point Likert-scale 6.5 where participants rated the memorability of their chosen passwords.

As Figure 19 shows, more than 50% (26 people) rated their password for the mail-interface as high as possible, while only one estimated a low memorability. In comparison, Figure 20 shows that there were less people compared to the first interface rating their second password

on the highest rank (20=39,22%). Also, only one person gave the lowest possible score. The self-rated memorability was higher for the mail-password (averagely 4,2) compared to the newspaper-password (averagely 3,9).

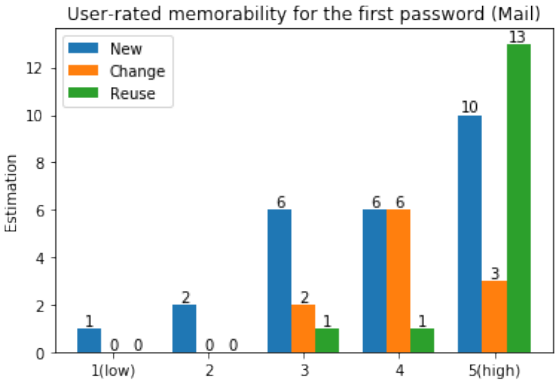

Figure 19: User-rated memorability for the first password (Mail)

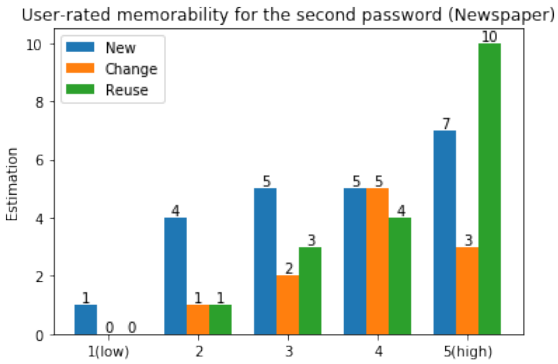

Figure 20: User-rated memorability for the second password (News)

From a comprehensive perspective, participants who reused a password rated the memorability higher than password-changing and new-creating participants. Only 2 of 15 password-reusing participants did not rate their choice the highest possible on the mail-interface. On the newspaper-interface 8 out of 18 password-reusing participants rated the memorability lower than 5, but none estimated a low memorability.

The single person who rated a low memorability created a new password. Overall new created passwords were rated as less memorable than changed and reused ones. While on the mail-interface new passwords were rated more on the memorable side with 16 people (64%) the memorability was divided almost equally in the range from 2 to 5 on the newspaper-interface.

Calculated Levenshtein Distance

In addition to the self-estimated memorability we decided to implement a scientific approach to verify memorability. To do so, we evaluated the Levenshtein-distance. The Levenshtein-Distance is calculated by the number of changes a sequence of characters has to make to become another one. Allowed changes are replacement, deletion and creation of a sequence character. For example the distance between 'amigo' and 'mugou' is 3 (remove 'a', change 'i' to 'u', add a new 'u' at the end). In our case it counts the changes between the entered password and a re-entered one at the end of the study. The code to calculate this distance was already available online [13] and did not have to be implemented. The distance had to be calculated in the end of the study during the final questionnaire, where participants had to re-enter their password. It compared both created passwords, mail and news, to the entries at the end and automatically calculated the Levenshtein distance. When the calculation was done, the results were stored in a CSV file.

9 people could not re-enter the password perfectly at the end of the study. The distance for their entries was 2 or lower for 4 participants and ranged from 4 to 8 for the remaining 5. 4 participants failed to re-enter the mail-, 2 the newspaper-password and one failed both. The passwords that were repeated incorrectly were new in 6 cases, reused in 2 cases and changed in 1 case.

Altogether participants were able to re-enter their passwords very well at the end.

Different Choice of Passwords between Mail- and Newspaper-Interface

In the last two questions of the questionnaire, participants were asked if both created passwords

differ from another and were asked for the reason for that outcome. The results displayed in Figure 21 show that both possibilities were chosen nearly equally. 26 participants chose different passwords, while 25 used the same password for both interfaces. Answers to the open question ('Why did you choose a different password?/Why not?') could be categorized into 4 groups:

1. Many of the participants who did not choose a different password did not elaborate their choice in the open entry field. The ones who did said it was for simplicity reasons. One of the participants stated that he/she made the same choice because *'The remembering is easy'* and another one used the same password *'So you do not forget it'*.

2. Participants who changed their password used the open answer to describe their changes. The answers show that just minor changes between both passwords were made (*'Capital versus small letters'*, *'I changed a number'*, *'I added a number for the newspaper'*)

3. Some people who used two completely different passwords did it for relevance reasons. They rated the two websites differently, like a participant that stated: *'I have to login more often to the the mail-account, so I chose a more simple password'*, or another one: *'Specifically adjusted for the newspaper account'* or: *'First a standard password, then a completely new one'*. One participant chose a different password *'For security reasons'*

4. Other people who had distinct passwords described the difference, like participants that changed their password (two participants wrote: *'Completely new combination'*,one wrote: *'Length and combination of characters and letters'*)

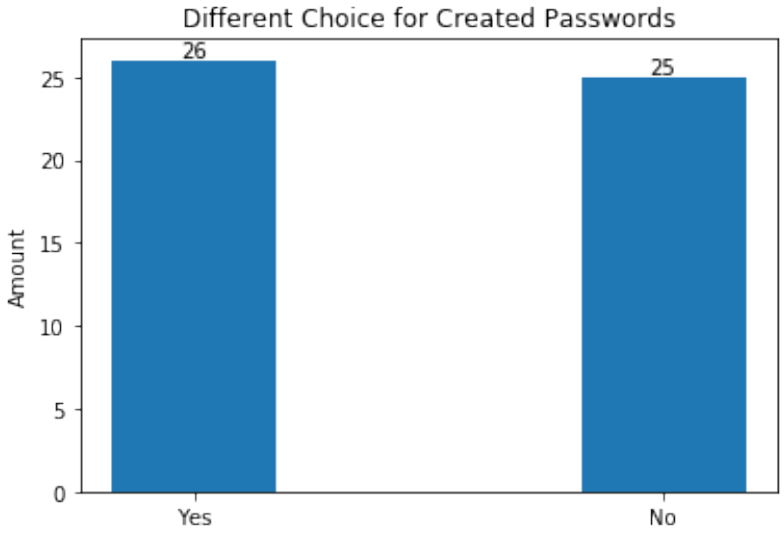

Figure 21: Did your choice of the two passwords differ?

7.1.2 Password Characteristics

Even though we did not collect plain-text passwords we evaluated characteristics of submitted entries to get an insight into user passwords without violating their privacy. We evaluated the total length of the password and the used character groups. We distinguished lower case letters, capital letters, digits and special characters. The evaluation covered only the final password created, so unsuccessful login attempts were not considered (for example if a user tried to submit a password that was too short in regards to the restrictions of the password policy). For this reason it was connected to the 'submit'-button and its monitoring-function. Figure 22 shows that most people used 2 character classes on the mail-interface (15=31,25%). On the

newspaper-interface most participants (17=35,42%) used 2 character classes as well, shown in Figure 23. Comparing both interfaces it can be said that on the mail-interface more character classes were used (average of 2,52) than on the newspaper-interface (average of 2,35).

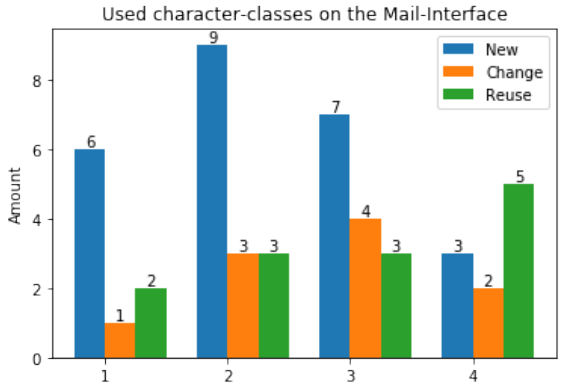

Figure 22: Used character classes for the first password (Mail)

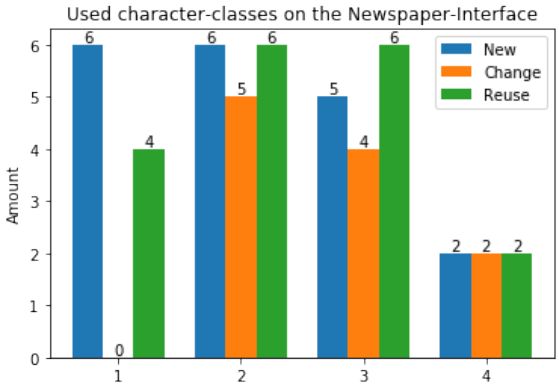

Figure 23: Used character classes for the second password (News)

New passwords often consisted of fewer character classes than changed and reused ones. An amount of 15 people (60%) out of 25 who created a new password used one or two character classes on the mail-interface. In comparison, the biggest part of participants who reused a password on the same interface (5 out of 13) involved the maximum of possible character classes, which was 4. On the newspaper-interface password-reusing people shifted to a usage of less character classes with only 2 using the maximum of 4, while the distribution for used character classes of new passwords stayed very similar.

7.2 Gaze Data

From all described gaze metrics 2.4 we decided to focus on fixations. They are easy to calculate from the raw gaze data and can provide a lot of information like areas of focus and interest of a person. Saccades and scanpaths provide information about searching behaviour, but are hard to calculate and refer more to the research of usability rather than the user's behaviour. The pupil diameter was not accessible because the eye tracker that was used was unable to capture this feature. Fixations were extracted in python. The code we used classified groups of data points as fixations when multiple following points with an added total duration of at least one hundred milliseconds stayed in a circular area with a diameter of 0,25 centimeter. With this method both CSVs (for mail and newspaper) of all participants were processed and the resulting fixations written to another CSV. Every row contained x- and y- position, the duration and the timestamp of beginning and ending of a fixation.

7.2.1 Fixation Patterns

To get an overview of the fixations they were visualized in a two-dimensional grid with charts from the matplotlib library [20]. For every participant we plotted two individual charts, one for each interface. An example for such a chart is shown in Figure 24. We also combined all fixations of all participants in order to get an overall-overview of the areas of interest, shown in Figure 26.

