MSc Thesis Interaction Technology

Measuring well-being with a conversational agent

Daniëlle Kwakkel

Supervisors:
dr. M. Theune (1st supervisor)
dr. J.T. Kraiss (bms-PGT)
dr. J.B. van Waterschoot

June 29, 2022

Department of Human Media Interaction
Faculty of Electrical Engineering,
Mathematics and Computer Science

UNIVERSITY OF TWENTE.

Acknowledgements

The research described in this thesis focuses on the area where humans and technology come together, an area that I find fascinating. Technology can influence the lives of humans in many different ways, both negatively and positively. For new technology to have a positive influence, it is crucial to keep the envisioned goal and the user in mind during all of the developments. I am happy and grateful that I got to work on a thesis for the past months which does that. Hopefully, this research will contribute to the development of technology which positively influences society.

This thesis would not have been possible without the people who have helped and supported me along the way. Firstly, I would like to thank my main supervisors, Mariët Theune and Jelte van Waterschoot, for guiding me through the whole process. Thank you for the frequent meetings, the quick replies to my emails, the helpful feedback, the useful references and all the encouragements you gave me. I would also like to thank my external supervisor, Jannis Kraiss, for bringing the expertise in psychology and well-being to this thesis.

I would like to thank the experts, Ernst Bohlmeijer and Enny Das, who made space in their busy schedules to discuss the design of the conversation with me. For the people who participated in the pilot test and the data collection, thank you for participating! Without you there would be no dataset to work with.

Lastly, I would like to thank my family and friends who supported me along the way. Thank you for never ceasing to believe in me! And for telling me, time and time again, to be kind to myself and to persevere.

I couldn't have done it without you!

Daniëlle Kwakkel

Enschede, June 2022

Abstract

Conversational agents can be used to get more insights in a person's well-being, something that is currently done by the BLISS project. As part of that project, this research aims at creating a system which can automatically extract information about someone's well-being through a spoken conversation with a conversational agent in Dutch. To achieve this aim, I defined well-being theories and described their measures, designed a conversation implementing these theories, gathered a dataset using the designed conversation, analyzed the gathered dataset and created a program for automatic information extraction and classification.

This research starts with introducing the BLISS project, identifying related work and stating the research questions. Thereafter, the concept of well-being is explored through a literature study. This study starts with describing the two main theories on well-being: hedonic and eudaimonic, corresponding to the psychological models of subjective well-being and psychological well-being. Different models and measurements were described, concluding with the two models used in this research: subjective well-being and the Flourishing Scale. These well-being measures were translated into a conversation for a conversational agent. The design of the conversation was done using an iterative process with expert interviews and a pilot test as input. After implementing the conversation into the conversational agent software called WhappBot, a dataset was gathered consisting of transcriptions of spoken data in Dutch. Two datasets were gathered, one where the participants filled it in as themselves and one where the participants answered in the role of a persona with lower well-being. The data from these dataset was analyzed, looking for common indicators in the answers of low, neutral and high well-being. For some questions clear commonalities were found, like for the question about competence in doing daily activities. Others did not show any difference between low, neutral and high well-being, like the answers about felt emotions or social relationships. The results from this analysis were implemented in an information extraction algorithm and in well-being classifiers. For the information extraction, a combination of the LIWC and WordNet lexicons was used, with which social relationships and different activity types were successfully extracted from the text. Although the dataset was found to be too small to yield good results for the classification, the results do show that the information extraction algorithm and well-being classifiers have potential to perform autonomous well-being classification in a bigger dataset.

Besides the obvious improvement of gathering a larger dataset, there are a few more ways in which the results could be improved. For example, already existing datasets could be used as training data for the classification. Also, future conversations could be designed to be more focused on a specific subject to make the data collection shorter and to formulate the questions in a way which facilitates information extraction. All in all, this research serves as a proof-of-concept of a method to design a conversation based on well-being models and to automatically extract information about the person's well-being from this conversation.

Table of contents

1 | Introduction

The percentage of elderly people in modern-day society has been increasing for years, and is expected to rise even more. According to He et al. [33] 17 percent of the world's population will be older than 65 in 2050, compared to 8,5 percent when the report came out in 2015. This vast change in population brings new challenges in healthcare with people living at home for a longer time [30]. Currently, interventions are being designed to help the elderly with adapting to circumstances and self-managing their health. One of these interventions is the BLISS project [90], of which this thesis is a part. This introduction will start with a brief overview of the BLISS project. Next, the related work will be described, which is grouped into *conversational agents for the elderly* and *conversational agents for mental health*. Finally, the research gap that this thesis will focus on will be identified.

1.1 | The BLISS project

A team from University of Twente and Radboud University has been working on the BLISS project[1], which stands for Behaviour-based Language-Interactive Speaking Systems. This personalized spoken dialogue is implemented into an artificial intelligent agent with the purpose of facilitating self/joint-management of health and well-being. This means that the user and their possible caretakers can use the system to get more insight into their well-being and in what way they can improve this. The targeted use for this system is in elderly care, where it can give precious insights into the well-being of elderly persons to their caretakers. These insights could be any change in aspects influencing the user's well-being, like changes in social relationships, decrease of engagement in activities, or optimism for the future. The current focus of the project is on extracting information from long-term conversational interaction through text mining of Dutch text. And on using this information for the generation of engaging questions which yield personal information which could be helpful to their caretakers [90].

1.2 | Related work

1.2.1 Conversational agents for the elderly

Conversation based technologies have been popular when designing for the elderly because this age group often has difficulty with new technologies and are not used to text entry on smartphones, but conversations come natural for them [94]. Most conversational agents in elderly care are used for helping the user in their daily tasks. Examples are Anne, a Dutch virtual assistant who facilitates video calls, reminds the user to take medication and helps managing the user's calendar [3]. And FitChat, a conversational AI intervention to motivate older adults for physical activity [94].

Some applications have been developed which focus specifically on holding conversations with elderly users. For example ElliQ [42], a companion robot which (among other things) holds normal conversations, tells jokes, answers questions, helps remembering important information and can help with physical exercise. The robot is still in the testing phase, which has shown that elderly users can feel connected to the robot and even express love for it.

A virtual agent called LISSA was created by Razavi et al. [68]. Their conversational agent is aimed at improving the communication skills of elderly users. The users in the study had 10 conversations with the virtual agent. In these conversations the agent observed the users' communicative behaviour and gave advice on how to improve their communication skills. The elderly participants of the study found that the system was easy to use and user-friendly.

Boumans et al. [10] studied whether social robots can be used to conduct medical questionnaires with older adults. In their study, the Pepper robot, which is a humanoid robot with a screen on its chest, conducted a questionnaire autonomously without any health professionals present. The robot asked the questions verbally and the participant was asked to answer and confirm their answer on the screen. Boumans et al. [10] found that 93% of the interactions were completed autonomously with valid results.

[1]see http://hstrik.ruhosting.nl/bliss/ for more information

They concluded that social robots can be used to autonomously interview older adults and collect data about their health.

Although these types of agents can hold conversations, they often have a different focus than gathering information and asking relevant questions (like the BLISS project). A concept that comes close to this focus is reminiscence, which is the recalling of past memories often used in the elderly population. Nikitina et al. [61] developed a conversational agent to stimulate reminiscence in an elderly user. Besides the reminiscence model, their chatbot also has a life model and a conversation model which could be useful for BLISS. The life model collects information about the user that describes them. Examples are habits, life events, values, beliefs, hobbies and relationships. The conversation model contains the information on questions, momentos and general information to hold a good conversation.

Finally, maybe the most related conversational agent for the elderly is a Swedish project called Memory Lane [83]. Memory Lane is a voice assistant which asks personal questions and talks about someone's life through a smart speaker. The gathered stories are compiled into a book and podcast. According to their website, the program *"understands the correlation between different answers, which triggers relevant follow-up questions."* [1]. It also creates long-term memory by using a memory graph to store a person's previous stories.

1.2.2 Conversational agents for mental health

There is a vastly growing body of research about applying conversational agents in mental health applications. Covering the whole body of research is too much for this paragraph. For a complete overview the review by Callejas and Griol [12] can be used. This paragraph will mention some applications and research which have one or more attributes similar to BLISS: (1) having voice-based conversations (2) about well-being (3) over a longer period of time.

Firstly, text-based conversational agents have been created who coach users through mental health problems and support self management. One example is Wysa, which is based on positive psychology and mental well-being techniques [40]. The self-help practices that are used in the conversations are *"CBT, dialectical behavior therapy, motivational interviewing, positive behavior support, behavioral reinforcement, mindfulness, and guided microactions and tools to encourage users to build emotional resilience skills"*[40, p. 3]. Another example is Woebot, a text-based conversational agent based on cognitive behavioural therapy. A study revealed that Woebot reduced depression symptoms and users formed a therapeutic bond with the agent [16].

Bickmore et al. [8] designed a theory-driven conversational agent for behaviour change where the agent has the role of a counsellor and holds multiple conversations over time. The conversations are built upon a reusable framework for health counseling. This framework is grounded in theories from behavioural medicine. It combines six different models: theory model (basic constructs and relationships), user model (user specific information), behaviour model (application of specific health behaviour theories), protocol model (information of particular interventions), external data model (data inputs and outputs) and a task model (enactment of intervention). All these data sources are combined into the counseling dialogue system, which is visualized in an OWL ontology. Bickmore et al. [8] conclude that such a model can be used and re-used in conversational behaviour change interventions and they hope that it will reduce the subjectivity of mapping theories in these interventions.

Conversational agents have also been used to administer health related questionnaires. One of these is described by Maharjan et al. [56]. Their conversational agent called Sofia administers the WHO-5 well-being scale through spoken conversation. Paper questionnaire answers were compared to discrete (meaning a limited number of possible answers e.g. answering either yes or no) spoken dialogue and open-ended dialogue. They found that open-ended dialogue gave a broader understanding of respondent's health and well-being and worked in the majority of the cases. However, the given answer did not always map to an answer in the questionnaire, so the coherence was lower. They recommend using discrete follow-up questions to be able to map an answer to the scale and confirm a user's intended meaning. *"If a user responds to a question by commenting "half of the time", "not as many days as I wanted" or "I tried to do things that interest me", for example, a follow up intent might ask "Ok, so how would you rate your experience on a scale of 0-5?" in order to confirm the user's intended meaning."* [56, p. 10] Philip et al. [62] compared a diagnostic conversation by a health professional with a conversational agent. They found that the conversational agent was able to diagnose major depression with high accuracy, although the accuracy with mild depressive symptoms was lower. A similar study used a voice based conversational agent for pain monitoring [51]. Users did multiple sessions over the course of two weeks. They found that the user's experience improved over the sessions because they got used to the shortcomings of the agent. This resulted in longer sessions and less troublesome dialogue situations over time.

A recent PhD thesis by Wu [96] describes a long term well-being agent called WellBe which purpose is very similar to the purpose of BLISS. *"WellBe can talk to the user about their recent activities, feelings and desires, provide empathetic responses, reflect on what the user has said, and make recommendations to the user of activities that they can schedule that are predicted to improve their overall well being."* [96, p. 143]. In the thesis, user affect and activity classifiers were created which had high accuracy. The recognized activities were classified in 5 well-being categories[2]: competence, connection, savouring, obligation, incompetence. Research was also done on engaging dialogue strategies, which are *acknowledgement, empathy, reflection* and *recommendation*, to use in well-being applications. It was found that responses using the user's affect and activity information were perceived as more engaging.

1.3 | Research gaps

The literature in the previous sections shows that conversational agents for the elderly that measure well-being do not exist yet. But also in a broader sense, there are hardly any conversational agents that measure well-being except for the study by Maharjan et al. [56] and the thesis by Wu [96].The other conversational agents in mental health often focus on detecting (the lack of) depressive symptoms or combine self-management theories like CBT. To measure well-being, Maharjan et al. [56] used the questions from the *World Health Organization-Five Well-Being Index (WHO-5)* directly. The WHO-5 measures a person's subjective psychological well-being in five questions. Wu [96] chose to use a combination of three well-being theories. Different than Maharjan et al. [56], Wu [96] does not base the conversation on these well-being theories but uses them as categories to classify the user's description of a experience.

1.4 | Research questions

The previous section showed a research gap of measuring well-being through conversational agents. This thesis is aimed at filling this gap by creating a voice-based conversational agent which design is based on existing well-being measurements and by estimating a person's well-being based on the given responses.

The research question for this thesis is:

How can we measure a person's well-being through spoken conversation with an agent?.

This question will be divided into four sub-questions:

1. How can well-being be defined and measured?
2. How to design a conversation for a conversational agent which implements the chosen well-being measures?
3. What useful information can be extracted from a dataset gathered with the conversation that was designed?
4. Can transcribed conversation data be used to automatically extract information about someone's well-being?

The first sub-question will be answered in Chapter 2. This chapter will start with the Positive Health model which is currently used in the BLISS project. Then it will dive deeper into positive psychology, hedonic and eudaimonic well-being, related psychological theories and relevant psychological measures. At the end of this chapter the well-being measures used in the rest of the thesis will be chosen. The second sub-question will be answered in Chapter 3 about the conversation design. In this chapter speech-specific design heuristics and social conversations with agents are explored through literature. With this input, a conversation is designed which covers the original questions from subjective well-being and the Flourishing Scale, as chosen in Section 2.8.2. The conversation design was done iteratively using expert interviews and a pilot test as input. The final conversation is used in Chapter 4 to gather a dataset. This chapter also answers the third sub-question, performing an analysis of the data and looking into what useful information the data contains. Chapter 5 builds upon the data analysis by creating information extraction algorithms and text classifiers, all aimed at automatically extracting information about someone's well-being. In the last chapter, Chapter 6, the whole thesis is summarized and the results are discussed.

[2]The categories are based on Self-Determination theory [71], Appraisal theory [76] and Savoring theory [11]

2 | Well-being models

In order to measure a person's well-being with a conversational agent it's vital to first determine how a person's well-being can be defined. Currently the BLISS project uses the model of Positive Health developed by Machteld Huber[1] as the definition of well-being. Although the model for Positive Health is widely used in practice, the next section will show that the model still lacks a sound basis from literature and a valid measurement tool. Since there are many different views and models on defining well-being in both philosophy and psychology, this chapter will dive into this literature to answer the first sub-question: *How can well-being be defined and measured?*.

This chapter will start with describing the Positive Health model which is currently used in the BLISS Project. Thereafter, the foundations of different definitions of well-being will be explored and their corresponding psychological models will be identified in Section 2.3. The psychological models implementing these different definitions will be addressed in Section 2.4. Section 2.5 describes research that focuses on the influence of older age on the models of well-being. Thereafter in Section 2.6 different measurement tools for these models will be set out. Finally, the Positive Health model will be compared to the found literature and a model will be chosen to be used further in this thesis.

2.1 | Positive Health

In 2011 Huber et al. published a paper which argued for a redefinition of health. From *a state of complete physical, mental and social well-being* [95] to *the ability to adapt and to self manage* [36]. From this initial redefinition general practitioner Machteld Huber has developed the concept of Positive Health. Positive Health is a model to measure health by rating someone's personal situation in six pillars: *bodily functions, mental well-being, meaningfulness, quality of life, participation, and daily functioning.* Each of these pillars has multiple sub-categories and a corresponding list of questions for health practitioners to use. These pillars are visualized in a spiderweb-diagram which can be seen in Figure 2.1.

The model was created using the bottom-up approach of grounded theory [84]. Participants of a conference were asked to define health. From his wide variety of characterisations of health the six main dimensions and 32 aspects of health were synthesized [37].

The model of Positive Health was received enthusiastically and many organisations in The Netherlands have started to implement this model into practice [50]. Although the model was received well, there are also some concerns in the use of the model. Even though the Institute for Positive Health clearly states that the model should be used as a conversational tool rather than a measurement tool, it is found that this is not clear for executing professionals causing them to use the tool in the wrong way [50]. If a measurement tool is desired, the Institute for Positive Health has suggested a list of validated measurement tools that cover one or multiple dimensions of the Positive Health model [41]. Prinsen and Terwee [66] have attempted, set in motion by the Institute for Positive Health, to combine validated measurement tools into a sound questionnaire to be used as a measurement tool specific for the Positive Health model. A group of stakeholders and experts identified 40 existing questionnaires, which were combined into a questionnaire covering all six pillars of the Positive Health model. The final questionnaire was evaluated on relevance, comprehensiveness and comprehensibility by an expert group. Unfortunately, the expert group judged the questionnaire to be inadequate in content validity and raised questions about the validity of the Positive Health concept. One of the questions raised was *"whether the present conceptual model of 'positive health' is an adequate reflection of health or rather a reflection of aspects of life that influence health"* [66, p. 75]. Although an indication of aspects that influence health could be useful in many cases, it carries the risk of missing important information regarding a person's health. Furthermore, the names used for the different aspects of the model were sometimes found not to match the given patient definitions or the medical definition of that dimension. Additionally, the coverage of the model was questioned. *"Important aspects, such as aspects on competency, being hopeful, long-lasting social contacts, self-acceptance, stress, vitality and worries, were considered to be missing from the conceptual model"* [66, p. 73]. The authors of the research conclude that the conceptual model of Positive Health has not been fully worked out yet for its intended application. They recommend users to use existing tools that cover the different dimensions and to clearly keep the purpose of measurement and the target population in mind.

[1] see http://www.iph.nl/

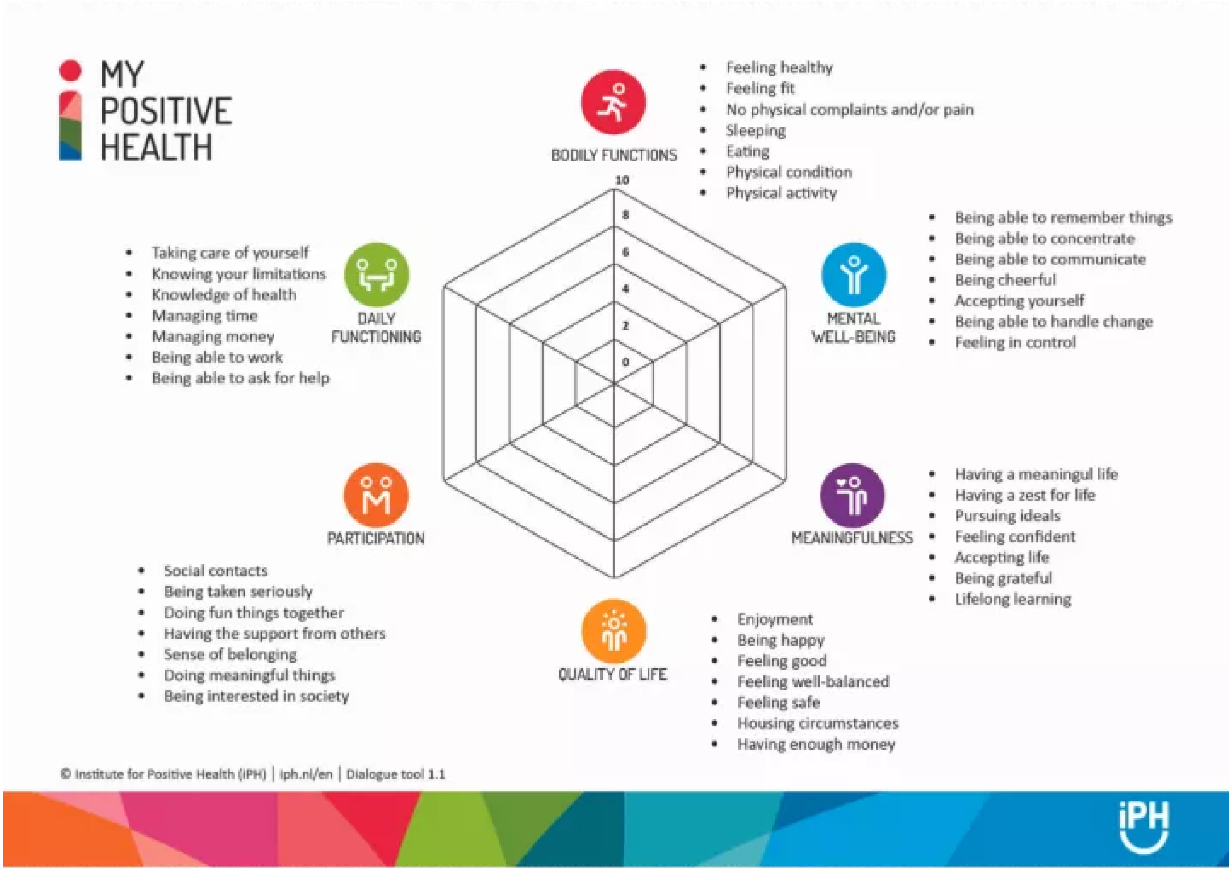

Figure 2.1: Spiderweb-diagram of positive health (source: http://www.iph.nl/)

2.2 | Positive Psychology

The model of Positive Health described in the previous section defines health as a broader concept than just the absence of illness. A similar movement in psychology started around the year 2000 by a psychologist called Martin E.P. Seligman. He started the psychological movement of Positive Psychology which is focused on building the best qualities in life rather than only repairing the bad situations in life [80]. Seligman [80] summarizes the movement as identifying, amplifying and concentrating on positive human traits. Gable and Haidt [27] define Positive Psychology as *"Positive psychology is the study of the conditions and processes that contribute to the flourishing or optimal functioning of people, groups, and institutions"*[27, p. 104]. With the focus on these positive conditions and processes a person can be taught to increase their own well-being. Building competency in these aspects is expected to work as a prevention technique acting as a buffer against mental illness [80]. Thus, focusing on positive qualities will help people, groups and institutions flourish. The question arises what these qualities are that cause flourishing. For simplicity this paper only focuses on flourishing of people and not on groups and institutions. Personal flourishing is often defined as well-being. The next sections will explore the concept of well-being through studies on philosophy, psychology and social studies.

2.3 | Hedonic and eudaimonic well-being

The prominent literature on psychological well-being defines two main traditions on well-being: hedonic and eudaimonic well-being. Both traditions date back to early philosophers trying to define how to live a good life. And more specifically, what motivates us to do things or what makes things worth doing. This section will discuss both of these traditions and the well-being theories originating from them.

2.3.1 Hedonism and the concept of subjective well-being

The first tradition, hedonism, focuses on pleasure and says that only pleasure and pain motivates us or makes things worth doing [58]. The definition of happiness (or well-being) that follows is that happiness is the total sum of someone's hedonic moments [72].

More recent literature about the hedonic psychology broadens the definition of pleasure by saying that well-being is about the subjective experience of one's happiness and the subjective experience of one's pleasures and displeasures [72]. Practically this means that most research uses the concept of subjective well-being (or SWB) as an assessment of hedonic well-being. Subjective well-being includes how someone is feeling, his satisfactions with things in his life like work and family, and how satisfied someone is with his life [20]. In other words, it can be defined as a combination of positive affect, negative affect and life satisfaction [39].

2.3.2 Eudaimonia and the concept of psychological well-being

The second tradition, eudaimonic well-being, views happiness as something more than just pleasure or desires. Aristotle was one of the first and most famous philosophers who defined the term *"eudaimonia"* in his theory on virtue ethics. He defines happiness as *"activity of the soul in accordance with virtue"* [75, p. 16]. *"Eudaimonia"* goes beyond the definition of subjective happiness and talks about true happiness, a flourishing life that is worth seeking or having [38]. This flourishing life can be reached by living synchronized with one's true self according to one's most important values [72]. Or, in other words, to *know thyself* and to *become what you are* [75]. Eudaimonic well-being is more difficult to practically define than hedonic well-being, since it covers many facets of one's life. Ryff [73] created a model to define and measure eudaimonic well-being which she calls Psychological Well-being. Ryff and Singer [74] combined the theoretical literature of that time on positive psychological functioning and found six core dimensions where the theories of psychological well-being converge. These dimensions and their conceptual foundations are visualized in Figure 2.2.

As can be seen in the figure, the six dimensions of psychological well-being [74] are:

Self-Acceptance
Holding positive attitudes towards oneself

Positive Relations with Others
Having warm, trusting interpersonal relations

Autonomy
Self-determination, independence and the regulation of behavior from within

Environmental Mastery
Able to choose or create environments suitable to his or her psychic conditions

Purpose in Life
Has goals, intentions and a sense of direction

Personal Growth
Continued personal growth and self-realization

2.4 | Combining hedonic and eudaimonic well-being

As can be concluded from the previous section, hedonic and eudaimonic well-being are fundamentally different concepts on how to view well-being. While some research chooses to focus on one of the concepts, others have investigated how a combination of the two can be used to measure all facets of well-being.

2.4.1 Models that explicitly combine hedonic and eudaimonic well-being

Huta and Ryan [39] executed a large research, consisting of four studies (two correlational, one experience-sampling, and one intervention study), about the effect of both hedonic and eudaimonic motivations on well-being. For this, well-being was broken down into different factors contributing to well-being which are positive affect, negative affect, life satisfaction, carefreeness, meaning, elevating experience and vitality.

