**Gottfried Wilhelm
Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut für Praktische Informatik
Fachgebiet Software Engineering**

Developing Personas for Explainability Requirements by Means of a User Study

Bachelorarbeit

im Studiengang Informatik

von

Joshua Giuseppe Puglisi

**Prüfer: Prof. Dr. rer. nat. Kurt Schneider
Zweitprüferin: Dr. rer. nat. Jil Klünder
Betreuer: M. Sc. Jakob Droste**

Hannover, 23.01.2023

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und ohne fremde Hilfe verfasst und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsamt vorgelegen.

Hannover, den 23.01.2023

Joshua Giuseppe Puglisi

Abstract

Explainability is the ability of a software system to explain itself to its users and is subjective in nature like other known non-functional requirements (NFRs). The evaluation process is difficult because not every user has the same needs for explanations. Moreover, different behaviors and personalities might lead to different needs for types of explanation. In this work, to mitigate the subjectivity of explainability, fictitious identities called personas are developed. Users with similar goals and behaviors are grouped and represented by these personas. The personas serve as representations of the users in raising explainability requirements to assist software developers. A concurrent mixed method research involving 70 survey respondents, 10 of which were interview respondents, is conducted as the first steps towards this goal. The interview data is supposed to help in providing more detailed answers from the respondents. At the end of the user study, 4 general explainability persona are created from the collected and evaluated data.

The evaluation of personas is known to be very difficult, thus these personas are put to test through hypothetical software demonstrations in order to validate their ability in raising explainability requirements. However, testing the personas alone will not guarantee their credibility. To prove that the personas developed can relate to the users they represent, an additional hands-on software demonstration similar to the persona test, but with real users, is necessary. The goal is to observe and test the resemblance between the respondents' behaviors and their personas. The end goal is to investigate if the general explainability personas developed in this bachelor's thesis are applicable to specific software systems.

Zusammenfassung

Erklärbarkeit ist die Fähigkeit eines Softwaresystems, sich seinen Benutzern selbst zu erklären, und ist wie andere nichtfunktionale Anforderungen (NFRs) bekanntermaßen subjektiv. Der Bewertungsprozess ist schwierig, da nicht jeder Nutzer denselben Erklärungsbedarf hat. Außerdem können unterschiedliche Verhaltensweisen und Persönlichkeiten zu unterschiedlichen Erklärungsbedürfnissen führen. Um die Subjektivität der Erklärbarkeit abzumildern, werden in dieser Arbeit fiktive Identitäten, sogenannte Personas, entwickelt. Benutzer mit ähnlichen Zielen und Verhaltensweisen werden gruppiert und durch diese Personas repräsentiert. Die Personas dienen als Repräsentationen der Benutzer bei der Erhebung von Erklärbarkeitsanforderungen zur Unterstützung von Softwareentwicklern. Als erster Schritt in Richtung dieses Ziels wird eine parallele Mixed-Method-Forschung mit 70 Umfrageteilnehmern, von denen 10 Interviewteilnehmer waren, durchgeführt. Die Interviewdaten sollen dabei helfen, detailliertere Antworten der Befragten zu erhalten. Am Ende der Nutzerstudie werden aus den erhobenen und ausgewerteten Daten 4 allgemeine Erklärbarkeits-Personas erstellt.

Die Bewertung von Personas ist notorisch schwierig, daher werden diese Personas durch hypothetische Softwaredemonstrationen getestet, um ihre Fähigkeit zu validieren, Erklärbarkeitsanforderungen zu erheben. Das Testen der Personas alleine garantiert jedoch nicht ihre Glaubwürdigkeit. Um zu beweisen, dass die entwickelten Personas sich auf die Benutzer beziehen können, die sie repräsentieren, ist eine zusätzliche praktische Softwaredemonstration ähnlich dem Persona-Test, aber mit echten Benutzern, notwendig. Ziel ist es, die Ähnlichkeit zwischen den Verhaltensweisen der Befragten und ihren Personas zu beobachten und zu testen. Das Endziel ist es zu untersuchen, ob die in dieser Bachelorarbeit entwickelten, allgemeinen Erklärbarkeits-Personas auf bestimmte Softwaresysteme anwendbar sind.

Contents

Chapter 1

Introduction

1.1 Motivation

The development of software has come to point where it is the norm that a system is able to perform complicated tasks simultaneously while being convenient and easy to use. Complicated process of decision making is often hidden behind a black box. This reduces the transparency of the system, which then might lead to possible bias, questions of justice and fairness in decision making, and lowered user acceptance and satisfaction [7].

An explanation might be needed for this matter. A good explanation can help increase users trust, satisfaction, and also help them in making decisions [29]. However, according to Chazette et. al [5], just like other NFRs, explainability is difficult to measure and validate. They argued that the difficulty comes from the nature of NFR being subjective and relative. In other words, the need for explanation differs for every user. The main purpose of this study is to investigate if personas could serve as a tool in raising explainability requirements.

1.2 Problem Statement

Personas have been widely used in many software development processes. As mentioned above, the need for explanations is different for every user, because explainability affects users differently. One user may need more explanations, another may need less. Finding the balance between the two extremes is not an easy task, hence the use of personas as representations of the users. During the development of personas, using only quantitative measurements leaves out the important detailing of the users' needs, which would make personas difficult to use and filled with the developers' bias and assumptions [1]. Furthermore, measuring them qualitatively through a number of interviews without quantitative data will only reduce the persona's credibility [25]. Combining both qualitative and quantitative measurements could prove to be useful in developing personas, which will be further discussed in this thesis.

1.3 Solution Approach

The first thing to consider before developing personas is the possible user groups that will later be represented by the personas. In order to find out how many user groups exist in a user base, a user study is created based on their experience in using different types of softwares and applications. The user study method used in this thesis was concurrent mixed method research, combining survey and interview, which both were conducted at the same time. The users were given a set of questions asking if they need an explanation in certain situations they might find themselves in during the use of the applications. Because the questions are made based on trend observations, it is possible to create assumptions concerning the users' need for explanations.

From the total of 70 respondents, 10 were interviewed with the same questions from the survey. The interviews are done in order to get a more detailed testimonies. As mentioned in the previous section, quantitative data alone does not suffice. By combining the survey and interview data, the aim is to increase the credibility of the persona, and prevent developer's bias.

1.4 Thesis Structure

This thesis consists of 6 chapters. The motivation on why this thesis is created, the problem encountered, and the approach to reach the solution were explained in the first chapter, in addition to the brief explanation of the thesis' structure. The related works and fundamental background knowledge will be the main focus of chapter 2. Chapter 3 will focus on the details of the research design used in this thesis. The evaluation of the study results is explained in chapter 4. Chapter 5 discuss the points that answers the research questions concerning the design and the evaluation results. Lastly, chapter 6 contains the conclusion and an outlook on possible future works.

Chapter 2

Background and Related Work

In this chapter, related works that are fundamental to this thesis and the experiment conducted will be explained in detail. The topics covered will be explainability, personas, and user study. The explainability section focuses on how it may impact other quality aspects. The persona section will explain the concept of personas and also their strengths and weaknesses. Lastly, the user study section will discuss the chosen method of data collection and evaluation.

2.1 Explainability

By definition, explainability is the ability of a software system to explain itself. It is classified as non-functional requirement or NFR. Following the definition by Chazette et al. [5], explainability can be understood as follows:

> *A system **S** is explainable with respect to an aspect **X** of **S** relative to an addressee **A** in context **C** if and only if there is an entity **E** (the explainer) who, by giving a corpus of information **I** (the explanation of **X**), enables **A** to understand **X** of **S** in **C**. [5]*

In other words, a system can be said to be explainable if at the very least it can successfully explain a certain aspect or feature of the system to the user, so that the user will be able to understand that part of the system. This definition has taken the user's knowledge capacity into consideration, in which the user doesn't have to be able to understand the system as whole, but only parts of it. That way, it may help developers as well in raising requirements, because there is a clear focus. From the definition above, it can be seen that explainability relates closely to understandability. However, after conducting their extensive reviews of related literature, Chazzete et al. [5] found that explainability impacts not only understandability, but other requirements as well. They concluded that explainability and other requirements have a causal relationship. It could be said that an impact on one of these requirements may lead to impacting other requirements. This will be further explained in the following subsections.

2.1.1 Explainability on Understandability and Confidence

Based on the definition of explainability by Chazzete et al. [5], understandability has been one of the main goals of explanations. Explanations may help users understand better how and why a decision was made by the system [6]. By achieving this understanding, users should feel confident in their decision making, which is supposed to increase their acceptance in the results delivered by the system.

Providing explanations may seem to be the definitive solution, but it comes with a number of challenges as well. As explained by Chazette, Brunotte, and Speith [5], due to NFR's nature being relative and subjective, it is hard to validate and evaluate such requirements. The information provided as explanations may not be needed by some users, or worse, they might not be helpful explanations, but a general information that is already known by many stakeholders [15]. Simple and watered down explanations leave out important information that may help the users. Overly complicated explanations also hinder the user from truly understanding how the system works, and could impact the confidence during decision making [30]. Therefore, explanations have to be carefully designed so that they can actually support the users in understanding the system [6].

2.1.2 Explainability on Trustworthiness and Supporting Decision Making

Explainability can have many advantages if applied correctly. It can help inspire trust and support users making good and accurate decisions [29, 15]. On the other hand, a lack of explainability could impact users' willingness to make decision on a software system. According to Rosenfeld and Richardson [24], users are less likely to adopt systems that they do not understand and trust. It is humane for users to avoid making mistakes and wanting to be in control over their decisions [23]. Software obscurity leads to users distrusting the system and creating their own understanding [2, 4]. In such case of self-reliance, users might deviate from the system's constraints. However, in line with the arguments from Tsai and Brusilovsky [30], too much explanation could negatively impact a user's confidence, which then according to Bussone et al. [4], leads to over-reliance on the system. Over-reliance means the user will risk navigating through the software without truly understanding how it works. In case of errors, the feeling of distrust would be instilled in the user, making them no longer wanting to depend on the system, thus resulting in a self-reliance issue.

2.1.3 Explainability on Persuasiveness and Supporting Decision Making

Rosenfeld and Richardson [24] stated that explainability should provide information that could persuade the user into taking certain actions. That is the persuasiveness of explanations. Going by the causal relationship between the quality aspects impacted by explainability, in order for explanations to be able to persuade the users, they have to instill trust first. This trust can be achieved if the explanations are focused on the transparency of the software system [24]. Supporting Rosenfeld's and Richardson's argument, Tsai and Brusilovsky [30] explained that a useful explanation would help users to understand the system's working process, which allows the users to make a better decision or persuade them to accept the suggestions provided by the system.

2.1.4 Explainability on Learnability

Explainability also has many advantages in relation to users' learnability. By increasing users' understandability concerning the inner workings of the system, the user should be able to learn the capabilities of the system when it was making the correct decision, or even wrong ones. According to Köhl et al. [18], explanations do not only enable understanding, instill trust and improve usability, but can also help minimize the chance of human error. Minimizing chance of human error does not only apply to the current situation, but also to future encounters. When a user has learned the capabilities of a system, they should be able to make better decisions, which correlates to previous subsections. Hoffman et al. [14] also mentioned that good explainability should enable the users to efficiently and effectively apply the system in their work, for the purposes that the system is intended to serve. This is possible if the users have achieved a good understanding of the system.

2.1.5 Conclusion of Findings

From these findings, it can be concluded that users' confidence in decision making is greatly influenced by how far their understanding of a system is. When users are able to understand the system, they will be more persuaded into making good decisions. Good decisions leads to users' confidence, which then leads to positive acceptance of the judgements made by the system. In the end, users are able to learn through their interaction with the system, and they will trust the system as well in future encounters.

Enabling this through explanations, however, is not an easy task. An appropriate amount of information has to be provided so that it could create a balanced reliance on the system. The relevance of the explanation must also be taken into consideration, so that user would not be burdened with abundance of information. Evaluating NFRs like explainability to find the perfect balance could be done through the use of personas, which their effectiveness will be investigated in this thesis. Before that, the next section will explain the general concept of personas.

2.2 Personas

Personas are fictitious people created to enable developers to better understand who their users are [22]. Although they are not real, they are very specific and concrete representations [28]. They serve as specifications that must be filled by the developers in order to raise requirements. However, they do not appear like formal specifications; they have names, jobs, feelings, goals they want to achieve that sometimes are not aligned to what the system is designed for [11]. However, that is what makes personas unique, as the goal is to make personas appear as real as possible [1]. The following subsections will discuss the advantages and challenges of using personas based on the findings of our literature reviews.

2.2.1 Advantages of Using Personas

Personas represent user groups with similar goals in mind [31]. In this case, the need for explanations may differs for most users, and chances are that we can categorize users according to these needs. Users with similar needs for explanations are grouped together, and a persona will represent them. Using personas is seen as a solution to counter developer's bias according to Faily and Flechais [11]. When developers design something without the users in mind, the development process will be filled with assumptions and the developers' own ideals, which most likely will not align with what the actual users have in mind, hence the bias. With that in mind, personas can also be seen as a communication tool between users and developers, which could help developers design ideas more suitable for the users [31].

Furthermore, personas are written in a common language, which could enable easy understanding and communication between developers [26]. When personas have been successfully used in aiding developers raise requirements, and if the personas are well written, they can be reused for future projects. This is possible because personas are representations of user groups in the software's user base.

2.2.2 Challenges of Using Personas

Though personas may have many advantages, especially when used to evaluate NFRs, they are known to be very difficult to validate [25]. Personas are made up people, meaning there is no concrete validation method to measure the credibility of personas. They are purely subjective depending on what the persona designers and developers think.

To minimize the subjectivity of personas, Ramos et al. [22] created a list of criteria that may help developers in evaluating personas. They are defined as follows:

- **Credibility**: How realistic is the persona?
- **Consistency**: The information in the description is consistent.
- **Completeness**: The personas capture essential information about the described users.
- **Clarity**: The information is presented clearly.
- **Likability**: How nice the persona appears to be.
- **Empathy**: How much the evaluator empathizes with the persona.
- **Similarity**: How similar is the persona to the evaluator.
- **Willingness**: Measures the evaluator's willingness to learn more about the persona.

Ramos et al. [22] noted, that not all 8 criteria should be used at the same time, but instead just a few that fit the need of the specific research purposes. In line with the criteria set by Ramos and other writers, Salminen et al. [25] came up with 4 main criteria, which helped in narrowing down on which criteria this thesis should focus on:

- **Credibility**
- **Consistency**
- **Completeness**
- **Usefulness and Willingness to Use**

The problem with the first criterion is that in order to make a persona appear real, an extensive research must be done. According to Salminen et al. [25] providing only a number of qualitative interviews with users would not be able to convince development team members who are more quantitative data oriented, and the persona would only be seen as "nice narratives". Furthermore, to avoid it being "nice narratives", the persona designer might add their ideals and assumptions into the supposedly abstract persona, making it biased.