All charts for each individual had similarities but deviated in some details. All participants created fixations in the area of the entry fields for username, password and password confirmation, the first area of interest. Except for one participant everyone looked to the keyboard, the second area of interest, as well. What differed between individuals was the amount of fixations

on the keyboard and the mouse. Some people just had very few fixations in this area, while others not only had many, but also broad spans of fixations. A specific area on the keyboard that was focused more often than others seemed to be the *CAPS*-key. Another feature most participants matched was that no fixations in the third area of interest were recorded. Only very few people looked at the area outside of the screen.

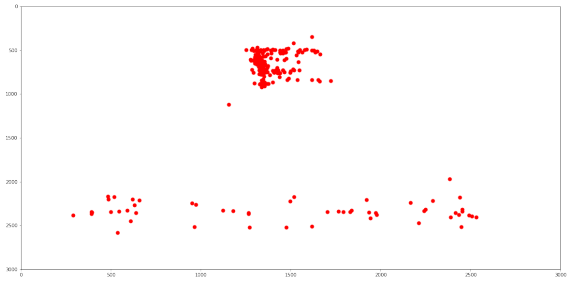

Figure 24: Fixation-pattern for the Mail-Interface of participant number 45

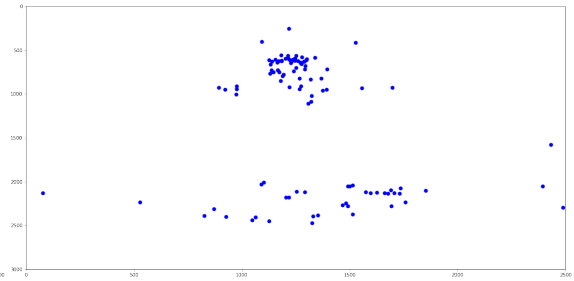

Figure 25: Fixation-Pattern for the Newspaper-Interface of participant number 12

After combining all fixations together (Fig. 26) in one big picture some elements of the setup are clearly visible. The entry fields stand out with a high concentration of fixations in a small area. The keyboard area is also clearly visible. Fixations here are not highly concentrated but spread all across the area evenly. There is also a thin visible gap at the bottom which resembles the gap between mouse and keyboard. As the mouse was placed at the same height as the keyboard the fixations for both are at the same level. Comparing to the keyboard, the mouse had less fixations. The chart also shows scattered fixations across the whole screen, which may be noise.

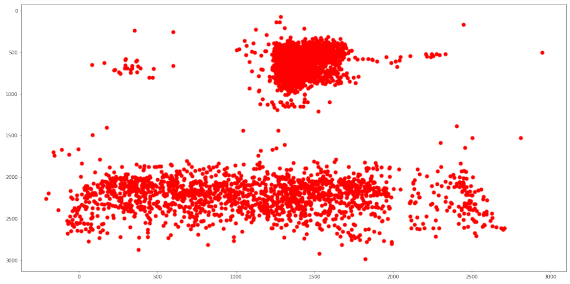

Figure 26: Combined fixations for the Mail-Interface

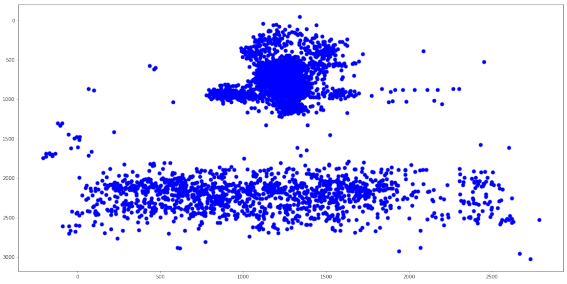

Figure 27: Combined fixations for the Newspaper-Interface

7.2.2 Time Spans

The collected CSVs did not only cover the eye movements but also the amounts of characters in the different entry fields. Together with the timestamps (4 to 6 per second) it was possible to calculate different durations (4.2.2).

Total Duration

The total duration was calculated by subtracting the last and the first timestamp of each interface. As shown in Figure 28, the overall duration varied strongly for participants who created a new password. While some needed less than 30 seconds, others needed more than one hundred. The average time was 52,98 seconds. This value is a little higher then the average total duration of changed passwords (50,54s). Again, all values are widely spread from

30 to 90 seconds for changed passwords. Only the total duration of participants who reused passwords was different. Not only the average duration was lower here (36,56s), but also some participants needed less time (20s) than in both other cases. The longest duration was below 70 seconds.

On the newspaper-interface the total duration decreased with the use of a known password resource as shown in Figure 29. The average duration was the highest when a new password was chosen (44,66s) and the lowest, when a password was reused (25,99s). In between these two values, the average time for a changed password was 37,37s seconds. For new and reused passwords the minimum duration was the same with nearly 20 seconds. For changed passwords it was nearly 30 seconds. The longest duration was similar for new and changed passwords (75s), but lower for reused passwords (55s).

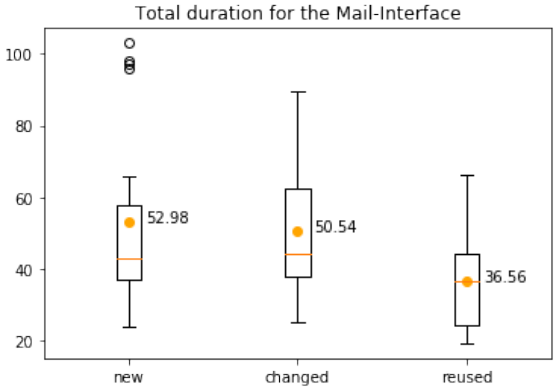

Figure 28: Overall duration in seconds (Mail-Interface)

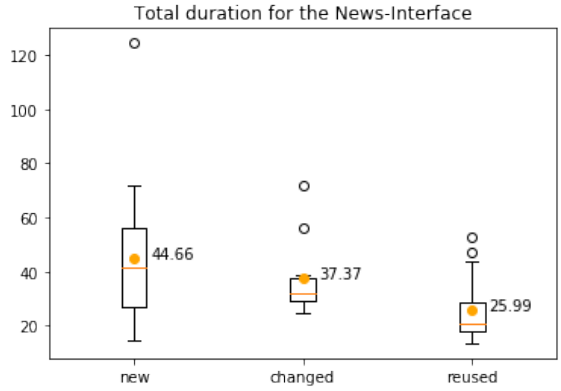

Figure 29: Overall duration in seconds (Newspaper-Interface)

Thinking Time

The initial duration or thinking time was calculated by subtracting the timestamp of the last keystroke in the username field from the timestamp of the first keystroke in the password field. In total, the thinking times were quite similar between all conditions.

As Figure 30 shows, the thinking time did not vary that much between different password conditions. The average time was similar for new (6,67s) and changed passwords (6,46s), and only a little bit lower for reused passwords (5,87s). Values had a similar distribution with 75% between 3,75 seconds and 7,5 seconds. The median was the same for all three with 5,3 seconds.

The same pattern also applied to the thinking time on the newspaper-interface (31), where 75% of all values were distributed similar in the range of 3 to 6 seconds. Values ranged from 2 to 15 seconds with one outlier of nearly 40 seconds for new created passwords. In contrast, the values had a very small span when an existing password was chosen (2 to 5 seconds).

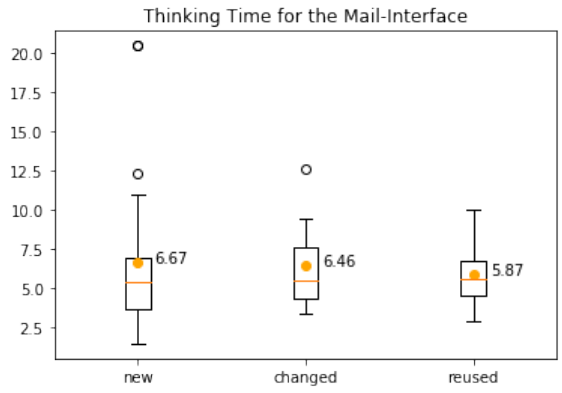

Figure 30: Thinking Time in seconds(Mail-Interface)

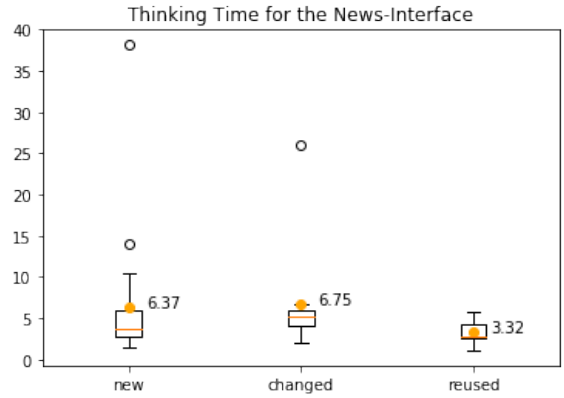

Figure 31: Thinking Time in seconds (Newspaper-Interface)

7.2.3 Ratios

After extracting fixations and different durations, we calculated the ratios (4.2.2) used to compare users.

Fixations per Second

Fixations per second (fixps) were calculated by dividing the number of fixations with the total duration of a user.

On the Mail-Interface 32 more fixations per second occurred when a new password was created. 75% of all values were between 2,3 and 2,6 fixps with an average (2,51) close to the median (2,5). For changed and reused passwords the range including 75% of all values were similar from 2,9 to 2,4 fixps. The difference was a bigger range of values with the lowest ratio of 1,25 to the highest of 2,75 for reused passwords compared to a lowest ratio of 1,5 to the highest of 2,5 for changed passwords.

For the Newspaper-Interface the values were very different (33). New and reused passwords shared the area where 75% of the values were found (2,3 to 2,6 fixps) and had a similar average (2.39 and 2,48 fixps). However, new created passwords had some outliers with a low rate of about 1,3 fixps, while reused created some in the higher area of about 3 fixps.