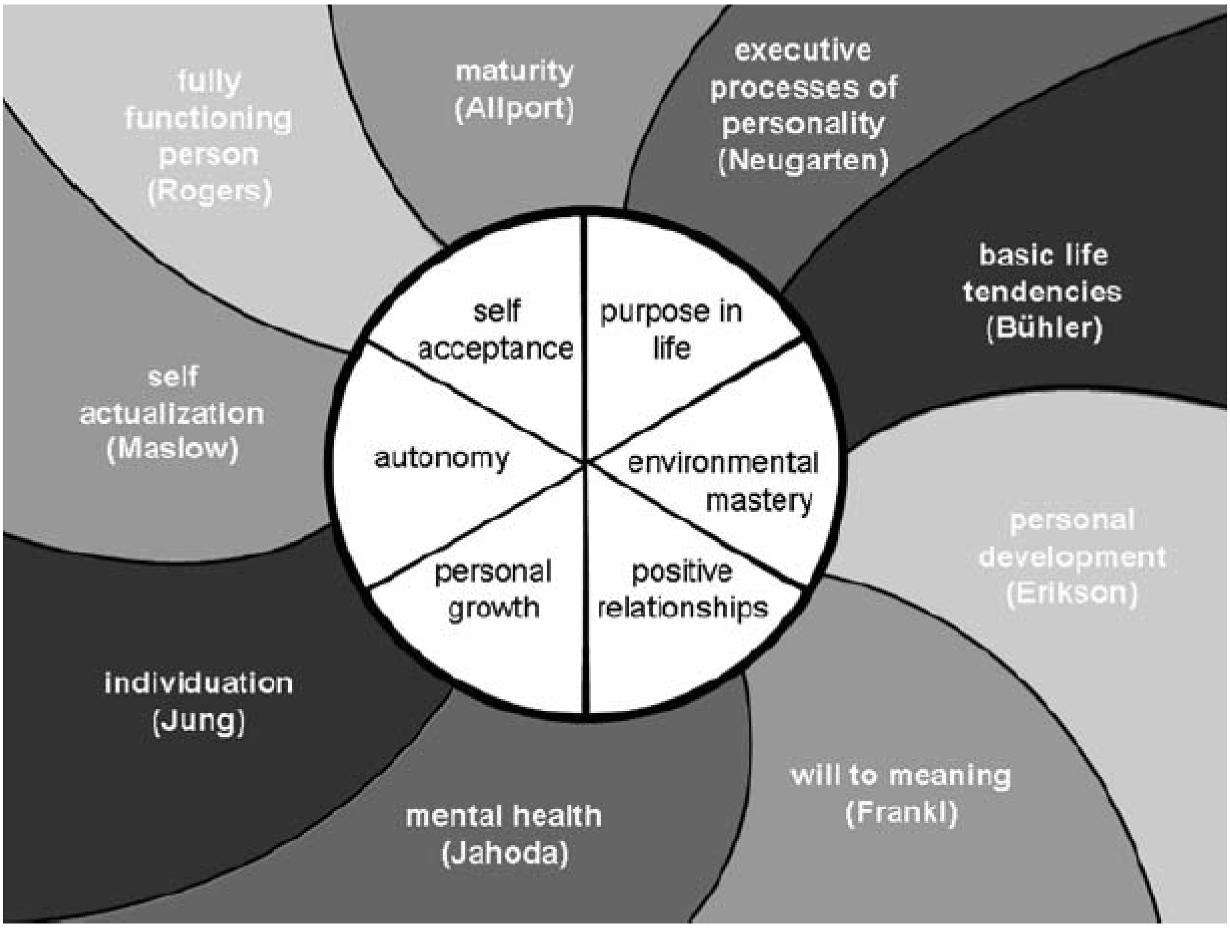

Figure 2.2: Core dimensions of psychological well-being and their theoretical foundations [75]

The first three of these factors come from the subjective well-being theory as defined by Diener et al. [20] (see Section 2.6.1). The others were selected from a multitude of other theories about psychological well-being. Combining the results from the four studies showed that hedonia influences positive affect greatly but for a short time, whereas eudaimonia somewhat causes positive affect but only after some time has passed. Practically this means that hedonia is part of self-regulating emotions and eudaimonic positive affect occurs when the result of the euadaimonic behaviour occurs such as the achievement of a goal. For negative effect the inverse relation holds as for positive affect. Carefreeness was strongly related to hedonia. Both meaning and elevating experience were strongly related to eudaimonia and less to hedonia. Huta and Ryan [39] concluded that eudaimonia and meaning are intimately linked and that eudaimonia can increase a person's baseline for elevating experiences over time. Both eudaimonia and hedonia are equally related to vitality, or the heightened sense of feeling alive. Similarly, both eudaimonia and hedonia are linked to life satisfaction, but hedonia in a greater way. From this information the authors conclude that both hedonia and eudaimonia relate to well-being in different ways. Hedonia in an affective, carefree and immediate way. Eudaimonia on a more cognitive, engaging with a broader whole, long-term level. Thus, hedonic and eudaimonic motives influence psychological well-being in a complementary way.

Henderson et al. [34] uses the study by Huta and Ryan [39] as background for their study, except they focus on hedonic and eudaimonic behaviour instead of motives. Their focus was to investigate the time people typically spend on hedonic/eudaimonic behaviours and what results come from that. It was found that people spend significantly more time on hedonic behaviours than eudaimonic behaviours. Also, there was a larger variance in eudaimonic time spent than in hedonic time spent. Furthermore, Henderson et al. [34] found that eudaimonic activities are often also experienced as hedonic, but that relationship does not work the other way around. Therefore, targeting to increase eudaimonic behaviour is likely more successful for improving a person's well-being than targeting hedonic behaviour. Hedonic activities were mostly leisure activities, like watching TV or shopping for non-essential items, chosen because they bring positive emotional results like enjoying oneself. Eudaimonic activities were often connected to something bigger than the activity itself, like long-term goals, values or spiritual beliefs. Examples of activities are studying, working, helping someone or comforting a child. These results correspond to the conclusions by Huta and Ryan [39]. They also asked for activities that were both hedonic and eudaimonic, and they found that those activities were often shared with others. From this it can be concluded that social connection and sharing experiences is a characteristic of an activity overlapping both categories. Intrinsic motivation was also an indicator of activities being both hedonic and eudaimonic. Another finding in the study is that hedonic behaviour is more effective to relieve psychological distress than eudaimonic behaviour, since hedonic behaviour is linked to vitality, positive affect, carefreeness and satisfaction. To summarize, the factors Henderson et al. [34] found to influence well-being are competence, relatedness, self-acceptance, self-esteem, engagement, optimism, having a sense of purpose and contributing to the happiness and well-being of others. And, just like Huta and Ryan [39] hedonic and eudaimonic behaviour were found to both contribute equally to flourishing.

2.4.2 Further models that combine hedonic and eudaimonic well-being

The studies mentioned earlier focus distinctly on the combination of hedonic and eudaimonic well-being. There are numerous other well-being models that also combine these two definitions of well-being, though not always stated as obviously.

One prominent model on well-being comes from an author that was mentioned before. Seligman, one of the early prominent figures in positive psychology, also writes about well-being since it is in essence the topic of positive psychology. He defines five measurable elements in a person's well-being and calls it the PERMA model. Every element of the model should contribute to a person's well-being, independently of the other elements, and be pursued for its own sake and not for gaining other elements [81]. The five elements that he defines are listed below. The descriptions are retrieved from Khaw and Kern [45, p. 4-5].

Positive emotion
: *"Positive emotion encompasses hedonic feelings such as happiness, pleasure and comfort."*

Engagement
: *"Deep psychological connection to a particular activity, organisation or cause."*

Relationships
: *"Relationships include feelings of integration with society or a community, feelings of being cared for by loved ones, and being satisfied with one's social network."*

Meaning
: *"Meaning refers to having a sense of purpose and direction in life, and feeling connected to something larger than the self."*

Achievement
: *"(...) making progress towards one's goals and achieving superior results can lead to both external recognition and a personal sense of accomplishment. Although accomplishment can be defined in objective terms, it is also subject to personal ambition, drive, and personality differences"*

Seligman [81] states that measuring *positive emotion* is purely subjective, and the four other elements have both subjective and objective components. This corresponds to the hedonic and eudaimonic definitions of well-being, where *positive emotion* is the hedonic component of the model and the other four elements together form compose the eudaimonic component. Although the definition by Khaw and Kern [45] of *positive emotion* seems to only include affect, Seligman [81] defines *positive emotion* as *"the pleasant life"* which also includes life satisfaction.

In their Self-Determination Theory (SDT), Ryan and Deci [71] expand their hedonic view on happiness to include eudaimonia and autonomy. SDT defines *"basic psychological needs, which are defined as those supports and satisfactions that are essential and necessary for psychological growth, integrity, and wellness."* [70, p. 1]. The three defined psychological needs are autonomy, relatedness and competence. According to SDT the result of fulfilling these three psychological needs will bring high well-being. Besides well-being, the three needs are also the basis for intrinsic motivation for behaviour. Practically this means that people are more likely to adopt a good behaviour when they feel attached or related to others, when they perceive themselves to be competent in doing those behaviours and when they have the autonomous choice for those behaviours.

In Section 2.4.1, it was established that activities combining hedonic and eudaimonic behaviour often include a social factor. Keyes and Lopez [44] add this category of activities as a separate dimension in their theory about positive mental health (viewing mental illness through the theory of positive psychology). They define positive mental health as the combination of high emotional well-being, high psychological well-being, and high social well-being. In the paper of Keyes and Lopez [44], the model of Ryff [73] is used to define psychological well-being, which is focused on eudaimonic well-being. For emotional well-being the terms satisfaction and happiness are used, which corresponds to hedonic well-being. For social well-being they use the model previously defined by one of the authors. Keyes [43] defines social well-being in a model with five dimensions.

Social integration
Feeling part of society and a community.

Social contribution
Believing one is a vital member of society, contributing something of value to the world.

Social coherence
Caring about the world one lives in and feeling one understands what is happening around them.

Social actualization
Being hopeful about the condition and future of society and recognizing society's potential.

Social acceptance
Believing others are trustworthy, kind and industrious.

According to Keyes [43], these five components together paint a picture of high social well-being. The whole model is simpler summarized:

> *"In short, insofar as the new scales measure social well-being, socially healthier individuals should not regard society and its custodians as unsavory, should perceive themselves as social resources, should care for and feel safe in their communities, and should lead coherent personal lives."* [43, p. 124].

2.5 | Well-being of older adults

Having discussed the general definition of well-being, it is time to look at more specific studies regarding factors influencing the well-being of older adults. Therefore, this section will dive into literature studies on well-being of older adults. The aim is to bring the specific factors into view which come into play when dealing with this user group. Like the previous sections, this section addresses subjective well-being, psychological well-being and social well-being.

Studies show that subjective well-being does not decrease with age [63]. In a study by Diener and Suh [18] it was found that from the three aspects of subjective well-being (positive affect, negative affect, and life satisfaction) only positive affect decreased with age. One explanation of the decrease in positive affect might have to do with the intensity of the positive affect variables that are often used in measures, since emotional intensity declines with age [18].

Ryff and Singer [75] have studied the influence of age on the different components of their Psychological Well-being model. Their results are shown in Figure 2.3. They found that *autonomy*, *environmental mastery* and *self-acceptance (men)* increased with age, whereas *purpose in life* and *personal growth* decreased with age. *Positive relations with others* and *self-acceptance (women)* showed little change in age groups. Therefore, Ryff and Singer [75] conclude that older adults should be provided with opportunities for personal growth and meaningful roles in order to improve their psychological well-being. One way this could be done is by doing volunteer work since this has shown to increase a person's purpose in life [29].

Baldassare et al. [5] looked into the different types of social relations and their influence on predicting well-being in the elderly population. The social relations that were considered are: subjective evaluations of social relations; objective social ties; and structural properties of social networks. The study found that social relations are an important indicator of unhappiness in older adults. The most significant social relation influencing unhappiness was found to be the perceived lack of companionship.

Pinquart and Sörensen [63] conducted a meta-analysis on the association between socioeconomic status, social network, and competence and subjective well-being in older adults. The results of the study show

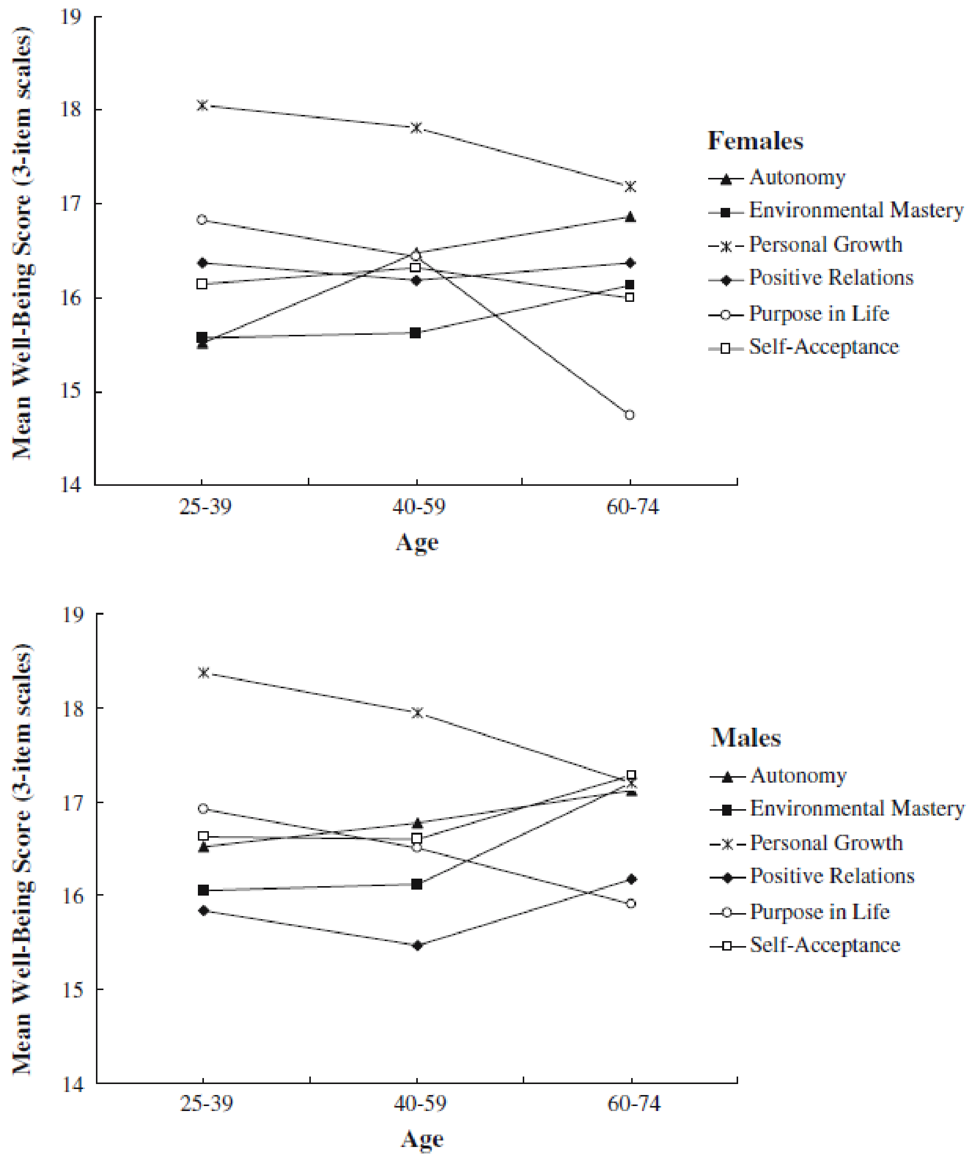

Figure 2.3: Age differences in Psychological Well-being [75]

that higher socioeconomic status positively influences subjective well-being. Regarding the influence of contact with family members and friends they found that contact with friends always positively influences subjective well-being. But contact with family members could both be positive and negative. From this leads that *"the quantity of social contact with friends was more closely related to subjective well-being than the quantity of contact with adult children."*[63, p. 197] However, looking at the quality of the contact showed that quality of contact with adult children was more important for life satisfaction than quality of contact with friends. As for competence, the study concluded that *"As long as cognitive deficits do not markedly influence everyday life (...) they are likely to have a lesser influence on subjective well-being in old age."* [63, p. 198]

2.6 | Measurements of well-being

This section will discuss the most common measurement scales for the well-being definitions mentioned in the previous sections. Even though most of the measurement scales are self-report they can still be used for measurement in conversations with a conversational agent. As mentioned by Maharjan et al. [56] conversational agents can be used as a substitute for traditional pen-paper methods for administering a questionnaire. Thus, although talking with a conversational agent might look like an interview instead of a self-report, the validated measurement tools should still be valid. As a reminder, subjective well-being is equivalent to hedonic well-being and psychological well-being is equivalent to eudaimonic well-being.

2.6.1 Subjective well-being

Various subjective well-being measurements have been developed, both for affective state and for life satisfaction. Since there have been an abundance of scales measuring almost the same things, this section presents a selection of the available scales. The scales are selected based on the well-being measures used by Diener et al. [21] to determine convergence of their new scales.

Measuring life satisfaction can be done by simply asking people how satisfied they are with their life. However, to get an honest and complete overview of someone's life satisfaction Diener et al. [19] developed a multi-item scale called *The Satisfaction with Life Scale*. This scale consists of five questions about the overall judgement of someone's life, which are rated with a 7 point measure ranging from *strongly disagree* to *strongly agree*. The sum of these questions represents the overall life satisfaction of the person. The developed scale has high internal consistency and high temporal liability.

Measuring someone's affective experience is less straightforward, since affective states fluctuate heavily through time and are subjective in essence. There are two main ways to measure the affective experience of someone's daily life, by prompting the person randomly through the day to ask about their emotions (Experience Sampling Method) or by asking them to describe their affect of the previous day (Day Reconstruction Method) [49]. Instead of the previous day, a measure can also ask about a longer period of time.

The most thoroughly analyzed well-being measure is the *Happiness Measures* [26]. These two simple questions about the quality and the frequency of someone's happiness have been used for a long time and give a good representation of someone's self-reported happiness. The first question asks how happy or unhappy someone generally feels, and the second asks about the percentage of time that someone feels happy/unhappy/neutral. The scale has been validated numerous times and shows no sex or age differences.

Lyubomirsky and Lepper [54] constructed a similar measure with more questions to allow for an assessment of internal consistency. The first two questions ask the person how happy he considers himself to be, and how happy he is in comparison to his peers. The last two questions describe a typical happy/unhappy person and ask the person to what extent the characterization describes him. All four questions are answered on a 7-digit scale. Like the previously mentioned scale, this measure has been validated thoroughly.

There are also affective measures which use larger scales. A well-known scale to measure someone's affective experience is the PANAS scale, which stands for *Positive and Negative Affect Schedule* [91]. PANAS consists of two 10-item scales, one measuring positive affect and one measuring negative affect. The scales are words describing different feelings and emotions, and the person is asked to rate to what extent he has experienced these feelings/emotions in the past time period. The rating is done on a 5-level scale ranging from *very slightly or not at all* to *extremely*. The time period can be changed between, for example, the current moment, day, week, year, or in general. The scale was found to be internally consistent and to be stable over a 2-month period. The study by Watson et al. [91] also showed that subjective well-being is sensitive to fluctuations in mood when asked over a short period of time, but is fairly stable when long-term instructions were asked, like in the past year or in general.

Similar scales have been developed with a different number and formulation of the used feelings and emotions. *The Scale of Positive and Negative Experience (SPANE)* is a measure described in the paper by Diener et al. [21] which measures positive and negative experiences over the past 4 weeks. The subject indicates how much they have experienced a certain emotion/feeling over the past month, indicating on a scale from 1 to 5. This scale was validated with a study in which a sample of students participated. This scale has been validated in Germany [67], Portugal [82], and China [52], but no validation has been executed yet for the Dutch version of this scale, or in the elderly population.

2.6.2 Psychological well-being

One author that was frequently mentioned in Section 2.3.2 and Section 2.5 is Ryff, who defined six dimensions of Psychological Well-being [73]. As a reminder, the six defined dimensions are: self-acceptance, positive relations with others, autonomy, environmental mastery, purpose in life, and personal growth. These six dimensions have been translated into assessment tools for measuring psychological well-being. To accomplish this, a high and low is defined for every dimension. Then, scales are constructed in this range to measure a person's position in a dimension. The high and low points are shown in Figure 2.4, which are the definitions by Ryff [73]. This scale is widely known as a reliable and valid tool to measure psychological well-being and there are different versions for various situations.

Table 1

Definitions of Theory-Guided Dimensions of Well-Being

Self-acceptance

High scorer: Possesses a positive attitude toward the self; acknowledges and accepts multiple aspects of self including good and bad qualities; feels positive about past life.

Low scorer: Feels dissatisfied with self; is disappointed with what has occurred in past life; is troubled about certain personal qualities; wishes to be different than what he or she is.

Positive relations with others

High scorer: Has warm, satisfying, trusting relationships with others; is concerned about the welfare of others; capable of strong empathy, affection, and intimacy; understands give and take of human relationships.

Low scorer: Has few close, trusting relationships with others; finds it difficult to be warm, open, and concerned about others; is isolated and frustrated in interpersonal relationships; not willing to make compromises to sustain important ties with others.

Autonomy

High scorer: Is self-determining and independent; able to resist social pressures to think and act in certain ways; regulates behavior from within; evaluates self by personal standards.

Low scorer: Is concerned about the expectations and evaluations of others; relies on judgments of others to make important decisions; conforms to social pressures to think and act in certain ways.

Environmental mastery

High scorer: Has a sense of mastery and competence in managing the environment; controls complex array of external activities; makes effective use of surrounding opportunities; able to choose or create contexts suitable to personal needs and values.

Low scorer: Has difficulty managing everyday affairs; feels unable to change or improve surrounding context; is unaware of surrounding opportunities; lacks sense of control over external world.

Purpose in life

High scorer: Has goals in life and a sense of directedness; feels there is meaning to present and past life; holds beliefs that give life purpose; has aims and objectives for living.

Low scorer: Lacks a sense of meaning in life; has few goals or aims, lacks sense of direction; does not see purpose of past life; has no outlook or beliefs that give life meaning.

Personal growth

High scorer: Has a feeling of continued development; sees self as growing and expanding; is open to new experiences; has sense of realizing his or her potential; sees improvement in self and behavior over time; is changing in ways that reflect more self-knowledge and effectiveness.

Low scorer: Has a sense of personal stagnation; lacks sense of improvement or expansion over time; feels bored and uninterested with life; feels unable to develop new attitudes or behaviors.

Figure 2.4: Table from [73] with the high and low points for each dimension

Another scale that is often used in measuring well-being is called the Flourishing Scale (previously known as the Psychological Well-being scale) by Diener et al. [21]. This scale combines previous theories about psychological and social well-being. Remarkable about this scale is that it captures the essence of psychological and social well-being in only 8 questions. These questions are:

- I lead a purposeful and meaningful life
- My social relationships are supportive and rewarding
- I am engaged and interested in my daily activities
- I actively contribute to the happiness and well-being of others
- I am competent and capable in the activities that are important to me
- I am a good person and live a good life

- I am optimistic about my future
- People respect me

Each question is answered on a 1-7 scale varying from Strong Disagreement to Strong Agreement. The scale was tested on a large sample of college students by the authors themselves. They concluded that the Flourishing Scale performs well in comparison with similar well-being scales. The scale has since been tested in other nations and populations to validate it. A Dutch version has been validated in a sample of adults with low/moderate levels of well-being in The Netherlands [77]. It was found that the (Dutch) Flourishing Scale *"seems a reliable and valid instrument for measuring social-psychological function in adults with suboptimal well-being"* [77, p. 1]. The reliability of the Flourishing Scale has also been validated in the elderly population in Iran [25]. In this population it was also concluded that the (Persian) Flourishing Scale is valid and reliable to use with older participants.

2.6.3 Comparison of scales

To summarize, well-being can be measured using two types of scales: subjective well-being (hedonic) and psychological well-being (eudaimonic). Subjective well-being scales can be of two kinds: life satisfaction [19] or positive/negative affect. Affective experience can be measured with scales of different sizes, like with 2 simple questions [26], 4 questions [54] or emotion frequency scales over a longer time [21, 91].

For psychological well-being two scales were discussed: the Psychological Well-being dimensions by Ryff [73] and the Flourishing Scale by Diener et al. [21]. Although these two scales are not completely measuring the same things, some overlap can be observed as shown in Table 2.1. It can be observed that the Flourishing Scale implements many factors from the Psychological Well-being dimensions. However, since the descriptions of psychological well-being are rather broad, each dimension including various sub-points, the Flourishing Scale does not fully encompass all Psychological Well-being dimensions.

Flourishing Scale item	Psychological Well-being dimension
I lead a purposeful and meaningful life	Purpose in life
My social relationships are supportive and rewarding	Positive relations with others
I am engaged and interested in my daily activities	Environmental mastery
I actively contribute to the happiness and well-being of others	Positive relations with others
I am competent and capable in the activities that are important to me	Autonomy
I am a good person and live a good life	Self-acceptance
I am optimistic about my future	Purpose in life
People respect me	Positive relations with others

Table 2.1: Comparison of Flourishing Scale [21] and Psychological Well-being [73]

2.7 | The Positive Health model compared to the discussed well-being models

This chapter started with the Positive Health model that is currently used in the BLISS project. This paragraph is focused on comparing the Positive Health model to the prominent literature about well-being. To reiterate, the Positive Health model was developed by Huber et al. [36] and redefines health as *the ability to adapt and to self manage.* For this definition health is evaluated in six pillars: *bodily functions, mental well-being, meaningfulness, quality of life, participation, and daily functioning.* Each pillar contains more specific sub-points, which can be seen in Figure 2.1.

The models from well-being literature that are used in this comparison are Subjective well-being [20], Psychological well-being [73], PERMA [81] and Self-determination theory [71]. The complete comparison can be found in Appendix A. Several sub-points from Positive Health can be recognized in the described well-being theories. The *mental well-being* pillar has strong overlap with *autonomy* and *environmental mastery* from psychological well-being. The *meaningfulness* pillar corresponds with *meaning* from the PERMA model and *life purpose* from psychological well-being. The *quality of life* pillar contains various items regarding affect, thus corresponds to *positive affect* from subjective well-being and *positive emotion* from PERMA. The *participation* pillar corresponds to *relationships* from PERMA, *positive relations with others* from psychological well-being and *relatedness* from self-determination theory. On top of that, it has strong overlap with the model of social well-being by Keyes [43].

Although there is much overlap, there are also points that are dissonant. *Achievement*, *competence* and *personal growth* have important roles in PERMA, self-determination theory and psychological well-being respectively. However, analogous terms are not found in the Positive Health model. Furthermore, many practical points from Positive Health (like *bodily functions* or *daily functioning*) are not found in other well-being theories. One could argue that these practical points help to feel autonomous or competent. However, in the theories about autonomy and competence (self-determination theory, PERMA and psychological well-being), the focus is more on choices of behaviour and not on practical necessities to be able to make choices in general. So, the thoughts behind *bodily functions* and *daily functioning* are not represented in the well-being models.