The second criterion speaks for itself. A persona would lose its value if the information contained within it does not add up [25]. The persona will not appear real, and developers would not be able to relate to it at all. This is comparable to a puzzle where all pieces do not fit with each other.

The third criterion is about completeness, meaning that not just any kind of information should be added into persona. The persona could have too much irrelevant information, which could distract the developers from the relevant features of the persona [25].

The fourth criterion is the usefulness of persona. As stated by Salminen et al. [25], there are some cases where persona had little to no impact during development process. The authors did not discuss this problem further, presumably the persona may have failed to meet the previous 3 criteria. Another disrupting factor is that the interpretation of personas always differs. Developers relate themselves with personas in many different ways, which has proven to be true in several heated discussions with the thesis supervisor during persona development.

To bridge the gap between developers, a table of measurement was developed by Salminen et al. [25] in order to help developers relate to persona better. This table was also used as a guideline in developing the personas for this thesis.

Clarity	Persona information is clearly presented.
Empathy	Personas are sympathized by respondents.
Familiarity	Personas remind the respondent of people they know.
Friendliness	Personas are perceived as friendly by respondents.
Interpersonal attraction	Personas are perceived as attractive by respondents.
Liking	Personas are liked by respondents.
Similarity	The respondents feel like the persona is like them.

Table 2.1: Table of measurement to help developers relate to personas [25]

Judging based on these findings, the application of personas in development stages may be beneficial in helping developers understand their user base. However, there is a number of challenges that must be considered during the development of personas. Not only they are difficult to evaluate and validate due to them being subjective, but they also come with the risks of them not being as useful in late development stages. Therefore, evaluation criteria and a relevant source of data are needed in order to minimize these risks. The first step towards developing a functioning persona is choosing the correct method of data collection through user studies. This will be discussed in the next section.

2.3 User Studies

As explained in the previous sections, personas are fictitious entities representing user groups. In order to obtain information from these user groups, a user study has to be conducted. After reviewing literatures with personas as the point of focus, the most effective method of collecting data from the users is combining both quantitative and qualitative research [1, 31]. In the following subsections, each type of data collection, their strengths and weaknesses will be explained.

2.3.1 Qualitative vs Quantitative Data Collection

There are two types of data commonly known, namely quantitative and qualitative data. The source of the data, however, can also be classified into two types: primary and secondary source of data [21]. According to Pruitt and Aldin [21], primary data is any data that can be collected through directly observing the users; secondary data is data that comes from a third party, or in other word, is indirectly obtained. Secondary data may not be as accurate as the primary data, because information based upon memory can be falsified, even if not on purpose. It is safe to say that the primary data source is more reliable and widely preferred, due to the direct interaction between researcher and respondent.

As for the types of data, quantitative data usually comes from a large number of respondents and collected through efficient methods such as surveys or questionnaires [21]. On the other hand, qualitative data includes a smaller number of respondents and collected through methods that promote deep understanding such as interviews [21]. Through interviews, more detailed answers can be achieved.

Applying only one of them when developing persona has been proven to not be effective, and it may negatively impact the personas' credibility. Though quantitative methods are fast and enable developers to collect a large amount of data, the collected information can be misleading [31]. Underlined by Adler [1], basing a persona only on quantitative data will make it wrongly seen as stereotype representing the developers' "assumptions and biases".

The same demerit applies when only qualitative data is used. Salminen et al. [25] pointed out that creating a persona based on only a number of qualitative interviews will make the persona less credible, as there is no concrete data backing it up that they usually end up as "nice narratives". Moreover, conducting more interviews to make up for the lack of data will consume more time and energy from the developers [31].

Since both approaches are not sufficient without the other, combining both quantitative and qualitative data [1, 31] enables developers to create convincing and reliable personas. The combination of both types of data can be collected through mixed method research.

2.3.2 Concurrent Mixed Method Research

There are two types of mixed method research according to Creswell et al. [10], sequential and concurrent mixed method research. The method used during the user study in this thesis is the concurrent mixed method. How they are structured can be observed in the figures below.

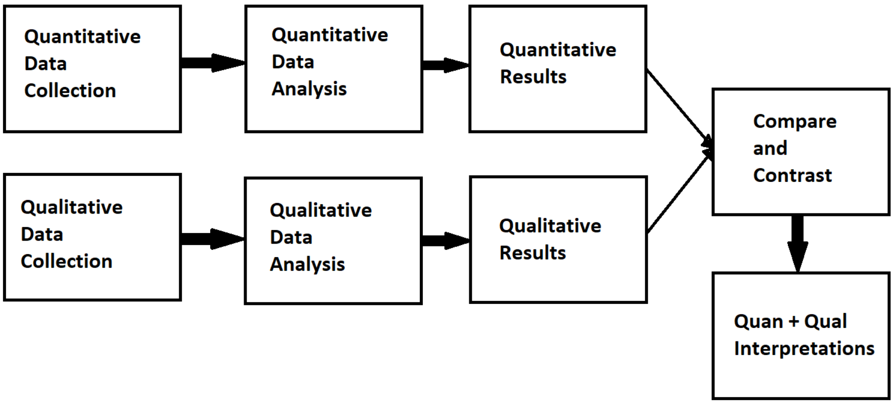

Quantitative Data Collection
Quantitative Data Analysis
Quantitative Results
Compare and Contrast
Qualitative Data Collection
Qualitative Data Analysis
Qualitative Results
Quan + Qual Interpretations

Figure 2.1: Concurrent mixed method research [10]

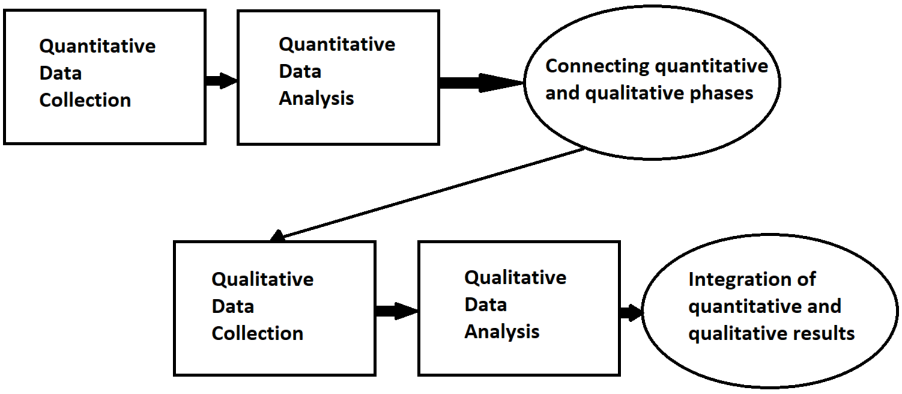

Quantitative Data Collection
Quantitative Data Analysis
Connecting quantitative and qualitative phases
Qualitative Data Collection
Qualitative Data Analysis
Integration of quantitative and qualitative results

Figure 2.2: Sequential mixed method research [17]

In comparison to sequential mixed method, concurrent mixed method takes shorter amount of time due to fact that the interview and survey are conducted at the same time. In sequential mixed method research [17], the quantitative data collection is conducted first. After that, before moving on to the qualitative data collection, a couple of respondents from the previous stage of data collection will be hand picked, and interview questions will be prepared for them. Then starts the qualitative data collection. In the end, both sets of data are integrated and interpreted.

Judging from the comparison above, and considering the limited time frame available during the making of this thesis, the concurrent mixed method research is more preferable and therefore used.

Chapter 3

Research Design

This chapter explains the concepts for the evaluations in chapter 4. The fundamentals are summed up into 2 questions containing the idea of using mixed method research and the application of developed personas on a real software. The details of survey questions and the design of the user study are explained in the following section and subsections.

3.1 Research Questions

As explained in chapter 2.2 and 2.3, applying personas in software development may provide some benefits in understanding the user base, thus increasing the overall quality of the software. Despite the positive impacts, the development of personas comes with a number of difficulties that needs to be taken into consideration in order to perfect the end results. Including the time frame available into the equation as well, the choice of user study method plays a significant role. In finding the answers to these problems, 2 research questions were defined:

RQ1: Is the mixed method research suitable for developing personas?

RQ2: Are the general explainability personas applicable for specific software?

The details for each question will be explained in the subsections below.

3.1.1 RQ1: Mixed Method Research

Mixed method research is conducted through combining 2 different methods of collecting data, as explained in chapter 2.3. The factors that need to be considered are the duration of the user study and the questions. The reason why concurrent mixed method research was used and how the examples for the questions were chosen will be explored in more detail later in this chapter.

3.1.2 RQ2: Application of Persona

Due to the characteristics of personas being fictitious, it is difficult to tell if a persona can be deemed as “real”, as personas might be perceived differently by persona users. In chapter 4.4, the validity and usability of personas will be tested to see if the personas created based on general explainability requirements from the users can be used to raise requirements for specific software. As an example for software that could need explanations, we will use the SEnti-Analyzer.

3.2 Study Design

This section will explore the design and the structuring of the user study conducted for the purpose of this bachelor’s thesis. The first subsection will briefly explain the platform and other tools used for the user study. The second subsection explains the question design as well as the reasoning behind the examples chosen and why certain questions are needed. The third subsection elaborates the process of collecting data.

3.2.1 Online Survey and Interview

The platform used to create the survey questions is LimeSurvey[1], provided by thesis supervisor. The survey questions also serve as a script for the structured interview.

3.2.2 Designing the Questions

There are 6 question groups in the survey, each with different themes. The themes are based on trend observations, personal experience, and discussions with peers from computer science as well.

The respondents are first asked to rate their experience with the features highlighted in each section. Before they reach the end of the section, respondents are asked if they need an explanation for the feature they use in their daily lives based on the examples set in the survey. The interviewees are given more questions for more detailed answers on why they need or don’t need an explanation. This is the highlight of each question group, and whether the respondents need or don’t need explanations will be essential for designing explainability personas.

Both, survey and interview respondents are given the option to leave further comments on the highlighted features at the end of each section of

[1] `https://survey.se.uni-hannover.de/index.php/admin/index`

the survey. This allows the survey respondents to give detailed answers that may support their previous answer choices, which later serve as supporting data, in addition to the interview answers.

Question Group 1: Content Recommendation

The content recommendation feature is one of the most encountered features in our daily lives. Through observations, it can be seen that recommendation features are used in applications such as social media, streaming sites, e-commerce sites, and many more. The reason why this theme was chosen as the first question group is that it may help the respondent to engage in the survey better if they start with something they are more familiar with, so that they can get a picture of what's to come in the next question groups.

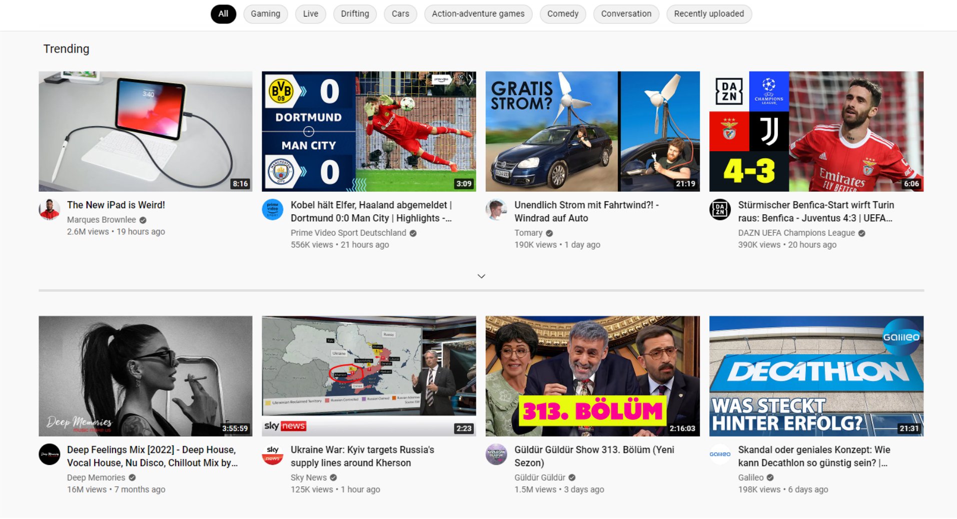

All
Gaming
Live
Drifting
Cars
Action-adventure games
Comedy
Conversation
Recently uploaded
Trending
8:16
The New iPad is Weird!
Marques Brownlee
2.6M views • 19 hours ago
DORTMUND
MAN CITY
0
0
3:09
Kobel hält Elfer, Haaland abgemeldet | Dortmund 0:0 Man City | Highlights -...
Prime Video Sport Deutschland
556K views • 21 hours ago
GRATIS STROM?
21:19
Unendlich Strom mit Fahrtwind?! - Windrad auf Auto
Tomary
190K views • 1 day ago
DAZN
4-3
6:06
Stürmischer Benfica-Start wirft Turin raus: Benfica - Juventus 4:3 | UEFA...
DAZN UEFA Champions League
390K views • 20 hours ago
Deep Memories
3:55:59
Deep Feelings Mix [2022] - Deep House, Vocal House, Nu Disco, Chillout Mix by...
Deep Memories
16M views • 7 months ago
sky news
2:23
Ukraine War: Kyiv targets Russia's supply lines around Kherson
Sky News
125K views • 1 hour ago
313. BÖLÜM
2:16:03
Güldür Güldür Show 313. Bölüm (Yeni Sezon)
Güldür Güldür
1.5M views • 3 days ago
Galileo
DECATHLON
WAS STECKT HINTER ERFOLG?
21:31
Skandal oder geniales Konzept: Wie kann Decathlon so günstig sein? |...
Galileo
198K views • 6 days ago

Figure 3.1: YouTube Homepage

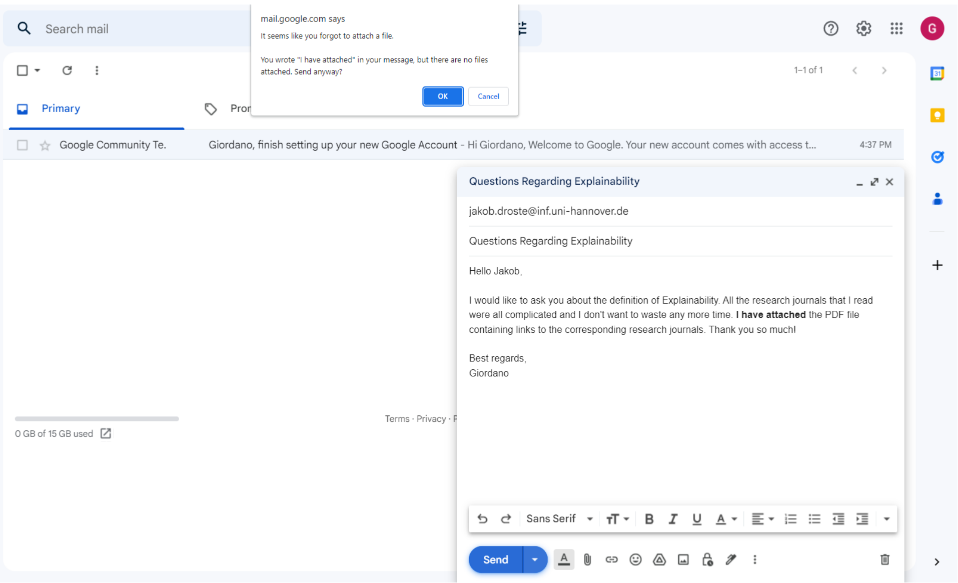
From many examples of websites or applications that use this feature, such as Instagram, Amazon, Facebook, etc., YouTube was chosen. Not everyone uses Instagram, even though it's popular. Instagram is targeted more at the younger respondents. The same with Facebook, it may be popular only to the older respondents. The survey is targeted at all respondents of all age groups. Amazon may not be available for respondents outside of EU. Therefore, YouTube is the most fitting choice of example.