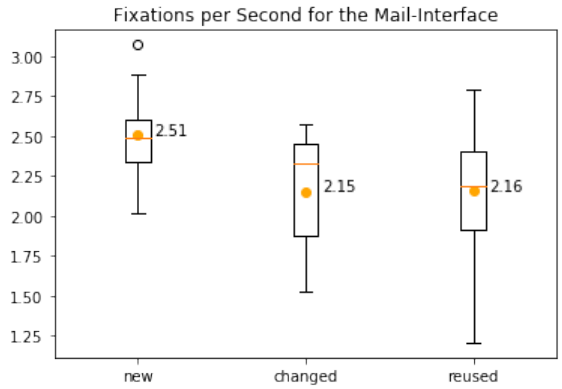

Figure 32: Fixations per Second (Mail-Interface)

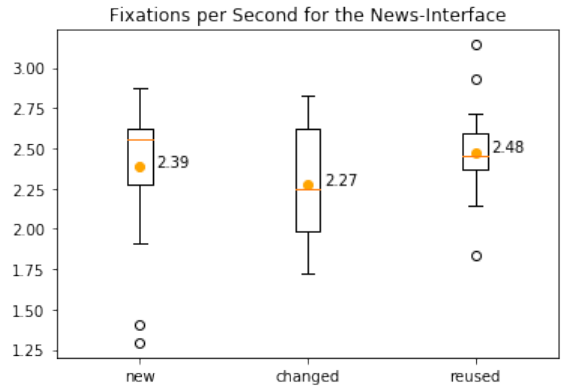

Figure 33: Fixations per Second (Newspaper-Interface)

Average Time between Keystrokes

The average time between keystrokes was calculated by dividing the total duration by the

number of keystrokes. Figure 34 shows that the average time between keystrokes was the lowest for reused passwords (averagely 0,98s) and the highest for new passwords (averagely 1,47s). The lowest ratio was similar for all conditions (nearly 0,7s), only the highest values were different. While three outliers with a maximum time of almost 4,5 seconds were found for new created passwords, the highest value was just a third for reused passwords (1,5s). 75% of all values were distributed in a lower area (0,75s to 1,25s) if a participant reused a password compared to if a new password was created (1s to 1,5s).

Similar characteristics seem to apply to the newspaper. Figure 35 shows that the average time between keystrokes was the lowest for reused passwords (average of 0,82s). It was the highest for new passwords (average of 1,24s). While the lowest time between keystrokes was nearly the same for all three conditions (about 0,5s), the longest time for a new password was much longer than for a reused password (1,7s).

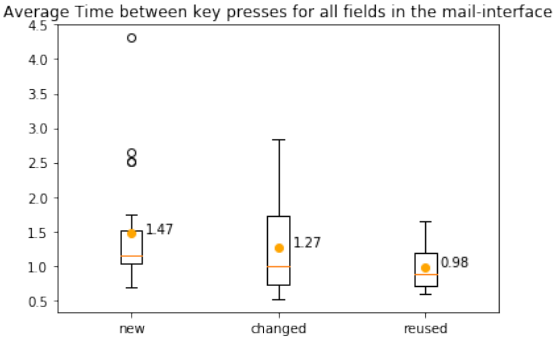

Figure 34: Average time between Keystrokes in seconds (Mail-Interface)

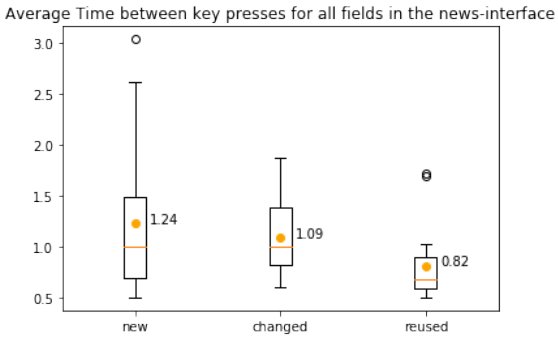

Figure 35: Average time between Keystrokes in seconds (Newspaper-Interface)

From the average time between keystrokes on the whole interface we went into detail and calculated the average time between keystrokes for single entry fields as well. For example we estimated a higher average for time in regards to the first password field if users created a new password. As seen in Figure 36 and 37 this was the case on both interfaces. New passwords took the most time in creation, while reused ones required the least time. Comparing both interfaces participants took more time filling out the password field of the mail-interface than they did on the newspaper-interface.

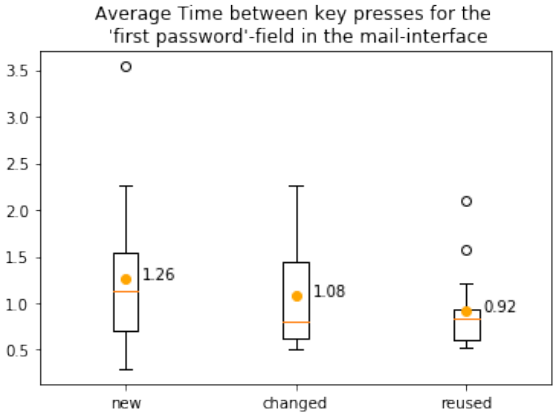

Figure 36: Average time between Keystrokes in seconds (Mail-Interface, password field)

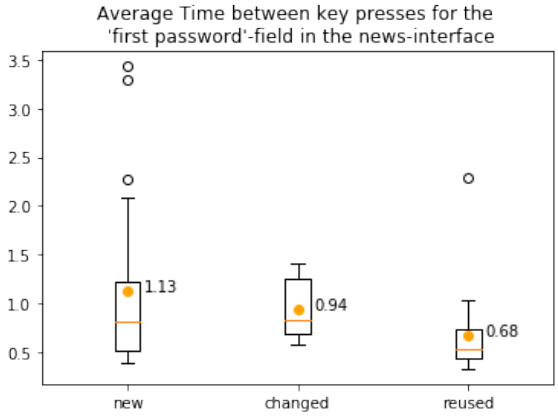

Figure 37: Average time between Keystrokes in seconds (Newspaper-Interface, password field)

As training features for the machine learning algorithm we also extracted the average time between keystrokes on the remaining fields of both interfaces, which included the username and the password re-entry field.

Screen to Keyboard Ratio

To get the screen to keyboard ratio we classified fixations based on their y-positions of the coordinates. We knew that the screen reached to a y-position of 1500 units, so all fixations were separated by comparing their y-value with this one. Figure 38 shows the lowest average ratio for reused passwords (1,86). In contrast, the highest one can be found for new passwords (3,07). In between lays the average ratio for changed passwords (2,01). For all three conditions 75% of all values can be found in a small range of 2, but all three have some outliers. While these outliers are rated higher (almost 14), for new passwords they can be found very low (more than -1) for reused passwords. The negative values found for all three conditions mean that there were users who looked more at the keyboard than at the screen.

In comparison, Figure 39 shows that for the second interface the ratios were very similar for all conditions (nearly 2,4). Again, the range where 75% of all values can be found was small (nearly 2) with outliers. In contrast to the first interface the outliers are distributed equally above and below the average value for all three conditions. Also, on the second interface few people looked more at the keyboard than at the screen, which is shown by negative values for all three conditions.

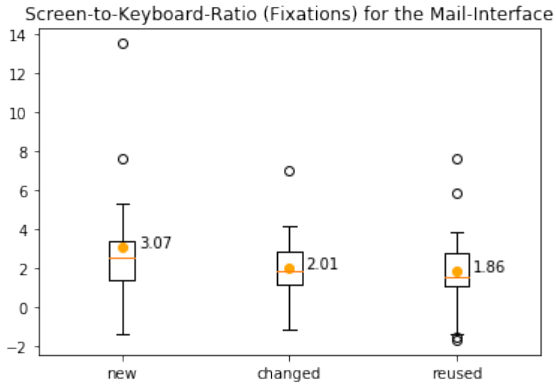

Figure 38: Screen to keyboard ratio (Mail-Interface)

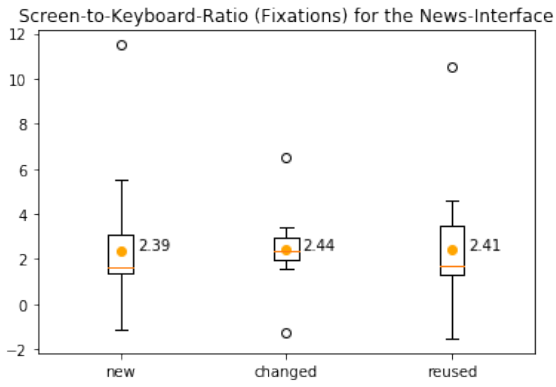

Figure 39: Screen to keyboard ratio (Newspaper-Interface)

7.3 Quantitative Analysis

In this section, we will report the statistical analysis made on the dependent variables.

7.3.1 News Vs. Mail Passwords

First, we divide the data into the interface/security level. The data reflects on two main aspects: password characteristics and gaze data.

Password Characteristics

A repeated measures ANOVA showed significant difference between the memorability of the mail passwords(Mean = 4.17; SD = 1.02) vs the news passwords(Mean = 3.94; SD = 1.14), $F(1,47) = 5.785$, $P = 0.02$. Consequently, the failed logins significantly differed between the mail (Mean = 0.35; SD = 0.635) and the news (Mean = 0.04, SD = 0.202) interfaces, $F(1,47) = 9.874$, $P = 0.003$. The results also showed statistically significant difference between the length of the

created passwords for the email interface (Mean = 10.71; 2.568) and the newspaper-interface (Mean = 9.96; SD = 2.843), F(1,47) = 4.615, P = 0.037. The overall duration of the passwords generations showed statistically significant difference in the ANOVA test between the mail (Mean = 48.39; SD = 20.678) and the newspaper-interface (Mean = 36.04; SD = 20.434), F(1,47) = 12.136 , P = 0.001. The use of upper case letters, digits, special characters did not show any significance for the generated passwords for both interfaces. Also, the same was noticed for the levenshtein distance (the distance between the created password and the login passwords) for mail password(Mean = 0.50; SD = 1.65) and news password(Mean = 0.29; SD = 1.35) F(1,47) = 0.433, P = 0.52 as seen in Table 2.