2.8 | Conclusions

This chapter dived into the literature on the philosophical and psychological theories on well-being. The aim of the chapter was to answer the first sub-question: *How can well-being be defined and measured?*. Two main philosophies on well-being were presented, hedonic and eudaimonic. Hedonic well-being is grounded in pleasure, and in psychological theories it is represented as subjective well-being, which consists of life satisfaction and experienced positive/negative affect. Eudaimonic well-being is grounded in virtue ethics and represents more profound well-being like social relationships, meaning and autonomy. For this well-being philosophy, also described as psychological well-being, there are many different psychological theories of which a few were described in this chapter. On top of that, the influence of age on these models was briefly discussed. Validated measurements of these theories were described. For subjective well-being many similar measurements were shortly discussed. For psychological well-being two measurements were discussed: the Psychological Well-being dimensions and the Flourishing Scale. The chapter concluded with a comparison of the Positive Health model with the described theories.

2.8.1 Model recommendation for the BLISS project

Since not all factors from Positive Health can be found in literature about well-being and the model does not fully include all factors from the literature, I would recommend switching from the Positive Health model to a more theoretically founded model. One of the reasons the Positive Health model was chosen by BLISS was the clear questionnaire that the conversations could be based on. Therefore, I would recommend basing the conversations on models described in Section 2.6 about the *Measurements of well-being.* This would practically mean using one of the many subjective well-being measures in combination with the Psychological Well-being dimensions by Ryff [73] or the Flourishing Scale by Diener et al. [21]. Using these models as a substitute for the Positive Health model would give both the conversational agent and the research done with it a much more sound theoretical foundation. On top of that, it would provide a clear questionnaire to base the conversations on. Conveniently, it would not change much for the current pilot implementation of BLISS since the current focus on social activities is present in almost all well-being theories.

2.8.2 Model selection for further use in this thesis

This chapter has discussed various well-being models, their philosophical backgrounds and their measurements. From the mentioned well-being measurements a combination of subjective well-being and the Flourishing Scale [21] is chosen for this thesis.

As described in Chapter 2, subjective well-being is often divided into affect and life satisfaction. Subjective well-being can be measured in various different ways which are all very similar. As mentioned before, the Flourishing Scale has been validated in Dutch [77] and for older adults [25], thus for the intended use in the BLISS project all facets have been validated. On top of that, the Flourishing Scale is, with only 8 questions, a small scale which still covers many facets of psychological well-being. The small size makes it feasible for implementing the full scale into this thesis. The Dutch version of the Flourishing Scale can be found in Appendix A.

3 | Conversation design

For the implementation of subjective well-being (see Section 2.6.1) and the flourishing scale [21] into the BLISS project, the questions of the measurements need to be converted into a conversation. As a reminder, subjective well-being measures a person's affect over a period of time and life satisfaction. The Flourishing Scale is aimed at measuring eudaimonic (long term) well-being and covers the areas of daily activities, social relationships and meaning. The questions of the Flourishing Scale can be found in Table 2.1 and in Appendix B. When converting measurement questions into a conversation, it is important to find a good balance between psychological reliability and conversational finesse. Finding this balance is the goal of this chapter.

This chapter starts with a short literature study on conversation design. Then it describes the process of designing this conversation, the design choices made, and the evaluation of the conversation. The chapter concludes with the final version of the designed conversation (see Appendix C) which is implemented in the conversational agent software and used for the data collection. All in all, this chapter aims at answering the second sub-question as defined in Section 1.4: *How to design a conversation for a conversational agent which implements the chosen well-being measures?*

3.1 | Literature background on conversation design

This section dives into speech-specific design heuristics and best practices which can be used for designing the conversation. Next, some literature on social conversations with an agent is discussed. In the conclusions of this section a list is presented with literature findings which are taken into account during the design of the conversation in Section 3.2.

3.1.1 Speech-specific design heuristics

When designing anything, a good starting point is the design heuristics for the particular field. At this time, heuristics for the development of user centered speech interactions, like conversational agents, are not established yet [14]. The most closely related field is Voice User Interfaces, or VUIs. Since this field is also relatively new, only a small amount of literature on speech-specific heuristics can be found [60]. Suhm [85] created design heuristics on telephone-based dialogue systems. These heuristics are:

1. Keep it simple
2. Carefully control the amount of spoken output
3. Word options the way users think
4. Minimize acoustic confusability of vocabulary
5. Provide carefully designed feedback
6. Abide by natural turn-taking protocol
7. Coach a little at a time
8. Yes/no queries (generally) are very robust
9. Offer alternative input modalities
10. Choose persona judiciously

More recently, Wei and Landay [93] combined the heuristics by Suhm [85], other literature with design guidelines and guidelines by big tech companies into a set of 17 usability heuristics for speech-based smart devices. The resulting list of heuristics are presented below.

1. Give the agent a persona through language, sounds, and other styles
2. Make the system status clear

3. Speak the user's language
4. Start and stop conversations
5. Pay attention to what the user said and respect the user's context
6. Use spoken language characteristics
7. Make conversation a back-and-forth exchange
8. Adapt agent style to who users are, how they speak, and how they are feeling
9. Guide users through a conversation so they are not easily lost
10. Use responses to help users discover what is possible
11. Keep feedback and prompts short
12. Confirm input intelligently
13. Use speech-recognition system confidence to drive feedback style
14. Use multi modal feedback when available
15. Avoid cascading correction errors
16. Use normal language in communicating errors
17. Allow users to exit from errors or a mistaken conversation

Murad et al. [60] comment on these heuristics from Suhm [85] and Wei and Landay [93]. They show that the proposed heuristics are grounded in fundamental Graphical User Interface (GUI) principles. Murad et al. [60] argue that any further attempt for speech-based heuristics should be kept grounded in GUI principles, because these are familiar to current usability designers and will help them to adapt the new guidelines.

3.1.2 Social conversations with an agent

The previously mentioned heuristics are applicable to every kind of speech-based user interaction. This section looks more specifically into speech-based interactions with conversational agents. As previously mentioned, heuristics specific for conversational agents do currently not exist. Therefore, this section discusses literature focused on conversational features that people find important in (agent-) conversations.

Clark et al. [15] identify two primary types of conversations: transactional (or task-based) and social (or interactional).

> *"Transactional conversation pursues a practical goal, often fulfilled during the course of one interaction. (...) The aim of more social conversation is not to complete a task as such, but to build, maintain and strengthen positive relations with one or more interlocutors"* [15, p. 2]

In their research about these two types of conversations between humans and agents they found that people conceptualize agent-human conversations almost exclusively as transactional. Consequently, it is likely that the participants of this research will also assume the conversations to be transactional. However, for the envisioned conversation about well-being it is important that the user feels comfortable and safe talking to the agent. Therefore, some research about social conversation techniques in human-agent interaction is identified to implement in the conversations for this research.

Clark et al. [15] identified characteristics people find important for the quality of human-human conversations, and how these characteristics apply to conversations with artificial agents. Attributes found important for human-human conversation are *mutual understanding & common ground*, *trustworthiness*, *active listening* and *humour*. When applied to conversational agents, the user perceived there to be common ground when the agent remembered information about the user. Trustworthiness came forward in aspects giving practical trust (not emotional), like privacy, efficiency and reliability. Active listening was perceived when the agent quickly understood the user. And finally, humour was perceived by the users as a novelty feature which could help make conversations with an agent more interesting. It should be noted that these findings were found through interviews, so the participants might have been prejudiced by their image of how a conversational agent looks and behaves.

Bickmore and Cassell [7] studied the building of trust in interactions with conversational agents. Two strategies are described to build trust: facework and establishing common ground. Facework is described as *"the positive social value a person effectively claims for himself by the line others assume he has taken during a particular contact"*[7, p. 398]. In other words, it is important to socially behave in a way that people expect of you. When this does not happen a loss of face can occur. An example of a loss of face would be when an event occurs which is incompatible with the image we wish others have of us. In speech this can occur when someone talks about a topic which has a much deeper level of familiarity than is appropriate in the current situation, like a stranger asking about your financial situation. The second strategy, establishing common ground, is mostly done by having small talk. Bickmore and Cassell [7] states that small talk can help diminish any power imbalance between an agent and a user.

3.1.3 Scope of the current research

After discussing speech-based design heuristics and literature about speech-based conversations, this paragraph considers this information in the context of this research project. Not all the heuristics can be implemented in the scope of this research. An overview of the design heuristics can be found in Table 3.1. The heuristics that can be implemented are written in black. The heuristics that cannot be implemented in this scope are red. The orange heuristics can only be implemented partially or in a simulated way. The following paragraphs will explain why these choices were made.

The focus of this research is to gather data on a person's well-being and to use this as input for a classification system. For classification and statistical analysis it is useful for the dataset to be homogeneous, or in other words to have the same agent input for every conversation. Based on this reason the choice was made to make the conversation fully scripted. This means that the agent does not reply to the input of the user and does not use the user's answers as input for a new question. Having a fully-scripted conversation means that quite a few heuristics cannot be included in the design of the conversation, like *adapt agent style to users* and *active listening* (5, 8, 13, 14 [93]; 3 [15]). Another downside of a scripted conversation is that the conversation might feel more transactional than social (as defined by Clark et al. [15]) because *common ground* and *active listening* can only be implemented in a simulated way. There are some heuristics that can only be implemented through simulated responses, like *confirm input intelligently* and *provide carefully designed feedback* (5 [85]; 6, 12 [93]; 1 [15]; 1,2 [7]). Even though some scripted replies will be implemented into the conversation, it is likely that these do not fit for all the users and loss of face [15] will happen. However, since the goal of this conversation is to gather information, and not to maximize user satisfaction, the agent utterances of the conversation are kept fully-scripted.

For the creation of the conversational agent an already existing software will be used, which is used throughout the whole BLISS project. This software created by Game Solutions Lab[1] is called WhappBot[2]. It is an interface on which agent conversations can be held, both through spoken and through written text. This research builds on conversational agents created earlier for the BLISS project. The WhappBot software has a built-in text-to-speech synthesizer with a few options to customize the voice. It also possesses its own error handling system. The user interaction with the agent is done through sending short audio recordings, otherwise known as voice memos. The user receives the agent's replies through these voice memos and can record their own by holding down a microphone button (not unlike the way voice memos are recorded in most contemporary chat software). The software runs online in any browser, so the conversation can be held anywhere at any time.

The use of the WhappBot software brings about a few limitations in the design of the conversational agent, as well as in the implementation of the heuristics mentioned in the previous sections. Firstly, since the software already has error handling built in, the heuristics about error handling (15-17 [93]) cannot be implemented in the design process. Secondly, since the text-to-speech system is already in place, the design choices that can be made about spoken language characteristics like tone of voice, emphasis and speed are limited (6 [93]). Lastly, the system status (2 [93]) and practical trust in the system (e.g. privacy) (2 [15]) are also controlled by the WhappBot software.

[1] https://gamesolutionslab.com/
[2] app.whappbot.com

Design heuristic	***Reason to not (fully) implement***
Suhm [85]	
1. Keep it simple 2. Carefully control the amount of spoken output 3. Word options the way users think 4. Minimize acoustic confusability of vocabulary 5. Provide carefully designed feedback 6. Abide by natural turn-taking protocol 7. Coach a little at a time 8. Yes/no queries (generally) are very robust 9. Offer alternative input modalities 10. Choose persona judiciously	 Only simulated Only spoken input
Wei and Landay [93]	
1. Give the agent a persona through language, sounds, and other styles 2. Make the system status clear 3. Speak the user's language 4. Start and stop conversations 5. Pay attention to what the user said and respect the user's context 6. Use spoken language characteristics 7. Make conversation a back-and-forth exchange 8. Adapt agent style to who users are, how they speak, and how they are feeling 9. Guide users through a conversation so they are not easily lost 10. Use responses to help users discover what is possible 11. Keep feedback and prompts short 12. Confirm input intelligently 13. Use speech-recognition system confidence to drive feedback style 14. Use multi modal feedback when available 15. Avoid cascading correction errors 16. Use normal language in communicating errors 17. Allow users to exit from errors or a mistaken conversation	 Controlled by software Fully scripted conversation Only simulated Fully scripted conversation Only simulated Fully scripted conversation Fully scripted conversation Controlled by software Controlled by software Controlled by software
Clark et al. [15]	
1. Mutual understanding and common ground (remembering information about the user) 2. Trustworthiness (practical trust) 3. Active listening (understanding the user) 4. humour	Only simulated Ensuring privacy, partially controlled by software Fully scripted conversation
Bickmore and Cassell [7]	
1. Establish common ground 2. Facework	Only simulated Only simulated

Table 3.1: Restrictions for the implementation of the design heuristics from literature. Black = implemented; Orange = implemented partially/simulated; Red = not implemented

3.1.4 Conclusions

This section has identified design heuristics and facets of social talk from literature that can help the design of the conversation for this research. As described in Section 3.1.3 and Table 3.1, not all the heuristics found in literature can be applied in this design process. The heuristics from literature that can be implemented are related to the persona of the agent and the text of the conversation. Similar heuristics are grouped below and labeled with a more generic term. The number citations relate to the numbering in Table 3.1, where the first number is the heuristic and the number between brackets is the reference.

Persona

Choose persona judiciously (10 [85]), give the agent a fitting persona (1 [93])

Simplicity

Keep it simple, carefully control the amount of spoken output, minimize acoustic confusability of vocabulary, yes/no queries (generally) are very robust (1, 2, 4, 8 [85]), keep feedback and prompts short (11 [93])

Writing style

Word options the way users think (3 [85]), speak the user's language (3 [93])

Turn taking

Abide by natural turn-taking protocol (6 [85]), make conversation a back-and-forth exchange (7 [93])

Start and stop

Start and stop conversations (4 [93])

Guide the user

Coach a little at a time (7 [85]) guide users through a conversation so they are not easily lost, use responses to help users discover what is possible (9, 10 [93])

Humour

Humour (4 [15])

As described in Section 3.1.3 there are also a few heuristics that are implemented in a simulated way.

Agent replies

Provide carefully designed feedback (5 [85]), confirm input intelligently (12 [93]), facework (2 [7])

Common ground

Establish common ground through mutual understanding, humour and small talk (1 [15]; 1 [7])

These design heuristics will be taken into the design process for the conversation in the next section.

3.2 | Design of the conversation

This section continues with the found heuristics from the previous section to design the conversation. The first version of the conversation was created from the well-being models with the heuristics from Section 3.1.4 in mind. How the well-being models were implemented in the first draft is explained in Section 3.2.1. The implementation of the design heuristics is discussed in Section 3.2.2. The resulting conversation was then discussed with experts in psychology, conversational agents, and communication, eventually leading to the final version of the conversation. The design of the conversations was done through an iterative process, meaning that the conversation was adapted with the feedback received after every expert conversation. The adaptations for every iteration are discussed in Section 3.2.3. All this combined leads to the final version of the conversation which can be found in Section 3.4.

3.2.1 Implementation of well-being models

The conversation is structured around the well-being models chosen in Section 2.8.2, which are described earlier in Chapter 2. As described in Section 2.8.2, the chosen well-being models for implementation in the conversations are subjective well-being, defined as affect and life satisfaction, and (the Dutch translation of) the Flourishing Scale [19] (Appendix B). For both aspects of subjective well-being, a formulation resembling the Likert scale from the flourishing model was chosen. Affect was asked through answering to the statement *"My emotions were the past week.."* with a scale ranging from *always negative* to *almost positive*. This statement is based on the Happiness Measures by Fordyce [26]. Life satisfaction was asked through a simple statement *"I'm satisfied with my life"*, with the same answers as used in the Flourishing Scale, ranging from *Fully agree* to *Fully disagree*.

The biggest challenge in the implementation of these models was the formulation of the questions. As the design heuristics from literature showed in Section 3.1.4 it is important to adapt the style of the conversation to the user, to keep it simple and understandable. Yet, the Dutch translation of the Flourishing Scale (Appendix B) uses complex formulations of the questions with words like *geëngageerd* and *capabel*. The original statements of the scale needed to be translated to conversational questions in a way that is simpler to understand but which still captures the essence of the statement. A draft for this was created and discussed with an expert to end with the best formulation (Section 3.2.3).

The questions from the well-being models are grouped into thematic subjects (in this specific order): affect, activities, social, and meaning. For every original statement one or more conversational questions are designed which capture the essence of the statement, but which are also suitable for a normal conversation. After asking these questions the original statement is presented through button input. The verification values coming from this are used during the classification as output values, meaning that the textual answers to the statement questions can be mapped to the numerical verification.

3.2.2 Implementation of literature findings

To make the conversation engaging and user friendly the design heuristics from Section 3.1.4 were taken into account for designing the conversation. This section explains for each general heuristic how it is implemented in the final conversation.

Persona

Kim et al. [46] have constructed a framework for designing agent personalities. They define three categories of personality traits: *common traits, distinctive traits* and *neutral traits*. *Common traits* are traits derived from the service domain of the agent. For this research that domain is healthcare, just as the domain in the study by Kim et al. [46] where the relevant *common traits* are identified as *"empathizing, trustworthy, submissive and smart-yet-modest"*. *Distinctive traits* are traits that differentiate one agent from another and give the agent more personality. Examples are entertaining, kind-hearted or witty. The last kind of traits, *neutral traits*, should be left abstract. Examples are things like *"race, politics, gender or economic level"*.

As recommended the *neutral traits* will be avoided in the conversation, except for the male voice of the agent as explained later. For this research the common traits of *empathizing* and *trustworthy* are the most important, since we want people to feel comfortable to share personal information about their well-being. This is implemented by scripted replies of the agent like *"Thank you for your openness"* and *"That is nicely formulated"*. The *distinctive traits* are more difficult since the population for the data collection is Dutch adults from all ages. The traits should be relatable to all the users and should thus stay quite neutral. Additional agent replies are added to the interaction so the user can relate to the agent more and common ground is created. For example, when talking about social contacts the agent shares that he sometimes finds talking with people draining and that he is relieved when they are gone. He also shares that he does not feel respected when people keep interrupting him when he talks. These bits are relatable for almost everyone and are designed to give the agent more personality. As mentioned before in Section 3.1.3, the used software has a built in text-to-speech system which means that there were only two options for the voice, male and female. The male voice was chosen because the female voice often pronounced words incorrectly. Nevertheless, the agent is given a name that can be both male and female, Marli, as to not stress its gender.

Simplicity & Writing style

The design heuristics combined into *simplicity* and *writing style* are implemented as described in the previous section about the implementation of the well-being model: the difficult words are changed to more understandable words. On top of that, longer sentences of the agent are cut into smaller voice memos to keep them comprehensible and simple. Questions are always asked at the end of the voice memo from the agent to make it clear what the user is supposed to respond to.

Start and stop

The design heuristic to *start and stop conversations* is implemented by clearly stating when the conversation has started and when it has stopped. For example, the last sentence is *"The conversation is now over. You can close this window."*

Guide the user

Guiding the user through the conversation is done by giving a small introduction on the topic of the coming questions. Like for the transition between questions about activities and questions about social contacts: *"Besides what you do in a day, I am also curious about who you do these activities with."* For some questions examples are given to help the user discover what is possible and to guide them in the right direction. For example, when asking about emotions the agent introduces examples of easily defined emotions (happy, angry, scared) and hard to define emotions (jealous, ashamed, loved). A similar situation can be found when talking about social relationships, where the examples of parents, family, colleagues, and friends are given.

Agent replies

Creating fitting agent replies is challenging to implement because the whole conversation is scripted, meaning no user input is used in the replies of the agent. Still, the agent replies are carefully designed to fit many situations. For example, the agent replies to the user's input by saying things like *"Thank you for your answer!"* and *"How nice to learn more about who you are!"*. These replies are kept very general to prevent situations where the agent says things which are unsuitable.

Common ground and humor

Just like the heuristic about personality, this heuristic is challenging to implement because it should be relatable for all Dutch adults. So very specific references to pop culture or hobbies should not be implemented. To implement this heuristic the agent sometimes talks about its (imagined) own experiences like watching cat videos and finding contact with people draining. These (imagined) experiences are kept quite general so it is engaging and understandable for all ages and subgroups of society. To add some humor, when talking about fun activities Marli shares that he likes to watch cat videos on the internet because their fluffiness makes him feel good.

3.2.3 Expert input

As discussed in Section 3.2.1, the content of the conversations is the content of the Flourishing Scale [21] and subjective well-being. Since the well-being measures do not use language that is appropriate to use in regular conversations, the statements are rewritten into other questions. These questions are designed to still capture the essence of the original question from the well-being model. To verify this, a meeting was set up with Prof. Dr. Ernst Bohlmeijer, a professor in mental well-being from the Psychology, Health and Technology department at the University of Twente. In this meeting the essence of every well-being question was discussed and changes were made to the formulation of the conversation to capture this essence. For example, the statement *I am optimistic about my future* was reconstructed as first asking about the desires someone has for their future, followed by if they believe these desires will come true. The question that we discussed the most is *I am a good person and live a good life.* Being a good person and living a good life are philosophical concepts which are difficult to translate into a conversation. Simply asking someone whether they find that they lead a good life would probably yield short answers without motivation. In the end we settled for five questions which address a person's characteristics and the influence they have on the people around them. The second version of the conversation was established by implementing the results from this meeting.

Next, this version of the conversation was discussed with the supervisors of this thesis, Dr. Mariët Theune and Dr. Jelte van Waterschoot, both specialists in the field of Conversational Agents at the University of Twente. The focus was on the establishment of agent character, appropriate agent feedback and humor. For example, it was advised to ask the name of the participant at the beginning of the conversation to establish a connection between the conversational agent and the user. Also, many replies were rewritten to make the agent more neutral and fitting for every situation. Like the situation where the agent reacts to the user telling about their activities. The reply used to be *That does sound fun!* but was changed to *How nice to learn more about who you are!*. With the new formulation it would also be fitting if people did not have many activities to tell about, or if they would reply negatively. Combining all of this feedback resulted in the third version of the conversation.

The last expert meeting was with Prof. Dr. Enny Das, a professor in health communication at the Radboud University in Nijmegen. With her expertise she identified a number of questions which would not generate open answers. Most importantly, in many cases the questions were rewritten so they started with *to what extent* (Dutch: *in hoeverre / in welke mate*), in this way users would give a more elaborate answer and it would prevent yes/no answers. The formulation of certain questions was also discussed. For example, the question *What do you think your future will look like?* was reformulated into *When you think about your future, what image do you see?*. Another question we discussed about was being a good person. It was decided to give examples of characteristics deemed as good or bad, and then asking the user how he would rate himself. With the feedback of this last expert meeting the final draft version of the conversation was created.

3.3 | Pilot test of the conversation

As described before, the initial evaluation of the conversation was done iteratively through four expert conversations in the domains of psychology, conversational agents, and health communication. With help of these conversations the final conversation was constructed, to make sure that the quality of the conversation is good. All experts gave their own kind of feedback and many questions in the designed conversation were rephrased after each conversation.

To further prevent problems in the data collection a pilot study was executed. Since the participants of the data collection can hold the conversation online at home (see Section 3.1.3 for an explanation of the software), the data collection is unsupervised with no chance of changing things halfway through. Therefore, it is vital that all the questions in the final conversation are clear and interpreted in the intended way. For this reason, a pilot test was used to test the final version of the conversation for any bugs, unclear statements or miscommunications.

3.3.1 Method

The pilot test was held with three participants who were asked to hold the designed conversation with the agent. The participants were three students from other study programs, found through personal connections. The pilot test was held in person on campus, keeping to the relevant COVID-19 regulations at that time. While holding the conversation with the agent, the participants were observed by the researcher. After the observation the researcher asked the participant four questions to verify their observations. The questions for the observation and for the semi-structured interview are listed below.

Observation questions for the researcher

- Are there places in the conversation where the user does not know how to answer the question?
- Are any questions unclear?
- Are questions interpreted in the intended way?
- Do the questions yield answers longer than a single word?

Debrief questions to ask the participant

- What did you think of this conversation?
- Which questions did you find unclear?
- Which questions did you find hard to answer? Why?
- Do you have any tips or improvements for me?

3.3.2 Results

The participants all completed the conversation. All questions were interpreted in the intended way, and there were no questions which structurally yielded single word answers. There were some questions where the participant did not know how to answer the question, or took a long time before answering. These questions were *"To what extent do your social contacts give your life meaning?"*, *"How do you think the other person experiences the contact with you?"*, *"Which emotions did you mainly feel?"*. In the debrief the participants remarked that they thought longer about these questions because they were difficult to answer. The hesitation was not caused by the formulation of the questions but by the philosophical or complicated nature of the question. One participant mentioned that it is difficult to talk about the emotions that you felt the past week when those emotions were primarily negative. Another participant commented that it is always difficult to talk about one's own good or bad qualities and one's look on the future, especially when you are in the last phase of university and the future is uncertain.

The participants spent 31, 29 and 26 minutes on the conversation, which was close to the estimated 30 minutes. The conversation with the agent was perceived as amusing, although all participants did mention that it did feel like talking to a robot. When asking more about the robotic feel the participants mentioned that it did not feel like a real conversation as they had expected, but more like an interview. The main reason for this was that the agent did not respond to the things that they said. This finding confirms that people perceive user-agent conversation often as transactional, which Clark et al. [15] concluded in their research (see Section 3.1.2). The participants' remarks also show that the perception of the participant starting the conversation did not match the actual conversation. A solution for this could be to describe the conversation as an interview. However, for the real representation of the data it is relevant that the user sees the agent as a conversational partner, not as an interviewer. So, Marli should still be introduced as a conversational partner, but the user's expectations should be lowered. Therefore, the choice was made to change the introduction of Marli by removing the sentence where he says he is a *"virtual friend"*, since *friend* implies a stronger social bond than a regular conversation partner. On top of that, in the introduction text for the data gathering the focus is placed more on data gathering and answering questions and less on it being a conversation.

Finally, some minor bugs came up in the pilot test, like an error in the consent form and a verification logging that did not come through. These bugs were fixed for the final version.

3.4 Final conversation

The process described above finally led to the final version of the designed conversation, of which the English translation is shown in Table 3.2. The original Dutch text can be found in Appendix C.