YouTube is one of the most visited websites in the internet, and the biggest streaming platform among others like Twitch, Netflix, and many more. The first thing users see when opening YouTube is video recommendations on what is trending. Using YouTube as an example, respondents are then persuaded to relate their experience with other applications that apply the same recommendation feature.

Recommendation feature benefits both users and software companies. Users are able to easily browse their preferred contents, or even discover new contents. At the same time, company owners are able to know users' preferences and this could help them plan their marketing better. However, this is where the link is usually broken. Company owners or service providers often "hide" the process of managing the users data. This means that they purposefully do not explicitly explain to the users what they do with the collected data [19].

This may pose a threat to the trustworthiness of the application, and raise the question of users' privacy. Therefore, using YouTube as an example to ask respondents if they need an explanation or not plays a significant role in finding out whether users care about their privacy, and at the same time measuring their reliance on these types of applications.

Question Group 2: System Notification - Reminder



mail.google.com says

It seems like you forgot to attach a file.

You wrote "I have attached" in your message, but there are no files attached. Send anyway?

Google Community Te. Giordano, finish setting up your new Google Account - Hi Giordano, Welcome to Google. Your new account comes with access t... 4:37 PM

Questions Regarding Explainability

jakob.droste@inf.uni-hannover.de

Questions Regarding Explainability

Hello Jakob,

I would like to ask you about the definition of Explainability. All the research journals that I read were all complicated and I don't want to waste any more time. **I have attached** the PDF file containing links to the corresponding research journals. Thank you so much!

Best regards,
Giordano

Figure 3.2: Google Mail attachment reminder.

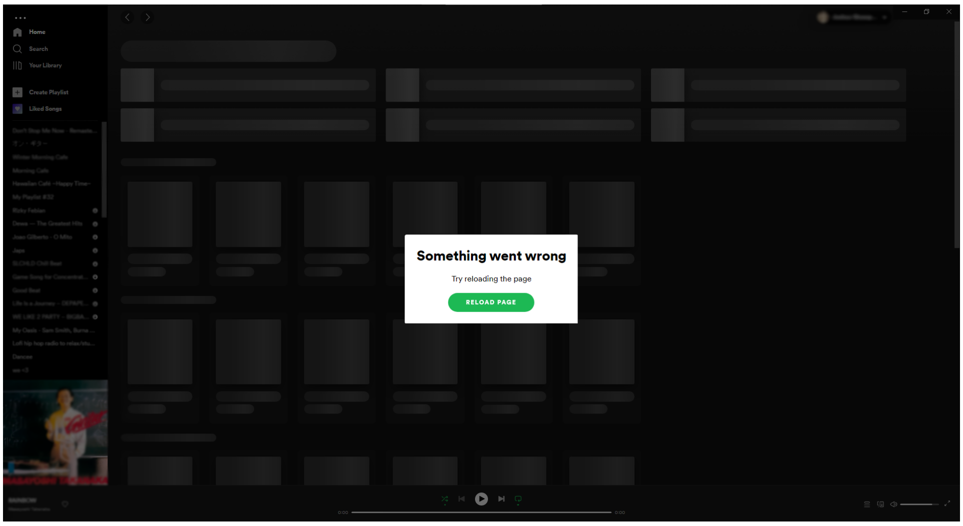
This section of the survey focuses on explanations given in the form of system notification. Explanations as a reminder are designed to help users avoid making mistakes. There are many examples that come to mind, such as reminders for software updates, or calendar reminders, etc. The example chosen for this section is an e-mail attachment reminder. This may seem like a specific example, compared to software update reminder example. However, electronic mailing services are also among the most used tools, whether it is for work or for education.

In users' daily lives, when they are about to write a formal email with attachments, the first thing they should do before sending the email is to double check if the attachment is there or not. This is usually the case because users don't want to waste time sending another email to fix the wrongly sent one, which could affect professionalism in working environment, or sometimes it could also cause embarrassment when the recipient is someone important.

However, oftentimes users are still vulnerable to making mistakes. To prevent this from happening, Google Mail developed a feature where a pop up notification will show up if triggered by the phrase "I have attached...", but no attachment is found in the email. Before the notification shows up, the user is prevented from sending the email. The notification acts as an explanation to the user that they forgot about the attachment, in the case that they are confused as to why they were not able to send the email.

This is a useful case of using explanation to maintain users' confidence in using Google's mailing service, and persuade them to make better decisions by giving them control.

Question Group 3: System Notification - Error Report



Something went wrong

Try reloading the page

Figure 3.3: Spotify, a music streaming application

The third section is the same theme as the previous, which is system notifications. This time it focuses on error reports. Based on observations and personal experience, the first reaction a user shows when encountering an error is usually confused, and curious of what might causing it. However, how users handle errors might differ. Some may be looking for the solution in the software, some may browse the internet for answers. Both cases are the same: the users are looking for explanations on how to solve the error. This is, however, just an assumption and may not apply to every type of users.

To test that, a hypothesis was made. If there exists a user who needs explanations, there has to be at least a user who doesn't need an explanation. That is the reason this question group was created using one of the popular music streaming applications, "Spotify", as an example. Spotify was chosen because the error notification was straightforward, and the chance that users may have encountered such errors in the application is appropriate to be taken into consideration.

Referring to figure 3.3, the error notification shown by Spotify is brief and doesn't include further information other than "Something went wrong". Referring back to subsection 2.1.1, the amount of information provided and its relevance is important, so that users are able to understand what is happening. Since it is difficult to measure what is the "right" amount of information needed as an explanation, it may be worth asking the respondents how they feel about these kinds of explanations and whether they need more or less information.

Question Group 4: Feature Transparency

Transparency means that the system is designed so that the user is made aware of how it behaves and operates. The example used in this section is a phone cleaner application. Phone cleaner applications, anti-virus programs, and other software alike are usually advertised to be able to speed up mobile devices, clean files off viruses, delete old and unused files to free up storage space and so on. At first glance, these types of applications may be deemed as convenient. However, some of them do not provide a clear explanation on the features they advertised.

Based on a casual conversation with a phone cleaner application user, they have had an experience where their important files were accidentally deleted. They explained that they didn't mean to delete the files, and just followed the suggestions given by the application.

In order to find out if other users have had the same problem or not, this question group was created. Firstly the respondents will be asked to rate their understanding of the phone cleaner application. Then the respondents are asked if they need more explanations, which could help them understand the application better. This question group is made with the assumption of the users having at least encountered or used such applications once.

This section of the survey tackles the transparency problem, which could

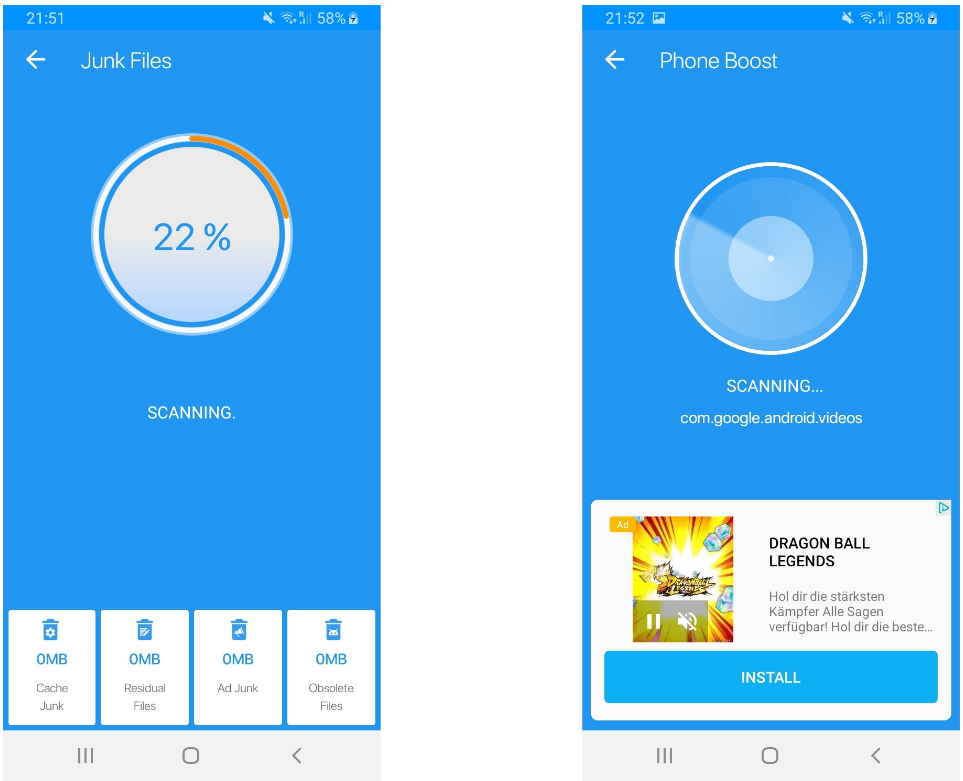
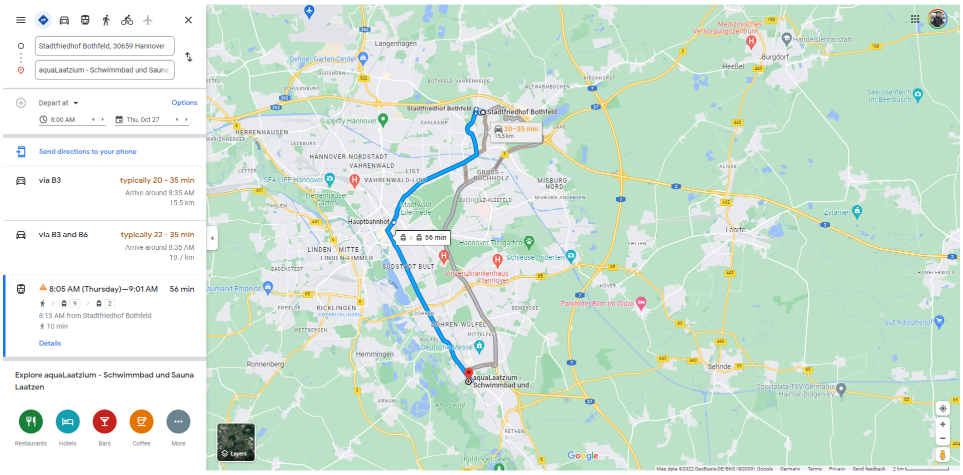


Figure 3.4: Phone cleaner application

affect the trustworthiness of said application and reduce the likeliness of users trying this application. In the worst case scenario, it could also enable users to create their own understanding of the application, which then leads to misuse or unwanted errors.

Question Group 5: Algorithm Explanation

The fifth section is about algorithm explanations. Complex algorithms are used in trip planner applications in order to calculate the shortest routes, provide users with information regarding traffic conditions, or explaining the users how to take certain routes. Trip planner applications have been one of the necessary tools used by users, whether they are used to plan train travels or flights, commuting to work, or even travelling by foot. This section is created using Google Maps as an example, because it is one of the most accessible and practical trip planner applications on the application market. The survey respondents will be asked if they need explanations in specific scenarios. The scenarios chosen will be the most encountered ones, such as delayed schedules and route instructions. The goal here is to investigate if



Stadtfriedhof Bothfeld, 30659 Hannover

aquaLaatzium - Schwimmbad und Sauna

Depart at ▾ Options

8:00 AM Thu, Oct 27

Send directions to your phone

via B3 typically 20 - 35 min
Arrive around 8:35 AM
15.5 km

via B3 and B6 typically 22 - 35 min
Arrive around 8:35 AM
19.7 km

⚠ 8:05 AM (Thursday)—9:01 AM 56 min
9 > 2
8:13 AM from Stadtfriedhof Bothfeld
10 min
Details

Explore aquaLaatzium - Schwimmbad und Sauna Laatzen

Figure 3.5: Google Maps

the users actually need the explanations provided by Google Maps to help them make better decisions when planning their trips.

Question Group 6: Demographic Questions

The demographic questions in this section serve as an additional biographic information in order to help developers visualize the personas in late development stage. Biographic backgrounds, according to Olsen [20], are important to humanize personas, as long as they don't divert our attention due to the amount of information provided. The main focus of the whole survey is whether users need explanations or not. To avoid inserting unrelated information that may hinder the persona development (refer to subsection 2.2.2), the questions are limited to age and occupation.

3.2.3 Conducting the Research

Mixed method research has been used in many persona developments. Referring to section 2.3, combining both questionnaire and interview has proven to be the most effective way to create a reliable persona. According to the comparison table made by Tu et al. [31], from 3 user study methods they used, conducting surveys delivers a great amount of data in the shortest time, but may not be precise and could be misleading. On the other hand, interviews deliver more accurate and detailed data from the respondents, but they consume more time and energy from the developers.

One advantage of using concurrent mixed method research over the sequential mixed method is the flexibility. The time frame available to conduct the user study for this bachelor thesis was 4 weeks. The amount

	Three User Study Methods		
	Survey	Interview	Observation
Data analysis method	Quantitative	Quantitative	Quantitative
Data gathering speed	Fast	Very slow	Slow
Data reliability	So-so	Very good	Good

Table 3.1: Comparison of three user study methods [31]

of data collected in total was 70 submissions, 10 of which were results from interviews. Based on figure 2.1, the survey was left running until the set deadline, and during this time, interviews with 10 respondents were conducted. This was done with the consideration of the interview respondents' availability.

If the user study was conducted using the sequential mixed method, it would be very difficult, especially in the given time frame of 4 weeks. According to figure 2.2, after the survey is conducted, there is a phase where the developer has to track back and select a few survey respondents to be interviewed. This is in contrary to the anonymity of the survey, and will take longer time to contact the respondents again and prepare them the interview questions. However, in the case where more time and freedom are given to research the user base, sequential mixed method may deliver high quality results as well.

After conducting the survey and interviews, the results were sorted and compared. After evaluating the data based on answer patterns and possible behavior similarities, they are then divided into user groups. This division, however, is still based on data observation and general assumptions. To prove if the division is fair and balanced, an agreement calculation was done. This will be further explained in chapter 4.1.2. The interview recordings are also reviewed again to clear possible misunderstandings, and also to make sure that there are no more assumptions in the data and that the answers are representing the user groups they belong to.