Table 2: News vs Mail Passwords Characteristics

	Memorability		Length		Duration		Failed Logins		levenshtein	
	Mail	News	Mail	News	Mail	News	Mail	News	Mail	News
Mean	4.17	3.94	10.71	9.96	48.39	36.04	0.35	0.04	0.50	0.29
SD	1.02	1.14	2.568	2.843	20.678	20.434	0.635	0.202	1.65	1.35

Gaze Behaviour

For the gaze data as seen in Table 3 and Table 4, repeated measures ANOVA showed statistical significance in the number of fixations between the news (Mean = 83.15; SD = 38.548) and the mail (Mean = 112.23; SD = 50.50) interfaces, F(1,47) = 13.073, P = 0.001. ANOVA has also shown significant difference of the number of screen fixations between the mail (Mean = 76.06; SD = 38.805) and the newspaper-interface (Mean = 55.56; SD = 29.314), F(1,47) = 10.031, P = 0.003. The duration of screen fixations also showed statistical significance between the mail(Mean = 18265.61; SD = 10366.261) and the news(Mean = 12835.82; SD = 6905.902) interface for ANOVA test, F(1,47) = 11.293, P = 0.002. Also, the number of keyboard fixations showed statistical significance between the mail (Mean = 36.17; SD = 22.744)and the news(Mean = 27.58; SD = 16.889) interfaces F(1,47) = 7.417, P = 0.009. The keyboard fixation duration has also showed statistical significance between the mail (Mean = 6570.10; SD = 4636.839) and the newspaper-interface (Mean = 4976.69; SD = 3301.785), F(1,47) = 6.372, P = 0.015.

ANOVA has also reviled statistical significance for the duration from when the interface popped up until the first keystroke is pressed (Thinking time) between the mail (Mean = 12.97; SD = 9.884) and the news (Mean = 6.17; SD = 3.708) interfaces, F(1,47) = 19.468, P < 0.001. The interface has also shown significant mean effect on the username field duration for both the mail (Mean = 17.22; SD = 14.334) and the news (Mean = 10.96; SD = 7.801) interface, F(1,47) = 7.269, P = 0.010. ANOVA test also shows statistical significance for the ratio for the keystrokes and duration for the user name field between the mail (Mean = 2.23; SD = 2.431) and the news interfaces (Mean = 1.20; SD = 0.744), F(1,47) = 7.834, P = 0.007. The interface showed statistically significant difference on the password field keystrokes for the mail (Mean = 18.33; SD = 12.798) and the news (Mean = 13.46; SD = 8.582) interface, F(1,47) = 7.460, P = 0.009. The same applies for the password keystroke duration, repeated measure ANOVA shows statistically significant difference of the password keystroke duration on the mail (Mean = 20.41; SD = 17.550) and the news (Mean = 12.51; SD = 11.613) interface, F(1,47) = 7.418, P = 0.009.

Lastly, the interface has a significant mean effect on the total registration duration in both mail (Mean = 55.00; SD = 41.028) and news (Mean = 36.55; SD = 25.220), F(1,47) = 8.212, P = 0.006. Finally, ANOVA did not reveal any significant difference between the news and email interfaces for the number of fixations per second, screen to keyboard fixation ratio, time writing, username field keystrokes, password key duration ratio, re-entry keystrokes, re-entry key duration ratio, total keystrokes or total key duration ratio.

Table 3: News vs Mail Passwords Gaze Behaviour I

	Num. Fix		Num. Screen Fix		Screen Fix Dur.		Num. keyboard Fix		Keyboard Fix Dur.		Thinking Time	
	Mail	News	Mail	News	Mail	News	Mail	News	Mail	News	Mail	News
Mean	112.23	83.15	76.06	55.56	18265.61	12835.82	36.17	27.58	6570.10	4976.69	12.97	6.17
SD	50.50	38.548	38.805	29.314	10366.261	6905.902	22.744	16.889	4636.839	3301.785	9.884	3.708

Table 4: News vs Mail Passwords Gaze Behaviour II

	Username Dur.		Username Keystroke-Dur. Ratio		Password Keystrokes		Password keystroke Dur.		Total Dur.	
	Mail	News	Mail	News	Mail	News	Mail	News	Mail	News
Mean	17.22	10.96	2.23	1.20	18.33	13.46	20.41	12.51	55.00	36.55
SD	14.334	7.801	2.431	.744	12.798	8.582	17.550	11.613	41.028	25.220

7.3.2 New Vs. Changed Vs. Reuse Passwords

In this section, we divide the data into the categories newly generated passwords, reused passwords and changed passwords. As previously discussed, the data reflects on two main aspects: password characteristics and gaze data.

Password Characteristics

Repeated measure ANOVA did not show statistically significant effect of the password schemes on the memorability, levenshtein distance, length, number of capital letters, number of digits, number of special characters, and number of failed logins (see Table 5 and Table 6).

Table 5: Reuse vs changed vs new Password Characteristics Stats I

	Memorability			Levenshtein Distance			Length		
	New	Reuse	Changed	New	Reuse	Changed	New	Reuse	Changed
Mean	3.76	4.38	4.00	.24	.43	.19	9.48	10.48	11.05
SD	1.338	.921	.837	1.091	1.748	.873	3.156	2.421	2.958

Table 6: Reuse vs changed vs new Password Characteristics Stats II

	# capital letters			# of special characters			# of digits		
	New	Reuse	Changed	New	Reuse	Changed	New	Reuse	Changed
Mean	.52	.76	.90	.24	.19	.38	1.90	3.62	3.48
SD	.750	.995	1.179	.625	.402	.498	2.071	2.559	2.462

Gaze Behaviour

Table 7 and 8, shows the statistical significance induced from password schemes on the gaze behaviour. Repeated measure ANOVA reviled significant mean effect of the password scheme on the total duration of password generation $F(2,40) = 5.113$, $P < 0.001$. Pairwise comparisons with Bonferroni correction showed significant differences between new (Mean = 44.49; SD = 24.553) and reuse (Mean = 27.55; SD = 12.067), $P = 0.029$ also between reuse (Mean = 27.55; SD = 12.067) and changed (Mean = 43.64; SD = 17.853), $P = 0.001$. ANOVA also shows significant mean effect of the total number of fixations on the password schemes, $F(2,40) = 5.281$, $P = 0.004$. Pairwise comparisons with Bonferroni correction showed significant differences between new (Mean = 99.33; SD = 38.271) and reuse (Mean = 67.24; SD = 28.933), $P = 0.013$ also between reuse (Mean = 67.24; SD = 28.933) and changed (Mean = 95.14; SD = 36.748), $P = 0.028$. We also found significant effect of the password scheme on

screen fixations F(2,40) = 4.719, P = 0.001. Pairwise comparisons with Bonferroni correction showed significant differences between new(Mean = 66.48; SD = 31.324) and reuse(Mean = 42.38; SD = 18.538), P = 0.014 also between reuse(Mean = 42.38; SD = 18.538) and changed (Mean = 65.57; SD = 32.040), P = 0.010.

ANOVA also showed significant effect of the password scheme on initial duration F(2,40) = 3.582, P = 0.006. Pairwise comparisons with Bonferroni correction showed significant differences between new (Mean = 9.31; SD = 6.568) and reuse (Mean = 4.54; SD = 2.659), P = 0.029. We found significant effect of the password scheme on username typing duration F(2,40) = 3.757, P < 0.001. Pairwise comparisons with Bonferroni correction showed significant differences between reuse (Mean = 8.05; SD = 5.764) and change (Mean = 16.32; SD = 12.362), P = 0.007. Lastly, we found statistically significant difference for the username keystrokes duration ratio between the three passwords schemes, F(2,40) = 3.766, P = 0.004. Pairwise comparisons with Bonferroni correction showed significant differences between reuse (Mean = .977; SD = .096) and change (Mean = 1.878; SD = .303), P = 0.007. Finally, we also found statistically significant effect of the password scheme on the total duration of password creation, F(2,40) = 2.797, P = 0.007. Pairwise comparisons with Bonferroni correction showed significant differences between reuse (Mean = 28.29; SD = 18.352) and change (Mean = 47.48; SD = 29.664), P = 0.019.

No statistically significant difference was found for the effect of password scheme on the number of fixations per second, keyboard fixations, keyboard fixation duration, keyboard fixation duration, keyboard to screen fixations ratio, thinking time, username keystrokes, password keystrokes, password entry duration, password keystrokes-duration ratio, password re-entry keystrokes, password re-entry duration, password re-entry keystrokes-duration ratio and total keystrokes.

Table 7: New vs Changed vs Reuse Gaze Data Stats I

	Total Duration			Total Fixations			Screen Fixations		
	New	Reuse	Changed	New	Reuse	Changed	New	Reuse	Changed
Mean	44.49	27.55	43.64	99.33	67.24	95.14	66.48	42.38	65.57
SD	24.553	12.067	17.853	38.271	28.933	36.748	31.324	18.538	32.040

Table 8: New vs Changed vs Reuse Gaze Data Stats II

	Initial Duration			Username Duration			Username keystrokes Dur. ratio			Total Duration		
	New	Reuse	Changed	New	Reuse	Changed	New	Reuse	Changed	New	reuse	Change
Mean	9.31	4.54	9.54	13.25	8.05	16.32	1.366	.977	1.878	43.86	28.29	47.48
SD	6.568	2.659	8.569	9.873	5.764	12.362	.234	.096	.303	30.943	18.352	29.664

7.4 Machine Learning

The machine learning algorithms described in section 4.3 used a CSV-file including all found and calculated values described in the result subsection. Mail- and news-interface columns were divided, then the resulting data again split into the password-choice column (including new, change, reuse) and the remaining columns containing collected and calculated values. Because the range of our values differed between features, all values were scaled first. These two data sets per password were input for the machine learning algorithms.

Both machine learning algorithms were trained with 80% of all data. The remaining 20% were used to test the calculated decision function. We did not use 100% of the data because in this case the decision function would have been biased. It would have already known the data we used to test it. Since there is no optimal way to implement the classifiers, a good configuration had to be found by experimenting with parameters. We also did not know if some features were more suitable than others to classify password choices, so the combination of features was modified as well.