Agent	***User***
Introduction	
Hello! My name is Marli. I would like to have a conversation to get to know you better, and to learn how you live your life. What can I call you?	
	Text input
Hello <naam>, so nice to meet you!	
First a few practical points. Don't be afraid to talk too much, the more you talk the more I can learn!	
Since I'm only a chatbot I sometimes find it hard to understand people. So I would like to ask you to speak slowly and clearly.	
Sometimes I will ask you to answer a question through choosing one of the multiple choice options. This helps me to double-check your answer.	
If you don't understand a question or you find it too hard then you're free not to answer the question by not recording anything. And of course you are always free to stop the conversation. I won't get angry.	
Are all things clear?	
	Yes/no buttons
If yes was pressed:	
Good! Then we can start.	
If no was pressed:	
I'm sorry to hear that! Please contact the researchers.	
You can send an email to d.j.kwakkel@student.utwente.nl with your question.	
The conversation has ended. You can close this window.	
Affect	
Let's start the way people always start... How are you doing?	
	Spoken reply

In a week many emotions and many situations causing certain emotions take place. Some emotions are easily identified like happy, angry or sad. Others can be more difficulty to identify, like when you feel jealous, ashamed or loved. Looking at the past week, which emotions did you mainly feel?	Spoken reply
On a scale from 1 to 7, from always negative to always positive, how would you finish this sentence? My emotions were the past week...	Button verification
Some people are very happy with how their life is, others would like to see some things differently. How satisfied are you with the way your life is right now?	Spoken reply
On a scale from 1 to 7, how much do you agree with the following statement: I am satisfied with my life?	Button verification
Activities	
The following questions are about activities you do. Can you describe to me which things are a regular part of your week?	Spoken reply
Thank you for your answer! Some people say that they find their daily activities a rut, but others are interested and motivated for the things they do. How motivated are you for your daily activities?	Spoken reply
On a scale from 1 to 7, how much do you agree with the following statement: I am engaged and interested in my daily activities?	Button verification
Okay. And do you have enough activities that you like to do? Which activities are those?	Spoken reply
How nice to learn more about who you are! Maybe I'll give those activities a try, who knows, maybe they'll work for me. In my spare time I like to browse the collection of cat videos on the internet. Such a sweet soft animal always makes my day better. But back to you. I already have a good idea of what your days look like. I still wonder if you manage to fill in your days the way you would like. To what extent are you able to fill your day with activities you want to do?	Spoken reply
On a scale from 1 to 7, how much do you agree with the following statement: I am competent and capable in the activities that are important to me?	Button verification

Social	
Besides what you do in a day, I am also curious about who you do these activities with. This is different for everyone. Some people have a lot of contact with their parents, children or other family. Others mainly seek out contact with their friends or colleagues. Which people do you have contact with in a normal week?	
	Spoken reply
It is always nice to be in contact with others. Although as a chatbot I sometimes find people very exhausting. Then I'm glad when they're gone. Looking at the people you just mentioned, how much do you enjoy interacting with them?	
	Spoken reply
And which of these people can you turn to if you need help?	
	Spoken reply
On a scale from 1 to 7, how much do you agree with the following statement: my social relationships are supportive and rewarding?	
	Button verification
Besides that the contacts can give you pleasure and help, how do you think the other person experiences the contact with you?	
	Spoken reply
Why do you think others experience contact with you this way?	
	Spoken reply
On a scale from 1 to 7, how much do you agree with the following statement: I actively contribute to the happiness and well-being of others?	
	Button verification
Not all contact is always easy. Sometimes I have moments when I try to make something clear, but the other person keeps interrupting me. Or doesn't even seem to make an effort to understand me. Then, I don't feel respected by them. To what extent do you feel respected in your social relationships?	
	Spoken reply
On a scale from 1 to 7, how much do you agree with the following statement: people respect me?	
	Button verification
Meaning	
We have discussed many things that seem important from the outside. But it is also important how things feel from the inside. The social contacts we were just talking about, to what extent do they give your life meaning?	
	Spoken reply
Can you name any other things that make life worth living for you?	
	Spoken reply
That was nicely said.	

As a virtual being, my creators have given me a clear purpose. I can't change that. Fortunately, people can decide for themselves what they want to do and achieve, and why they want it. Can you give an example of a goal you have for your life?	
	Spoken reply
On a scale from 1 to 7, how much do you agree with the following statement: I lead a purposeful and meaningful life?	
	Button verification
I often think I am very funny, but others don't always agree with me on that. What do you think are good qualities you have?	
	Spoken reply
Qualities like being kind, caring or generous are often attributed to a good person. Things like being selfish, reckless or secretive are often said about a bad person. When you look at yourself, to what extent do you consider yourself a good person?	
	Spoken reply
Thank you for being open.	
You humans are social being and are continuously influencing each other and your environment. Sometimes in a positive way and sometimes in a negative way. Think back to those traits you just mentioned, in what way do they affect the people around you?	
	Spoken reply
Can you give an example of how you influence the people around you in a positive or negative way?	
	Spoken reply
Nicely said. How satisfied are you with the influence you have?	
	Spoken reply
On a scale from 1 to 7, how much do you agree with the following statement: I am a good person and live a good life?	
	Button verification
And finally, a few questions about the future. When you think about your future, what image do you see?	
	Spoken reply
What wishes do you have for your future?	
	Spoken reply
How much confidence do you have that those wishes will actually come true?	
	Spoken reply
On a scale from 1 to 7, how much do you agree with the following statement: I am optimistic about my future?	
	Button verification
Wrapping up	
That was it. Thank you for all your answers and your openness!	

If you found any of these questions difficult and you would like to talk to someone, you can click on the following link. https://www.rijksoverheid.nl/onderwerpen/geestelijke-gezondheidszorg/vraag-en-antwoord/waar-vind-ik-hulp-bij-psychische-problemen. It was nice to meet you! Enjoy the rest of your day! The conversation is now over. You can close this window.	

Table 3.2: Text of the final conversation

3.5 | Conclusions

This chapter answered the second sub-question from Section 1.4: *How to design a conversation for a conversational agent which implements a well-being measure?* To answer this question, first the relevant literature on speech heuristics and social agent conversations was looked into. The used application and fully-scripted design of the conversation restricted the use of a few heuristics. This literature study resulted in several relevant heuristics to take into account while designing the conversation, like *"Adapt style to the user"* and *"start and stop conversations"*. Then, with these heuristics in mind, the conversation was designed. First, the implementation of the well-being models and design heuristics was explained which were used to construct the first version of the conversation. Then the expert interviews were discussed which each led to an adapted version of the conversation. The final conversation was evaluated through a pilot test. The results from the pilot test were mostly very positive. The only negative feedback was that it felt more like an interview than a conversation. The introduction of the experiment was adapted and some minor changes were made to the conversation to reduce this effect. This chapter concluded with the final conversation.

4 Data collection and analysis

This chapter is aimed at answering the third sub-question: *What useful information can be extracted from a dataset gathered with the conversation that was designed?*. To answer this question, first in Section 4.1 the method of the data collection is explained, including the design choices, personas used, software implementation, and structure of the dataset. Next in Section 4.2, characteristics of the original dataset as a whole are discussed, such as the size and the distribution of the dataset. The characteristics of the answers to each statement in the original dataset are analyzed in Section 4.3. The dataset gathered with the personas is analyzed separately from the *"original"* dataset in Section 4.4. This was done because the quality of the data from the personas experiment did not correspond with the quality of the original dataset, which is further explained in that section. This chapter concludes with the answer to the sub-question in Section 4.5.

4.1 Data collection

The designed conversation from Section 3.4 was actualized in the conversational agent software and used to gather a dataset. This section starts with discussing the implementation in the software, together with more details on the interface and user interaction. Afterwards, the data collection method is discussed, followed by the personas used in the data collection for the additional experiment. Lastly, this section describes the creation and structure of the final dataset, and the system used to reference a statement and its questions.

4.1.1 Implementation in the software

As already briefly explained in Section 3.1.3, the data collection was done using the software used previously in the BLISS project. This software is called WhappBot and makes it possible to program conversations with a conversational agent. The conversation was programmed in Structured Conversation Language (.scl), which is a kind of mark-up language. The code for the conversation was based on earlier conversations made for the BLISS project. The interface, text-to-speech synthesis and transcription are handled by the WhappBot software.

In the WhappBot software, users can have a conversation with the conversational agent by recording voice memos. The interface of the software before recording and during recording are shown in Figure 4.1 and Figure 4.2. The white voice memos on the left are the replies from the agent, the green memos on the right are from the user. The user starts a voice recording by clicking the blue button in the interface shown in Figure 4.1. The recording can be deleted using the button with the red cross, or send using the green button, as can be seen in Figure 4.2. For the verification questions, verification buttons are used instead of voice memos. An example of this interface is shown in Figure 4.3, where each blue square is a separate button.

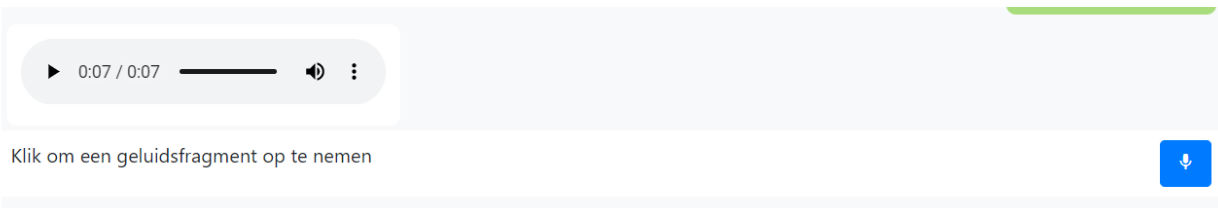

Figure 4.1: WhappBot interface before recording
translation: Press to record audio

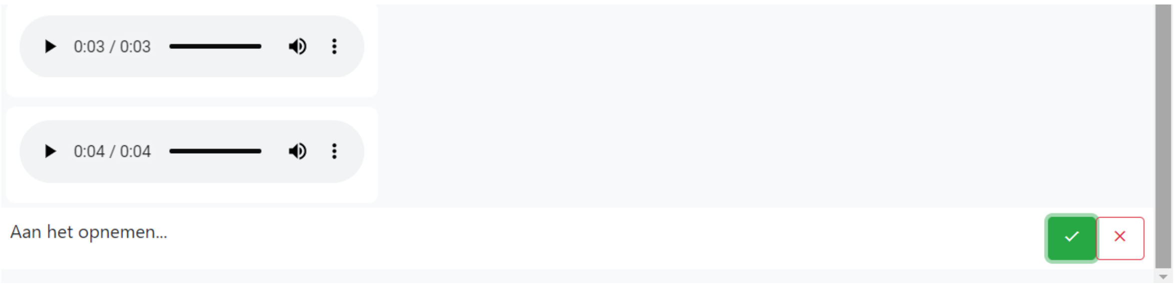

Figure 4.2: WhappBot interface while recording
translation: Recording...

Figure 4.3: WhappBot interface during a verification question
translation: I am engaged and interested in my daily activities.
1. Strongly disagree; 2. Disagree; 3. Slightly disagree; 4. Neither agree nor disagree; 5. Slightly agree; 6. Agree; 7. Strongly agree

4.1.2 Method of collection

The implementation of the conversation in the WhappBot software makes it possible to hold the conversation online on any device. The data gathering was set up for people to hold the conversation with the agent by themselves anywhere and anytime. Participants were contacted through personal networks and social media. Since the conversation was about well-being, resulting in very personal data, effort was put into the prevention of suffering and the protection of the participant. For this, extensive ethical approval was acquired by the ethical board of the EEMCS faculty at University of Twente. The participant was first send to a Microsoft form containing the (shortened) information brochure and consent form. Once they consented to all the statements they were given the link to the conversational agent. The information brochure and informed consent forms can be found in Appendix E. The form did not gather any information about the user, so the conversation was mostly anonymous. As mentioned earlier, the participant was asked their name in the beginning of the conversation to establish trust, but the given name was not transferred into the dataset. More information about the creation and structure of the dataset is given in Section 4.1.4.

In the end 35 people participated in the data collection. Initially the experiment was aimed at gathering more participants, but due to it being difficult to find participants and some technical problems this result was not realized. The experiment was closed with only 35 participants due to time constraints of this thesis.

When analyzing the gathered data, it became clear that the dataset was very skewed towards the positive verification replies (*slightly agree* and *agree*), with hardly any data entries for the lower verification replies. A reason for this could be that people were more likely to talk about their well-being when they felt good than when they were having a bad day. When discussing this with friends and family, often people mentioned that they waited until they felt well before participating in the experiment. For classification based on this dataset a more balanced dataset is needed, which includes data entries for the lower well-being levels. To accomplish this, an additional experiment was set up which used personas.

4.1.3 Experiment with personas

To gather data entries with lower verification replies (negative responses to the Likert scale), a additional experiment was set up which used personas of people with lower well-being. The participants were asked to hold the conversation in the role of the person from the persona. Every participant could choose one of the personas or fill it in as themselves, whatever they were most comfortable with. The participants for this experiment were gathered through a system where students from the bachelor of Psychology get credits for participating in research. Six different personas were created, whose stories were based on personal experiences and stories from friends and family. The images used in the personas were AI-generated[1] (the images can be seen in the personas document presented to the participants in Appendix D). The six personas that could be chosen were:

- **Wilke Hendriks, female, 43**
  Wilke is a woman her friends and family would call caring and social. Yet it is also clear that she has a lot on her plate. After her divorce 4 years ago, the care of the 3 children has completely fallen on her. Although she loves doing this, it is difficult to combine it with her full-time job at an insurance company. Because life is already so full, she feels she has little space for her own needs and has lost sight of her own desires. There is also hardly any time in her calendar for her hobbies and friends.

- **Wout Blankema, male, 39**
  Wout has had his life on track for a while. The children are almost grown up, his job is steady and his marriage is going well. In fact, every day is the same: getting up early, reading the newspaper, going to work, chatting with colleagues, returning home at 5 o'clock where dinner is ready, and watching TV in the evening on the couch. In the last few months, Wout has started to think more and more about his life. He became a father when he was 20 and in the years that followed it feels like his life was lived for him. Now he sits on the couch and wonders if this is it. Everything in his life is going well, but does he want this life? Where has the hope and energy gone that overflowed him when he was younger?

- **Steffan Molenaar, male, 34**
  Steffan worked as a nurse in the hospital until he became ill in the first weeks that Corona came to our country. Now, 2 years later, his symptoms are called *long covid*. For Steffan this mainly manifests itself through fatigue and shortness of breath. Where, before he got ill, he enjoyed walking or exercising all afternoon, he is now exhausted after only 20 minutes. Due to these fatigue complaints, he cannot go to work. Because Steffan lives alone, he also does not have much contact with others and feels like life has been put on hold. The doctors cannot say how long this will last.

- **Nicole van der Velden, female, 31**
  To outsiders, Nicole seems like a very ordinary woman. She has been working as an administrative assistant for several years, chatting with colleagues and sometimes meeting with friends. But not many people know that, for a number of years now, she has also been suffering from depression. As a result, Nicole gets no joy or energy from the things she does. She struggles to get herself out of bed in the morning. Often without the energy to eat breakfast, she goes to work on an empty stomach. When she comes home she flops back on the couch, where she fills her time with playing video games and watching TV-series. She does not fall asleep until early in the morning, because only then she is tired enough to dispel all the thoughts that vex her throughout the day.

- **Brechje Maas, female, 68**
  Brechje taught for years at a secondary school not far from her home. She enjoyed her days surrounded by colleagues and students. Since she retired a year ago she has fallen into a hole. Where her days used to be completely filled, now they are almost fully empty. With a few hobbies and a bit of messing around at home, she can fill some hours a day. But the remaining hours she usually spends waiting for the evening so she can go back to sleep. She also no longer has the social contact she had at work. She misses her husband, who passed away 20 years ago, more and more. Only when you no longer have it do you realize how full your life was, and how empty it can be.

[1] https://this-person-does-not-exist.com/en

- **Tom van Dijk, male, 20**
  Tom is a second-year psychology student and is enjoying his studies very much. Only navigating his life is a lot more difficult for Tom than for most people. Tom suffers from anxiety and perfectionism. As a result the smallest things cause a battle in his head. Handing in something for a grade, asking people for help, unfamiliar situations... all these situations cause Tom hours of stress. It's like he's dragging a big concrete rock behind him and it takes him a lot of effort before something is done. All these fears often leave Tom feeling exhausted, his social contacts and school results suffering as a result.

4.1.4 Dataset creation

The data from the finished conversations could be downloaded from the WhappBot server. The data was saved in audio files and a JSON file with data entries for each audio message containing a time stamp, the name of the audio file, and other information that the programmer could add during the conversation design. For this conversation, information about the well-being statement (1-10 for each statement in the measures from Section 3.2.1) and the question that the audio file belongs to were added. The precise statements with the corresponding statement number and question number are further explained in Section 4.1.5. The provided JSON file was quite extensive, cluttered and contained lots of unimportant information. To make the data more accessible, a Python script was written which changes every user conversation into a separate CSV file. This CSV file contains six columns:

- **id**: the id of the conversation, generated by the software
- **statement nr**: the statement number of the statement that the question belongs to
- **question number**: the number of the question about the statement
- **audio id**: the id of the audio file that was recorded by the user
- **text**: the text transcribed from the audio file
- **verification**: the selected answer to the verification question (1-7 representing *strongly disagree - strongly agree*)

The software provided by BLISS has built-in automatic speech recognition, however it was found that the speech recognition contained a lot of errors and did not work when the audio contained background noise. Therefore, the Python library *SpeechRecognition*[2] was used to get a more accurate transcription from the audio files. This transcription is also saved in the CSV file. An example of a data entry for the two questions regarding statement 3 is shown in Table 4.1.

id	**statement nr**	**question nr**
RECGo5q0ynaC5LUsOGRfiVIrZ2M2	3	1
RECGo5q0ynaC5LUsOGRfiVIrZ2M2	3	2
audio id	**text**	**verification**
623477e088c73a40698e9766	This is the transcribed answer to the first question	5
623e1754820cbcb4d177c3d3	This is the transcribed answer to the second question	5

Table 4.1: Example of a data entry

[2]https://pypi.org/project/SpeechRecognition/

4.1.5 Well-being statements belonging to the statement ids

For convenience, every well-being statement is represented by a statement number, and the corresponding questions from the designed conversation (see Section 3.4) are also given a number. These statement numbers will be used in the data analysis described in this chapter, and the classification in the next chapter. Here is a list describing which statement is represented by what number, and what abbreviation will be used when referring to the statement. The abbreviations correspond to the categories of the statement as described in Section 3.2.1 and Section 3.4: SWB = Subjective well-being; ACT = Activities; SOC = Social; MNG = Meaning.

1. **My emotions were the past week (SWB-1)**
   - 1.1 How are you doing?
   - 1.2 Looking at the past week, which emotions did you mainly feel?
2. **I am satisfied with my life (SWB-2)**
   - 2.1. How satisfied are you with the way your life is right now?
3. **I am engaged and interested in my daily activities (ACT-1)**
   - 3.1. Can you describe to me which things are a regular part of your week?
   - 3.2. How motivated are you for your daily activities?
4. **I am competent and capable in the activities that are important to me (ACT-2)**
   - 4.1. Do you have enough activities that you like to do? Which activities are those?
   - 4.2. To what extent are you able to fill your day with activities you want to do?
5. **My social relationships are supportive and rewarding (SOC-1)**
   - 5.1. Which people do you have contact with in a normal week?
   - 5.2. Looking at the people you just mentioned, how much do you enjoy interacting with them?
   - 5.3. And which of these people can you turn to if you need help?
6. **I actively contribute to the happiness and well-being of others (SOC-2)**
   - 6.1. How do you think the other person experiences the contact with you?
   - 6.2. Why do you think others experience contact with you this way?
7. **People respect me (SOC-3)**
   - 7.1. To what extent do you feel respected in your social relationships?
8. **I lead a purposeful and meaningful life (MNG-1)**
   - 8.1. The social contacts we were just talking about, to what extent do they give your life meaning?
   - 8.2. Can you name any other things that make life worth living for you?
   - 8.3. Can you give an example of a goal you have for your life?
9. **I am a good person and live a good life (MNG-2)**
   - 9.1. What do you think are good qualities you have?
   - 9.2. When you look at yourself, to what extent do you consider yourself a good person?
   - 9.3. Think back to those traits you just mentioned, in what way do they affect the people around you?
   - 9.4. Can you give an example of how you influence the people around you in a positive or negative way?
   - 9.5. How satisfied are you with the influence you have?
10. **I am optimistic about my future (MNG-3)**
    - 10.1. When you think about your future, what image do you see?
    - 10.2. What wishes do you have for your future?
    - 10.3. How much confidence do you have that those wishes will actually come true?

4.2 | Characteristics of the original dataset

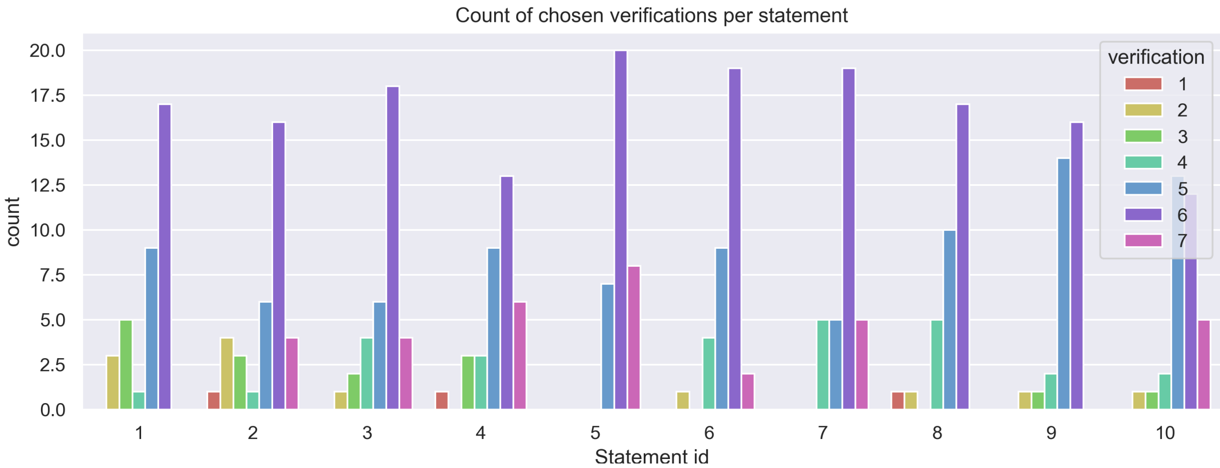

Figure 4.4: Distribution of chosen verification in the dataset

In this section the original dataset, meaning the dataset without the data from the additional experiment with personas, is analyzed. The original dataset consist of 35 answers from people about their own well-being. During one of the conversations the software crashed, so for statements 7-10 there are only 34 data entries. The data entries used are the automatically transcribed text of the participants' recorded audio files. This has influence on the data quality because the speech-to-text algorithm sometimes makes mistakes and does not add punctuation. So sometimes words do not make sense in a sentence, which is probably a transcription error. And the longer answers are long pieces of text which sometimes makes it difficult to easily see the meaning when reading them. Nevertheless, the quality of the transcribed text is generally quite good and these errors do not occur very often. And when they occur it is often possible to figure out the true word by reading the context and searching for phonetically similar words.

As briefly mentioned before, the distribution of the verification values is very skewed, which can be seen in the plot in Figure 4.4. This figure shows the number of times a certain verification value was chosen for each well-being statement. It becomes clear that especially verification value 6 is highly over-represented, and verification 1-4 have hardly any data entries. It also shows that for some statements there are more low verifications than for others. The mean and standard deviation for every statement are shown in Table 4.2. Note that the first two statements represent subjective well-being and the last eight are from the Flourishing Scale.

The over-representation of positive answers is supported by literature. Similar values were found in earlier studies validating the Flourishing Scale. In the study by Silva and Caetano [82] who validated the Flourishing Scale in Portugal, they found that: *"The mean values of the items ranged from 4.81 to 5.93, suggesting that all participants have positive perceptions of themselves in the main areas of their functioning"* [82, p. 473]. For these measurements the standard deviations were 0.76 and 0.60 respectively in the two studies. Fassih-Ramandi et al. [25] validated the Flourishing Scale in Iran and found the mean values of the statements to be between 5.60 and 6.41 with standard deviation between 1.12 and 1.61. The Dutch validation of the Flourishing Scale by Schotanus-Dijkstra et al. [77] found a mean of the individual items between 4.7 and 5.5 with an average standard deviation of 0.81. Comparing the gathered dataset to the values found in literature it shows that the skewing of the data is expected. With mean values between 5.14 and 6.02 the results are in line with [25], but significantly more positive than [82] and [77]. That the gathered data is more positive could be explained by the fact that people with lower well-being did not want to participate in the study. Either because they were warned that the conversation with the agent could be triggering when the participant had low well-being, or because they didn't want their low well-being audio files to be heard by the researcher whom they knew personally.

statement	abbreviation	mean	std
1	SWB-1	4.914286	1.379928
2	SWB-2	5.028571	1.688816
3	ACT-1	5.428571	1.195229
4	ACT-2	5.342857	1.370763
5	SOC-1	6.028571	0.663578
6	SOC-2	5.457143	0.980482
7	SOC-3	5.705882	0.905519
8	MNG-1	5.147059	1.184044
9	MNG-2	5.264706	0.931237
10	MNG-3	5.441176	1.106213

Table 4.2: Mean and standard deviation of the different statements in the dataset

4.3 | Data analysis per well-being statement

This section will dive deeper into the answers of every well-being statement. For every statement, the answers are analyzed, looking for information which tells us something about the filled in verification. As a reminder, the verification statement was asked after the questions belonging to the well-being statement by means of buttons. These buttons represent the 7-step Likert scale from the Flourishing Scale [21] (see Appendix B), ranging from *1. Strongly disagree* to *7. Strongly agree.* So, this section looks for information in the spoken responses which tells something about the filled in verification. At the end of this section the data analysis is discussed and conclusions are drawn.