Using Fleiss' Kappa to Measure the Rate of Agreement

The inter-rater agreement coefficient used for this thesis is Fleiss' Kappa. After a number of attempts to calculate the agreement rate using Cohen's kappa [9], intraclass correlation coefficient [3], and Fleiss' kappa [12], Fleiss' kappa was the only one which showed decent results. Cohen's kappa was not suitable because it is only limited to two raters only, while there are more than 10 raters at least. Intraclass correlation was the next possible option, but the results were not as decent.

Fleiss' kappa is preferred and used because it allows multiple raters, and provide decent percentage as answers. However, this method is not without its shortcomings. The homogeneity of the survey results directly affected the Fleiss' kappa coefficient, which caused a paradox. This will be explained further in chapter 4.1.2.

Post-Persona-Development Interview

The personas are developed with the evaluation criteria mentioned in chapter 2.2 in mind to prevent developers' bias and to make the personas feel more human, while trying to ensure their usability in late development stages. However, general explainability personas are supposedly made from general explainability issues, which may or may not relate to a specific software. The SEnti-Analyzer is a specific software, which provides features that are mostly not presented as examples in the survey. By conducting the post-persona-development interview, which included using the SentiAnalyzer, with the same interviewees from the mixed method research stage, there is an opportunity to investigate if the personas developed can truly represent their corresponding user groups when faced with a specific software.

The post-development interview was a hands-on software demonstration, where the respondents were given control of the SEnti-Analyzer and had to follow the required tasks. The respondents were also allowed to freely explore the features of the software if they wished to do so.

To ensure comfort and flexibility during the hands-on demonstration sessions, the respondents were required to use TeamViewer[2], a remote desktop software, in order to access the interviewer's PC remotely. The goal of this interview was to observe the direct interactions between the respondents and the software.

[2] https://www.teamviewer.com/

Chapter 4

Evaluation

In this chapter, the results from the user study are explained in detail. The quantitative and qualitative results are presented to answer **RQ1**, and the persona experiments are used to answer **RQ2**, as stated in chapter 3.

4.1 Quantitative Data Analysis

First of all, the survey data are exported from LimeSurvey[1] into *.xls* files. The survey respondents reached a total of 70, 10 of which are interview respondents. The data collected were divided into 5 question groups, excluding the demographic questions section. From the questions in one section, the questions about the users' need for explanations were the main focus, while other questions were used as supporting argumentation for the initial assumptions.

Question Group 1: Content Recommendation

Question [CR4]: Do you feel the need for an explanation on why you are recommended the contents you see on your device?

CR4	Need explanations	Don't care	Don't need explanations
Answer count	45	12	13

Table 4.1: Answer count for Content Recommendation Feature.

The results from the first question group in section 3.2.2 indicate that the majority of the respondents needed explanations for content recommendation

[1] `https://survey.se.uni-hannover.de/index.php/admin/index`

feature. From the other 25 respondents, 12 did not really care if there were explanations or not, and 13 of them felt that no explanations are needed.

Question [CR2]: Are you aware why you are recommended the content you see on your device(s)?

CR2	Yes	No
Answer count	61	9

Table 4.2: Answer count for User Awareness.

Interestingly, a large number of respondents answered "yes" when they answered the question asking them if they know where the recommended contents come from. Our assumption was that most of the respondents knew because of the explanations given to them, and they still need explanations in order to regulate the contents they see.

Question [CR3]: If you are being recommended content you are not familiar with, how would you feel about it?

CR3	Like it	Don't care	Dislike it
Answer count	11	33	26

Table 4.3: Answer count for Unfamiliar Content Recommendation.

This was supported by the low number of respondents who answered that they "liked it" if they were recommended contents they are not familiar with, with total of 11 out of 70 answers only (refer to table 4.3). The amount of respondents who answered "don't care" and "dislike it" to the unfamiliar recommended contents question represents the majority.

Question Group 2: System Notification - Reminder

Question [SNR1]: How do you feel about receiving an explanation in situations like these?

Referring to the question in section 3.2.2, the majority of the respondents saw the benefit from the Google Mail attachment reminder feature. Since the difference between the other two answers are too big, an assumption was made that most of the users rely on reminder features and its explanations, as long as it helps them to avoid making mistakes.

SNR1	Need explanations	Don't care	Don't need explanations
Answer count	63	1	6

Table 4.4: Answer count for System Notification - Reminder.

There was one respondent who answered "don't need explanation", but in the next question they commented:

> *"I think it would be best if I can get reminded to put my attachment because sending email twice will be a big problem especially in a working environment, because it shows us how unprofessional we are for not really checking it again."*

This case was one of the concrete examples of how quantitative data can be misleading. Human errors could still happen in surveys and should be expected. So far only one respondent who presumably had answered this question incorrectly by mistake.

Question Group 3: System Notification - Error Report

Question [SNER1]: Do you feel the need for a more thorough and informative explanation?

SNER1	Need explanations	Don't care	Don't need explanations
Answer count	58	4	8

Table 4.5: Answer count for System Notification - Error Report.

Most of the respondents answered that they need informative explanations when an error message shows up, as shown in section 3.2.2. In line with the explanation from chapter 2.1, and using figure 3.3 as reference, it can be observed that users need informative explanations so that they can understand the cause of the errors. In the survey, the respondents were given the option to write comments concerning the error report feature as well. Quoting from the survey results, here are some of the answers written by the respondents:

- *"Better explanation makes me feel that the provider gives more attention to its user."*
- *"It would be better if I can get a information what causes the bug because if not we will never know if it is serious problem to our PC or laptop."*

- *"Adding an error code would be nice so we can look it up online easier. And it can also help user understand how others solve the problem."*

From some of the answers the respondents had written, we can assume that for them, explanations enable them to locate sources of error, promote learnability, and increase the feeling of trustworthiness.

Question Group 4: Feature Transparency

Question [FT2]: Do you feel like you need more informative explanation when using these types of applications?

FT2	Need explanations	Don't care	Don't need explanations
Answer count	60	7	3

Table 4.6: Answer count for Feature Transparency.

Based on these results, most of the respondents were concerned when they are faced with a complex application. The example used here is a phone cleaner application, which is shown in figure 3.4. This was also due to the fact that the majority of the respondents answered that they didn't really understand how these applications work in the previous question. The first assumption was that the users need explanations because they didn't know how the application works.

Question [FT1]: I am aware of how the application works, what it does to my device, and which files are deleted/cleaned.

FT1	Agree	Neutral	Disagree
Answer count	22	12	36

Table 4.7: Answer count for User Awareness.

However, the number of respondents who agreed, which was 22, is not that far of a difference with 36 of those who disagreed. The assumption was then changed, in order to avoid misleading results. To complete the new assumption, we reviewed results from question FT2 again. From 22 respondents, 15 answered that they still need explanations when using these types of application. Referring back to chapter 2.1, balanced reliance between the user and the system was mentioned. The assumption we arrived at in the end was that even though these 15 respondents are familiar with such complex applications, they still want to avoid self-reliance. They need to make sure they are making the right decisions, and they want the application to support them by giving them explanations. This assumption applies to the respondents who didn't really understand how these types of application work as well. The explanations they needed might aid them in learning the boundaries of the application, so that they know what to do and not to do.

Question Group 5: Algorithm Explanations

Question [SQ002]: If there are traffic jams or delayed schedules, I want the application to provide a relevant explanation concerning the problem.

SQ002	Agree	Neutral	Disagree
Answer count	64	4	2

Table 4.8: Answer count for Algorithm Explanations (1).

The results from this question group show the reliability of most trip planner applications, judging by the high number of respondents who answered that they needed explanations. The example used was Google Maps (figure 3.5). For the first scenario, it can be assumed that most of the respondents needed explanations so that they can plan their trip better.

Question [SQ003]: I want the application to provide an elaborate explanation on each route, for example which turns or which exits should I take, transit stations, etc.

SQ003	Agree	Neutral	Disagree
Answer count	59	8	3

Table 4.9: Answer count for Algorithm Explanations (2).

For the second scenario, the number of respondents who either didn't care or disagreed was slightly higher than the answers from the first scenario. Though the number of those who agreed is high, it can also be assumed that there are some respondents who didn't usually utilize these kinds of features, hence the "neutral" and "disagree" answers. However, upon further investigation, one of the respondents wrote a comment concerning the second scenario. They answered "neutral", but commented:

> *"Depends on what 'elaborate' means. Especially when driving, I want to be able to perceive the information quickly, so I can keep paying attention to the road - hence, the information has to be informative but also precise and short."*

The respondent neither agreed nor disagreed with the scenario, because they believed that it should depend on the amount of information provided by the application to the user. Referring back to chapter 2.1, some users might prefer precise and also concise explanations, so that they wouldn't be burdened with a lot of information.

4.1.1 Cluster Analysis

Cluster analysis was done through quantitative data observation. The goal is to divide the respondents into groups with similar goals and behaviors [31]. Since the focus of this thesis is to investigate whether users need explanations or not, only the results from questions **CR4**, **SNR1**, **SNER1**, **FT2**, and **SQ002** as well as **SQ003** are needed.

According to the quantitative data analysis, the results from question **CR4** (table 4.1) was deemed to be the suitable indicator to divide the respondents into 3 groups, because of the wide variety between the answers. After the respondents are divided into groups, they were analysed again to see if there are some respondents that may belong to other groups.

Group 1	Group 2	Group 3
44	13	13

Table 4.10: Divided user groups.

Group 1 represents the respondents who needed explanations, group 2 represents those who don't care, and group 3 consists of respondents who didn't need explanations. These data, however, didn't represent the final result, and needed to be evaluated further using an inter-rater agreement coefficient. This will be explained further in the next subsection.

4.1.2 Fleiss' Kappa

After the division of respondents into 3 user groups, we still needed to evaluate whether the behaviors of the respondents correspond with each other. This is called the *rate of agreement*. Each respondents may have answered the questions differently, thus resulting in different answer patterns. However, by using Fleiss' Kappa coefficient, we can find out how far the various answer patterns from the respondents agree with each other.

Formula

Fleiss' formula (1971) [12] is actually an extended version from Scott's pi coefficient (1955) [27]. Scott's pi coefficient was developed in order to calculate the inter-rater reliability between 2 raters, while Fleiss' kappa coefficient can be used when there are multiple raters.

	$A = 1$	$A = 0$	Total
$B = 1$	a	b	g_1
$B = 0$	c	d	g_2
Total	f_1	f_2	n

Table 4.11: Example data with two raters. [16]

For this example, the formula from Scott [27] would look like this:

$$p_c = \left(\frac{(f_1 + g_1)/2}{n}\right) + \left(\frac{(f_2 + g_2)/2}{n}\right)$$

Percentage of chance probability or chance agreement is symbolised by p_c. Chance agreement is simply the probability of agreement occurring by chance. To calculate kappa (k), we need to calculate the observed agreement, which is p_o. When all three are put together, the formula for kappa is as follow:

$$k = \frac{p_o - p_c}{1 - p_c}$$

Derived from Scott's formula for p_c, Fleiss' formula [12] can be defined as:

Step 1:

$$q_j = \frac{1}{nm} \sum_{i=1}^{n} x_{ij}$$

Step 2:

$$p_c = \sum_{j=1}^{k} q_j^2$$

And the formula for p_o is defined as:

$$p_o = \frac{1}{mn(m-1)} \left[\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}^2 - mn \right]$$

To make it clear which variables corresponds to which values, the table layout below is used to calculate the Fleiss' kappa coefficient using the data from the survey:

Question	5	4	3	2	1	Total
CR4	x_{11}	...	...	...	x_{1j}	m
SNR1	...	...	...	...	...	m
SNER1	...	...	...	...	...	m
FT2	...	...	...	...	...	m
SQ002	...	...	...	...	...	m
SQ003	x_{i1}	...	...	...	x_{ij}	m
Total	$\sum_{i=1}^{n} x_{ij}$	...	...	...	...	nm

Table 4.12: Table Layout

- The columns represent the scaling of the answer choices. From "Strongly agree" or "Insist on explanations" as 5, to "Strongly disagree" or "Insist on not having explanations" as 1.
- The rows represent the questions, indicated by the question codes.
- The x represents the amount of respondents who chose the answers labeled in the column.
- m is the number of respondents in the group.
- n is the number of questions, which in this case, 6.

The results from each group will be presented using the template layout in the following subsection. Additional graphs for visual representations of the data are provided in appendix A.

4.1.3 Fleiss' Kappa Application to Survey Data

User Group 1

In the first group, we have total of 45 respondents. Upon observation, it can be seen that most of the respondents answered that they need explanations on everything. However, after reading the comments written by the survey respondents, their opinions were mixed between needing explanations for personal satisfaction or out of curiosity, and for privacy concerns. The assumption made here is that some portions of the respondents need explanations because it could not only help them, but other users as well, and increase the overall quality of the software. The other portions are concerned about their own experience. Table 4.13 provides the answers given by the respondents.

Question	5	4	3	2	1	Total
CR4	5	39	0	0	0	44
SNR1	16	22	1	3	2	44
SNER1	12	25	2	3	2	44
FT2	9	31	3	1	0	44
SQ002	16	25	2	1	0	44
SQ003	14	26	2	2	0	44
Total	72	168	10	10	4	264

Table 4.13: User Group 1 Results.

Using the formulas to calculate p_c, p_o, and k, we calculated the rate of agreement as follows:

- $p_c = 0.48243$
- $p_o = 0.49682$
- $k = 0.027$ or 2.7%

The kappa coefficient for this group is unexpectedly low, despite the high observed agreement percentage. It can be seen that in total, the answer "4" has been chosen in this data group 168 times, while "5" has been chosen 72 times. These numbers alone implied that there should have been a great percentage of agreement. However, this wasn't the case, and the reason why will be discussed in section 4.1.4.

User Group 2

In the second group, we have total of 13 respondents. There were not as many comments as the first group, especially on question CR4. Looking at the survey data alone, we assumed that most of the respondents in this user group are not that concerned about their experience when using applications with recommendation features. Instead, they are more focused on applications and features that may aid them in their daily activities, hence the high frequency of "need explanations" answers in other questions except CR4. Table 4.14 provides the answers given by the respondents.

Using the formulas to calculate p_c, p_o, and k, we calculated the rate of agreement as follows:

- $p_c = 0.38297$
- $p_o = 0.52777$
- $k = 0.23468$ or 23.47%

Question	5	4	3	2	1	Total
CR4	0	1	12	0	0	13
SNR1	2	11	0	0	0	13
SNER1	2	9	2	0	0	13
FT2	3	8	1	1	0	13
SQ002	5	6	2	0	0	13
SQ003	6	6	1	0	0	13
Total	18	41	18	1	0	78

Table 4.14: User Group 2 Results.

The kappa value in this user group appears to be much higher than the kappa value from the group 1 (figure 4.13). Compared to the first user group, the p_c for second group is lower, while p_o is higher. In other words, the chance agreement probability is lower due to the spread out data in the group, and the observed agreement probability is higher, which resulted in higher kappa value. However still, because of the high frequency of the answer “4”, 23.47% agreement rate was the best we could achieve, despite it being only a “fair” rate of agreement.