For the SVM-classifier the best modification we found was with the kernel type 'linear' and the decision function type 'ovr'. The linear kernel separated the multidimensional data points (every feature was one dimension) by creating a hyperplane that divided the points based on their label. The decision function parameter 'ovr' meaning 'one-versus-rest' modified the SVM regarding the number of labels to classify. It classified a data point by comparing it to all remaining data points, treating them as same-labeled. For example, if a point belonged to a reused password the algorithm did not distinguish between new or changed points for the compared remaining data points. By default the parameter is more useful to classify two different label-groups, but we had three (new, change, reuse).

The 'Random Forest'-classifier needed less configuration. We limited the number of grown trees in every run to 30. Also, we scaled the data to get equally distributed data.

The test-results were shown in a confusion matrix showing which samples were classified in which way. We also used the classification report provided by the machine learning tool. The value we focused on was the accuracy. It showed the percentage of correctly classified data points.

It was found that results could vary depending on the nature of the training and testing sets. Therefore, we ran multiple runs of machine learning when we investigated a feature. Every iteration a new, randomly chosen set of data was used for training and evaluation (split 80/20). The number of iterations was set to 30. A confusion matrix and a classification report was shown after every run. At the end of the iteration, the summed up confusion matrix was presented together with the average classification report. We used the average accuracy as a comparative feature between different features. First we calculated the accuracy for every single feature and then for all features together. Table 9 shows the scores for these runs.

	SVM accuracy (%)		Random Forest accuracy (%)	
Feature	**mail**	**news**	**mail**	**news**
Total duration	46,2	51,9	44,2	33,5
Number of fixations	51,4	47,6	41,9	27,1
Fixations per second	53,8	31	40,6	29
Number of screen fixations	56,2	46,2	51,6	47,1
avg. screen-fixations duration	51,9	46,7	30,6	28,1
Number of keyboard fixations	50	39	31,9	32,3
avg. keyboard-fixations duration	44,3	42,9	42,6	27,7
Screen to keyboard ratio	52,4	29	31	39,7
Thinking time	50	37,6	47,4	41,6
avg. time between keystrokes (username field)	50	29	48,4	28,1
avg. time between keystrokes (1st pw field)	51	34,8	30,3	29,4
avg. time between keystrokes (2nd pw field)	45,2	36,2	40,6	39,7
avg. time between keystrokes (whole interface)	51,4	44,3	30,6	44,5
All features combined	**54,2**	**36,4**	**55,2**	**37,74**

Table 9: Average accuracy of 30 runs machine learning for every feature

7.4.1 SVM

The SVM had problems classifying single features for the mail-interface, probably because its strength is to classify in higher dimensions. By analysing the confusion matrices we found that it only classified passwords as 'new' for most cases and very few as 'reuse'. Only for the two features 'average screen-fixations duration' and 'average keyboard-fixations duration' few samples were classified as 'change' and 'reuse', even though they were classified wrong. Most of the times nearly all data points were classified as 'new'. This could be explained by the linear kernel. In a one-dimensional space a linear separator is just a border on a graph. As the algorithm found values for new passwords in a wide range, they might have covered the values belonging to changed and reused passwords. The accuracy-values can be found in a corridor of 45,2% (avg. time between keystrokes) to 56,2% (number of screen fixations). In order to be able to rate the performance of the machine learning we introduced a comparative value: if a password was always classified as the most common label-group, which was 'new', it would result in an accuracy score of 51%. Compared to this value, the accuracy for single features was low. Nevertheless, a higher accuracy was achieved when all features were combined to train the SVM. The score of 54,2% is better than the comparative value by 3,2%. In table 10 the confusion matrix for a run of machine learning on all features combined is shown. Similar to the classification of single features, the majority of data points was classified as 'new' (195=67,24%) with an accuracy of 74,86%. Changed passwords were classified correctly in just 8 out of 57 cases (14,03%) and reused passwords in 26 out of 64 (40,62%).

On the newspaper-interface accuracy-scores ranged from 29% to 51,9%. A higher range of values compared to the mail-interface can be found. The comparative value of 40% this time was even higher than the accuracy if all features were used to train the algorithm (36,4%).

The confusion matrix shown in Figure 11 explains this value by the wrong classification of new and reused passwords. 47 out of 100 (47%) data points were classified as 'new' even though they were labeled as 'reuse'. On the other hand, 47 out of 122 (38,52%) data points were classified as 'reuse' although they actually were 'new'. This confusion of data points resulted in an overall low accuracy score. Still, there were some single features for which a high accuracy was achieved like the total duration. A reason for this may be the clear difference between password choices, visualized in Figure 29. This cluster structure for every feature made it easier to classify. Regarding the classifications the SVM behaved differently compared to the mail-interface. On the newspaper-interface passwords were not classified as 'new' for most cases, but equally as 'new' or 'reused'. Similar to the first interface, nearly no choices were classified as 'change'.

		Prediction outcome		
		new	**change**	**reuse**
Actual value	**new**	134	20	25
	change	36	8	13
	reuse	25	13	26

Table 10: A confusion matrix of 30 runs SVM on all combined features (mail-interface)

		Prediction outcome		
		new	**change**	**reuse**
Actual value	**new**	45	27	36
	change	45	18	19
	reuse	49	11	50

Table 11: A confusion matrix of 30 runs SVM on all combined features (newspaper-interface)

Overall, the SVM classified single features with a low accuracy score, which we thought was caused by its weakness classifying low dimensional data. When using all features combined on the mail-interface, it scored a slightly higher accuracy than the comparative value. On the newspaper-interface the SVM with all features combined as an input could not classify labels with a higher accuracy than the comparative value.

7.4.2 Random Forest

The 'Random Forest Classifier' scored lower accuracy scores for the mail-interface while analysing single features. Scores ranged from 30,3% (average time between keystrokes in the first password field) to 48,4%. In contrast to the SVM, it did not only classify one feature, but all three equally distributed. Still, it categorized many data points wrongly and stayed below the comparative value of 51% in all cases, except for one (number of screen fixations with 51,6%). The low scores might again be caused by the low dimension of the input data. With only one feature to categorize into three label groups, the decision trees were very small and therefore left little leeway. Nevertheless, when combining all features together, the Random Forest-Classifier outreached the comparative value by 4,2% and classified 1% better than the SVM-classifier. The confusion matrix for a classification with all features combined as input, shown in Figure 12 shows that as well as the SVM, the 'Random Forest' algorithm classified most data points as 'new' (138=46%). Similar to the SVM it had problems classifying changed passwords, as only 10 out of 72 (12,89%) data points were classified right. The majority of reused passwords (35=55,56%) was wrongly classified as 'new'. In total, the algorithm classified less new passwords wrongly, which resulted in a higher accuracy overall, compared to the SVM.

On the newspaper-interface most scores for single features were worse than the SVM accuracies with two noticable exceptions ('screen to keyboard-ratio' 10,7% higher and 'thinking time' 4% higher). Three features were classified with higher accuracy-scores ('number of screen fixations', 'thinking time' and 'average time between keystrokes for the whole interface') than the comparative value of 40%. When trained with all features combined, the 'Random Forest' scored a higher accuracy (37,74%) then the SVM-classifier but still stayed more than 2% below the comparative value. The confusion matrix shown in Figure 13 reveals that the majority of values (158=56,43%) was classified as 'new' but just one third of them belonged to this label group. Similar to the SVM many 'new' and 'reuse' data points were confused.

		Prediction outcome		
		new	**change**	**reuse**
Actual value	**new**	138	12	15
	change	45	10	17
	reuse	35	5	23

Table 12: Confusion matrix of 30 runs Random Foreston all combined features (mail-interface)

		Prediction outcome		
		new	**change**	**reuse**
Actual value	**new**	51	23	34
	change	53	23	13
	reuse	54	6	43

Table 13: Confusion matrix of 30 runs Random Foreston all combined features (newspaper-interface)

Compared to the SVM-classifier, the Random-Forest algorithm achieved higher accuracy scores on both interfaces. It was 1% better on the mail- and 3,9% better on the newspaper-interface. The best score for single features could not surpass an accuracy of 55,2%.

8 Discussion

In this section we evaluate the meaning of our results and what they could indicate. We also discuss limitations. We discuss how and why users chose their passwords, what the found gaze data and features could mean and how the results of machine learning could be interpreted.

8.1 Passwords

The results regarding the choice, the nature and the memorability of passwords seem to correspond with literature and our hypothesis about different passwords for websites with different sensibilities. Participants used methods to manage passwords more easily. This included password-reuse and simplicity of passwords.

8.1.1 Password Choices

The results of user's password choices indicate a high number of reuse. Even though participants were not asked to reuse passwords, more than 50% did so independently, which confirmed our hypothesis (6.5). This seems to correspond with the literature 2.1 as well. People were faced with the challenge of creating a new password and tend to reuse already existing ones. This does not mean that the same password is used for all accounts but it indicates possible personal preferences and groups of password behaviour for different application fields (for example e-mail, bank accounts or less sensitive accounts like newspaper). In addition, there were still almost 50% who created a new password for the known e-mail-interface. When comparing both interfaces it is noticeable that the number of people who created a new password decreased and a reuse increased from mail- to newspaper-interface. This could have two reasons. First, it could be due to the repetition of the task. Since participants already thought of a new password, some may have decided to simply reuse it on the second interface with the same task, as they had a password resource available. A second explanation could be that the newspaper website was rated lower regarding sensibility and therefore a weaker (reused) password resource was chosen.

8.1.2 Password Creation Scheme

The scheme by which passwords were created could indicate an easy-to-memorize approach. Very few participants (4 for the first and 2 for the second interface) used a completely random combination of different characters. Everyone else determined their choice based on a memory aid. Most people relied on one or multiple words or a meaningful combination of characters to build their password.

These results again could correlate with the behaviour described in literature (2.3) as people mostly rely on their memory to keep track of their logins. Since random combinations are harder to remember, people base their creation process around known things like words, meaningful phrases or a well defined personal scheme.

Comparing the different password choices with each other, different statements can be supported as stated below.