4.3.1 My emotions were the past week (SWB-1)

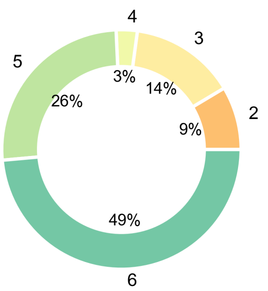

Figure 4.5: Verification distribution for statement 1

This first statement measures the affect-component of a person's subjective well-being (see Section 2.6.1). This is done through two questions *"How are you doing?"* and *"Looking at the past week, which emotions did you mainly feel?"*. The distribution of the verification values for this statement is shown in Figure 4.5. Half of the participants indicated that their emotions the last week were *mostly positive* (verification 6). Nobody indicated their emotions to be *always negative* or *always positive* (verification 1 and 7). For the other verification values there are data entries, ranging from 3% for verification 4, to 26% for verification 5.

For the sake of the conversation flow the first question is an introductory question, which was not meant to gather important information. This was confirmed by the answers to the first question, where most participants replied with some version of *I'm doing well (Het gaat goed)*. Some participants added some information on their life like having a holiday or being stressed because of a busy week. But since this is a pleasantry question in Dutch, we need to look at the second question for the true information about their well-being.

The second question asks which emotions someone has felt the past week. Almost all responses mention the emotion of happy (blij), regardless of the verification. But also negative emotions are mentioned both with lower verifications and with higher verifications. For example, people who indicated verification 6 also mention, more than once, feeling *irritated*, *frustrated*, and *sad.* From this data it can be concluded that the real important information about a person's emotions the past week is missing in the formulation of the question, namely the intensity or the duration of these emotions. Looking back at the subjective well-being measurements from Section 2.6.1 it indeed is evident that most scales not only ask about certain emotions, but also how often someone experienced that emotion. The scales most similar to this study were created by Watson et al. [91] and Diener et al. [21], since these also ask for emotions over a past time period. These scales ask people to indicate for a list of emotions how much they experienced that emotion, ranging from *Not at all/never* to *extremely/always*.

Learning from this data it is recommended that if a conversational agent wants to gather information about the affect-component of someone's subjective well-being, an additional question should be added which measures the duration of these emotions. An example could be *"Which of the emotions that you just mentioned did you feel the most often?"*. Or the participant could be asked to mention their emotions in the order of most frequent to hardly present.

4.3.2 I am satisfied with my life (SWB-2

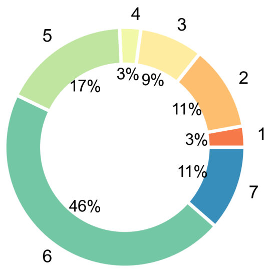

Figure 4.6: Verification distribution for statement 2

The second well-being statement measures a person's life satisfaction, which is part of the subjective well-being theory. Like for the first statement, verification 6 makes up (almost) half of the answers. Like the first statement, there was no one who answered with verification 1 or 7. The verification distribution is shown in Figure 4.6.

This well-being statement could be directly implemented in the conversation without the formulation being too vague or too formal. This is fortunate for the well-being extraction, since the responses are to the original formulation of the Flourishing Scale. Two types of answers can be distinguished from the data. Firstly, there are short answers that simply say something along the lines of *I am very satisfied (ik ben heel tevreden)* or *I am not satisfied (ik ben niet zo heel tevreden)*. Secondly, there are people who elaborate on why they are or are not satisfied. Responses that are given multiple times are *I can't complain but some things could be better* and *I'm surrounded by nice people and I'm enjoying my study.*

Regarding automatic information extraction, it is recommended to look at the adjectives for the word *'satisfied*. Someone's well-being regarding this statement can probably be extracted from the spoken answers quite well by looking at these adjectives. For example, *not (niet)* means low well-being, *somewhat (redelijk)* corresponds to medium well-being, and *very (heel erg)* means high well-being. And example of a research that used such a vocabulary system in English was done by Touset [89].

4.3.3 I am engaged and interested in my daily activities (ACT-1)

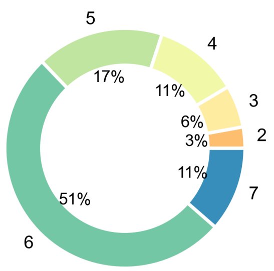

Figure 4.7: Verification distribution for statement 3

Statement three is the first statement from the Flourishing Scale [21] and measures how engaged and interested people are in their daily activities. There are two questions representing this statement. The first question, *"Can you describe to me which things are a regular part of your week?"*, is the introductory question into the topic. The second question, *"How motivated are you for your daily activities?"*, asks about the person's engagement with those activities. The verification distribution for this statement can be found in Figure 4.7. Again, verification 6, representing the option *agree (mee eens)*, makes up for half of the responses. Followed by verification 5 (17%), 4 (11%) and 7 (11%), 3 (6%) and 2 (3%).

The first question asks about the participant's activities which he/she does throughout the week. Activities that are often mentioned are work/study, sports like running and walking, meeting with friends, watching series, and household chores like grocery shopping and cleaning. Since this is the introductory question, the responses do not indicate much about the engagement of the participant for these activities. Both low, neutral and high well-being answers mention the same kind of activities. An activity that is only mentioned in the lower well-being answers is therapy appointments, unfortunately there are not enough low data points to draw valid conclusions at this point.

The second question asks about the person's motivation for their daily activities and is expected to give information about their well-being level. There are some differences from which you can recognise a low/neutral/high answer. The answers with lower well-being talked solely about lacking motivation. The neutral answers often say *that depends*. The higher well-being answers are often longer and mention more about why they are motivated. The higher answers often also mention that they enjoy the work they do. Even though these differences between the low/neutral/high answers exists, there is also a lot of overlap between the categories. For example, there are quite a few higher well-being answers which say that they are not very motivated for some things but motivated for others.

For information extraction, the answers to the first question could be useful. Although this question does not give information about the well-being level of the participant, it does give lots of information about the person. The activities mentioned in these answers could be extracted and used in later conversations, for example to establish common ground with someone (see Section 3.1.2). The second question could be used for classification,since there is some distinction between the different verification values. However, the responses are very diverse in language use which makes automatic classification a challenge.

4.3.4 I am competent and capable in the activities that are important to me (ACT-2)

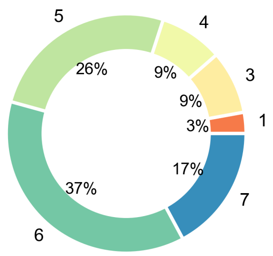

Figure 4.8: Verification distribution for statement 4

The fourth statement asks about the competence of people to do the activities that are important to them. This statement is covered by two questions: *Do you have enough activities that you like to do? Which activities are those?* and *To what extent are you able to fill your day with activities you want to do?* Just like for the previous statement, the first question is the introductory question and the second question provides the information that is relevant to the original well-being statement. The distribution of the verification for this statement, shown in Figure 4.8, is slightly more balanced than the previous ones, with 37% for verification 6, 26% for verification 5 and 17% for verification 7.

Although the first question only introduces the topic and does not directly target the user's well-being, it does tell a lot about the participant. All given answers include multiple activities which the participants liked to do. Frequently mentioned activity categories are: music, social contacts (friends, family, student associations), gaming, reading, sports, traveling, making puzzles, being creative (drawing, making videos), watching tv-series, walking outside, baking, and school/work. There is no difference in the answers for high, neutral or low well-being, which shows that people who have trouble executing the activities that they enjoy still have enough activities that they like to do.

The second question gives insight into the capability of people to do the activities they just mentioned. The answers of the people who indicated that they felt (very) capable (verification 6 and 7) are short and generally say that they are capable and enjoy the things they do. The answers with the low and neutral verifications generally give more information. People indicate that they are not capable because they are too tired, they lack motivation, or they feel like there are not enough hours in the day for the things they would like to do. Something that is also said multiple times is that they feel like their days are planned full without their influence, for example with school, work or other mandatory activities.

These two questions provide information that could be useful for later conversations with agents or care professionals. The answers to the first question give insight in a person's activities, and could be used as subjects to talk about in later conversations with agents and help establishing common ground (see Section 3.1.2). The negative answers to the second question could be useful for care professionals since it gives the reasons why someone is not capable to do things they would like to do. It could be useful to extract this information for care professionals and to ask more in depth about the reasons why people are not motivated, too tired, or their days are planned without their influence.

4.3.5 My social relationships are supportive and rewarding (SOC-1)

This statement is the first of the statements about social contacts and asks the participants if their social relationships are supportive and rewarding. Same as before, it starts with an introductory question: *Which people do you have contact with in a normal week?* Then the agent first asks *how much do you enjoy interacting with them?*, followed by *Which of these people can you turn to if you need help?* Interestingly, as can be seen in Figure 4.9, only the highest three verifications were chosen, which represent *slightly agree, agree* and *strongly agree*. This means that everybody found their social relationships (to some extent) supportive and rewarding.

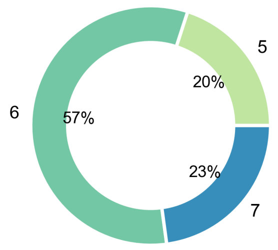

Figure 4.9: Verification distribution for statement 5

To answer the first question, the participants spoke about the people they come in contact with in a normal week. Frequently mentioned are family members, friends (sometimes with an explanation of how they know them), colleagues, neighbors, and people from church. This question does not indicate much about the degree of someone's well-being, except when someone would not have anyone to mention, which did not occur in this dataset. The given information to this question identifies the network of the person and could be used in future conversations with a care professional or with the agent. Furthermore, if the answer to this question would change over time, for example by mentioning less social contacts, it could indicate a change in the person's well-being.

The second question asks how much people enjoy interacting with the people they just mentioned. All responses were some form of *I enjoy the contact (ik haal er plezier uit)*. Most people explained with whom they enjoy the contact, or why the contact with people make their life better. Examples of responses to this question are: *"I enjoy it very much, being alone is boring (heel veel plezier anders is het maar een saaie boel in je eentje)"* and *"Sometimes I like being with others but sometimes I'd rather be alone (wisselend, meestal vind ik het wel fijn maar soms ben ik ook gewoon even liever alleen)"*. Frequently, people mentioned that they sometimes enjoy contact and sometimes they don't. This occurred more often in the answers of the people who indicated a verification of 5 (neutral). However, since there are only positive verifications (even though 5 is classified as neutral it still indicates a positive answer), automatic classifications would be difficult for this question.

For the third question people indicated who they can turn to when they need help. The most frequently mentioned people are family or parents, partner, and friends. These answers were the same for all the verification values. Even though this question cannot be used for automatic classification, there is a lot of useful information in these answers that can be used in future conversations or by care professionals. For example, if it is noticed that someone is not doing well, it could be suggested to contact the person that they have mentioned to the third question.

4.3.6 I actively contribute to the happiness and well-being of others (SOC-2)

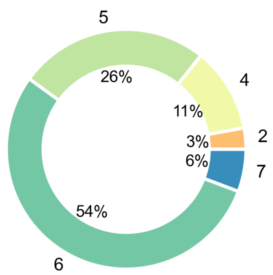

Figure 4.10: Verification distribution for statement 6

This statement asks how people think others experience them. Firstly, it is asked how the participant thinks others experience the contact with him or her. Afterwards, it is asked why they think this. To this question, a little more than half of the participants indicated verification 6 (see Figure 4.10). Also verification 5 (26%) and 4 (11%) are chosen quite a few times.

In the answers to the first question most people indicate that others think contact with them is fine, enjoyable or nice. It is difficult to find differences in the answers between the low/neutral/high verifications. The only entry with low verification mentions that they feel like a burden to others. Unfortunately there is not enough data to draw conclusions about low well-being in this category. In the neutral verifications (especially for 4) it seems like people are less sure about their answers. They mention things like *"I think it is fine (ik denk wel prima)"* and *"I hope it is good (ik hoop van wel)"*. The higher verifications generally show less doubt and are certain that at least their friends and family enjoy contact with them.

To the second question the participants answered with characteristics of themselves that they think others liked or disliked. Like, *being inconsistent in contact* as negative or *being social and friendly* as positive. Frequently mentioned is also that people keep seeking contact with them, so they must like it. It stands out that the really high verifications (7) often mention that people have told them that they enjoy contact with them. So it can be concluded that having people tell you that they enjoy contact with you helps to believe it. It is difficult to distinguish between answers for the different verifications, and the answers are very diverse. Therefore, automatic classification or information extraction would be difficult.

4.3.7 People respect me (SOC-3)

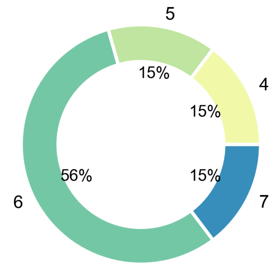

Figure 4.11: Verification distribution for statement 7

This last statement about social relationships asks people how respected they feel in their social relationships. As shown in Figure 4.11, the answers ranged from verification values 4 to 7, with verification 6 accounting for more than half of the answers. This means that all the participants felt neutral or respected in their social relationships.

Interestingly, in all the answers, except the ones that indicated a 7, there are people who indicated that they feel respected in some relationships and not respected in others. It is mentioned that they feel most respected in the relationships with friends and family. The neutral verification values, especially the ones who indicated 4, generally gave longer answers. Here participants indicated that sometimes in new situations they feel like they are not heard, or they don't always feel understood. Since the replies do not differ much between the different verifications, it would be difficult to automatically classify the verification level from the given text. It could maybe be done by looking at the word *respected* and see what adjectives are used to describe this word. Although not all replies used the word *respected* in their answer, and some used synonyms like *valued (gewaardeerd)* instead.

4.3.8 I lead a purposeful and meaningful life (MNG-1)

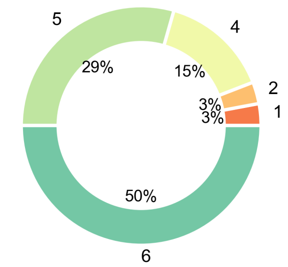

Figure 4.12: Verification distribution for statement 8

Statement 8 is the first of the three statements about meaning of life. This statement is covered by three questions: *"To what extent do the social contacts in your life give meaning?"* as a bridging question, *"Can you name any other things that make life worth living for you?"*, covering the word *meaningful* and *"Can you give an example of a goal you have for your life?"*, covering the word *purposeful*. The most common verification given is 6 *(agree (mee eens))* which is given by exactly half of the participants (see Figure 4.12). Verification 5 *(slightly agree (een beetje mee eens))* was given 29% of the time, 4 *(neither agree nor disagree (noch mee oneens noch mee eens))* 15% of the time and 2 *(disagree (mee oneens))* and 1 *(strongly disagree (helemaal niet mee eens))* both have one data entry, which means it has a value of 3%.

The first question functions as a bridging question between the questions about social relationships and about meaning. It asks how much meaning social contacts give to someone's life. Many answers include that social contacts give their life meaning by making life more pleasant and more fun. Other things that are mentioned frequently are that social relationships are an essential part in life and that they make you feel loved and valued. For all the verifications the answers are positive, so automatic classification based on this question is difficult.

The second question asks about other factors that give life meaning. Notably, the two data entries with low well-being (1 and 2) only mention one other thing that give their life meaning, being their dog and their family respectively. Most other replies are longer, mentioning multiple things that give life meaning to them. There are no differences between the neutral and the high verification answers. Themes that are mentioned the most are: faith in God, nature, helping others, music, practicing hobbies, family and good food. Other things that that are mentioned a few times are discovering new things, traveling, being thankful, and developing yourself. The frequent mention of God can be explained by the personal network that is used to gather responses. In datasets representing the whole of society this would probably be less present. The mentioned items could be extracted from the text and used as input for care professionals or future conversations with an agent. With only two answers and no difference between neutral and high, classification could be difficult.

This third question asks about a goal someone has for their life. The answers vary a lot and there are no differentiating factors between the low, neutral and high verifications. To illustrate, someone who gave a 6 as verification mentioned that he/she cannot mention any goals for their lives, whereas someone with verification 1 wants to have a positive influence in the lives of others. Generally there are two types of answers given, specific practical goals and general philosophical goals. Examples of the practical goals are finishing their studies, living abroad, getting a or another job, finishing a project about bees and insects, being successful at work, and finding a house. The goal to finish one's studies is mentioned a lot, which is probably a result from the personal network used to gather data in which contained quite a few students. Examples of more philosophical goals are being happy, improving the lives of others, spreading love, being happy with themselves. Improving the lives of others or helping others is mentioned most frequently, which was also mentioned a lot to the previous question. Since the answers are very diverse and do not differentiate for the different verifications, information extraction and classification would be difficult.

4.3.9 I am a good person and live a good life (MNG-2)

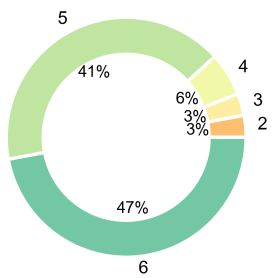

Figure 4.13: Verification distribution for statement 9

The ninth statement is covered by five questions. Being a good person is covered by asking about the good personal qualities someone has, and from these qualities if they see themselves as a good person. Living a good life is covered by asking how their personal qualities influence the people around them, an example of the influence they have, and how satisfied they are with this influence. This is the first statement where verification 5 and 6 are equally represented and make up almost all of the answers (see Figure 4.13). This means that people are less certain if they are a good person and live a good life.

The first question asks about which qualities the person has that they consider good qualities. This is an introductory question to the second question. The given answers all mention positive qualities, some very brief and some more elaborate. The answers do not differ between the higher and lower verifications given. Qualities that are mentioned frequently are, for example, humble, honest, perseverance, enthusiastic, interested in others, sensitive, funny. There are many more qualities mentioned but it is too much to mention all of them. Two participants said they cannot mention good qualities of themselves, these both gave a neutral verification (3 and 5).

The second question comes right to the point by directly asking the participants if they think of themselves as a good person. From the answers it is clear that this question was difficult to answer for some people. Various people say that they try to be a good person, others say something along the lines of *"I think I am a good person"* or *"I consider myself a good person"*. The lower verifications are less confident in their answers, and say that their negative and positive qualities are both equally present. The neutral verifications generally give longer answers, often saying that they try to be a good person but they also have negative qualities. The phrase *"I think I'm quite a good person (ik vind mijzelf best een goed persoon)"* is said multiple times. The higher verifications are generally shorter, so more confident in their answering. These answers mainly focus on their positive qualities that make them a good person. The difference in confidence could be used as information for automatic classification.

The third question is a bridging question, asking how the personal qualities they just mentioned influence the people around them. All answers say that their qualities influence the people around them in a positive way. A reason for this could be that the first question of this statement asked about positive qualities, so the participants were only focused on that. Even though there is no difference between the verifications, this question does give interesting answers of people describing the good parts of themselves. For example, *"people can come to me when they need help and I try to help them"*, *"I try to involve everyone in a group so I improve the solidarity"*, and *"I help people become more to who they are"*.

The fourth question continues with asking about an example of how they influence the people around them in a positive or negative way. Again, there are no differences between the verifications. Throughout all verifications some people mention that they find it hard to give an example. The other examples given are similar to the answers to question three, although here people also mention negative influences. Examples of the negative influences are being moody, being critical to others, or not doing things. However, most of the mentioned influences are positive, like making others laugh, driving a friend to the airport, listening to others, or baking a cake for family.

The fifth and last question for this statement asks if people are satisfied with the influence they have. This question has a clear difference between the low, neutral and high verifications. The lower verifications say *"not positive"* or don't say anything. The neutral verifications are less secure saying things like *"I would like to be there more for others"*, *"it is a start but I'm not finished yet"* and *"depends on the day"*. But the neutral answers also have people saying that they are satisfied. The high verifications are generally more confident in saying that they are satisfied with their influence. Examples are *"Yes that is enough for me"*, *"yes most of the time"* and *"yes I am satisfied with that"*. The difference in confidence for the different verifications could be used as input for a classifier.

4.3.10 I am optimistic about my future (MNG-3)

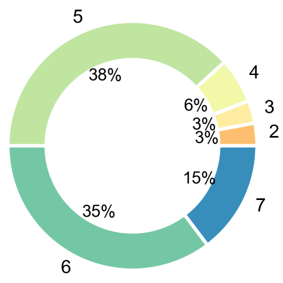

Figure 4.14: Verification distribution for statement 10

The last statement is about how optimistic someone feels about their future. This is covered in three questions: *"When you think about your future, what image do you see?"*, *"What wishes do you have for your future?"* and *"How much confidence do you have that those wishes will actually come true?"*. The distribution of the given verifications can be found in Figure 4.14. The verifications for this statement are very positive, with 15% for verification 7 *(strongly agree (heel erg mee eens))*, 35% for 6 *(agree (mee eens))*, and 38% for 5 *(slightly agree (een beetje mee eens))*. The remaining 12% is divided between verification values 2, 3 and 4. This means that a big majority of the participants is (very) optimistic about their future.

The first question asks what the participants think of when they think of their future. For all the verifications there are two kinds of answers. On the one hand there are people who indicated that their future is unsure and they do not know exactly what it will look like, and on the other hand people who give a clear plan for the future. Many responses include having a nice family and friends around them, a nice job or retiring, and a nice house. As before, these answers could be influenced by the over-representation of young people in the gathered participants.

The second question asks about wishes the participant has for their future. Again, the answers are a lot alike between the different verifications. Wishes that are frequently mentioned are: being healthy, being happy, having a nice family, and having a nice house. The answers are a lot like the first question, so this question might have been redundant. Although for people with lower well-being there could be a difference between their view of the future and their wishes for it. But with the current dataset there was only one person who indicated a low verification, so no conclusions can be drawn.

The third question asks about the confidence they have that those wishes come true. The answer with verification 2 says that he/she has low confidence it will come true and that they are loosing their hope for it. For the neutral and high verifications there is some difference in confidence in the future. The high verifications frequently say that they have a lot of confidence, or that they are optimistic. The neutral verifications are slightly less confident. Sayings that often occur are along the lines of: *"I'll try and then we'll see"*, *"that depends on many factors"*, and *"I think it will be okay"*. Although there is a difference between the neutral and high verifications, it is only a slight difference and there is also a lot of overlap so it might be difficult to automatically extract it.

4.3.11 Discussion

This section has analyzed every well-being statement and every question to look for common themes and similarities between the answers. The analysis of the dataset showed valuable information about the well-being of the participants and about the kind of answers one could expect with these kind of questions. There were many questions which did not show a clear difference between the low, neutral, or high verifications. This means that it cannot easily be used for automatic classifications. For questions that did have a distinction between the low, neutral and high verifications, the difference was in the certainty of the answer, the length of the answer, or in the adjective used to describe a certain word. This analysis also showed some questions which answers could be useful for information extraction. The person's hobbies, social network, helping persons, and things that give meaning could be extracted as information for care professionals or used in future conversations. The classification and information extraction will be done in the next chapter.

4.4 | Dataset from experiment with personas

As explained in Section 4.1, the experiment with personas was set up to gather more negative data. This dataset was kept separate from the original dataset because the method for the data gathering influenced the quality of the data (explained more in Section 4.4.1). This data was gathered through a subject recruitment pool from the study of Psychology at the University of Twente. This study was online for 23 days and gathered 13 responses. The participants could choose which persona to use, or could choose to have the conversation as themselves. The distribution of how many participants choose each persona is shown in Section 4.4. Most participants (5) choose the persona of Tom, who is a student with anxiety and perfectionism. 3 participants choose to fill it in as themselves. The others choose Wilke, the busy mother, Brechje, the lonely elderly, and Steffan, the chronically ill person. An explanation for the preference for Tom could be that this persona was a student, so the participants could identify with him the most. On top of that, this persona could be familiar since many students nowadays suffer from anxiety or depression. [3] With these numbers it is interesting that no one choose the persona of Nicole who is also young and suffers from depression.

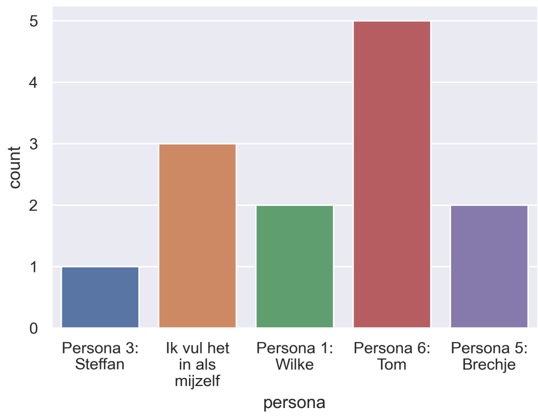

Figure 4.15: Distribution of chosen personas

The aim of this experiment was to gather more negative data which balances the dataset so it can be used for classification. Section 4.4 shows the distribution of the given verifications in the persona experiment. It shows that there are much more low and neutral answers than in the original experiment (for the distribution in the original experiment see Section 4.4). Especially verifications 2 and 3 are used more often. Even though the verifications are much lower than the original experiment, it does not balance the dataset out evenly. The experiment with personas only has 13 respondents, against the 35 respondents of the original experiment. This is not enough to evenly balance the dataset, although it does improve it. A more detailed analysis will be given in Chapter 5.