User Group 3

The third group consists of 13 respondents. A comment for question CR4 written by a survey respondent claiming that they are always aware of how the recommendation system works, therefore they didn’t need any explanations, gave the impression that the respondents in this group are confident in the way they use applications in their daily lives. However, the answers for questions other than CR4 indicated that they still need explanations when the situation is important for them. Table 4.15 provides the answers given by the respondents.

Question	5	4	3	2	1	Total
CR4	0	0	0	12	1	13
SNR1	2	10	0	0	1	13
SNER1	3	7	0	2	1	13
FT2	4	3	5	1	0	13
SQ002	2	10	0	0	1	13
SQ003	2	5	5	1	0	13
Total	13	37	8	16	4	78

Table 4.15: User Group 3 Results.

Using the formulas to calculate p_c, p_o, and k, we calculated the rate of agreement as follows:

- $p_c = 0.30802$
- $p_o = 0.47649$
- $k = 0.24346$ or 24.35%

Similar to user group 2, the data in the third user group is well spread out. Due to the high number of respondents answering “4”, we could only achieve a “fair” agreement rate, which was 24.35%.

4.1.4 Fleiss’ Kappa Paradox

As explained in subsection 4.1.2, Fleiss’ kappa coefficient is meant to look for agreement in a data set. The anomaly we encountered in the first user group is that high number of the same answers does not guarantee that the agreement rate would be high as well.

Based on the basic understanding of how kappa works, the more the data agree with each other, the higher the rate of agreement is [9]. However, when it comes to Fleiss’ kappa, which involves multiple raters at the same time, it caused a paradox. Data homogeneity affects the percent chance agreement. If more similar answers is found in the data, p_c will be higher. When the data is highly homogeneous, p_c will be high, but at the same time p_o will not be able to get any higher. In the case where p_c value approaches close enough to p_o, the kappa k value will be low [16].

The calculation in the first user group exemplifies this issue (table 4.13). It can be observed that there is not much difference between p_c and p_o in the first group, compared to that of the second (table 4.14) and third (table 4.15) user groups.

It is unfortunate that not much could be done to raise the kappa value any higher than 2.7%. The results from the survey data cannot be controlled or manipulated. The best result achieved from user group 1 is a kappa value of 4.23%. That was only achievable by excluding survey answers that are highly similar with one another, thus increasing the answer diversity. From 45 respondents, we had to cut down to 19 only.

Using the formulas to calculate p_c, p_o, and k, we get:

- $p_c = 0.44736$
- $p_o = 0.47076$
- $k = 0.04232$ or 4.23%

In this case, the p_c value is very close to p_o as well. Although the kappa value is higher because of the reduced homogeneity, excluding so many respondents’ data was not what we aimed for, especially when the excluded data was relevant. The end result was also not very satisfactory, as “fair”

Question	5	4	3	2	1	Total
CR4	4	15	0	0	0	19
SNR1	10	7	1	1	0	19
SNER1	7	12	0	0	0	19
FT2	5	14	0	0	0	19
SQ002	10	6	2	1	0	19
SQ003	8	8	2	1	0	19
Total	44	62	5	3	0	114

Table 4.16: User Group 1 New Results.

agreement rate could only be achieved if the kappa value is 20% at the very least. By the end of the user study, we decided to use the original data from table 4.13.

4.2 Qualitative Data Analysis

This section focuses on reviewing the interview results and how they can be used to either counter or support any assumptions that have been made during the quantitative data analysis. These are the IDs of the 10 interview respondents and the groups they belonged to before the review:

	Group 1	Group 2	Group 3
ID	11, 12, 16, 45, 104, 108	13, 44	107, 111

Table 4.17: Interview respondents and their corresponding groups.

4.2.1 User Group 1 Interview Review

In the first group, there are 6 interview respondents whose answers based on the survey data, fit in the category. After reviewing all the recordings from the interview sessions, similarities in how the respondents explained their answers were found. They were categorized into 2 answer types:

Positive	Negative
11, 45, 104, 108	12, 16

Table 4.18: Interview respondents sorted after recording review.

The interviewees in the positive groups have similar ways in explaining their answers. How they reacted when given bad examples of explanations, and how they explained what can be improved instead, gave a positive impression. On the other hand, interviewees in the negative group expressed their answers in a negative light. For example, interviewee 12 and 16 explained that they need explanations in content recommendation features because they are worried about their privacy.

After this stage, the interviewees from the positive group will be used to create a persona with a character trait "Critical Thinker". The negative group will be used to create a persona with a character trait "Overthinker". Both characters have a very different way of thinking, but they have a similar need for explanations.

4.2.2 User Group 2 Interview Review

The second group consists of interviewees 13 and 44, and the rest are survey respondents. According to their answers in question CR4, interviewee 13 answered that they did not care if there is an explanation or not for recommendation features. They explained that they don't really feel that it is important to know, but nevertheless it is something that is nice to have as long as they are not wordy and understandable. Interviewee 44 answered that they did need explanations, but then clarified that actually they don't care that much. Similar to interviewee 13, interviewee 44 stated that a little bit of explanation is better than no explanations at all.

Another similarity between the two of the respondents is that in some cases, they relied on themselves, rather than explanations for most of their application usage. Interviewee 13 explained that they don't need further explanations when the application already showed them what they need to know in question FT2. They referred to the on-screen system log loading screen shown by the application in the example figure 3.4. As for interviewee 44, they explained for their answer in question SQ003, that they didn't really care for explanations when using trip planner applications for detailed navigation, because they have a habit of relying on memory instead of the application when travelling through familiar areas. In other cases, they still need explanations from the application if they are in an unfamiliar area.

After reviewing both of the respondents, a character with the trait "Utilitarian" was created to represent this group, because of the way they use applications to fulfill their goals. They only need explanations when they need to make important decisions.

4.2.3 User Group 3 Interview Review

The interview respondents in the third user group are interviewee 107 and 111. Similarities found during the recording reviews between both of the respondents are their confidence in using content recommendation features to fit their needs, and how much information they needed for it to become a good explanation. In question SNER1, interviewee 107 explained that they don't need further informative explanations if the application can just solve the problem automatically before notifying the users. Using figure 3.3 as their example, if the error was about connection issues, the software should just reload itself rather than notifying the user with explanations that don't actually help them. Interviewee 107 also explained their answer in question SQ002 that they do not need more explanations on the cause of traffic conditions or delayed schedule. They stated that they only need to know if they are able to commute or not.

Interviewee 111 also explained their answer in question SQ003 that it depends on what "elaborate" means in the question. They emphasized that elaborate explanations may be distracting for users who are already occupied with other activities such as driving. They added that it would be best if the explanations are short but direct so there wouldn't be information overload. From these findings, a character with the trait "Minimalist" was created to represent this user group.

4.2.4 Summary of Findings

To summarize the findings so far, we need to connect the results from the interview review with the survey data. The first user group (table 4.13) consists of respondents who in general need explanations for their daily application use. After conducting the interview reviews, we are able to understand that some of the respondents need explanations because they are curious and think it could help the users and also increase the overall quality of the software. Meanwhile, other respondents expressed that they need explanations because otherwise they wouldn't be able to trust the developers' decisions, which leads them to being concerned about their privacy and comfort.

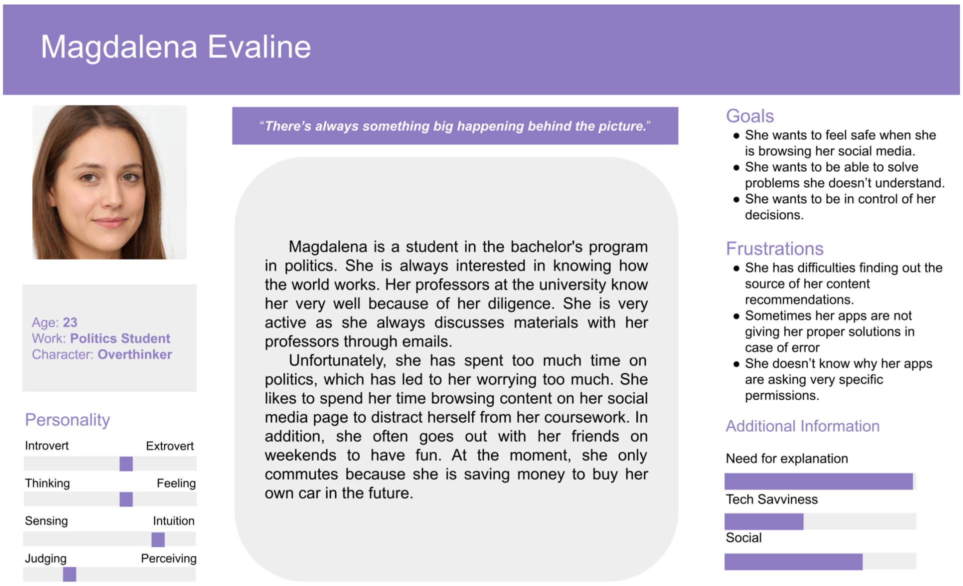
The second user group (table 4.14) consists of respondents who also need explanations. However, it is noticeable that almost all of the respondents in this group didn't care for explanations when they are about content recommendations. When the interviewees were asked why they thought that way, they explained that they simply enjoy the contents as they were. It can be implied that they are more focused on other software features that might actually benefit them. This has been proven by the large frequency of the answers "4" and "5". Another noteworthy testimony from one of the interviewees is that they were showing a balanced user-system-reliance when they are using a trip planner application. This supports the assumption that the users in this particular group are more focused on using the application for their own benefits like email reminder, phone manager application, and important system notifications, rather than entertainment.

Similar to the second user group, the third user group (table 4.15) consists of respondents who need explanations in their daily application use. In this user group, most of the respondents answered that they didn't need explanations for content recommendation features. The early assumption is that they are already aware of the inner workings of content recommendation features. This was then confirmed after reviewing the interview recordings of interviewee 107 and 111, as well as written comment from respondent 10. They showed confidence and self-reliance. However, when they were given examples of features that may help them in important scenarios such as email reminder, data management software, error notifications, and navigation software, they actually needed explanations. The differentiating factor is that the interviewees explained that while they do need explanations, it is important to them that it has to be brief and to the point that the users wouldn't be burdened with too much information.

4.3 Personas

This section will showcase all the personas developed using the data from both survey and interviews. Images of the personas and their biographic data will be shown, as well as the details concerning their daily application use related to the examples from the survey.

4.3.1 Overthinker Persona: Magdalena Evaline



Magdalena Evaline

"There's always something big happening behind the picture."

Age: **23**
Work: **Politics Student**
Character: **Overthinker**

Personality

Introvert Extrovert

Thinking Feeling

Sensing Intuition

Judging Perceiving

Magdalena is a student in the bachelor's program in politics. She is always interested in knowing how the world works. Her professors at the university know her very well because of her diligence. She is very active as she always discusses materials with her professors through emails.

Unfortunately, she has spent too much time on politics, which has led to her worrying too much. She likes to spend her time browsing content on her social media page to distract herself from her coursework. In addition, she often goes out with her friends on weekends to have fun. At the moment, she only commutes because she is saving money to buy her own car in the future.

Goals

- She wants to feel safe when she is browsing her social media.
- She wants to be able to solve problems she doesn't understand.
- She wants to be in control of her decisions.

Frustrations

- She has difficulties finding out the source of her content recommendations.
- Sometimes her apps are not giving her proper solutions in case of error
- She doesn't know why her apps are asking very specific permissions.

Additional Information

Need for explanation

Tech Savviness

Social

Figure 4.1: Persona Magdalena Evaline

Details on Magdalena Evaline

Magdalena often spends her time on social media to pass her free time. While she does enjoy the contents she thinks are appropriate, she has this constant fear that the developers are recording her activities and selling her data among the advertising companies. Her fear sometimes escalated when she encounters recommended content that she had never interacted with before in the application, but had seen on others. This leads to her believing that explanations are essential so that the users are not left in the dark with how the data are manipulated inside the app.

For every application or software she uses, she has to know at the very least how the system came up with certain outputs or decisions. She hates it when an application asks her for permission, but doesn't explain to her what it needs the permission for. The same applies when an error occurred in one of her apps, and the system just tells her to try again later. The lack of explanations in such situations makes her think about how the developers and company owners of the app see her and other users.

However, she doesn't completely think that all developers are evil masterminds. She actually benefits from the features in her mailing software, where she is reminded in case she forgot to attach a file to an email. She often sends formal emails to her professors and lecturers. For her, the reminder feature could prevent users from making mistakes, especially when it explains to the user what they may have forgotten. She also thinks that trip planner apps provide a good amount of explanations concerning traffic conditions, which help her move around in her busy schedule at the university.

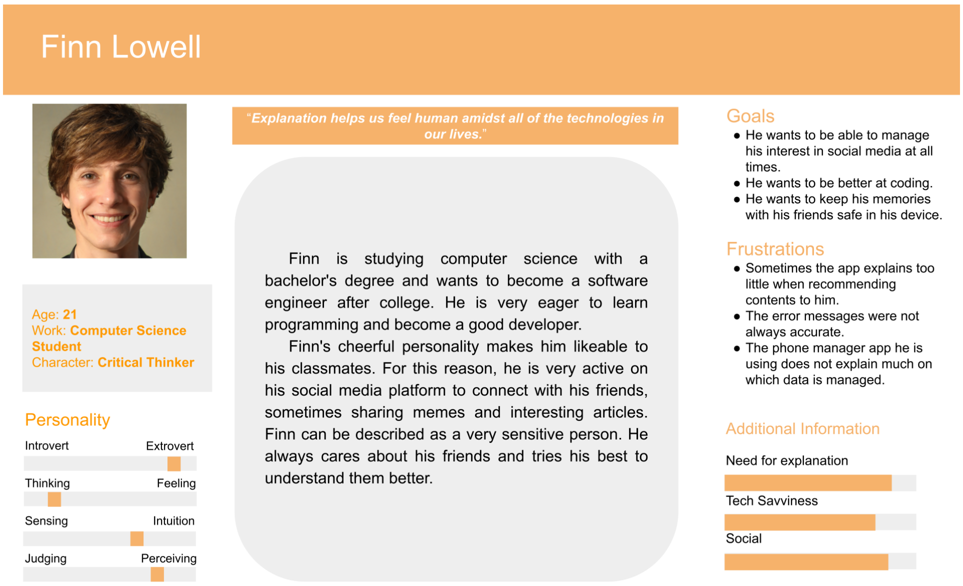
Magdalena believes that even though explanations are good to increase the user experience, the most important thing is that explanations could increase the trust between users and developers and product owners.

4.3.2 Critical Thinker Persona: Finn Lowell

Details on Finn Lowell

Finn likes to spend his free time catching up with his friends and trends in social media. He always makes sure that the contents he sees fit to his interest. He is able to do so because of the explanation provided by the app as to why he is recommended the content he sees. However, Finn claims that he is always open to exploring new contents, as long as they are appropriate and he knows why he gets such recommendations. The same could be applied to one of his hobbies, watching drama series online.