1. People who change their password seemingly own a password resource consisting of a personal scheme or a meaningful combination of characters. Based on a foundation they once determined and can remember well, they use a personal algorithm to change it just slightly. The foundation seems like something memorable, as no one who changed a password changed a random combination and also only one changed one or multiple words.

2. Participants who created new passwords almost always chose to base them on words or a meaningful combination of characters. This could mean that users either still are unaware

that strong passwords consist of random combinations or they consciously use passwords that are easily memorable.

3. Reused passwords occurred in all categories. This could mean that the type of password that is reused depends on the user. While some reuse a strong password they once created, others reuse a simple one they can memorize well for multiple accounts.

The nature of the password indicates a confirmation of our hypothesis about different sensibility of websites. Even though the number of passwords using random combinations was higher on the mail-interface, there was just a minor difference of one user (2,08%). The distribution of different schemes was similar on both interfaces. Nonetheless, there was a difference regarding the password choice between both interfaces. The fact that less participants created a new password on the newspaper-interface and more reused a password could indicate that they saw no need to create a new password which would be more time-consuming.

8.1.3 Memorability

The hypothesis established in the sections before, saying users would choose well memorable passwords, can be supported by the memorability of passwords. Not only the self-appraisal of participants but also the Levenshtein distance attested a good memorability. Participants created passwords they could remember without password aids. The memorability was also an indicator for the participant's seriousness towards this study. In other studies participants have shown behaviour of mindlessly creating random password combinations in order to reduce the time consuming to complete the study. The fact that most of the participants could reenter their password at the end perfectly and only two reported to have chosen a password particularly simple for the study seems to indicate that this was not the case.

Comparing the memorability between the different password choices, reused passwords were most likely to be memorized well. A reason might be the frequent use of these passwords. The more often it has to be typed the better it is memorized. Even though new passwords were created well-memorable too, they have not been repeated enough yet. This could explain why new passwords were rated with a lower memorability in total. Another salient finding is the user-rated memorability of changed passwords. The peak at the second highest score could indicate a behaviour reported by literature (2.3), describing users choosing a resource and modifying it to get different passwords. They always remember this password resource but may forget on which interface they applied which change.

The different average memorability could be a confirmation for our hypothesis about different types of interfaces. The higher value on the mail-interface could be caused by people choosing a good memorable password for a website they had to visit frequently. On the other side, the newspaper website probably was not as important to users. For that, another reason for creating a less memorable password could be that those websites would be visited less frequently.

8.1.4 Password Characteristics

The used character classes can give an insight to password features and composition, even though they can not tell the concrete structure of a password.

On the mail-interface the password choice seems to correlate with the amount of used character classes. Most reused passwords involved all possible classes (letters, capital letters, digits, special characters). Together with the nature of passwords (section 8.1.2) this could mean that participants involved all classes into their personal scheme or meaningful combination. This way a conflict between two security goals may be visible: on the one hand side users create a strong password containing multiple character classes, on the other side they may embed

it into an obvious combination and in addition reuse the combination. On the newspaper-interface reused passwords included less character classes. Less than half of the total amount of participants compared to the first interface used all possible character classes. This could indicate that more simple passwords were reused. Nevertheless reused passwords still contained 3 or 4 character classes for almost 50% of all participants.

In contrast to reused passwords newly created ones oftentimes contained two or less character classes. This, together with the nature of passwords, might be an indicator for users choosing a simple password based on words. A reason for these simple passwords might be that words are easier to remember, a phenomenon described in researched literature.

Comparing both interfaces it can be detected that on the mail-interface more character classes were used than on the newspaper-interface. The usage of less classes could be an indicator for more simple passwords which in turn could support the hypothesis about different relevance of websites.

8.2 Gaze Data

The collected gaze data opened a new approach to investigate user behaviour regarding password creation. We wanted to explore what gaze behaviour looked like during password creation and if features existed that were more suitable than others to classify user behaviour in the three categories 'new', 'change', and 'reuse'.

8.2.1 Fixation Patterns

The fixation patterns were a good way to get an overview over participant's gaze behaviour. They revealed that our hypothesis regarding the three areas of interest (4.2.2) could be falsified or in need of modification. Users only looked at either the screen or the keyboard and mouse, but not at the outside area of the screen. The reason for this might be that users were focused on the interface, because they were in an unknown environment and not at home, where they would be able to find inspiration from objects on their desk for example. But it could also trace back to technical restrictions by the eye tracker as the pupils had to be visible the whole time. If for example a user turned his or her head too far, no gaze data could be collected. An identification of areas of interest on the side would therefore be impossible.

What the fixation patterns could not reveal was a repeating pattern correlating with the password choice. Participants varied strongly regarding their gaze behaviour. This may be caused by different levels of experience with the operation of computers and keyboard-typing. Also, the individual behaviour could be a remarkable influence. Some participants may extract information faster than others, resulting in less fixations. Hypotheses like 'few fixations on the keyboard if a password was reused because participants did not need to look for keys they wanted to include' applied in some cases, but could not be generalized and therefore had to be dropped.

All combined gaze data points clearly reveal the login fields and the keyboard/mouse as areas of interest. Nearly all fixations happened in these areas with very few data points somewhere else. The reason for this might be that the interfaces were created very simple and the few elements attracted all gaze.

8.2.2 Time Spans

The time spans we calculated supported our hypothesises for the total duration, but did not align with the ones for the thinking time. The total duration revealed differences comparing the three password choices and thus could be a candidate for a classification of user behaviour.

Total Duration

Even though the ranges overlapped, new passwords had a longer duration than reused ones in general. This corresponds with our hypothesis, that new passwords take more time to create compared to reused ones. The more a password differs from existing ones, the more time has to be spent thinking of the combination of characters. This could also explain why the values for changed passwords can be found exactly between new and reuse, as a password resource was already given, but still changes had to be thought of.

Compared to the news-interface, the mail-interface was visited for a longer duration. We assumed two reasons for this. First, the instruction was new. Participants faced a new task and took some time to explore the page they saw on the mail-interface. After completion, they knew what to do and could finish the same task faster on the second interface. Second, it might have taken participants longer to create a password on the mail-interface because of the password policy. While it obliged a minimum effort on the mail-interface, users were free to create passwords without any restrictions on the second interface.

Thinking Time

Contrary to the hypothesis we made about thinking time, there were little to no differences between the three conditions. On both interfaces the thinking time occurred in a small range with nearly the same mean and average. The few outliers can be explained by some participants talking to study-personnel. We hypothesized that users who used a new password might think of a new password after entering their username and before entering their password, however this was not measurable. The single value that supported this hypothesis was the thinking time deviation for the reused passwords on the newspaper-interface. Times occurred in an even smaller range than other thinking times and the average was significantly lower. Password-reusing participants on the newspaper-interface seemed to be very sure and quick about what they chose as a password. Overall, the time difference between 'new' and 'reuse' we observed for the total duration (8.2.2) seems not to be caused by differences regarding the thinking time. Participants might have thought about their password at some other point during the study. This could have taken place at the beginning, even before the username was typed, or during the creation of the first password.

8.2.3 Ratios

With calculated ratios we wanted to compare users among themselves. One ratio aligned with our hypothesises about user behaviour while another one partially disagreed and a third one even seems to show the opposite effect. The average time between keystrokes and the screen-to-keyboard ratio might be good metrics to classify password reuse, while the fixations per second had some flaws.

Fixations per Second

In our hypothesis about fixations per second (4.2.2) we assumed that there would be more fixations for a new compared to a reused password. On the mail-interface this seems to be the case. Users who created new passwords had a high ratio of fixations per second as opposed to this changed and reused passwords. An explanation could be that people who created a new password searched for characters on the keyboard. Also they would have to think of the new password and meanwhile rested their gaze somewhere to concentrate. Users who had a password resource available knew where to look for the elements of a password and created less fixations but probably more saccades while quickly switching between elements. This includes not only reused passwords but also changed ones, where a basis for a password was already given.

On the newspaper-interface this pattern does not seem to apply. Even though the mean for new passwords was higher than for reused ones, there were some outliers in unexpected directions. Some people created very few fixations per second for new passwords. This might be explained with simple passwords. As we did not apply a password policy on the second interface, people could have chosen a very simple password which could be typed quickly. Therefore, less overall time would have been needed. With a shorter duration the fixation rate would go up.

Comparing both interfaces, the fixations per second rate seems to be a good metric for the mail-interface but not for the newspaper-interface.

Average Time between Keystrokes

The average time between keystrokes shows differences between all three conditions. An explanation for the longer duration between keystrokes could be the total time users needed to create a new password. This corresponds with our hypothesis for the total duration in 8.2.2. Reused passwords were already known and there was no additional time required to think of characters. They could be typed faster and resulted in a low ratio. The ratio for changed passwords in between the ratio for new and reused passwords could indicate a mixture of both behaviours. A new change might have been applied to a reused component. A higher ratio for new and a lower one for reused passwords could be observed similarly on both interfaces, mail and newspaper. This could again confirm the hypothesis of additional time needed for a new password. What was different between both interfaces were the actual values. On the mail-interface ratios for all three password choices were higher than on the newspaper-interface. An explanation could be that users needed less overall time on the newspaper-interface. They already knew what the task was and could also make up shorter passwords.

All in all, the time between keystrokes could be a good metric to distinguish between password choices in addition to other gaze related metrics.

The average time between keystrokes on the first password field on both interfaces indicates differences in user behaviours as well. Participants who reused passwords took significantly less time between keystrokes than password-changing or new creating participants. This in combination with our findings about thinking time might be caused by users thinking about new passwords. As the thinking time did not vary between conditions but the total duration showed differences between these, we supposed that users must have thought of changes at other points. The higher average time between keystrokes for new and changed password could indicate the fact that users took time thinking of their choice on top of typing it down.

Screen to Keyboard Ratio

Our hypothesis that people would look more at the keyboard to search for keys if they created a new password could not be approved by the ratio. Nearly all users had more fixations on the screen than on the keyboard, which might be evidence for users extracting more content on this part of the setup. On the newspaper-interface, the ratio for all three password choices was very similar. All values were located in a similar area and the outliers also matched across conditions. We thought of two explanations why the ratio was not as expected:

1. Participants might not look at the keyboard while creating a new password. They could be quite sure about the layout of the keyboard and focus their gaze not on the keys they pressed but somewhere else. The keyboard as an area of interest then would be useless to distinguish between the different passwords. A task would be to identify the area of interest for this behaviour.