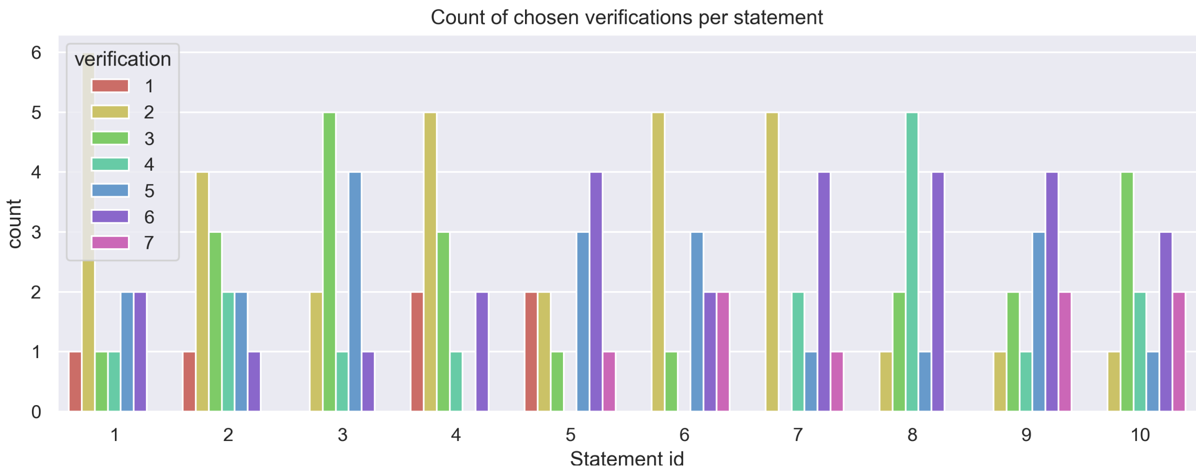

Figure 4.16: Distribution of chosen verifications in the persona dataset

[3] https://www.trimbos.nl/actueel/nieuws/hoger-onderwijs-moet-werk-maken-van-mentale-gezondheid-studenten/

4.4.1 Data analysis of dataset from experiment with personas

The dataset gathered from the experiment with personas has 13 responses. The biggest differences between the data in the original dataset and from this experiment are two things: the answers are short, and the answers miss nuance and depth. This was true for the participants who filled it in from a persona and who filled it in as themselves. The short answers could be explained by the fact that the participants filled it in to get credits, and wanted to get it done as soon as possible. As an illustration, someone simply answered to the question about activities they like to do:*"no I can't"*. There are also entries that are meant as funny answers. For example, for statement 3 question 1 someone answered that an activity that they like to do is *"cutting pieces of bread perfectly"*.

The shallowness of the answers can be found in almost all questions. Take the first statement as an example. The answers to the questions of the first statement were all very short or added some reasons why they felt this way which were a direct copy of the persona. The emotions mentioned were only negative, for example *stressed* and *fearful* were mentioned a lot. In the original dataset it showed that people mostly give a mix of positive and negative emotions, even when they did not feel good. So the data collected from the persona misses this variation that is present in the original dataset.

Some questions were easier to answer from the personas and give quite good answers resembling the original dataset. For example, the questions about social relationships (statement 5) often mention family and friends, just like the original dataset. Also for statement 6 about how people perceive them, the answers are very similar to the original dataset. To the question why they think people perceive them that way there are answers like *"there is a nice vibe"* and *"you know it from someone's body language"*.

There are also a few questions that were more difficult to answer, which can be seen by the general, unspecific answers. Especially the questions about meaning (statement 8,9 and 10) gave this kind of answers. Question 1 from statement 8 asked how much meaning social contacts give to their life. Here some unrealistic answers were given, such as *"social contacts are important but not for me"* and *"social contacts are not important, studying is the most important for me"*. For the questions of statement 9, some participants did not give an answer at all. This especially occurred for question 3 and 5, which ask about the influence they have. To question 3, which asks how their qualities affect the people around them, the answers only say something like *positively* or *not much*, and do not have a lot of depth. Same for question 5, which asks how satisfied they are with the influence they have, many answers simply say *yes*.

4.4.2 Discussion

In this experiment it is difficult to verify how much the data represents a real scenario because there is not enough original data with low well-being to compare it to. However, it can be concluded that using personas to gather answers with lower well-being give answers that lack length, variation and detail. Answers are shorter than in the real world and the mentioned reasoning is often directly copied from the persona. Nevertheless, for most questions the answers have similarities to the answers in the original dataset. So it can be assumed that they do represent a real-world scenario to some extent. It should be noted that it is not clear if the shortness of the answers is caused by the use of personas or the use of payment of the participants (through credits). For further research, another study could be set up where people participate out of intrinsic motivation instead of payment, to measure if this gives more realistic and elaborate answers. Furthermore, it could be studied if the content of the persona influences the kind of answers that are given. The length and detail of the personas could be tweaked to see if participants add more information that is not a direct copy from the persona.

4.5 | Conclusions

This chapter was aimed at answering the question: *"What useful information can be extracted from a dataset gathered with the conversation that was designed?"*. The chapter started by explaining the method used for the dataset collection. The original dataset was gathered through personal contacts, resulting in a dataset with 35 entries. Since this dataset was too unbalanced for classification, an additional experiment was set up which used six personas with lower well-being. This additional experiment was held with psychology students and resulted in 13 additional data entries. The data was gathered through the WhappBot software, which was also further illustrated in this chapter. The generation methods and structure of the final dataset were also described.

Thereafter, the original dataset was analyzed in depth. For every well-being statement and every question the answers were analyzed, looking for themes and similarities between the answers. The analysis of the answers gave many insights into the well-being of the participants, and into the kind of answers people give to these kinds of questions. For example, it was found that not only past emotions should be identified, but also their frequency or intensity. Also, the data showed that people with lower well-being do have enough fun activities to do, but lack the time, energy or motivation to do these. The data analysis also showed that all participants had fulfilling social relationships. When asked what gave their life meaning, answers that occurred frequently were: faith in God, nature, helping others, music, practicing hobbies, family and good food. The answers are used in the next chapter in two ways, for automatic classification of someone's well-being, and for information extraction. Information that can be extracted and used in further conversations by an agent or a care professional are, for example, the person's hobbies, social network, helpful persons, and things that give meaning to their life.

After the analysis of the original dataset, the dataset from the experiment with personas was examined. These answers were generally shorter with less variation and details than the answers from the original experiment. But for most of the questions the answers did resemble the original dataset so it is expected that they can be used for improving the classification. For future research it could be investigated what causes the short and superficial answers from the personas, and if better answers can be gathered by changing the personas or the experiment in a certain way.

5 | Information extraction and classification

In the last chapter the gathered dataset was analyzed. This chapter will use the same dataset to answer the last sub-question: *"Can transcribed conversation data be used to automatically extract information about someone's well-being?"* To answer this question, the chapter is split into two parts: information extraction and classification. Information extraction can be the goal itself, for example to use the extracted information for further conversations with an agent, but it can also be used to improve text classification results. Therefore, the results from the information extraction implementation will be used in the classification section. For the information extraction, a subset of the dataset is used, which consists of the questions found to give useful information for future conversations in Chapter 4. These questions were the ones about daily activities, social contacts and facets that give meaning in life. The second part of this chapter is dedicated to automatic classification of the well-being level of the participant. Using three machine learning algorithms and methods found in literature the classification is performed. Before discussing the implementation, a literature background is provided about information extraction and classification with text and/or small datasets.

5.1 | Literature background

Most information extraction and classification techniques depend on statistical relations in large datasets. Research about classification with small datasets is also being done, for example by Kou et al. [48], who define *small datasets* as datasets with a size of 1000. The dataset gathered in Chapter 4 only has 48 entries (35 if you only use the original dataset). This extremely small size of the gathered dataset means that the techniques that require large datasets, like neural networks, unsupervised learning, or creating domain specific dictionaries [69], cannot be used. Alternative techniques need to be used which do not require such large datasets. For example, as an alternative to creating domain specific dictionaries through data mining, one could focus on using already existing dictionaries that contain the relevant information (if such a dictionary exists). Instead of focusing on common machine learning techniques, this section will focus on information extraction techniques which work with smaller datasets. The body of literature on this topic is large, so this section cannot give a complete overview. Therefore, this literature background will start with identifying techniques that are used in studies about well-being classification. Then, research papers using similar techniques are described in Section 5.1.2. Section 5.1.3 describes how these techniques can also be used for classification problems.

5.1.1 Related work on well-being classification

Most research on automatic well-being classification uses social media posts as the dataset. Chen et al. [13] worked on extracting a person's subjective well-being based on language use in Facebook status updates. Their dataset consisted of 470k posts from 2,6k users. For extracting the affect of a user they used sentiment analysis established by a pre-defined list of valance words. For determining the life satisfaction of the user, topic clusters were used. They found that using a combination of topics automatically extracted from the data (LDA topic modeling) and topics from the *Linguistic Inquiry and Word Count (LIWC)* dictionary gave the best results for the model predicting subjective well-being. The machine generated topics are gathered by a word count approach for topic words, which results in a list of probabilistic relationships between words, topics and documents. The number of topics was reduced to prevent over-fitting. The model predicting subjective well-being was combined with user affect found through sentiment analysis and used as input for the classification algorithm for which they used the random forest algorithm with manual optimization of the trees. The study concluded that machine learning techniques can indeed be used to identify subjective well-being of social media users.

Schwartz et al. [78] also studied the prediction of well-being through natural language use in social media posts. He predicted the well-being of a social media user based on lexical and topical features in status updates on Twitter and Facebook. For this, the well-being models *satisfaction with life* [19] and PERMA [81] were used. The lexical features used for well-being prediction were unigrams and bigrams, topic extraction, and lexicons. The used lexicons are the LIWC and the weighted lexica from Dodd's hedonometer [23]. Each category in those lexicons was used as a separate input feature, included as binary (mentioned at least once or not) and as relative frequency. Two models were created, message level models (from 5k randomly selected status updates) and user level models (from 260k updates from 2,2k users), which were combined to predict the user's well-being. Schwartz et al. [78] concluded that well-being can be predicted by using language-based analyses. Additionally, he stated that the analyses should not end at prediction, but the language-based analyses can be used to understand the determining topics in social media posts that are related to a users' well-being.

Preotiuc-Pietro et al. [65] worked on a task to classify self-reported depression and PTSD from Tweets from 1k Twitter users. User features (like number of followers and gender) and textual features were used as input for the classification algorithms. The textual features used were unigrams forming bag-of-word representations, and several topic cluster algorithms which give clusters of semantic or syntactic similar words. These topic clusters were created from large unrelated datasets. Topic clustering is especially useful to prevent over-fitting of data. The various clustering algorithms techniques are compared to find the best method for this task. For the classification they use two different linear classifiers: logistic regression and a linear support vector machine (SVM). The tuning of the SVM was done by using 10 cross-fold validation and 10 random restarts on the training set. They acquired a precision of approximately 85% with several combinations of features.

A well-being classification study that used another kind of data then social media posts was done by Adiga et al. [2]. They conducted a study on extracting well-being information from daily journals. They did this by extracting activity phrases and their sentiment from those 750 journal entries, together with related public datasets (Twitter, airline reviews, and HappyDB[4]). The sentiment classifier was created with machine learning models from large public datasets, which gave a 72% accuracy of classification on the validation dataset. The activity phrases are extracted through pre-defined grammar rules, which are different combinations of part-of-speech tags that form an activity phrase. These grammar rules defined a verb (with its supporting verb if applicable), a noun phrase, and a possible preposition. From the resulting activity phrases the ones describing actions from the author himself were filtered. Then, these activity phrases were categorized in five activity classes, which were determined by the researchers based on categories with enough samples in the different datasets: *family and friends, interest and leisure, health and benefits, official tasks* and *habits and preferences*. The classification was done using different machine learning models. For the activity phrases extraction an accuracy of approximately 74% was achieved.

5.1.2 Research using similar information extraction techniques

The studies above identified ways to extract well-being information from text using various lexical based techniques (e.g. LIWC lexicon, WordNet, part-of-speech tagging, topic clustering). This section will elaborate on a few papers that use text classification techniques similar to the ones mentioned in the previous section. Adding these papers will give a broader view on the implementations of these techniques, which could be useful in the remaining of this chapter. The body of research on these techniques is too extensive to give a complete overview, therefore only a few relevant papers are discussed in this section.

Scott and Matwin [79] explored the influence of using part-of-speech information and a clustering technique called hypernym density on text classification. Hypernyms are more general words to which other words belong. For example, the hypernym of rabbit and fish is animal. The lexical database called WordNet is used to find these hypernyms, which stores information in combinations of a word and a meaning called a *synset*. Subsets of three large public datasets were used, consisting of approximately 200-400 texts per category and six combinations of similar and unrelated topics. For example, from the Reuters corpus they collected the texts from the classes *livestock and gold* as a subset and the classes *corn and wheat* as another subset. The algorithm started with finding the part-of-speech tag for each word in the text. Then the verbs and nouns were extracted and their hypernyms were looked up in WordNet. Infrequent synsets were discarded and for the rest of the synsets their density was calculated. This density was calculated as *"the number of occurences of a synset in the WordNet output divided by the number of words in the document"* [79, p. 47]. The resulting numerical feature vectors were used as input for the classifier. They found that hypernym density can greatly improve classification accuracy in certain datasets. Especially for datasets with an extended or unusual vocabulary this could give significant improvements.

Another way to complement training information for text classification by using WordNet was done by de Buenaga Rodríguez et al. [17]. They used the semantic relations for a word in WordNet to place the word in a pre-defined group. These groups consisted of words related in meaning. They found that this approach improves text classification results, even for categories which have few training documents. They used the Reuters dataset to train and test their algorithm, with 13,6k training documents and 6k test documents.

Tanawongsuwan [86] also used part-of-speech tags from a text, which they used to determine the helpfulness of 1,4k book reviews from Amazon. The content of each review was tagged by a part-of-speech tagger. From this tagged text various features were extracted and used as input for the classification. The extracted features are the number and percentage of tokens associated with each tag (nr of tokens with certain tag/total nr of tokens), and the occurrences of positive and negative adjectives. They found that adjectives appeared the same rate in positive and negative reviews, and only a few adjectives were found that influenced the classification significantly. Three classification models were compared: C4.5, a Neural Network and Logistic Model Trees. All three models gave good accuracies between 80% and 90%. This study showed that occurrences of certain adjectives and the count of part-of-speech tags can be used to improve text classification.

5.1.3 Classification

Most of the previously described research not only used information extraction techniques as a goal, but as a means to improve text classification, especially in small datasets. Kou et al. [48] did research on feature selection methods for text classification with small datasets. They used ten feature selection models on ten small datasets, all with 1k samples and between 2,4k and 28k features. Although they did not find a feature selection method that outperformed the others, they did conclude that classification can be improved by reducing the number of input features. This can be done by statistical feature reducing techniques [48] or by clustering the existing features. As we have seen in the previous sections, feature clustering in text classification is often done through topic clustering. Liu et al. [53] also used linguistic features to improve neural text classification. They used word embeddings by WordNet to enrich their data. They performed feature reduction by clustering words by the lexical categories given by WordNet. Preotiuc-Pietro et al. [65] also showed that topic clustering can improve the classification of text considerably.

Hartmann et al. [32] did a large scale comparison of different methods for automated text classification of social media datasets. Five lexicon-based and five machine learning algorithms were used on 41 datasets. They found that the performance of each classification method was dependent on the application, and none outperformed the others in all cases. In general, they found that the naive bayes and the random forest algorithms gave the best performances. Also for smaller sample sizes, the performance of the naive bayes did not decrease. The support vector machine (SVM), an algorithm which is used predominantly in their targeted application field, scored good on some datasets bus less on others. An advantage of SVMs is that they are considered more resilient against over-fitting [32], which is a common problem in text classification.

5.1.4 Conclusions from information extraction and classification literature

The relating literature has shown that there are various ways to extract information from text based on lexical information, and to use that information in text classification in smaller datasets. Methods that are frequently used in the found literature are:

- using plain text representations in ngrams or bag-of-words representations
- using lexicons like the LIWC and WordNet
- clustering words in topic clusters
- sentiment analysis
- using part-of-speech tags of the sentence

There are also various algorithms to use on text classification, with performance dependent on the dataset and problem. The SVM is often used in text classification and less prune to over-fitting, and random forest and naive bayes showed good performances across datasets from various domains [32].

5.2 | Method

The literature analysis has shown various information extraction and classification techniques that could be useful for this dataset. However, as also found in literature, which technique yields the best results is dependent on the dataset used. This chapter describes the approach for discovering the best techniques for the gathered dataset. The continuation of this chapter will be divided into two separate parts, information extraction and classification.

5.2.1 Information extraction

As stated in the introduction of this chapter, three types of information were identified in the gathered dataset that could be useful for further conversations with agents or with care professionals. This information is covered in several questions which were identified during the data analysis in the previous chapter. These questions are:

- 3.1 - Can you describe to me which things are a regular part of your week? *(activities)*
- 4.1 - Do you have enough activities that you like to do? Which activities are those? *(activities)*
- 5.1 - Which people do you have contact with in a normal week? *(social relationships)*
- 5.3 - Which of these people can you turn to if you need help? *(social relationships)*
- 8.2 - Can you name any other things that make life worth living for you? *(meaning)*

The three types of information (daily and fun activities, social relationships, and things that give life meaning) are each handled as its own dataset. For each, the best approach for information extraction is researched by using the approaches described in the literature section. For this task, only the original dataset is used and the persona dataset is not included. This decision was made because the quality of the data with personas cannot be guaranteed, and the extra data is not needed for this task to yield good results.

5.2.2 Classification

For the classification task, both the original dataset and the persona dataset are used. As stated in Section 4.1, the persona dataset was gathered with the purpose of balancing the original dataset so that classification is possible. The original dataset on its own has too little data for the lower well-being levels to perform classification. Adding the persona dataset to the original dataset gives a better balance. The best possible balance is achieved by splitting the data between verification 1-5 and 6-7. The resulting distribution is shown in Figure 5.1.

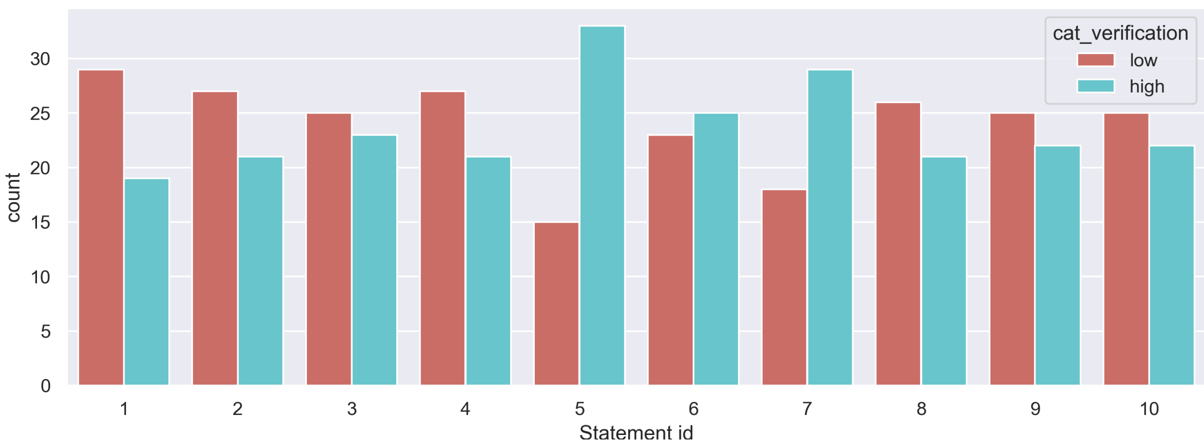

Figure 5.1: Balance of the dataset with two categories

Although for this division the data is most balanced, it does not make sense when looking at what the data represents. This division would place the answers ranging from *strongly disagree* to *slightly agree* in the same class, and both *agree* and *strongly agree* in the other class. Depending on the use case this might be useful, for example when looking for a decrease in well-being when the well-being starts high. However, such a decrease could not be detected when the person changes between neutral and low. A division which distinguishes the three categories of the used Likert scale (low, neutral and high) makes more sense, and represents the used scale better. This division was also already used in the text of Chapter 4. Using these three categories results in the distribution shown in Figure 5.2.

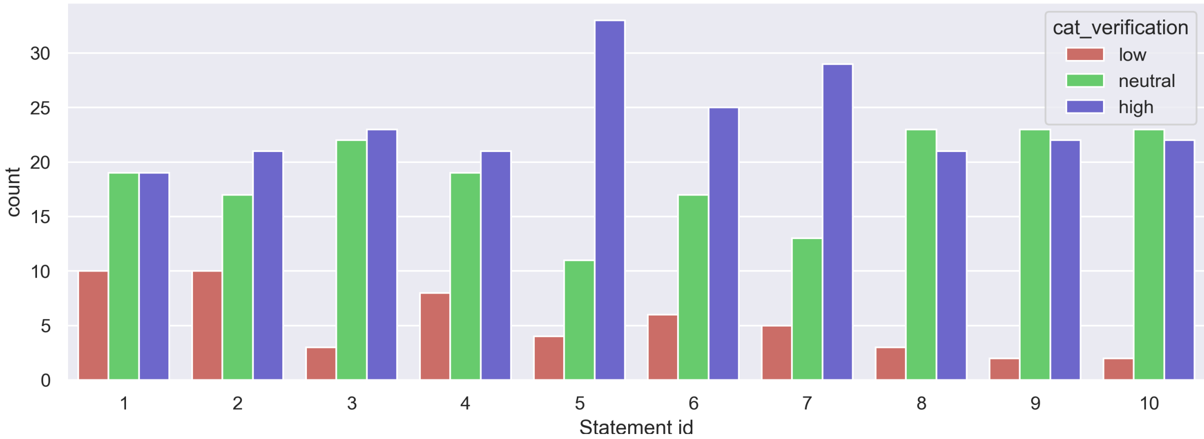

Figure 5.2: Balance of the dataset with three categories

With this division, the neutral and high verification are almost evenly represented. The divisions of two classes and seven classes are disregarded in the current scope. The classification in this chapter only uses these three categories for the classification. A support vector machine, random forest classifier and naive bayes classifier are used, as they show good results in the comparative study on text classification methods by Hartmann et al. [32]. Since the verification is given per well-being statement, ten separate classifiers are used.

5.2.3 Software

Both the information extraction and classification were done using Python. Below the used libraries and lexicons are explained briefly.

Part-of-speech tagging

As part-of-speech tagger the NLP service called Spacy[1] is used. This service can be imported as a Python library and is multilingual, so it can also be used with Dutch text. Besides the part-of-speech tagging, Spacy also includes, among others, a dependency parser and a lemmatizer.

Sentiment analysis

There is only one Dutch sentiment analyser in Python called Pattern[2]. However, as Gatti and van Stegeren [28] showed, it is biased to words used in reviews and does not perform well on more general or emotional texts. Therefore, the more commonly used algorithm NLTK vader[3] is used instead. To use this algorithm, the text is first translated into English. It should be noted that this method also has some limitations since some sentiment can be lost in machine translation [35].

[1] https://spacy.io/
[2] https://github.com/clips/pattern
[3] https://www.nltk.org/_modules/nltk/sentiment/vader.html

Classification and parameter optimization

The classification models (*SVC, RandomForestClassifier, MultinomialNB*), vectorizers (*TfidfVectorizer*) and optimization algorithms (*GridSearchCV*) are all imported from the library called scikit-learn[4].

LIWC

A database that was mentioned multiple times in Section 5.1, and can be used to get more affective information about a word is the LIWC (Linguistic Inquiry and Word Count) [87]. This lexicon maps frequently used words in psychologically meaningful categories. An example of words and their categories of this Dutch version is shown in Table 5.1. These categories can say something about the linguistic function of the word (e.g. pronoun or number), psychological processes (e.g. social or negative emotion), thematic category (e.g. health or time), or personal concerns (e.g. work or religion). Most words in the lexicon correspond to multiple classes. In 2007 a Dutch version of this lexicon was created by Boot et al. [9].

word	**LIWC categories**
overgestoken	Past, Motion
overgeven	Present, Physcal, Body
overgewicht	Physcal, Body, Eating
overgrootvader	Social, Family
overhalen	Cogmech, Insight, Social, Comm, Present
overheers	Affect, Negemo, Anger, Present

Table 5.1: Example of a slice from the Dutch LIWC lexicon [9]

WordNet

WordNet is a lexical database created by Miller [57]. It works as an online dictionary, consisting of words and corresponding meanings. Each word can have multiple meanings, just like regular dictionaries. The words and meaning are represented in synsets, which are of the form *word.n.01*, where *word* is the given word or synonym that it belongs to and the number represents different meanings. Take for example the word *friend*. The synsets corresponding to this word are *friend.n.01, ally.n.02, acquaintance.n.03, supporter.n.01, friend.n.05.* Each synset includes the definition, hypernyms, hyponyms, synonyms, antonyms and examples of the word. On top of that, similarity between two synsets can be calculated to find how similar their definitions are. The Dutch version of WordNet, called *Open Dutch WordNet*, was created by Postma et al. [64].

5.3 | Implementation and results - Information extraction

As described in the method, for every information type (activities, social relationships, meaning) the best method for information extraction is searched for using the approaches found in literature. This section describes the various approaches that were tried, the choices made, the difficulties and shortcomings of each approach, and the approach with the best result.

5.3.1 Part-of-speech tagging in transcribed data

Before going into extracting the three types of information, there is one limiting factor in all of the data that should be discussed: part-of-speech tagging in transcribed data. Literature has shown that part-of-speech tagging and the corresponding dependencies between words can be very useful in text analysis. Unfortunately, it proved to be rather difficult for the dependency parser, which is the algorithm that does

[4] scikit-learn.org

the part-of-speech tagging, to come up with the right parse. A dependency parser looks at a complete sentence and combines the words into the most likely parse. However, the transcribed data in the gathered dataset does not have clear sentences.
The vocal information about start, stop and pauses in the speech of the participants is not represented in the data. Multiple sentences are simply glued together. Where we as people can realise that words stand on their own, the dependency parser tries to combine the words into a sentence and gives a parse which combines unrelated words. This happens mostly with answers where people do not give whole sentences but only keywords. For example in the parse in Figure 5.3, where the word *bijbaan (part-time job)* is seen as the verb to combine the words *school* and *muziek (music)*. Another example is the parse in Figure 5.4, where *natuur (nature)* and *muziek (music)* are seen as nouns belonging together just like *office chair* are two nouns belonging together. But the participant was likely not talking about music about nature, but about two separate activities.