Unfortunately, because of the assignments he got from the university, he began to rarely finish the series he likes. Most of the time, Finn forgets where he left off. Thanks to the reminder feature provided by the streaming software, Finn is able to catch up with ease. The reminder explained to Finn which episode he had previously watched at which timestamp. He also likes



Finn Lowell

"Explanation helps us feel human amidst all of the technologies in our lives."

Age: **21**
Work: **Computer Science Student**
Character: **Critical Thinker**

Finn is studying computer science with a bachelor's degree and wants to become a software engineer after college. He is very eager to learn programming and become a good developer.

Finn's cheerful personality makes him likeable to his classmates. For this reason, he is very active on his social media platform to connect with his friends, sometimes sharing memes and interesting articles. Finn can be described as a very sensitive person. He always cares about his friends and tries his best to understand them better.

Goals

- He wants to be able to manage his interest in social media at all times.
- He wants to be better at coding.
- He wants to keep his memories with his friends safe in his device.

Frustrations

- Sometimes the app explains too little when recommending contents to him.
- The error messages were not always accurate.
- The phone manager app he is using does not explain much on which data is managed.

Figure 4.2: Persona Finn Lowell

it when the code editor he always uses reports an error in his syntax directly to the source of the error (e.g. which line contains the error).

Finn also benefits from trip planner applications, because it helps him commute to his classes on time or when he is meeting his friends. He doesn't like to be tardy, and the explanation about the traffic conditions helps him plan better. When Finn is taking a holiday trip, he also benefits from the instructions given by the app on which stations he should transit at. That way, he could go anywhere without feeling lost.

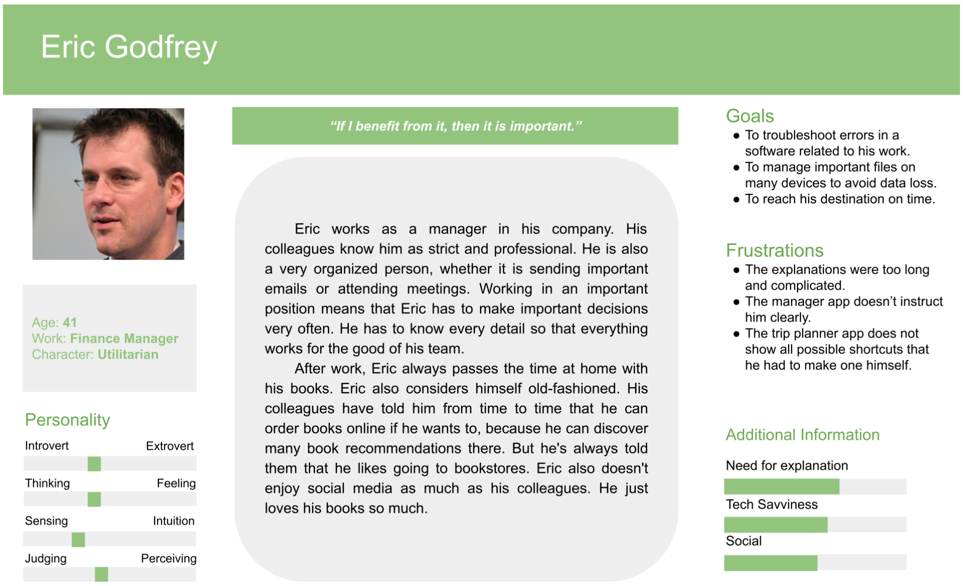
Overall, Finn believes that a well informed explanation is important, because it is fair for the users and helps them feel more confident when making decisions.

4.3.3 Utilitarian Persona: Eric Godfrey

Details on Eric Godfrey

Eric is a busy worker. He has to send countless emails everyday to his coworkers and clients. Dealing with so many important recipients leaves him no room for mistakes. That's when he fully utilized the reminder feature in his mailing software. He always goes with the similar wording template to trigger the reminder, in case he made human mistakes such as forgetting an attachment. When the reminder is triggered, he then proceeds to attach the forgotten file. This has helped him handle such matters professionally.

For other applications outside of his mailing software, he uses a phone



Eric Godfrey

"If I benefit from it, then it is important."

Age: **41**
Work: **Finance Manager**
Character: **Utilitarian**

Personality

Introvert – Extrovert

Thinking – Feeling

Sensing – Intuition

Judging – Perceiving

Eric works as a manager in his company. His colleagues know him as strict and professional. He is also a very organized person, whether it is sending important emails or attending meetings. Working in an important position means that Eric has to make important decisions very often. He has to know every detail so that everything works for the good of his team.

After work, Eric always passes the time at home with his books. Eric also considers himself old-fashioned. His colleagues have told him from time to time that he can order books online if he wants to, because he can discover many book recommendations there. But he's always told them that he likes going to bookstores. Eric also doesn't enjoy social media as much as his colleagues. He just loves his books so much.

Goals

- To troubleshoot errors in a software related to his work.
- To manage important files on many devices to avoid data loss.
- To reach his destination on time.

Frustrations

- The explanations were too long and complicated.
- The manager app doesn't instruct him clearly.
- The trip planner app does not show all possible shortcuts that he had to make one himself.

Additional Information

Need for explanation

Tech Savviness

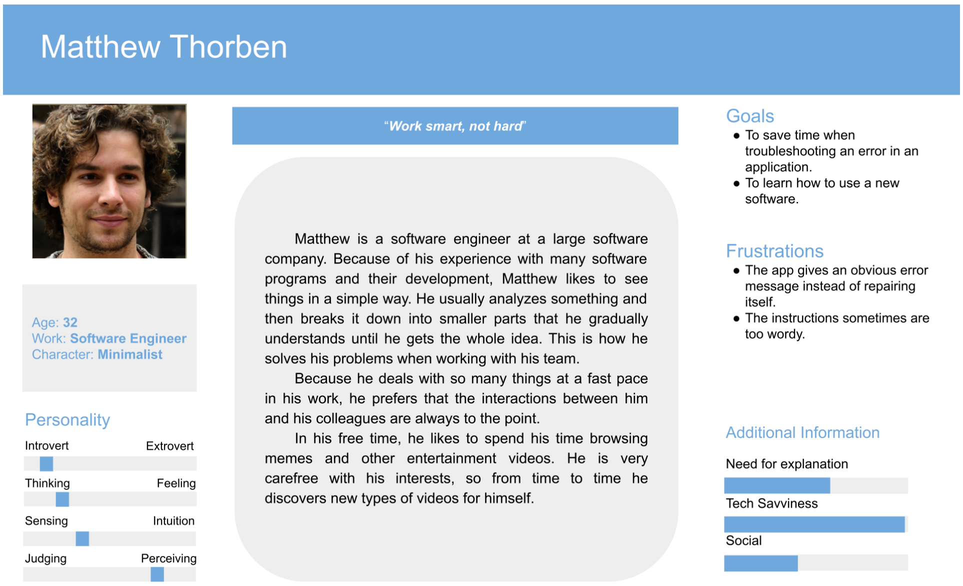
Social

Figure 4.3: Persona Eric Godfrey

manager application, to keep his phone tidy from junk files and to keep it running smoothly. Although he trusts the decision made by the app, he still needs some explanations on which files are being deleted, because there are also important files saved on his phone.

He uses a trip planner application on a daily basis, because he needs to avoid traffic jams on his way to work, or delayed schedules when using public transport. Though he relies on the app like normal people would, Eric sometimes chooses to do things his way. If he is in a familiar area and he believes that his known shortcut is more reliable than the app, he will choose his shortcut any day.

Modern entertainment is never his main focus. Eric may check his social media or some streaming sites if he is done reading his books for the day. He never really cares about the recommender system used by apps or sites, as to why he got the content he sees. He just enjoys them as they are, and when the time is up, he's back to work again. For him, any kind of explanation is fine, as long as they are understandable and not too wordy. That is much better than no explanations at all. He believes that an explanation could still help him in important times.



Matthew Thorben

"*Work smart, not hard*"

Age: **32**
Work: **Software Engineer**
Character: **Minimalist**

Matthew is a software engineer at a large software company. Because of his experience with many software programs and their development, Matthew likes to see things in a simple way. He usually analyzes something and then breaks it down into smaller parts that he gradually understands until he gets the whole idea. This is how he solves his problems when working with his team.

Because he deals with so many things at a fast pace in his work, he prefers that the interactions between him and his colleagues are always to the point.

In his free time, he likes to spend his time browsing memes and other entertainment videos. He is very carefree with his interests, so from time to time he discovers new types of videos for himself.

Goals

- To save time when troubleshooting an error in an application.
- To learn how to use a new software.

Frustrations

- The app gives an obvious error message instead of repairing itself.
- The instructions sometimes are too wordy.

Figure 4.4: Persona Matthew Thorben

4.3.4 Minimalist Persona: Matthew Thorben

Details on Matthew Thorben

Matthew works with online repositories on a daily basis. Managing these repositories with his team means that they have to be careful about conflicting file changes. Matthew and his team are able to prevent unwanted file losses because of the merge reminder feature of the repository software. Matthew also likes it when the reminder is brief and gets to the point. When he encounters an error while testing a code, he always relies on the error log, because it tells him straight away where the problem is. He prefers it that way with other apps as well. He believes that it would be faster to just show the stop or error code, and then let the users learn what they mean by looking it up themselves.

Matthew also has this concern regarding over-reliance on technology. He thinks that every "reliable" software or application on the market has its "side effects". Unsuspecting users will not be aware of these side effects, and when something unexpected happens, it would be too late. That's why he believes that informative explanation may still be useful for most people.

Matthew uses a trip planner app on his way to work to avoid traffic jams. Because Matthew always rushes to work and doesn't have time to plan his trip too long, he likes that the application is just showing him the overall trip duration, sometimes with a warning sign if there's a traffic jam. He couldn't care less what may have caused the traffic jam. However, he is glad that the user is still given an option to expand the warning sign to get the detailed report.

Matthew is confident enough in his ways with technologies, especially his social media, where he is in total control of the content he sees everyday. He feels that a short explanation text explaining to him why he got recommended the content he sees is more than enough. He believes that any more would be an abundance of information. Matthew also thinks that an explanations are good for the users who are not familiar with technology, as long as they are informative and easy to understand. Otherwise it would be difficult for the users to identify what is happening.

4.4 Application of Personas

To test if the personas can be applied to raise requirements for a specific software, each of them will be put in a scenario where they have to run a software and raise requirements like a real user would. The software used for this experiment is SEnti-Analyzer. The details of the experiment will be explained in the following subsections.

4.4.1 SEnti-Analyzer

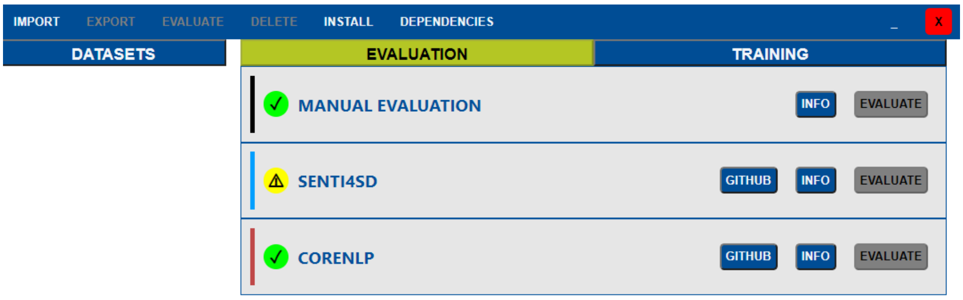


Figure 4.5: SEnti-Analyzer by Herrmann et al. [13]

The SEnti-Analyzer is a tool that allows developers to apply sentiment analysis to verbal communication in meetings for software projects [13]. It can be used to evaluate phrases manually or automatically through the use of AIs, such as SENTI4SD and CORENLP. In this experiment, SEnti-Analyzer will be used to analyse 10 sentences or passages of a text. The personas need to simply run the software and follow the instructions given as stated in the use case tables (table 4.19 and 4.20). The personas are expected to be able to raise requirements based on their interactions with the software (figure 4.5).

4.4.2 Use Case Tables

In this section, the personas will run the SEnti-Analyzer by following the use case tables provided. The use case tables are not complete and contain only the main functionalities of SEnti-Analyzer and direct procedures to run a sentiment analysis.

Use Case 1: Import Dataset

The personas have to follow the instructions listed in the first use case table (see table 4.19) and raise requirements that may fit to their needs. In this scenario, an error after importing the file is expected to happen, and the personas will have to react to it.

Use Case 2: Dataset Evaluation

Before running the instructions in the second use case table (see table 4.20), the personas are required to run a manual evaluation, which is listed as one of the preconditions. Otherwise, the SENTI4SD will only be able to deliver the AI's evaluation result.

4.4.3 Walk-through Summaries and Raising Requirements

In this stage of the experiment, we had to run the SEnti-Analyzer and raise requirements, while acting as the personas. The personas have different personalities and behaviors, which should affect the way they interacted with SEnti-Analyzer. The purpose of this hypothetical demonstration was to test the credibility and applicability of the personas.

Magdalena Evaline's Walk-through

During most of her run with the software, Magdalena (figure 4.1) was full of doubt in herself and the system as well. Magdalena looked for hints in the software multiple times. When she was asked to fill out the form to import the dataset, she was confused and wondered what the form is used for. She was surprised when the import failed and thought that maybe she made a mistake. After following the instruction to reopen the software, she was confused and annoyed, but thought maybe it might be just a bug.

During her run on the second use case, she had no problems. When she was confused about the SENTI4SD AI, she clicked the warning symbol and

Use Case 1	**Import Dataset**
Goal in Context	Before the SEnti-Analyzer can be used, the user must import the dataset.
Preconditions	The SEnti-Analyzer is installed and opened.
Success End Condition	The dataset is imported.
Primary Actors	User, SEnti-Analyzer
Trigger	The user clicks “Import” on the navigation bar.
Main Success Scenario	1. The user selects the dataset in the file explorer. 2. The SEnti-Analyzer displays a pop-up window for import confirmation. 3. The user selects “custom”, and enters “ID” in the first drop-down menu. 4. The user selects “rating” in the second drop-down menu. 5. The user selects “message” in the third drop-down menu. 6. The user clicks on “OK”. 7. The SEnti-Analyzer gives an error message “Something when wrong during insert of the dataSet. Click anywhere to exit”. 8. The user closes the SEnti-Analyzer and opens it again. 9. The SEnti-Analyzer displays the imported dataset on the main menu.
Extensions	4a. If “rating” or “message” is selected on more than one field, an error message will show up and the “OK” button will not be clickable.

Table 4.19: Use Case 1 Table based on the design by Cockburn [8].

the “info” button, which showed her error reports and information about the AI. She was satisfied with the evaluation result. Her only complaint was that she is not familiar with some of the terms shown in the result window, which were “Kappa”, “Kendalls tau”, and others.