2. The duration of gaze might be too short to be considered a fixation. If participants just wandered across keys they chose, it would only result in saccades, which we did not cover. Therefore, this behaviour would not have been covered in our calculation.

On the mail-interface the ratio did not stay equal like on the newspaper, but decreased with a reuse. This means that users looked more on the keyboard if they reused a password, while users who created a new one had more fixations on the screen. An explanation for this might be the first mentioned reason: if the area of interest while creating a new password was not the keyboard the ratio would rise instead of going down.

Apart from the fact that the ratio was not divided like we expected, it could be a metric to classify reuse, at least on the mail-interface. The values were different with different means and average values.

8.3 Machine Learning

If it was possible to classify password-choices with the aid of machine learning, it would be possible to transfer this model to other interfaces and detect password reuse.

8.3.1 First Approach

Unfortunately, our accuracy-scores on the first iteration of machine learning were not very high. If we calculate the accuracy that one would achieve if all data points were classified as belonging to the biggest group ('new'), we would achieve a score of 51% under the mail- and 40% under the newspaper-condition. Compared to these values the scores of the SVM and the Random-Forest-Classifier were similar. On the newspaper-interface the classifiers both were unable to exceed the comparative value. Although the Random Forestalgorithm predicted with a higher accuracy than the SVM, it was still more than 2% worse than guessing the most likely outcome. On the mail-interface the classifiers performed better than on the newspaper-interface and were able to exceed the comparative value by more than 3%. Still an accuracy score for the mail-password of 54,2% (SVM) and 55,2% (Random Forest) is too low to clearly distinguish between password choices.

The bad machine learning results could be caused by different reasons, single or combined:

Too few data Machine learning needs a lot of data. If we did not offer enough data points the algorithms may be incapable of finding patterns. It then classified based on wrong (random) characteristics.

Wrong features We evaluated features based on our hypotheses. Some may give more information about user behaviour than others (7). Also there may be useful features we did not include yet and that can still be found.

Differences between participants Eye movements are a very individual biometric. They differ among all people even though they share characteristics. But if these overlapped it would be impossible to tell behaviours apart. A user that created a new password could have the same values for features as a password-reusing user and would be impossible to be classified by machine learning. To compare users it could be necessary to first collect eye movements of every single user in multiple situations. Then the difference in his own behaviour could be set in relation to others.

Too little differences between behaviours Maybe behaviours were very close to each other and a transition was fluid. For example the 'change'-behaviour can be found between 'new' and 'reuse' (7), which means values could always be matched to one of those groups. Most of the time, no clear cuts between features can be found. This does not allow a strict separation and leaves miscalculations inevitable.

Overall the results did not match our expectation. Based on our findings it is not reliably predictable if a person reused a password, changed an existing one or created a new one. The

machine learning accuracy scores were lower and not much more helpful than just guessing the most common case. Even though we extracted multiple features and metrics a classification based on these was not possible by machine learning.

8.3.2 Second Approach

Which factors influenced the scores were not easy to find out afterwards. The advantage of machine learning algorithms, that they train themselves to find their result now has an impact on the evaluation of the results themselves. It is not described how the decision function looked like and how features influenced it. We still wanted to improve the score of the machine learning and started to modify the training.

We started by combining the five best scoring features together and retrained the algorithms for the mail-interface. The Random-Forest-Classifier scored a similar accuracy as before (55,16%), while the SVM classified almost 4% worse (50,6%) and below the comparative value. The selection of features we evaluated therefore could not improve the scores of machine learning. This could support our second error-hypothesis about wrong features. Unfortunately, we were limited to the features we extracted because of time constrains.

We also combined label-groups. Based on our fourth error-hypothesis about similar behaviours we decided to include the 'change' group into both others. Then we trained the machine learning algorithms with the data from the mail-interface again.

		Prediction outcome	
		new	**reuse**
Actual value	**new**	105	56
	reuse	53	86

Table 14: A confusion matrix of the SVM classifier if 'change' and 'reuse' labels are combined (mail-interface)

		Prediction outcome	
		new	**reuse**
Actual value	**new**	91	70
	reuse	32	107

Table 15: A confusion matrix of the Random Forest Classifier if 'change' and 'reuse' labels are combined (mail-interface)

If the 'change' and 'reuse' groups are combined and were used to train the algorithms the SVM scored an accuracy of 63,87% and the Random-Forest 61,61%. With an updated comparative value of 68,7% both classifiers predicted below it. In Table 14 and 15 the confusion matrices show that the SVM classified more data points as new, while the Random Forest classified more as reused, but since both classified multiple points the wrong way, they resulted in a similar low accuracy score.

		Prediction outcome	
		new	**reuse**
Actual value	**new**	184	48
	reuse	36	32

Table 16: A confusion matrix of the SVM classifier if 'new' and 'reuse' labels are combined (mail-interface)

		Prediction outcome	
		new	**reuse**
Actual value	**new**	212	16
	reuse	49	23

Table 17: A confusion matrix of the Random Forest Classifier if 'new' and 'reuse' labels are combined (mail-interface)

If in contrast 'new' and 'change' were grouped together the SVM predicted correctly with an accuracy of 69,67% and the Random-Forest with 75,8%. The confusion matrices in Table 16 and 17 show the way the classifiers predicted. In comparison to the SVM the Random Forest algorithm classified more passwords as new and therefore achieved a higher accuracy. Still, it was not really able to classify many data points belonging to the 'reuse' label group in a correct way. The comparative value of 71% was higher than the SVM score and lower than the Random Forest accuracy score. Still, a score 4,8% higher than the comparative value meant that a reliable prediction was not likely. Nevertheless, the results could reveal another interesting finding. That the score worsened when combining 'change' and 'reuse' and rose when combining 'change' and 'new' could mean, that the behaviour of changing a password might be closer to creating a new one.

modification (mail)	**SVM accuracy (%)**	**Random Forest accuracy (%)**	**Comparative Value (%)**
use 5 best features	50,6	55	51
combine 'change' and 'reuse'	63,87	61,61	68,7
combine 'change' and 'new'	69,67	75,8	71

Table 18: Accuracy scores for modified conditions

Lastly, based on the third error-hypothesis about different user behaviours we went through the results again and removed the data points of outliers for different features from the data set. We presumed that this would make it easier for the machine learning to classify because the range of values compressed. Unfortunately, the results were not better. For example if we removed outliers for the 'fixation per second'-feature the score dropped from 40,6% to 36,7% (Random Forest) and stayed at 53% (SVM) on the mail-interface. An explanation for the missing increase of accuracy could as well be our fourth error-hypothesis. If behaviours were too close to each other and overlapped, a removing of outliers would not have an influence. Maybe even the opposite effect would be caused because easily classifiable data points would be removed.

Overall, machine learning was not able to classify user behaviours with a high accuracy using the features we extracted as training data. However, we were able to classify better than prediction based on the most likely outcome, which could indicate that there still is a difference between behaviours. Based on the low accuracy scores we thought of multiple error-hypotheses. We tried to verify these as well and could improve our accuracy score in one case.

Comparing the machine learning algorithms we used, 'Random Forest' classified better when given all features at once. The SVM could only make better predictions for some single features but failed to exceed the 'Random Forest' in higher dimensions.

9 Conclusion and Future Work

In this thesis, we conducted a study to research gaze behaviour of people while creating passwords. Through the use of machine learning we wanted to find out whether we could classify the password choice of users based on their gaze behaviour. If this is possible, it would allow interactive feedback for users while they are creating passwords. For example, users could be warned that they reused a password, which is a threat regarding IT-security-goals.

The study consisted of two interfaces and a questionnaire about users' password choices. We decided to use two different interfaces. One was the university's mail-interface, a well known, frequently visited website with access to sensitive data. The other one was a popular newspaper website, which we estimated was rated as less sensitive. With the help of a focus group we designed the interfaces and determined features to extract from the raw gaze data. In order to be able to classify users regarding their password choice, we hypothesised how gaze behaviour would look like if participants thought of a new, changed an old or reused an existing password. The final study took place in the university canteen, where 52 participants were recruited. Gaze data was collected by an eye tracker and saved to later process it. As we included a questionnaire about users' password choices, we could also verify some hypotheses about password use in general.

Although participants were instructed to create new passwords, many reused existing ones. Furthermore, the used passwords often were of simple nature, consisting of words or meaningful combinations. The findings seem to support the hypothesis that people mainly use their memory to keep track of passwords and therefore create simple passwords. Also, the behaviour of password-reusage to cope with many logins was observed by us. Comparing both interfaces, the results indicate that users used weaker and reused passwords on the newspaper interface. This could indicate that people behave differently on 'less important' websites.

Based on the collected gaze data we wanted to find out if features existed that indicated a specific user behaviour and which ones were more suitable. We presumed that users who created new passwords would take longer altogether, compared to users who reused an existing one. The durations we extracted seem to confirm this thesis. A hypothesis that could not be confirmed was that the area the gaze was directed to would be influenced by the behaviour. We could not find evidence for 'new-password-creating' users to look less at the screen and more at the keyboard to search for keys they included in their passwords. Also, the patterns of fixations did not allow a classification of users. Different ratios and values we calculated showed differences between the password choices but the transitions were fluent and did not allow clear separation.

To further investigate the data, we implemented two machine learning algorithms ('Support Vector Machine', 'Random Forest') and trained them to classify password choices. The accuracy scores of both machine learning algorithms show that it was difficult to distinguish between behaviours. The best predictions were just slightly better than guessing the most probable outcome. One classifier even predicted less accurately than this comparative value. In general, we hypothesized that gaze behaviour overlapped between conditions and made it hard to strictly classify a specific behaviour. In addition, eye movements are very individual and may be hardly comparable between people. Still, we assumed that gaze behaviour shows different characteristics comparing different conditions, which could be confirmed by modifying the data. The machine learning algorithms scored a higher accuracy when changed and new created passwords were treated as one class compared to when changed and reused passwords were treated as one class. This might indicate that gaze behaviour for new and changed passwords are similar and could be useful for future research.