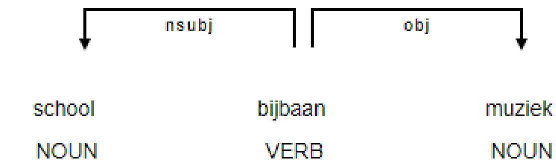

Figure 5.3: Example of an incorrectly parsed answer by Spacy - 1

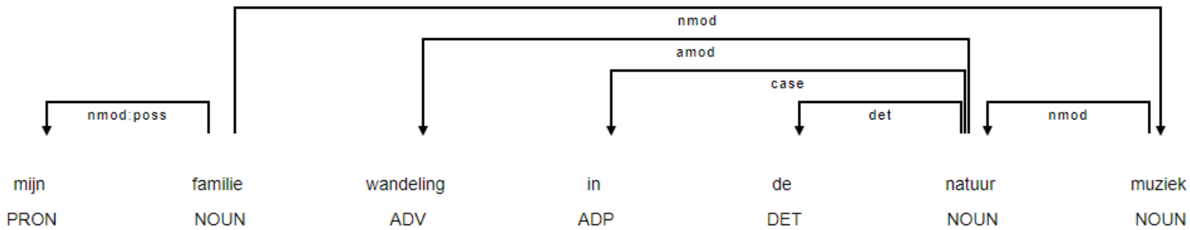

Figure 5.4: Example of an incorrectly parsed answer by Spacy - 2

This low performance of a dependency parser on spoken language is also a subject of research. For example, Dobrovoljc and Martinc [22] and Bechet et al. [6] looked into ways to improve dependency parses of spoken language (French/Slovenian respectively). The mistakes in the dependency parses makes some forms of information extraction more difficult. For example, simply extracting nouns does not always give all the nouns, since some nouns are classified as verbs or adjectives. Furthermore, noun phrases or verb phrases contain many errors. The sections below will discuss more about the way in which this causes difficulties for the extraction.

5.3.2 Extraction of social relationships

For the extraction of social relationships, the answers to two questions were analysed: *"5.1 - Which people do you have contact with in a normal week?"* and *"5.3 - Which of these people can you turn to if you need help?"* Words that are often mentioned are *parents, friends* and *colleges*. But there are also more complex words like *supermarket workers (supermarktmedewerkers)* and *fellow students (studiegenoten)*. Some words were missed because of their part-of-speech tag. This especially occurs in the second question, where people respond with *all*: the word *iedereen* is seen as a pronoun, and *allemaal* as an adverb. For a correct extraction of the social relationships, the relationships of the first question should be returned when someone refers to all the previously mentioned contacts in the second questions. This is currently not implemented.

First it was tried to extract the social relationships using the LIWC. This database has a label *Social* which covers a great part of the mentioned contacts. To accomplish this the text was first processed by Spacy. The lemma of the word was used to look up in the LIWC, since *nicht* is known but its diminitive *nichtje* is not. If the lemma was not present in the LIWC, the original word was tried because sometimes the lemma was not in the LIWC while the original word was. For example, *vriendin* was not in it but *vriendinnen* was. Besides using the label *Social (social processes)*, the label *Comm (communication)* was excluded because words like *contact* are labeled as social but describe the communication and not the relationship. Extracting all words with the *Social* label and without the *Comm* label from the LIWC returns the common relationships quite well. But more complicated ones are missed, like *schoonouders*, *hulpverleners*, *medestudenten* or *huisgenoten*. Also, information was lost with association of words, like *people from my student association* or *people from hockey* were reduced to *people*.

As a second approach the WordNet dictionary was used to extract the social relationships. As mentioned before, the words in WordNet have hypernyms which is a more generic word than the original word. For words describing a person there is a hypernym with the synset *person.n.01* somewhere. To extract this, a recursive algorithm was made which checks all hypernyms, and the hypernyms of those hypernyms, to see if this word has a synset including *person*. With this method, more words were found as people that the LIWC missed, for example: *leden, huisgenoten, bezoekers, leerlingen, hulpverleners*. To also include social groups like *family*, the synset *social_group.n.01* was also looked for as hypernym. This returned words such as *familie, gemeente, universiteit, bestuur*. Although this method included a lot of the social relationships, there were also some errors. For example, the word *stuk* is seen as a person, since it can be used to describe a very pretty person. The word *land* was seen as a social group with the definition *a unit with political responsibilities*. Technically, these words could be used to describe social relationships, but it is not likely in this context. It was tried to only use the first definition of a word to filter out the unlikely ones. This caused problems for other words, since the first definition is not always the best applicable definition. This is for example the case with the word *kinderen*, for which the first synset is *young.n.01* and not *child.n.01*. In general, WordNet returned more correct words than the LIWC. The one word that the LIWC extracted that WordNet missed was *buurt*.

The method with the best results was to combine the two approaches above, with only the first definition from WordNet. This approach gets rid of the small mistakes that both libraries make and returned almost all social relationships. The words that are missed are compound words like *supermarktmedewerkers* or *wandelmaatjes*. The decomposed words *medewerker* and *maatje* are identified as persons by WordNet. For future work, one could look into splitting compound words into separate words, like was done in the research by Macken and Tezcan [55].

5.3.3 Extraction of activities

The extraction of activities was done using the data gathered through two questions: *"3.1 - Can you describe to me which things are a regular part of your week?"* and *"4.1 - Do you have enough activities that you like to do? Which activities are those?"* The discussed literature already shortly discussed activity extraction, which was done through looking at combinations of part-of-speech tags. Adiga et al. [2] extracted noun phrases and classified the activities into categories. For our implementation it can also be useful to perform topic clustering and categorize the activities into more general categories, like *sports* or *social*. This makes it easier to use it for implementation in further conversations with a conversational agent. Reading through all the answers given to the questions, there are five clear topics that are often mentioned: *work, music, sports, social* and *housekeeping*. These, together with an additional *other* category, are used to cluster the activities.

As described in Section 5.3.1, it is not possible to simply extract all noun phrases due to the structure of the gathered data. Therefore, the two lexicons, the LIWC and WordNet, are used to classify all words labeled as noun, proper noun or verb. The words were looked up in the lexicons in the same way as was done for the social relationships, but with different labels or hypernyms. For some topics it was straightforward which category or ancestor to choose, for others it was a lot more difficult (as explained below). Table 5.2 displays the best selection of LIWC categories or WordNet ancestors when using either LIWC or WordNet (so not combined). The optimal combination of both LIWC and WordNet together is shown in Table 5.3 Thereafter, for every category it is explained what the difficulties are, and what considerations were made.

Work

The labels *Occup (occupation)* and *School* give all work/school related words from the LIWC. This also included words like *onvoldoende (insufficient)* and *overtreffen (surpass)*, which are no activities. By excluding the label *Affect* these kinds of words were filtered out. In WordNet the work or education related words often had an ancestor containing *work* or *education*. There were also work unrelated words that had an ancestor containing *work*, like the words *koken (to cook)* and *boodschappen (grocery shopping)*. This last one is included because the lemma of *boodschappen* is *boodschap*, which has another meaning of an

topic	LIWC categories	WordNet ancestor
work	*include:* Occup, School *exclude:* Affect	work, education
music	*include:* Music	musical
sports	*include:* Sport, Motion	sport, travel
social	*include:* Social *exclude:* Sport	person, social
housekeeping	*include:* Home *exclude:* TV, Sport	commodity, clean, housework
other	*include:*: Leisure, Sleep, Eating, TV, Relig *exclude:* Sport, Home, Social, Music, Occup	religious, food, make

Table 5.2: Topics and their corresponding categories/ancestors with the best results using either LIWC or WordNet

assignment, which then is classified as work. So by using the WordNet hypernyms, a lot of false positives were given. For the best results, only the returned words from the LIWC were used.

Music

The label *Music* in the LIWC includes words like *singing, instrument, guitar*. However, it does not include many different musical instruments. Fortunately, these can be extracted through WordNet by looking for the ancestor *musical_instrument*. WordNet does not classify verbs, like *to sing*, as musical. And also the word *muziek (music)* is not classified as something musical, but instead as a sense. For the best results in extracting all music-related words, the results from the LIWC and WordNet were combined.

Sports

The LIWC contains many sports-related words. By looking at the labels *Sport* and *Motion* most sport-related words can be extracted. The label *Motion* is needed to include the words *hardlopen (running)* and *lopen (walking)*. Unfortunately this also gives some false positives. The words *ga (to go)* and *komen (to come)* have the exact same labels as *hardlopen (to run)*, so they are also returned as sports activities. Another problem with the LIWC is that it does not contain the word *sporten (to sport)*, which is one of the key words people mention in their answers.

The WordNet ancestors are similar to the LIWC categories: sport and travel. With *travel* including words like *hardlopen (to run)* and *wandelen (to walk)*. The ancestor *travel* does give a lot of false positives. For example, the word *boodschappen (grocery schopping)* has a meaning of errands, which has the ancestor of travel. For the best results, the LIWC words were combined with WordNet words with only the ancestor *sport*.

Social

The choice for these categories and ancestors was the same as for the social extraction, which was explained in Section 5.3.2. The difference is that for activity extraction, also the verbs need to be extracted. On top of that, the answers are a lot more diverse than in the questions about social contacts. This causes many words to be part that do not belong in this cluster, like *avond (evening)* is classified by WordNet as a social gathering. For the best results, only the results from the LIWC were used.

Housekeeping

This label should include all household tasks someone does, like doing groceries, cleaning, cooking. The LIWC has the label of *Home* which includes words like *schoonmaken (to clean)*. However, it also contains many words related to a house that have nothing to do with household activities and that cannot be separated from it, like *garage* and *dakterras (roof terrace)*. For this topic, WordNet gives much better results. With the ancestors of *commodity housework* and *clean*, it extracts most household activities.

Other

The *other* topic contains activities that do not fit to the other topics. For the LIWC, the labels *Leisure, Sleep, Eating, TV* and *Relig* are used, while the earlier labels like *Sport* are excluded. For WordNet it is a lot more difficult to extract *other* words. With the ancestors *religious, food* and *make*, most other activities are extracted. But this also gives many words that belong to other categories or that are unrelated. Therefore, the best results were achieved by only using the LIWC.

Difficulties

For some words it was impossible to extract them. Creative verbs like *tekenen (to draw)* and *schilderen (to paint)* are not in the LIWC and do not have a common ancestor in WordNet. Pets are also not in the LIWC, although it can be discussed if these should be extracted as an activity. The word *lezen (to read)* is not classified in the LIWC as *Leasure* but as *Cogmech, sense, see* and *present*. Therefore, it is not possible to extract this word using the LIWC. The word TV is also difficult to extract, since it contains the label *Home* in the LIWC. However, this label needs to be excluded, otherwise many other words are included that are not relevant.

topic	**LIWC categories**	**WordNet ancestor**
work	*include:* Occup, School *exclude:* Affect	
music	*include:* Music	musical
sports	*include:* Sport, Motion	sport
social	*include:* Social *exclude:* Sport	
housekeeping		commodity, clean, housework
other	*include:*: Leisure, Sleep, Eating, TV, Relig *exclude:* Sport, Home, Social, Music, Occup	

Table 5.3: The best combination of LIWC categories and WordNet ancestors per topic

5.3.4 Extraction of meaning

The extraction of meaning was done through analyzing the answers to the question: *"8.2 - Can you name any other things that make life worth living for you?"* As described in Section 4.3, the answers mention diverging topics, e.g.: faith in God, nature, helping others, music, practicing hobbies, family and good food. Since these topics do not share a common topic, the most straight forward approach is to extract noun phrases. But as already explained in Section 5.3.1, many answers are parsed incorrectly. Another incorrect parse is shown in Figure 5.5. Still, there is no good way to extract all meaningful things from the text besides extracting the nouns. However, there are a few approaches that could work, regardless of the wrongly parsed results. Firstly, if you know what you look for this can be extracted. For example, if you would want to know if the person is religious or believes in God, this can be extracted using the LIWC and WordNet with the same approach as in the previous sections. Secondly, one could create a domain specific list of keywords to look for. This was done in some papers in discussed in Section 5.1, but a bigger dataset is needed than the current one. For the current dataset with such a small size no good approach was found.

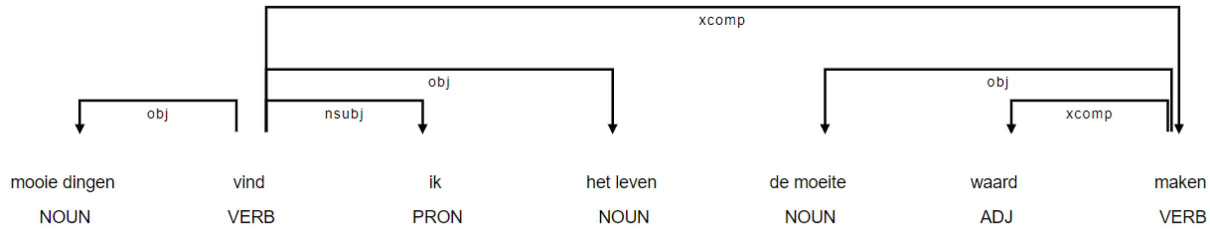

Figure 5.5: Example of an incorrectly parsed answer by Spacy - 3

5.4 Implementation and results - Classification

This section describes the classification methods and results for the original dataset combined with the persona dataset. As described in the method of this chapter, three classification algorithms are used, support vector machine, random forest classifier and naive bayes classifier, based on the paper by Hartmann et al. [32]. During the implementation it showed that the performance of the classifier was strongly influenced by the split of the data. For one split the accuracy could be 20%, for another 70%. To deal with this, two steps were taken. Firstly, each classifier was run 10 times and the mean results of those 10 runs were taken. Secondly, during the splitting of the data in a train and a test set the imbalance of the dataset was taken into account. Each category (low, neutral, high) was represented with the same percentage in the train and test set. This was done with the *stratisfy* function in Python's train-test-split algorithm.

5.4.1 Classification results without added features

The first classifications were created by applying the three classifiers to the data for every statement. As input for the classifiers, the vector representations of the cleaned text are used. The text is cleaned by removing the stop-words and non-alphabetic characters, and by lemmatizing the resulting words. A TF-IDF vectorizer was used to transform the text into use-able input for the classifiers. The data was split into a train set of 80% of the data and a test set of 20%. Before each classification (10 per statement for each classifier), the optimal parameters were found by running a grid search on the data and using the resulting parameters as input for the classifier. As with the performance, the optimal parameters were highly dependent on the split of the data in train and test set. For the SVM classification, the parameters *C*, *gamma* and *kernel* were optimized. For the naive bayes algorithm, these were *alpha* and *fit prior*. And for the random forest algorithm only the number of estimators was optimized, because optimization of this algorithm takes very long, and the criterion *entropy* and max-features *auto* gave good results during the trial runs. As results, the weighted average of the accuracy, precision, recall and F1 score were used. The results of all the classifiers with the vectorized text as input are shown in Table 5.4.

As the numbers in Table 5.4 show, the classification results are not very good. The studies described in Section 5.1 generally obtained an accuracy of at least 80%. None of the classifiers reached an accuracy that high. It is interesting that the more unbalanced the given verifications, the higher the classifiers accuracy. The resulting accuracies are almost directly corresponding with the percentage of data in the largest class. Take for example statement 5, which has the highest accuracy (around 70%) and the highest unbalance with 68% of the data in the highest class. In other words, if the classifier would only predict the highest class this would give the same results.

The low results can be explained by the extremely small dataset that is used, combined with the high number of features in this dataset because it is all textual data.

5.4.2 Adding features

In literature it was found that adding certain features could improve classification results, even with small datasets. To repeat, these mentioned features were: part-of-speech frequencies, sentence length, sentiment, and topic clusters. Nevertheless, not too many features should be added to prevent over-fitting of the model. Therefore, several features were added. For part-of-speech frequencies, the word types *noun*, *verb* and *adjective* were used, since these generally include most of the important information of the text. For every sentence the percentage of this part-of-speech tag compared to the whole sentence (e.g. #nouns/#words) was added as a feature. For topic clusters, the activity clustering from Section 5.3.3 were used, giving six added features which correspond to the count of words belonging to a certain topic. With sentence length and sentiment a total of 11 features were added as input to the classifiers. The classification was done the same way as before, and the results are shown in Table 5.5.

Even though the F1 score improved a little for some statements, from these results it cannot be concluded that adding these features improves the classification. However, this could again be caused by the extremely small size of this dataset. For more reliable change in the results a bigger dataset is needed. Something can be said about the influence of the added features on the classification. The random forest classifier provides a list of *feature importances* representing the statistical significance of the feature in the classification. Looking at these feature importances showed that the part-of-speech frequencies, sentence length, and sentiment had a positive relationship with the given verification. So this implementation indicates that adding these features can improve the classification. Because of the size of the used dataset these improvements in classifications are not visible, so to verify this claim a bigger dataset is needed.

statement	classifier	accuracy	precision	recall	F1 score
1	SVM	0.43	0.379333	0.43	0.376196
	RF	0.31	0.278857	0.31	0.277059
	NB	**0.49**	**0.505810**	**0.49**	**0.459965**
2	SVM	0.45	0.418254	0.45	0.401830
	RF	**0.55**	**0.507714**	**0.55**	**0.494912**
	NB	0.52	0.483841	0.52	0.464581
3	SVM	0.37	0.321063	0.37	0.322695
	RF	**0.43**	**0.363976**	**0.43**	**0.388329**
	NB	0.40	0.373310	0.40	0.368144
4	**SVM**	**0.58**	**0.495143**	**0.58**	**0.519986**
	RF	0.49	0.401667	0.49	0.423835
	NB	0.51	0.432524	0.51	0.454984
5	SVM	0.65	0.552556	0.65	0.589441
	RF	**0.72**	**0.577500**	**0.72**	**0.630485**
	NB	0.67	0.547278	0.67	0.597025
6	SVM	0.41	0.222222	0.41	0.285836
	RF	**0.48**	**0.250000**	**0.48**	**0.328571**
	NB	0.44	0.261984	0.44	0.319103
7	SVM	0.60	0.430286	0.60	0.480923
	RF	**0.65**	**0.517500**	**0.65**	**0.544286**
	NB	0.56	0.378786	0.56	0.446780
8	SVM	0.51	0.363500	0.51	0.408272
	RF	**0.49**	**0.456976**	**0.49**	**0.455753**
	NB	0.50	0.432139	0.50	0.447937
9	SVM	0.59	0.593135	0.59	0.555289
	RF	**0.58**	**0.596726**	**0.58**	**0.572589**
	NB	0.54	0.550179	0.54	0.529053
10	**SVM**	**0.55**	**0.482837**	**0.55**	**0.463964**
	RF	0.36	0.345833	0.36	0.321967
	NB	0.43	0.426667	0.43	0.421217

Table 5.4: Mean classification results per statement using cleaned text as input

statement	classifier	accuracy	precision	recall	F1 score
1	**SVM**	**0.53**	**0.544190**	**0.53**	**0.507535**
	RF	0.43	0.420571	0.43	0.389163
	NB	0.48	0.463492	0.48	0.444480
2	**SVM**	**0.46**	**0.424048**	**0.46**	**0.399472**
	RF	0.43	0.390317	0.43	0.372177
	NB	0.44	0.403286	0.44	0.397065
3	SVM	0.40	0.372405	0.40	0.375255
	RF	**0.51**	**0.436532**	**0.51**	**0.463687**
	NB	0.41	0.351270	0.41	0.332090
4	SVM	0.52	0.476333	0.52	0.483333
	RF	**0.54**	**0.447333**	**0.54**	**0.478889**
	NB	**0.54**	**0.451810**	**0.54**	**0.470644**
5	**SVM**	**0.69**	**0.622889**	**0.69**	**0.642660**
	RF	**0**.71	0.561667	0.71	0.616985
	NB	0.68	0.513417	0.68	0.580645
6	**SVM**	**0.43**	**0.360528**	**0.43**	**0.380681**
	RF	0.43	0.232639	0.43	0.300783
	NB	0.47	0.290556	0.47	0.348831
7	SVM	0.50	0.414190	0.50	0.445751
	RF	**0.62**	**0.471500**	**0.62**	**0.525143**
	NB	0.60	0.512500	0.60	0.536006
8	SVM	0.40	0.384476	0.40	0.385318
	RF	**0.55**	**0.507714**	**0.55**	**0.515805**
	NB	0.41	0.252127	0.41	0.306595
9	SVM	0.61	0.627560	0.61	0.596835
	RF	**0.62**	**0.646230**	**0.62**	**0.604633**
	NB	0.48	0.335317	0.48	0.358745
10	SVM	0.28	0.265536	0.28	0.270350
	RF	**0.44**	**0.424107**	**0.44**	**0.421183**
	NB	0.50	0.373095	0.50	0.413662

Table 5.5: Mean classification results per statement using a combination of cleaned text and meta data as input

5.5 | Discussion

5.5.1 Summary

The aim of this chapter was to answer the question: *"Can transcribed conversation data be used to automatically gather information about someone's well-being?"* Before the practical implementation of this question, literature was studied in the same domain. This literature review showed that various studies have managed to automatically classify someone's well-being, primarily from social media posts. During the literature review, some techniques were found that are useful for information extraction and text classification with small datasets. These techniques were: plain text representations, using lexicons like the LIWC and WordNet, topic clusters, sentiment analysis, and using part-of-speech tags.

The practical implementation started with information extraction from the dataset, which was divided into extracting social relationships, activities and things that give life meaning. The social relationships were extracted using the part-of-speech tags for nouns and verbs, and looking up those words in the LIWC and WordNet. The best results were found by using the *social* label from the LIWC and the first definition of a word from WordNet. This combination gave very good results. Next, the activities were extracted using the same lexicons. The topics in which the activities were clustered are: work, sports, music, social, household and other. Since the text in which these activities were found had a lot of variation, tweaking in the LIWC and WordNet needed to be done to find the best results. In the end the LIWC was used as a baseline and WordNet was used for some topics. For most topics this extracted most activities correctly, although it had more mistakes than the social relationships. As a third things that give meaning were extracted. This was approached by using noun phrases from the sentence. However, the structure of the transcribed text caused many errors in the sentence parses making it impossible to simply extract noun phrases. It was concluded that this kind of extraction is only possible if you know what to look for, or with a question specific list of keywords to search for.

After the information extraction, automatic classification to predict someone's well-being from the transcription of the spoken text was implemented. This was done with three different classifiers (SVM, RF, NB), using only the text, and text with added features as input. Due to the extremely small size of the dataset, both methods did not give good results. Feature importance from the random forest classifier suggested that part-of-speech frequencies, sentence length and sentiment did have a positive influence on the classification. However, larger datasets are needed to verify these findings.

5.5.2 Conclusions

Taking everything into account, it can be concluded that answering the question: *"Can transcribed conversation data be used to automatically extract information about someone's well-being?"* is not clear-cut. Yes, transcribed conversation data can be used to gather information about things that influence someone's well-being, such as activities and social relationships. Especially if it is known what information is wanted for future conversations, this information can be extracted using the illustrated techniques. Nevertheless, the classification section showed that, with the size of the current dataset, it is not possible to automatically gather information about their well-being directly.

There are various ways in which the information extraction and classification could be improved. During the information extraction it was found that information extraction is easier when you know what kind of answers will be given. For example, the extraction of social relationships gave good results because the mentioned relationships could be grouped into a few LIWC categories and WordNet ancestors. Contrarily, for the extraction of meaning no such grouping could be made and information extraction was not possible. When the questions would be made more specific, it would be easier to group the words into LIWC categories or WordNet ancestors. The good results for social relationships show that the presented method works well with certain questions. For future work it could be researched what kind of questions give the best answers for information extraction. Also, when the expected answers are know, a vocabulary could be created manually which includes expected words in combination with their well-being level. For example, for social relationships the words *social worker* or *therapist* could indicate a lower well-being. Such a vocabulary would need to be supported by well-being literature. Lastly, as already mentioned in Section 5.3.2, an algorithm could be added which splits the compound words into its separate words.

The results from this chapter suggested that part-of-speech frequencies, sentence length and sentiment did have a positive influence on the classification, however a bigger dataset was needed to verify this. The results from Table 5.4 and Table 5.5 showed that the more unbalanced the data, the better the results. This is a logical outcome, because if two-thirds of the data points have a high verification value, and the classifier just predicts everything to be a high verification, then the accuracy of the classifier is already 66%. The most straightforward way to work around this is to gather a balanced dataset. However, well-being data often has a positive bias in the answer. Section 4.2 showed this bias in literature, which found the mean of the Flourishing Scale statements to be around 5 or 6. This means that obtaining a balanced dataset for classification is practically infeasible. A well-functioning classifier could still be obtained by either creating a big dataset, or by training it with different training data. Kou et al. [48] showed good results are possible with datasets around 1000 textual data entries. To obtain such a number of participants, funds might be needed to encourage people to participate. Alternatively, the conversation could be made shorter to make it easier for people to participate. Also, gathered data could be augmented through data augmentation, as was (among others) done by Wei and Zou [92]. If this size of dataset cannot be gathered, another approach could be to gather training data through other sources. For example, typed responses with a conversational agent could be used as training data instead of spoken transcriptions. Alternatively, for every question other public/private datasets could be searched for which the data is similar to the expected answer. In English this was done, for example, by Preotiuc-Pietro et al. [65], Adiga et al. [2] and Wu [96].

All in all, the presented methods showed potential for well-being classification through spoken conversation with a conversational agent. Extracting social contacts, activities and meaningful factors give information about a person's well-being. With the improvements mentioned above, it should also be possible to automatically get an indication of a person's well-being level from the gathered data.

6 Conclusions

6.1 Summary

This research was aimed at answering the question: *"How can we measure a person's well-being through spoken conversation with an agent?"* This question was divided into four different sub-questions covering the definitions of well-being and their measures, the design of a conversation implementing these theories, the analysis of the dataset gathered with this conversation, and the information extraction and classification performed on this dataset.

The first sub-question, *"How can well-being be defined and measured?"*, was answered through a literature study on different definitions, models and measurements of well-being. Two main well-being philosophies were identified: hedonic and eudaimonic well-being. Hedonic well-being is grounded in pleasure, and is often referred to as subjective well-being in psychological theories. Subjective well-being consists of experienced positive/negative affect and life satisfaction. Eudaimonic well-being is grounded in virtue ethics and represents long-term well-being. This well-being philosophy is often implemented as psychological well-being, for which there are many different psychological theories. The Positive Health model [36] that is currently used in the BLISS project was not made as a measurement tool and not all aspects of it are supported by well-being literature. Therefore, I argued to use a combination of a subjective well-being measure (of which there are plenty) and the Psychological Well-being dimensions by Ryff [73] or the Flourishing Scale by Diener et al. [21] instead, since these measures are grounded in psychological theory and verified by various studies. For the current thesis the Flourishing Scale [21] was chosen as psychological well-being measure because it has been validated in Dutch [77] and with older adults [25].