Requirements Raised by Magdalena Evaline

Even though she knew the software is not for her, she still believes that the developers can do better by adding necessary explanations:

- Adding explanations when importing the file. A simple button that when clicked, it will show explanations on what the inputs are needed

Use Case 2	**Dataset Evaluation**
Goal in Context	Evaluation of the imported dataset using the AI-Bot SENTI4SD.
Precondition	The file is imported and the user has completed the "Manual Evaluation".
Success End Condition	The results window displays the answers of the user and the SENTI4SD AI-Bot.
Primary Actors	User, SEnti-Analyzer, SENTI4SD
Trigger	The user selects "Evaluate" on SENTI4SD's menu.
Main Success Scenario	1. The user waits until the AI has finished evaluating the dataset. 2. The SEnti-Analyzer displays the system log. 3. The user selects "Continue". 4. The SEnti-Analyzer shows a window containing the evaluation results.
Extensions	1a. If the user has not performed the manual evaluation beforehand, only the AI's answers will be displayed in the results window.

Table 4.20: Use Case 2 Table based on the design by Cockburn [8].

for and what the users should do.

- It could be better if there were more explanations regarding the values above the graphs in the result page.

Her thoughts on the software are that the importing process is a bit buggy, but it does not make the software unusable, just confusing. She felt that the software is not user friendly for those who don't have experiences in software engineering or AI related field.

Finn Lowell's Walk-through

Finn (figure 4.2) was very interested in trying out the SEnti-Analyzer, because he wants to be a software engineer. Finn decided to explore the software as he followed the instructions to test the features. He found a couple of things worth mentioning in the software concerning UI and bugs. When the instructions told him to fill out the import form, Finn wanted to test the error detection features. Finn was fascinated as the software was working as intended. He thought that it would have been better if the users were informed of what they should do during the import process. When Finn encountered the bug where the imported file didn't show up, Finn was confused, and went on with the instructions. He thought that it must have been a bug as well.

Following the second use case table was also not a problem for Finn. He noticed and clicked the yellow warning symbol, and a window popped up informing him that he should've started the software with administrator rights for optimal results. He checked out the "github" and "info" features, and he was able to get the general idea of the AI. Finn was satisfied and went ahead with the evaluation. On the result page, Finn understood the graphs and the tables. He didn't understand the numbers and terms showed on top of the page. He went on to google to look for the terms mentioned. All in all, Finn enjoyed his demonstration.

Requirements Raised by Finn Lowell

Finn has made a list of requirements that could make the software more user friendly to others, even to those who are not familiar with AI related materials:

- When the software opens for the first time, it would be great if there are a short explanation below the "Dataset" menu, instructing the users to import a file, so that the users know what they should do.
- When the users are filling the inputs in the mapping menu, it would be better if there were explanations for each field.
- On the result page, it would be great if there was a help button that when clicked, shows a number of simple explanations, so that users will be able to learn from the data as well.

Finn thought that if there are enough explanations provided, even users who are not likely to use these kinds of software will be interested in trying. For now, Finn understood that the SEnti-Analyzer is used for research purposes only.

Eric Godfrey's Walk-through

Eric (figure 4.3) ran the program as instructed. He followed the instructions, filled the import form with the supposed inputs, and clicked on "OK". When the error message showed up, Eric was confused why it didn't work. When he read the instructions again, he decided to follow them, and thought that it must have been a bug.

On to the next instructions from the second use case, Eric started to question what the purpose of him testing the software is. He only knew that SEnti-Analyzer is a sentiment analysis tool, but he did not know the purpose of the software. Fortunately, Eric was able to find out more about SEnti-Analyzer by clicking the "info" button. After he clicked "Evaluate" and got the results, Eric looked at the numbers above the graphs. He immediately recognized them because he uses correlation coefficients in his line of work.

However, Eric thought it could be better if there's a button where the users can see information about the graphs when they click on it. Otherwise, Eric was fascinated by the software, but he is not sure if the software is usable for people who are not working in IT related fields.

Requirements Raised by Eric Godfrey

Eric noticed some features in the app that could be improved:

- It would be great if the importing procedures are made as clear as possible with the help of written instructions. Without the use case table, he doubts that users would be able to directly know what they should do.
- On the result page, a help button where users can see information concerning the graphs would be helpful, especially if the data is important.

Eric also thought that had he not read the instructions beforehand, he would have thought the error during the import was caused by him. In the end, he thought that the SEnti-Analyzer was an interesting software, but he couldn't see himself using it in the near future.

Matthew Thorben's Walk-through

Matthew (figure 4.4) was interested in testing the SEnti-Analyzer. Following the first use case instructions, Matthew decided to try to mess with the mapping fields inputs. He saw an explanation just above the "OK" button, telling him that there has to be exactly one "message" value in the mapping fields. Matthew followed that, but he tried to not use a "rating" value, and then went on to set the two first fields to "custom" and typed in whatever. When he clicked "OK", the file was imported. Matthew was confused because he did not expect that to happen when he read the instructions.

Matthew proceeded to continue following the use case instructions. Matthew already knew that the SEnti-Analyzer is a sentiment analysis tool from the instructions, so he skipped the "info" button. However, Matthew was interested with the development of the SEnti-Analyzer, so he clicked on the "github" button in the SENTI4SD menu. Matthew then continued with the demonstration and clicked "evaluate" in the SENTI4SD menu. After he waited for the loading animation to finish, he saw a system log giving him details of what happened over the last 2 minutes. Matthew thought that he didn't need to see or know that, since he believed that most users would've just skipped it to see the results.

On the result page, Matthew tried to understand the graphs and other information that came with it. He thought that maybe a little bit of an explanation would be great, especially for the numbers above the graphs. He had to google some terms to know what they meant. In the end, Matthew was satisfied and also fascinated with the SEnti-Analyzer software.

Requirements Raised by Matthew Thorben

Matthew made a few notes on what could be improved in the software:

- The system log could be collapsed and not shown, unless the users want to see the logs. That way the users can directly go to the result page without having to be confused looking at the words they don't understand.

- It could be great if there were buttons on the result page that will show information related to the data shown.

Other than the improvements he mentioned, he didn't mind the bugs, as they would be fixed by the developers anyway. He was satisfied with the demonstration, and hoped that SEnti-Analyzer could fascinate more researchers and engineers.

4.5 Post-Persona-Development Interview Results

This section discusses the additional interviews conducted after the personas were developed and used to test the SEnti-Analyzer. The walk-throughs above may have made the personas believable, but they are still reactions made by fictitious people. These interviews served as supporting evidence that the personas actually represent the real respondents who participated in the initial user study.

In this interview, 1 respondent from each user group was asked to do a hands-on demonstration of SEnti-Analyzer, just like the personas did. The purpose of this experiment is to observe the respondents' reactions during their interaction with the software. The personas were made from compilation of data and testimonies, which are also still vulnerable to assumptions. Meanwhile it is important to note that the respondents were real people, whose documented needs for explanations were limited to the survey questions and their personal experience with familiar systems. Presumably, the results of this experiment will not be the same as what the personas have shown us.

After conducting the interviews, the results were different from what we predicted. Interestingly, the behavior of the respondents observed showed some resemblance to the personas. A total of 3 out of 4 respondents also expressed their curiosity in the SEnti-Analyzer. The results from the interviews will be summarized and listed in bullet points.

4.5.1 "Overthinker" Persona Respondent

- The respondent followed the instructions given, but was very timid and always sounded unsure of what they did.

- When they encountered the bug after importing the file, they were confused and the first thing they did was to try importing the files again. They stopped midway because they were uncertain, and waited for further instructions from the interviewer.

- The respondent asked if the result of the evaluation was important and if they had to memorize them.

From these results, the respondent's behaviors and personality closely resembled the persona Magdalena Evaline. The persona Magdalena Evaline was described to be overthinking and doubtful. This indicates that the persona Magdalena Evaline is relatable to the group of respondents she represents.

4.5.2 "Critical Thinker" Persona Respondent

- The respondent's first reaction when encountering the importing error was to import the files again. The respondent thought it was an error from the software's part, and it might work if they tried again.

- The respondent was fascinated with the results, and tried to change the graphs settings to see alternative results. They also asked questions about the graphs and some terms they didn't understand.

- Before the respondent ended the demonstration, they tried other options available in the SEnti-Analyzer, such as clicking on the "dependencies" menu and "info". They also commented on what could be improved in the SEnti-Analyzer.

These results show the similarities in the respondent's behaviors and the persona Finn Lowell throughout the demonstration session. The respondent asked about the purpose of the SEnti-Analyzer, and expressed their opinion on improvements, which could mean that they believe that the SEnti-Analyzer could be useful for many people.

4.5.3 "Utilitarian" Persona Respondent

- The respondent's first reaction on the importing error was to try importing the files again. But when instructed to reopen the program, they followed.

- The respondent asked how the SENTI4SD works. They had thought that SENTI4SD was an AI that evaluates the users' answers from the manual evaluation.

- After the evaluation, the respondent was intrigued to discuss the results shown on the page. They tried to explain their understanding of the purpose of SENTI4SD.

The respondent hinted a subtle resemblance to the persona of Eric Godfrey when they asked about the SEnti-Analyzer and guessed from what they understood basing on experience.

4.5.4 "Minimalist" Persona Respondent

- The respondent asked a lot of questions regarding the SEnti-Analyzer's functionality and about the evaluation results.

- The respondent explained their experience and background knowledge on natural language processing and proceeded to try out the CORENLP to evaluate the dataset.

- After the evaluation, the respondent went to check the "info" features from both SENTI4SD and CORENLP. They were fascinated and interested to learn more about SEnti-Analyzer.

During the demonstration session, the respondent showed interest and eagerness to learn more about SEnti-Analyzer. They were confident in their background knowledge on some features like CORENLP. Furthermore, they were the only respondent who tried running CORENLP to compare the results with their manual evaluation and with SENTI4SD's. Their confidence and willingness to learn the full capabilities of the software were in line with the personality of the persona Matthew Thorben.

Chapter 5

Discussion

This chapter contains explanations that may answer the research questions asked in section 3.1. It also discusses about the limitations of the research design and threats to the validity of the research results in chapter 4. Furthermore, any challenges we met during the user study and development of personas are included as well.

5.1 Answering the Research Questions

5.1.1 Mixed Method Research

RQ1: Is the mixed method research suitable for developing personas?

As explained in chapter 2.2 regarding the criteria to make a proper persona, there are many points needed to be taken into consideration concerning how the data is collected. Using only either qualitative interviews or quantitative data from surveys does not support the ideal persona development. Based on explanations from section 2.3 and section 3.2.3, combining both interview and survey data will cover each method's weaknesses. Survey method enables developers to collect a large amount of data in a short time, and interview data will complete and represent the survey data with detailed answers.

Including and reviewing the interview data helps in clearing up possible misunderstandings and assumptions from the survey data. Some of the interview respondents had answered "Need an explanation" on the survey. However, when they explained their answer further, they actually didn't care if they need one or not. They just explained that it is good if there was an explanation, but if there was not, it would not affect them that much.

Another respondent had answered "No need for explanations", but when they were asked to explain their reasoning, they stated that they actually need one, but not in certain scenarios, as that would overload them with information. They just wanted a brief but informative explanation. The details of the examples can be read in chapter 4.2

The results from the interviews helped clarify assumptions, reduce bias, and improve the overall quality of the data. They also enabled us in designing realistic personalities for the personas. Therefore, mixed method research is suitable for developing personas.

5.1.2 Persona Application

RQ2: Are the general explainability personas applicable for specific software?

The personas we developed have general characteristics. They have their goals and frustrations that don't necessarily align with the software they are reviewing, which was the SEnti-Analyzer. In chapter 4.3, the personas were showcased to have different habits and personalities. They were all created based on the data collected from the survey and the interviews. We tested the personas by running hypothetical demonstrations (see chapter 4.4.3) and raised requirements while acting as them.

However, reports from these hypothetical demonstrations were not sufficient evidence to prove the credibility and usability of the personas. To counter this problem, we used a post-persona-development interview method, which included some of the previous interview respondents. They went through a hands-on demonstration, where their interactions with SEnti-Analyzer were observed. This additional interview was a reenactment of the hypothetical demonstrations, but with real users.

From the results we have gathered, the behaviors shown by the respondents during their run with the software resembled the personas (see chapter 4.5). We were expecting a different result, because in the survey we used general examples that did not correlate to the SEnti-Analyzer's specific features. The data gathered from the survey and the details added from the interviews have built the personas' behaviors and personalities to the point that they were actually able to relate to the respondents. We concluded that, based on the study and experiments we have done so far, general explainability personas are applicable for specific software.

5.2 Limitations and Threat to Validity

Comparing Survey and Interview Results

In the span of 3 weeks, we have collected 70 survey submissions, 10 of which were from the interviews. When we first analysed the data based on only the survey submissions, there was a lack of diversity, which made the clustering process difficult. Referring to chapter 4.1.1, the respondents were divided based on the first question group's answer, because it was the only group with a wide variety of different answers.

The clustering has to be proven as well if the respondents are well distributed using inter-rater agreement coefficient. However, the link between respondents was very weak, and there were no supporting factors other than similar answer patterns. The clusters were still subject to assumptions and developers' bias. This weakness was then mitigated using the interview data. By including the answers recorded in the interviews, which could not be observed in the survey data, we were able to improve the overall quality of the data.

Proving the Credibility and Validity of Persona

Referring back to chapter 2.2, although the personas were developed using the evaluation criteria, it is not enough to guarantee their credibility and "realness". The personas could still be deemed subjective if other developers or persona users didn't have the same impression. To mitigate the subjectivity of the personas and at the same time increase their credibility and validity, additional interviews were conducted (refer to chapter 4.5). The interview was a hands-on demonstration, where respondents had to interact with the SEnti-Analyzer while they had their behaviors observed. The end results from the interviews showed us that the personas were applicable to a specific software.

5.3 Challenges

Fleiss' Kappa

Fleiss' Kappa has been used to find out if there's an agreement in multiple evaluations of data. When the values are homogeneous, however, it affects the variable in the calculation, lowering the percentage of agreement. After numerous attempts to rearrange the participants within the persona clusters, and additional research on why this has happened, it turned out to be a paradox (refer to chapter 4.1.4). The agreement percentage is low when there is no similarities in the data, but it should be decent when there is some similarities; it could, however, be even lower when the data is lacking in diversity [16].

Hands-On SEnti-Analyzer Demonstration

The personas we have developed may possess personalities and behaviors of a real person, but there was no solid ground for us to prove if they truly represent the respondents. There are risks of personas not being used in late development stages. Furthermore, they lose their credibility if the developers changed their opinions on the personas (refer to chapter 2.2.2). We needed to come up with another research method where we can confirm if the personas are relatable to the respondents. The method we came up with was a post-persona-development hands-on demonstration.

The concept itself was already a challenge (refer to chapter 3.2.3 and 4.5). We needed to trace back and contact the interview respondents, and asked them to participate in the demonstration. The time for writing the bachelor's thesis was at risk as well, because it was conducted when the personas have been developed and the thesis was already in progress. Fortunately, all the respondents were willing to participate in the demonstration. The next challenge was to have all the participants download TeamViewer[1] and instruct them on how to operate it, because none of the participants have any experience with the software. In the end, the demonstration was successfully conducted, and we managed to have supporting data for the personas.