Overall, we were not able to develop a model that predicts which password a user chose. Still, we got an insight into gaze behaviour during password creation which could be taken as foundation for future work.

In future works one could try to extract further features. Our set found some relevant features but it was limited in its extent. Some yet uncovered features lead to better classifications. From our findings we propose features like the 'number of fixations to a certain point of time'. The investigation to a certain point like the first keystroke in the password entry field could reveal that new-creating users had more fixations because of more time they needed to create an entry for example.

Furthermore, the effect of more data points on the machine learning could be investigated in future works. This means that more participants would have to be recruited. More data points may be able to crystallize areas of focus for different behaviours or in contrast reveal that gaze behaviour does not differ enough to get classified with a high enough accuracy.

To improve the machine learning results and also understand users' behaviour it could also be an approach to analyse single users. The data split of 80/20 we applied had multiple users in both groups. If in the testing only ever one participant was used, the ones that were hard to classify may be identifiable and could be studied separately. This might give access to further features and help to understand similarities and differences among gaze behaviours.

References

[1] N. Alkaldi and K. Renaud. Why do people adopt, or reject, smartphone password managers? *EuroUSEC 2016: The 1st European Workshop on Usable Security*, 2016.

[2] E. Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[3] S.-i. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

[4] R. Bednarik, T. Kinnunen, A. Mihaila, and P. Fränti. Eye-movements as a biometric. In *Scandinavian conference on image analysis*, pages 780–789. Springer, 2005.

[5] J. Bonneau, C. Herley, P. C. Van Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *2012 IEEE Symposium on Security and Privacy*, pages 553–567. IEEE, 2012.

[6] R. Dhamija, A. Perrig, et al. Deja vu-a user study: Using images for authentication. In *USENIX Security Symposium*, volume 9, pages 4–4, 2000.

[7] S. Drimer, S. J. Murdoch, and R. Anderson. Optimised to fail: Card readers for online banking. In *International Conference on Financial Cryptography and Data Security*, pages 184–200. Springer, 2009.

[8] G. B. Duggan, H. Johnson, and B. Grawemeyer. Rational security: Modelling everyday password use. *International journal of human-computer studies*, 70(6):415–431, 2012.

[9] ElNinja. Tkinter - python wiki, 3 2020. `https://wiki.python.org/moin/TkInter`.

[10] D. Florencio and C. Herley. A large-scale study of web password habits. In *Proceedings of the 16th international conference on World Wide Web*, pages 657–666. ACM, 2007.

[11] P. S. Foundation. tkinter - python interface to tcl/tk - python 3.8.2 documentation, 3 2020. `https://docs.python.org/3/library/tkinter.html`.

[12] S. Gaw and E. W. Felten. Password management strategies for online accounts. In *Proceedings of the second symposium on Usable privacy and security*, pages 44–55. ACM, 2006.

[13] gentaiscool Program Creek. Python levenshtein.distance() examples, 3 2020. `lhttps://www.programcreek.com/python/example/94974/Levenshtein.distance`, example 15.

[14] C. Ghaoui. *Encyclopedia of human computer interaction*. IGI Global, 2005.

[15] J. H. Goldberg and X. P. Kotval. Computer interface evaluation using eye movements: methods and constructs. *International journal of industrial ergonomics*, 24(6):631–645, 1999.

[16] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky. Eye tracking in web search tasks: design implications. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 51–58, 2002.

[17] J. H. Goldberg and A. M. Wichansky. Eye tracking in usability evaluation: A practitioner's guide. In *the Mind's Eye*, pages 493–516. Elsevier, 2003.

[18] B. Grawemeyer and H. Johnson. Using and managing multiple passwords: A week to a view. *Interacting with Computers*, 23(3):256–267, 2011.

[19] E. Hayashi and J. Hong. A diary study of password usage in daily life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2627–2630. ACM, 2011.

[20] J. Hunter and D. Michael. Pypi-matplotlib, 1 2020. `https://pypi.org/project/matplotlib/`.

[21] P. G. Inglesant and M. A. Sasse. The true cost of unusable password policies: password use in the wild. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 383–392, 2010.

[22] I. Ion, R. Reeder, and S. Consolvo. "... no one can hack my mind": Comparing expert and non-expert security practices. In *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)*, pages 327–346, 2015.

[23] B. Ives, K. R. Walsh, and H. Schneider. The domino effect of password reuse. *Communications of the ACM*, 47(4):75–78, 2004.

[24] ivw Statista GmbH. Ranking der auflagenstärksten überregionalen tageszeitungen in deutschland im 4.quartal 2019, 01 2020. `https://de.statista.com/statistik/daten/studie/73448/umfrage/auflage-der-ueberregionalen-tageszeitungen/`.

[25] R. J. Jacob and K. S. Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye*, pages 573–605. Elsevier, 2003.

[26] A. K. Jain, A. Ross, and S. Pankanti. Biometrics: a tool for information security. *IEEE transactions on information forensics and security*, 1(2):125–143, 2006.

[27] M. A. Just and P. A. Carpenter. The role of eye-fixation research in cognitive psychology. *Behavior Research Methods & Instrumentation*, 8(2):139–143, 1976.

[28] C. Katsini, Y. Abdrabou, G. E. Raptidis, M. Khamis, and F. Alt. The Role of Eye Gaze in Security and Privacy Applications:Survey and Future HCI Research Directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, New York, NY, USA, 2020. Association for Computing Machinery.

[29] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of passwords and people: measuring the effect of password-composition policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2595–2604. ACM, 2011.

[30] A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[31] S. Mare, M. Baker, and J. Gummeson. A study of authentication in daily life. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 189–206, 2016.

[32] S. P. Marshall. Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity, 7 2000. US Patent 6,090,051.

[33] L. Maughan, S. Gutnikov, and R. Stevens. Like more, look more. look more, like more: The evidence from eye-tracking. *Journal of Brand management*, 14(4):335–342, 2007.

[34] G. Notoatmodjo. *Exploring the'weakest link': A study of personal password security*. PhD thesis, Citeseer, 2007.

[35] N. I. of Standards and Technology. Nist special publication 800-63b, 1 2020. `https://pages.nist.gov/800-63-3/sp800-63b.html`.

[36] B. Parno, C. Kuo, and A. Perrig. Phoolproof phishing prevention. In *International conference on financial cryptography and data security*, pages 1–19. Springer, 2006.

[37] M. Pomplun and S. Sunkara. Pupil dilation as an indicator of cognitive workload in human-computer interaction. In *Proceedings of the International Conference on HCI*, volume 273, 2003.

[38] A. Poole and L. Ball. *Eye tracking in human-computer interaction and usability research: Current status and future prospects*, pages 211–219. Idea Group Reference, 01 2006.

[39] A. Poole, L. J. Ball, and P. Phillips. In search of salience: A response-time and eye-movement analysis of bookmark recognition. In *People and computers XVIII—Design for life*, pages 363–378. Springer, 2005.

[40] C. U. Press. Definition of machine learning from the cambridge advanced learner's dictionary & thesaurus.

[41] R. W. PROCTOR, M.-C. LIEN, K.-P. L. VU, E. E. SCHULTZ, and G. SALVENDY. Improving computer security for authentication of users: Influence of proactive password restrictions. *Behavior Research Methods, Instruments, & Computers*, 34(2):163–169, 2002.

[42] S. Ray. Commonly used machine learning algorithms, 3 2020. `https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/`.

[43] S. Ray. scikit-learn machine learning in python, 3 2020. `https://scikit-learn.org/stable/index.html`.

[44] D. Recordon and D. Reed. Openid 2.0: a platform for user-centric identity management. In *Proceedings of the second ACM workshop on Digital identity management*, pages 11–16, 2006.

[45] S. Riley. Password security: What users know and what they actually do. *Usability News*, 8(1):2833–2836, 2006.

[46] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[47] R. Shay, A. Bhargav-Spantzel, and E. Bertino. Password policy simulation and analysis. In *Proceedings of the 2007 ACM workshop on Digital identity management*, pages 1–10. ACM, 2007.

[48] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements: user attitudes and behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, page 2. ACM, 2010.

[49] J. L. Sibert, M. Gokturk, and R. A. Lavine. The reading assistant: eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*, pages 101–107, 2000.

[50] J. M. Stanton, K. R. Stam, P. Mastrangelo, and J. Jolton. Analysis of end user security behaviors. *Computers & security*, 24(2):124–133, 2005.

[51] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, et al. How does your password measure up? the effect of strength meters on password creation. In *Presented as part of the 21st {USENIX} Security Symposium ({USENIX} Security 12)*, pages 65–80, 2012.

[52] B. Ur, F. Noma, J. Bees, S. M. Segreti, R. Shay, L. Bauer, N. Christin, and L. F. Cranor. " i added'!'at the end to make it secure": Observing password creation in the lab. In *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)*, pages 123–140, 2015.

[53] R. Wash, E. Rader, R. Berman, and Z. Wellmer. Understanding password choices: How frequently entered passwords are re-used across websites. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 175–188, 2016.

[54] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben. Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research*, 4(1):23–45, 2016.

[55] H.-J. Yoon, T. R. Carmichael, and G. Tourassi. Gaze as a biometric. In *Medical Imaging 2014: Image Perception, Observer Performance, and Technology Assessment*, volume 9037, page 903707. International Society for Optics and Photonics, 2014.

[56] Y. Zhang, F. Monrose, and M. K. Reiter. The security of modern password expiration: An algorithmic framework and empirical analysis. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 176–186. ACM, 2010.

[57] M. Zviran and W. J. Haga. Password security: an empirical study. *Journal of Management Information Systems*, 15(4):161–185, 1999.

Attached Files

The attached Files on a CD contain:

1. The code to run the study (Python and c#)
2. All collected gaze data (CSV-files)
3. All answers to questionnaire (CSV-files)
4. Codes to extract features from gaze data (Python)
5. Calculated features (CSV-files)
6. Machine Learning code (Python)
7. Plotted Graphs and images of the study's Design
8. Pictures of the Setup and the 'Form of Consent'
9. Results of the statistical analysis (PDF)