[1] https://www.teamviewer.com/

Chapter 6

Conclusion and Future Works

This chapter concludes all the study results and findings throughout the thesis, with the addition of possible future endeavors on what could be improved.

6.1 Conclusion

The purpose of this work was to investigate if personas are useful to raise explainability requirements from a specific software. Explainability is known to be subjective like other NFRs. The evaluation process is difficult, because not every user has the same needs for explanations. It was observed as well, that different behaviors and personalities led to different needs for types of explanations. To mitigate the subjectivity of explainability, fictitious identities called personas are developed. Users with similar goals and behaviors are grouped and represented by these personas. A user study was conducted in order to understand what the users need and to categorize them accordingly.

The first step was to design the user study. Following related literature regarding persona development, a preferable method of study to effectively and efficiently collect data from users is the concurrent mixed method research. Combining both survey and interviews has proven to be effective in creating credible and reliable personas. The survey questions consisted of general examples of applications that most of the respondents are familiar with, in order to measure their needs for explanations. The interviews were conducted not only to provide the data with more detailed answers from the respondents, but to observe their behaviors when answering the questions as well.

The collected data was then evaluated, and the respondents were grouped based on their similarities in answers and behaviors. From these user groups, with the help from the interview data as well, the personas were designed and tested through hypothetical software demonstrations. However, the personas were still questionable in regards of credibility and validity. There was no guarantee that the personas developed were without bias.

To counter that problem, an additional interview was conducted after the personas were developed and tested. The interview was conducted in the form of a hands-on demonstration, just like the hypothetical demonstrations. The purpose of this real life demonstration is to observe the behaviors of the respondents when they test the software, and to see if their behaviors resemble the personas who represent them. From the results of the demonstrations, the behaviors of the respondents did resemble the personas. This indicates that not only the personas truly represented the users, thus eliminating any possible biases, they were also able to raise explainability requirements from a specific software.

6.2 Future Works

Developing personas by conducting mixed method research has its strengths and weaknesses, as explained in chapter 3.2. From the study design and reports gathered in this thesis, there are some points that could have been improved.

Improve Survey Questionnaire

The quality of the questions could have been improved. According to the comments written by one of the survey respondents, they mentioned the ambiguity of some of the questions. One of the interview respondents mentioned that there were repeating questions; similar questions but with different wording. A few questions were also not written clearly and caused misunderstandings for some of the respondents. The questions should be reviewed more carefully, content- and structure-wise, in order to achieve a higher quality of answers.

Try Sequential Mixed Method Research

Concurrent mixed method research has proven to be time efficient in this thesis. Due to the fact that both survey and interviews were conducted at the same time, there were no difficulties in tracing back the interview respondents, as their respondent IDs were noted differently from the survey respondents. However, the interpretation and integration processes were challenging and there were some risks of lingering assumptions even after successfully combining both data.

Hypothetically speaking, if there was more time for the experiment, using sequential mixed method might deliver higher quality results. Persona developers Tu et al. [31] implied in their journal that they used sequential mixed method research. Upon reviewing said literature, it can be seen that by focusing on the survey data first, the respondents could be clustered more properly, and interview respondents can be conveniently chosen from the clustered respondents. The data interpretation process are done sequentially, as opposed to the concurrent mixed method, which makes it easier to acquire high quality and controlled data. Important thing to note is that the survey data cannot be completely anonymous, otherwise it wouldn't be possible to trace back every survey respondents to invite them for interviews.

Appendix A

Additional Graphs

Quantitative data regarding the answer patterns from the respondents in chapter 4.1.1 are shown in tables. This section provides the visual presentation of the data.

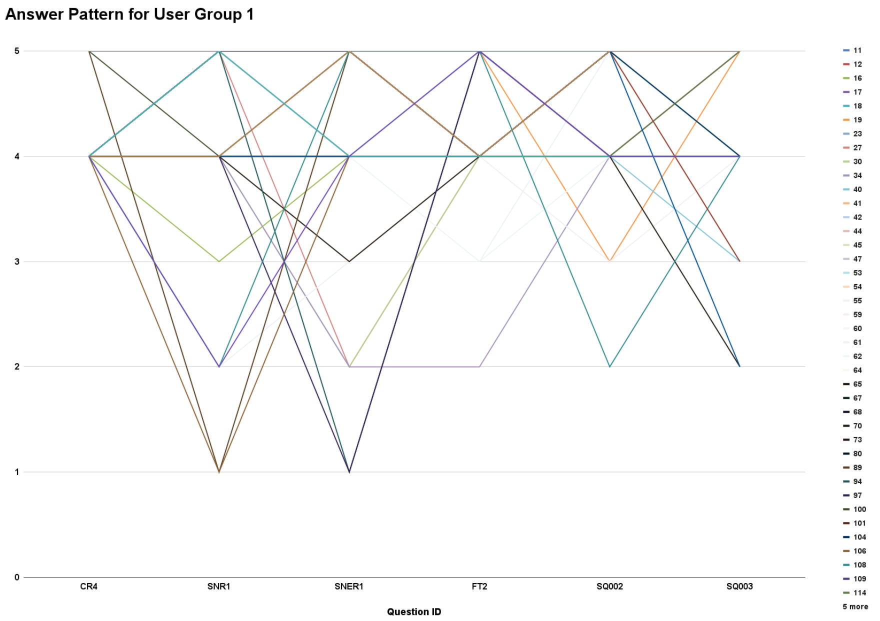


Figure A.1: Answer Pattern for User Group 1 (Table 4.13).

There are many overlapping lines in the 3 graphs indicating the similarities in answers between respondents. However, in figure A.1, the answer count is too high and the diversity is low, resulting in a low kappa coefficient.

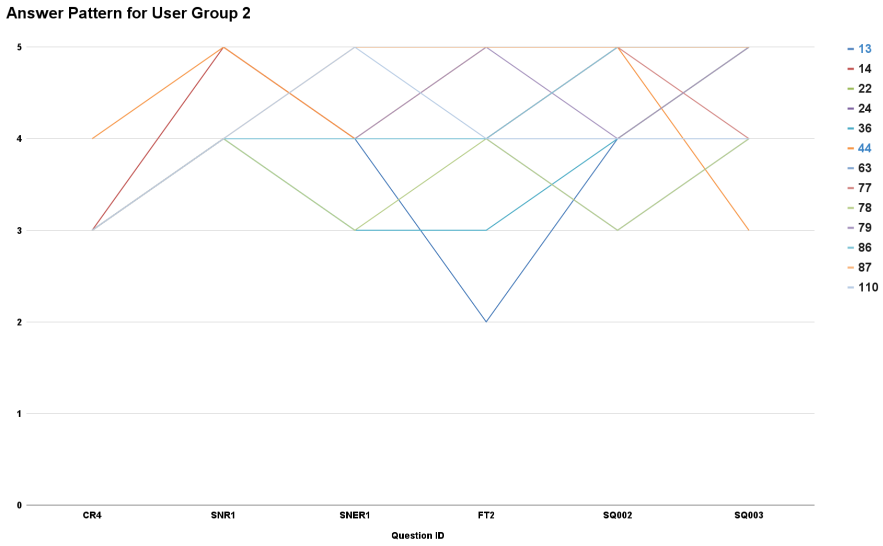


Figure A.2: Answer Pattern for User Group 2 (Table 4.14).

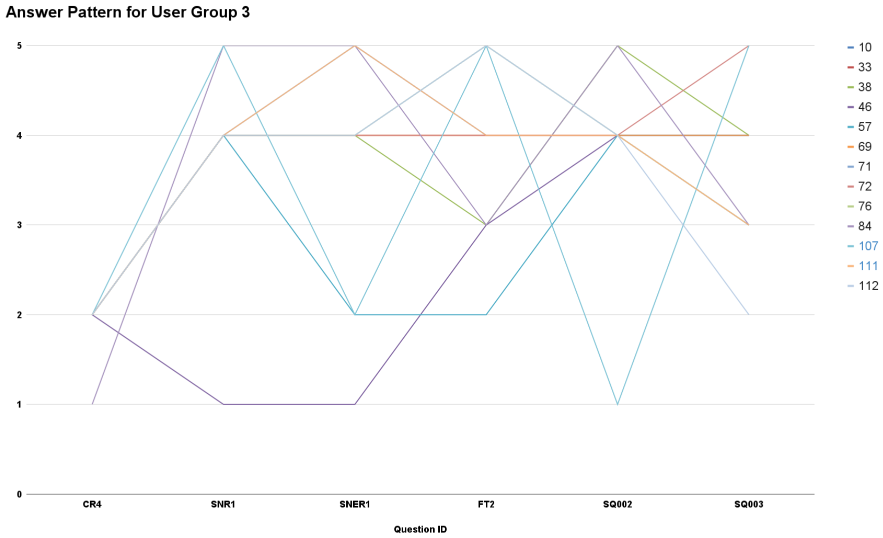


Figure A.3: Answer Pattern for User Group 3 (Table 4.15).

Appendix B

Contents on the USB Drive

The USB drive submitted along with the bachelor's thesis contains the following:

- This bachelor's thesis in *.pdf* format.
- Results of the data analysis in *.xls* format.
- Persona skeletons and drafts in *.pdf* format.
- File directory named "Explainability Literature" containing related literature regarding explainability.
- File directory named "Persona Literature" containing related literature regarding persona development.
- File directory named "User Study Literature" containing related literature regarding user study methods.
- File directory named "Notes" containing the following:
  - List of reviewed literature.
  - File named "BA Explainability", which contains summary and notes on reviewed literature.
  - File named "BA Persona", which contains summary and notes on reviewed literature.
  - File named "BA User Study", which contains summary and notes on reviewed literature.
  - Initial draft of the survey questions.
- File directory named "Interview Files" containing survey submissions from the interview respondents and their voice recordings.
- File directory named "SEnti-Analyzer Demo Videos" containing video recordings of the hands-on demonstrations.

- Git repository containing all the progress of this thesis.

Appendix C

Acknowledgement

I would like to express my gratitude to the people who supported and encouraged me in my work on the bachelor's thesis.

M. Sc. Jakob Droste, my thesis supervisor, for providing me with constructive feedback, encouragement and guidance throughout the working of this bachelor's thesis.

Prof. Dr. rer. nat. Kurt Schneider and Dr. rer. nat. Jil Klünder for examining and evaluating this thesis.

The survey and interview participants, for their time and for providing meaningful data for the user study.

My Fiancée, Felicia Celins, for being there for me, as well as supporting and caring for me throughout the journey. My family and friends, for their words of encouragement and support with the survey and interviews.

List of Figures

List of Tables

Bibliography

[1] P. J. Adler. Dealing with interviews when creating personas: A practical approach. In *Proceedings of Student Interaction Design Research Conference SIDER05*, pages 84–88. Citeseer, 2005.

[2] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[3] J. J. Bartko. The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, 19(1):3–11, 1966.

[4] A. Bussone, S. Stumpf, and D. O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE, 2015.

[5] L. Chazette, W. Brunotte, and T. Speith. Exploring explainability: a definition, a model, and a knowledge catalogue. In *2021 IEEE 29th international requirements engineering conference (RE)*, pages 197–208. IEEE, 2021.

[6] L. Chazette, O. Karras, and K. Schneider. Do end-users want explanations? analyzing the role of explainability as an emerging aspect of non-functional requirements. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 223–233. IEEE, 2019.

[7] L. Chazette and K. Schneider. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering*, 25(4):493–514, 2020.

[8] A. Cockburn. Basic use case template. *Humans and Technology, Technical Report*, 96:28, 1998.

[9] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[10] J. W. Creswell and V. L. P. Clark. *Designing and conducting mixed methods research*. Sage publications, 2017.

[11] S. Faily and I. Flechais. Persona cases: a technique for grounding personas. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2267–2270, 2011.

[12] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[13] M. Herrmann, M. Obaidi, and J. Klünder. Senti-analyzer: joint sentiment analysis for text-based and verbal communication in software projects. *arXiv preprint arXiv:2206.10993*, 2022.

[14] R. R. Hoffman, G. Klein, and S. T. Mueller. Explaining explanation for "explainable ai". In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 197–201. SAGE Publications Sage CA: Los Angeles, CA, 2018.

[15] M. Hosseini, A. Shahri, K. Phalp, and R. Ali. Crowdsourcing transparency requirements through structured feedback and social adaptation. In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, pages 1–6. IEEE, 2016.

[16] J. G. (https://stats.stackexchange.com/users/111380/jeffrey girard). Why does fleiss39;s κ decrease with increased response homogeneity? Cross Validated. URL:https://stats.stackexchange.com/q/207640 (version: 2016-04-16).

[17] N. V. Ivankova, J. W. Creswell, and S. L. Stick. Using mixed-methods sequential explanatory design: From theory to practice. *Field methods*, 18(1):3–20, 2006.

[18] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender. Explainability as a non-functional requirement. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 363–368. IEEE, 2019.

[19] A. Luma, B. Abazi, and A. Aliu. An approach to privacy on recommended systems. In *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–5. IEEE, 2019.

[20] G. Olsen. Persona creation and usage toolkit. *Retrieved March*, 25:2014, 2004.

[21] J. Pruitt and T. Adlin. *The persona lifecycle: keeping people in mind throughout product design*. Elsevier, 2010.

[22] H. Ramos, M. Fonseca, and L. Ponciano. Modeling and evaluating personas with software explainability requirements. In *Iberoamerican Workshop on Human-Computer Interaction*, pages 136–149. Springer, 2021.

[23] S. Robbins. A misdirected principle with a catch: explicability for ai. *Minds and Machines*, 29(4):495–514, 2019.

[24] A. Rosenfeld and A. Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6):673–705, 2019.

[25] J. Salminen, H. Kwak, J. M. Santos, S.-G. Jung, J. An, and B. J. Jansen. Persona perception scale: developing and validating an instrument for human-like representations of data. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2018.

[26] L. Schneidewind, S. Hörold, C. Mayas, H. Krömker, S. Falke, and T. Pucklitsch. How personas support requirements engineering. In *2012 First International Workshop on Usability and Accessibility Focused Requirements Engineering (UsARE)*, pages 1–5. IEEE, 2012.

[27] W. A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325, 1955.

[28] V. Thoma and B. Williams. Developing and validating personas in e-commerce: A heuristic approach. In *IFIP Conference on Human-Computer Interaction*, pages 524–527. Springer, 2009.

[29] N. Tintarev. Explanations of recommendations. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 203–206, 2007.

[30] C.-H. Tsai and P. Brusilovsky. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 391–396, 2019.

[31] N. Tu, Q. He, T. Zhang, H. Zhang, Y. Li, H. Xu, and Y. Xiang. Combine qualitative and quantitative methods to create persona. In *2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering*, volume 3, pages 597–603. IEEE, 2010.