Based on the chosen models, I designed a conversation to measure someone's well-being. To do this, I first conducted a literature study which identified speech-specific design heuristics and related literature on social conversations with an agent. This literature study resulted in a list of design heuristics that were taken into account in the design of the conversation. Examples are to establish common ground or to guide the user through the conversation. After the literature study, I adapted the questions from the well-being models (subjective well-being and the Flourishing Scale) into a conversation, which led to four topics that the conversation covered: affect, activities, social, and meaning. The original statements were changed into questions which were appropriate to use in a conversation and which made the conversation flow nicely. The literature findings were also implemented in this step, for example by giving the agent a name, adding funny replies like the agent talking about watching cat videos, and adding small topic introductions to guide the user through the conversation. All of this combined resulted in a draft conversation which was discussed with experts in three fields: mental well-being, conversational agents, and health communication. The conversation was changed after every expert interview. The resulting conversation was then tested in a pilot test, which led to some last changes and removal of bugs.

The designed conversation was implemented using the conversational software called WhappBot, and then used in two experiments to gather a dataset. The first, original dataset was gathered through personal contacts, resulting in a dataset with 35 entries. Since this dataset was too unbalanced for classification, I set up an additional experiment in which the participants answered in the role of people with a lower well-being, as were described in personas. This experiment resulted in 13 answers from psychology students. The gathered data was used to answer the third sub-question: *"What useful information can be extracted from a dataset gathered with the conversation that was designed?"*. I analyzed the dataset, describing the nature of the answers and looking for common themes in the answers of low, neutral and high well-being. For some questions clear distinctions were found between the different indications of well-being, like for the question about competence in doing daily activities. Others did not show any difference between low, neutral and high well-being, like the answers about experienced emotions or social relationships. The answers from the experiment with personas were generally shorter with less variation and details than the answers from the original experiment. But for most of the questions the answers resembled the original dataset, so the data from the two experiments could be used together as one dataset for the information extraction and classification.

The last sub-question, *"Can transcribed conversation data be used to automatically gather information about someone's well-being?"*, was answered using information extraction and classification techniques. The information extraction was focused on three types of information, which were identified previously during the data analysis: social relationships, activities, and meaningful things. To extract these, I used an approach combining the LIWC and WordNet lexicons. I found that the quality of the results depended on the kind of data extracted. The social relationships and some of the activities could be extracted well, but the answers about meaningful things were too broad to find fitting labels in the lexicons. I concluded that this approach works well if you know what kind of answers are expected. The results from the information extraction, together with the textual answers and information like length and sentiment, were used as input for classifiers. Three different classification algorithms were compared, and the influence of the added features was studied. Unfortunately, the classification results were not good enough to say that transcribed conversation data can be used to classify someone's well-being directly. Nevertheless, the transcribed data can successfully be used to gather information about things that influence someone's well-being, such as activities and social relationships.

6.2 | Discussion

The research in this thesis showed a way to gather information on a person's well-being through a spoken conversation with an agent. Even though all the phases of this research included a literature study to make informed decisions, still there are things that could have been done differently or in a better way. This section will discuss these points and give recommendations for future research on this topic.

To start, the choices made in the first stages of this research were not the most optimal for the later stages. From the literature on well-being (discussed in Chapter 2), several models were found that combine hedonic and eudaimonic well-being, thus encompassing the full spectrum of human well-being. I chose to design a conversation covering all of subjective well-being and psychological well-being, resulting in ten statements about well-being. Ten statements seemed not too extensive to cover in one conversation, but it turned out differently. The conversation became longer than I originally expected because most of the well-being statements could not be asked directly in the conversation with the agent. Instead, the statements were split up in questions targeting the separate facets of the statement, and introductory questions were added to make the conversation flow more naturally. The resulting conversation consisted of 24 questions and 10 verification statements, which took the participant around 30 minutes to finish. This time investment, together with the already sensitive nature of the experiment, made finding participants difficult.

Secondly, also the information extraction and classification were negatively influenced by the choices made in the first stages of this research. The formulations of the questions were designed to cover the well-being statements fully and to make the conversation flow nicely. However, the information extraction was not explicitly included in the conversation design. During the information extraction and classification phase, it turned out that the chosen formulations were not all suited for information extraction. It was found that information extraction gives good results when you know what kind of answers to expect. For example, the questions about social contacts yielded answers similar to each other, thus the social contacts could almost all be extracted. Contrarily, the question about meaningful things in life gave such diverging answers that information extraction was impossible (besides just extracting all noun phrases). So, for designing an agent conversation for information extraction, and by extension classification, the questions should be more explicitly designed to get certain types of answers.

Finally, it should be discussed whether classification is the best focus for a spoken conversation with an agent about well-being. The Positive Health model [36] (Section 2.1) was created as a conversational tool to talk about health, and not as a measurement tool. There already are studies that focused on reminiscence or mindfulness interventions to improve well-being. These concepts are based on the assumption that thinking about well-being related things can already improve the situation. Indeed, Harrington and Loffredo [31] have shown that a person's insight into their own well-being can be used as a direct prediction for both subjective and psychological well-being. With this information, the focus of well-being conversations with a conversational agent could be switched from measuring someone's well-being to improving someone's insight in their well-being. To do this, the literature research on well-being and the conversational design process could be used as a foundation.

6.3 | Future work

This research has shown different opportunities for future research. These opportunities were identified both in the conclusions of the separate chapters and in the overall discussion. This section discusses these opportunities in the order of the research.

To start, in the literature research on well-being a broad overview of the definitions was given and two well-being models were chosen. For future research the focus could be on implementing different well-being models in conversations with agents, like the PERMA model [81] or the Psychological Well-Being dimensions by Ryff [73]. Besides choosing another well-being model, future research could focus on one aspect of a well-being model. These aspects could be chosen from the ones used in this thesis, which are *activities, social relationships* and *meaning* for the Flourishing Scale, and *affect* and *life satisfaction* for subjective well-being. But if other models are chosen, these aspects could be *positive emotion, engagement, relationships, meaning* and *achievement* for PERMA [81] or *self-acceptance, positive relations with others, autonomy, environmental mastery, purpose in life* and *personal growth* for Psychological Well-Being [73]. Focusing on only one of these facets would shorten the conversation, most likely making it easier to find participants. On top of that, there is more room to change the questions, or add questions, without the conversation being too long. Then questions could be designed specifically for information extraction and classification.

For the conversation design, the iterative design process used for the design of the conversation worked very well. The input of the different experts was very helpful to translate the essence of the well-being models into the conversation, and to create the most optimal formulation for the conversation flow. I would recommend using this iterative approach with experts in future research. There are a number of research opportunities in the conversation design. The formulations of the questions could be studied to see what the influence of the formulation is on the kind of answers that are given. Examples of research already studying the influence of formulation and conversation style on given answers are the papers by Moroney and Cameron [59] and Kim et al. [47]. Furthermore, the persona of the agent could be changed in various different ways, like changing the voice, gender or adding an appearance. For each of these aspects, the influence on the perception of the agent could be studied. There have already been numerous studies on the perception of conversational agents. Examples are the influence of age, gender and role on the perception by ter Stal et al. [88], or the relationship between vocal dynamics and trust by Elkins and Derrick [24]. Relating these studies to conversational agents about well-being, it could be studied whether the previous findings hold when the conversation is about a very sensitive topic like personal well-being.

The data analysis mentions possibilities for future work in the use of personas in data collection. The data showed a decrease in length and detail of the answers when using personas. For future research it would be useful to see what caused this decrease in quality, for example whether it was caused by the use of personas or by the source of participants (students who got credits). If the personas are the cause, different versions could be made and it could be studied what version causes answers that are most similar to the real data. A design method for personas that yield answers similar to the real data could be very useful for gathering a balanced dataset. These versions could differ, for example, on length or number of details, or the persona could be created with the help of the participant.

During the information extraction and classification phase, various improvements were already identified (see Section 5.5). The main improvement point identified during this phase was the size of the training dataset. The gathered dataset was skewed towards values 5 and 6, which was also the case for previous studies using the Flourishing Scale [25, 77, 82]. This means that either lots of data needs to be gathered to have enough negative data, or other solutions need to be used to focus on gathering negative data. Besides using personas, as was done in this research, techniques like data augmentation [92] could be useful. More data could be gathered by making the conversation shorter, paying participants, or again by augmenting the data [92]. Besides increasing the size of the dataset, other sources could be used as training data, like using typed responses instead of spoken transcriptions or using other public/private datasets with similar data [2, 65, 96]. Also, technical improvements could be made, like splitting compound words into separate words [55], or creating a domain-specific vocabulary to capture the expected words and how they relate to well-being, for example the words *social worker* or *therapist* might indicate a lower well-being.

There are also opportunities for future research building upon this research more generally. Firstly, the choice of the well-being models and the design of the conversation could be discussed with care professionals to identify what information they would be interested in. Besides well-being classification, one could discuss information gathering, automatic conversation summaries or factors that give insights in someone's well-being. Then, the best implementation for this could be researched and evaluated with the care professionals. Secondly, future research could dive into the best way to use the extracted information in future conversations with the agent. Examples of this research could be to focus on real time information extraction and instant reply generation, long term data storage and access of the extracted information or selecting fitting follow-up conversations.

Bibliography

[1] Accenture. Memory Lane Conversational AI Solution, 2021. URL https://www.accenture.com/us-en/case-studies/interactive/memory-lane-conversational-ai-solution.

[2] D. T. Adiga, M. Bhavsar, U. Palan, and S. Patel. Daily journals: Extracting insights for well-being. *ACM International Conference Proceeding Series*, pages 305–315, 2020. doi: 10.1145/3421937.3421983.

[3] Anne4Care. Anne, 2020. URL https://anne4care.nl/.

[4] A. Asai, S. Evensen, B. Golshan, A. Halevy, V. Li, A. Lopatenko, D. Stepanov, Y. Suhara, W. C. Tan, and Y. Xu. HappyDB: A corpus of 100,000 crowdsourced happy moments. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 647–655, 2019.

[5] M. Baldassare, S. Rosenfield, and K. Rook. The Types of Social Relations Predicting Elderly Well-Being. *Research on Aging*, 6(4):549–559, 1984. ISSN 15527573. doi: 10.1177/0164027584006004006.

[6] F. Bechet, A. Nasr, and B. Favre. Adapting dependency parsing to spontaneous speech for open domain spoken language understanding. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (September):135–139, 2014. ISSN 19909772. doi: 10.21437/interspeech.2014-39.

[7] T. W. Bickmore and J. Cassell. Relational Agents: A Model and Implementation of Building User Trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 396–403, 2001.

[8] T. W. Bickmore, D. Schulman, and C. L. Sidner. A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. *Journal of Biomedical Informatics*, 44(2):183–197, 2011. ISSN 15320464. doi: 10.1016/j.jbi.2010.12.006. URL http://dx.doi.org/10.1016/j.jbi.2010.12.006.

[9] P. Boot, H. Zijlstra, and R. Geenen. The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, 6(1):65–76, 2017. ISSN 2211-7245. doi: 10.1075/dujal.6.1.04boo.

[10] R. Boumans, F. Van Meulen, K. Hindriks, M. Neerincx, and M. G. Rikkert. Robot for health data acquisition among older adults: A pilot randomised controlled cross-over trial. *BMJ Quality and Safety*, 28(10):793–799, 2019. ISSN 20445415. doi: 10.1136/bmjqs-2018-008977.

[11] F. B. Bryant, E. D. Chadwick, and K. Kluwe. Understanding the Processes that Regulate Positive Emotional Experience: Unsolved Problems and Future Directions for Theory and Research on Savoring. *International Journal of Wellbeing*, 1(1):107–126, 2011. doi: 10.5502/ijw.v1i1.18.

[12] Z. Callejas and D. Griol. Conversational Agents for Mental Health and Wellbeing. *Logic, Argumentation and Reasoning*, 22:219–244, 2021. ISSN 22149139. doi: 10.1007/978-3-030-61438-6{_}11.

[13] L. Chen, T. Gong, M. Kosinski, D. Stillwell, and R. L. Davidson. Building a profile of subjective well-being for social media users. *PLoS ONE*, 12(11):1–15, 2017. ISSN 19326203. doi: 10.1371/journal.pone.0187278.

[14] L. Clark, P. Doyle, D. Garaialde, E. Gilmartin, S. Schlögl, J. Edlund, M. Aylett, J. Cabral, C. Munteanu, J. Edwards, and B. R Cowan. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers*, 31(4):349–371, 2019. ISSN 09535438. doi: 10.1093/iwc/iwz016.

[15] L. Clark, N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, E. Gilmartin, C. Murad, C. Munteanu, V. Wade, and B. R. Cowan. What makes a good conversation? Challenges in designing truly conversational agents. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–12, 2019. doi: 10.1145/3290605.3300705.

[16] A. Darcy, J. Daniels, D. Salinger, P. Wicks, and A. Robinson. Evidence of Human-Level Bonds Established With a Digital Conversational Agent: Cross-sectional, Retrospective Observational Study. *JMIR Form Res*, 5(5), 2021.

[17] M. de Buenaga Rodríguez, J. M. G. Hidalgo, and B. Díaz-Agudo. Using WordNet to complement training information in text categorisation. *Proceedings of 2nd International Conference on Recent Advances in Natural Language Processing II: Selected Papers from RANLP*, 97:353–364, 2000. doi: 10.1075/cilt.189.35deb.

[18] E. Diener and M. E. Suh. Subjective well-being and age: An international analysis. In *Annual review of gerontology and geriatrics, Vol. 17: Focus on emotion and adult development.*, Annual review of gerontology and geriatrics., pages 304–324. Springer Publishing Company, New York, NY, US, 1998. ISBN 0-8261-6499-4 (Hardcover).

[19] E. Diener, R. A. Emmons, R. J. Larsen, and S. Griffin. The Satisfaction With Life Scale. *Journal of Personality Assessment*, 49(1):71–75, 1985.

[20] E. Diener, E. M. Suh, R. E. Lucas, and H. L. Smith. Subjective Well-being: Three Decades of Progress. *Psychological Bulletin*, 125(3):276–302, 1999. ISSN 00332909. doi: 10.1037/0033-2909.125.2.276.

[21] E. Diener, D. Wirtz, W. Tov, C. Kim-Prieto, D. w. Choi, S. Oishi, and R. Biswas-Diener. New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, 97(2):143–156, 2010. ISSN 03038300. doi: 10.1007/s11205-009-9493-y.

[22] K. Dobrovoljc and M. Martinc. Er... well, it matters, right? On the role of data representations in spoken language dependency parsing. *EMNLP 2018 - 2nd Workshop on Universal Dependencies, UDW 2018 - Proceedings of the Workshop*, (Udw):37–46, 2018. doi: 10.18653/v1/w18-6005.

[23] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, 6(12), 2011. ISSN 19326203. doi: 10.1371/journal.pone.0026752.

[24] A. C. Elkins and D. C. Derrick. The Sound of Trust: Voice as a Measurement of Trust During Interactions with Embodied Conversational Agents. *Group Decision and Negotiation*, 22(5):897–913, 2013. ISSN 09262644. doi: 10.1007/s10726-012-9339-x.

[25] Z. Fassih-Ramandi, M. A. Soleimani, K.-A. Allen, O. Gorgulu, and S. A. Motalebi. Validity and reliability of the flourishing scale in a sample of older adults in Iran. *Clinical Interventions in Aging*, 15:673–681, 2020. ISSN 11781998. doi: 10.2147/CIA.S268539.

[26] M. W. Fordyce. A review of research on the happiness measures: A sixty second index of happiness and mental health. *Social Indicators Research*, 20(1988):335–381, 1988.

[27] S. L. Gable and J. Haidt. What (and why) is positive psychology? *Review of General Psychology*, 9 (2):103–110, 2005. ISSN 10892680. doi: 10.1037/1089-2680.9.2.103.

[28] L. Gatti and J. van Stegeren. Improving Dutch sentiment analysis in Pattern. *Computational Linguistics in the Netherlands Journal*, 10:73–89, 2020. ISSN 22114009.

[29] E. A. Greenfield and N. F. Marks. Formal volunteering as a protective factor for older adults' psychological well-being. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, 59(5):258–264, 2004. ISSN 10795014. doi: 10.1093/geronb/59.5.S258.

[30] G. Grossi, R. Lanzarotti, P. Napoletano, N. Noceti, and F. Odone. Positive technology for elderly well-being: A review. *Pattern Recognition Letters*, 137:61–70, 2020. ISSN 01678655. doi: 10.1016/j.patrec.2019.03.016.

[31] R. Harrington and D. A. Loffredo. Insight, rumination, and Self-Reflection as predictors of well-being. *Journal of Psychology: Interdisciplinary and Applied*, 145(1):39–57, 2011. ISSN 00223980. doi: 10.1080/00223980.2010.528072.

[32] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann. Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20–38, 2019. ISSN 01678116. doi: 10.1016/j.ijresmar.2018.09.009.

[33] W. He, D. Goodkind, and P. Kowal. An Aging World : 2015 International Population Reports. *Aging*, (March):165, 2016. doi: 10.13140/RG.2.1.1088.9362.

[34] L. W. Henderson, T. Knight, and B. Richardson. An exploration of the well-being benefits of hedonic and eudaimonic behaviour. *Journal of Positive Psychology*, 8(4):322–336, 2013. ISSN 17439760. doi: 10.1080/17439760.2013.803596.

[35] A. Hogenboom, B. Heerschop, F. Frasincar, U. Kaymak, and F. De Jong. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. *Decision Support Systems*, 62:43–53, 2014. ISSN 01679236. doi: 10.1016/j.dss.2014.03.004.

[36] M. Huber, J. André Knottnerus, L. Green, H. Van Der Horst, A. R. Jadad, D. Kromhout, B. Leonard, K. Lorig, M. I. Loureiro, J. W. Van Der Meer, P. Schnabel, R. Smith, C. Van Weel, and H. Smid. How should we define health? *BMJ (Online)*, 343(7817):1–3, 2011. ISSN 09598146. doi: 10.1136/bmj.d4163.

[37] M. A. S. Huber. *Towards a new , dynamic concept of Health : its operationalisation and use in public health and healthcare, and in evaluating health effects of food.* Number 2014. 2014. ISBN 9789462594715. doi: 10.26481/dis.20141217mh.

[38] R. Hursthouse and G. Pettigrove. Virtue Ethics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 201 edition, 2018. URL https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/.

[39] V. Huta and R. M. Ryan. Pursuing Pleasure or Virtue: The Differential and Overlapping Well-Being Benefits of Hedonic and Eudaimonic Motives. *Journal of Happiness Studies*, 11(6):735–762, 2010. ISSN 13894978. doi: 10.1007/s10902-009-9171-4.

[40] B. Inkster, S. Sarda, and V. Subramanian. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):1–14, 2018. ISSN 22915222. doi: 10.2196/12106.

[41] Institute for Positive Health. Positieve gezondheid en meetinstrumenten. Technical report.

[42] Intuition Robotics. ElliQ, 2021. URL https://elliq.com/.

[43] C. L. M. Keyes. Social Well-Being Author (s): Corey Lee M . Keyes Source : Social Psychology Quarterly , Vol . 61 , No . 2 (Jun ., 1998), pp . 121-140 Published by : American Sociological Association Stable URL : http://www.jstor.org/stable/2787065 Accessed : 17-06-2. *American Sociological Association*, 61(2):121–140, 1998. URL https://www.jstor.org/stable/2787065.

[44] C. L. M. Keyes and S. J. Lopez. Toward a science of mental health. In *Handbook of Positive Psychology*, chapter Towards a, pages 89–95. Oxford University Press, 2001.

[45] D. Khaw and M. L. Kern. A cross-cultural comparison of the PERMA model of well-being. *Undergraduate Journal of Psychology at Berkeley*, 1:1–22, 2015. URL http://ujpb.org/a-cross-cultural-comparison-of-the-perma-model-of-well-being/.

[46] H. Kim, G. Lee, Y. K. Lim, D. Y. Koh, and J. M. Park. Designing personalities of conversational agents. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–6, 2019. doi: 10.1145/3290607.3312887.

[47] S. Kim, J. Lee, and G. Gweon. Comparing Data from Chatbot and Web Surveys. pages 1–12, 2019. doi: 10.1145/3290605.3300316.

[48] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing Journal*, 86:105836, 2020. ISSN 15684946. doi: 10.1016/j.asoc.2019.105836. URL https://doi.org/10.1016/j.asoc.2019.105836.

[49] A. B. Krueger and D. A. Schkade. The reliability of subjective well-being measures. *Journal of Public Economics*, 92(8-9):1833–1845, 2008. ISSN 00472727. doi: 10.1016/j.jpubeco.2007.12.015.

[50] L. Lemmens, S. de Bruin, M. Beijer, R. Hendrikx, and C. Baan. Het gebruik van brede gezondheidsconcepten : inspirerend en uitdagend voor de praktijk Samenvatting. Technical Report December, Rijksinstituut voor Volksgezondheid en Milieu, 2019.

[51] E. Levin and A. Levin. Evaluation of spoken dialogue technology for real-time health data collection. *Journal of Medical Internet Research*, 8(4), 2006. ISSN 14388871. doi: 10.2196/jmir.8.4.e30.

[52] F. Li, X. Bai, and Y. Wang. The Scale of Positive and Negative Experience (SPANE): Psychometric Properties and Normative Data in a Large Chinese Sample. *PLoS ONE*, 8(4), 2013. ISSN 19326203. doi: 10.1371/journal.pone.0061137.

[53] M. Liu, G. Haffari, W. Buntine, and M. R. Ananda-Rajah. Leveraging linguistic resources for improving neural text classification. *Australasian Language Technology Association Workshop*, 2017. URL http://www.aclweb.org/anthology/U17-1004.

[54] S. Lyubomirsky and H. S. Lepper. A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, 46(2):137–155, 1999. ISSN 03038300. doi: 10.1023/A:1006824100041.

[55] L. Macken and A. Tezcan. Dutch Compound Splitting for Bilingual Terminology Extraction. *Multiword Units in Machine Translation and Translation Technology*, 341:147–162, 2018.

[56] R. Maharjan, D. A. Rohani, P. Bækgaard, J. E. Bardram, K. Doherty, P. Baekgaard, J. E. Bardram, and K. Doherty. Can we talk? Design Implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent. *ACM International Conference Proceeding Series*, 2021. doi: 10.1145/3469595.3469600.

[57] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. doi: 10.1145/219717.219748.

[58] A. Moore. Hedonism. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 201 edition, 2019. URL https://plato.stanford.edu/archives/win2019/entries/hedonism/.

[59] W. F. Moroney and J. Cameron. The Questionnaire as Conversation: Time for a Paradigm Shift, or at Least a Paradigm Nudge? *Ergonomics in Design*, 24(2):10–15, 2016. ISSN 10648046. doi: 10.1177/1064804615623631.

[60] C. Murad, C. Munteanu, B. R. Cowan, and L. Clark. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing*, 18(2):33–45, 2019. ISSN 15582590. doi: 10.1109/MPRV.2019.2906991.

[61] S. Nikitina, S. Callaioli, and M. Baez. Smart conversational agents for reminiscence. *Proceedings - International Conference on Software Engineering*, pages 52–57, 2018. ISSN 02705257. doi: 10.1145/3195555.3195567.

[62] P. Philip, J. A. Micoulaud-Franchi, P. Sagaspe, E. D. Sevin, J. Olive, S. Bioulac, and A. Sauteraud. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific Reports*, 7(February):1–7, 2017. ISSN 20452322. doi: 10.1038/srep42656.

[63] M. Pinquart and S. Sörensen. Influences of Socioeconomic Status, Social Network, and Competence on Subjective Well-Being in Later Life: A Meta-Analysis Preparation for Future Care Needs-general interest area View project How Effective Are Interventions With Caregivers? View project. *Psychology and Aging*, 15(2):187–224, 2000. URL https://www.researchgate.net/publication/12438552.

[64] M. Postma, R. Segers, and P. Vossen. Open Dutch WordNet. 2013.

[65] D. Preotiuc-Pietro, M. Sap, H. A. Schwartz, and L. Ungar. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task. pages 40–45, 2015. URL http://www.wwbp.org/data.html.

[66] C. A. Prinsen and C. B. Terwee. Measuring positive health: for now, a bridge too far. *Public Health*, 170:70–77, 2019. ISSN 14765616. doi: 10.1016/j.puhe.2019.02.024. URL https://doi.org/10.1016/j.puhe.2019.02.024.

[67] T. Rahm, E. Heise, and M. Schuldt. Measuring the frequency of emotions-validation of the scale of positive and negative experience (SPANE) in Germany. *PLoS ONE*, 12(2):1–10, 2017. ISSN 19326203. doi: 10.1371/journal.pone.0171288.

[68] S. Z. Razavi, M. R. Ali, L. K. Schubert, K. A. Van Orden, B. Kane, and T. Ma. Dialogue design and management for multi-session casual conversation with older adults. *CEUR Workshop Proceedings*, 2327, 2019. ISSN 16130073.

[69] E. Riloff and W. Lehnert. Information Extraction as a Basis for High-Precision Text Classification. *ACM Transactions on Information Systems (TOIS)*, 12(3):296–333, 1994. ISSN 15582868. doi: 10.1145/183422.183428.

[70] R. M. Ryan. Self - determination Theory and Wellbeing. *Wellbeing in Developing Countries*, (June): 1–2, 2009.

[71] R. M. Ryan and E. L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000. ISSN 0003066X. doi: 10.1037/0003-066X.55.1.68.

[72] R. M. Ryan and E. L. Deci. On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology*, 52:141–166, 2001. ISSN 00664308. doi: 10.1146/annurev.psych.52.1.141.

[73] C. D. Ryff. Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of personality and social psychology*, 57(6):1069, 1989. ISSN 10880224. doi: 10.1037/0022-3514.57.6.1069.

[74] C. D. Ryff and B. H. Singer. Psychological Well-Being: Meaning, Measurement, and Implications for Psychotherapy Research. *Psychotherapy and psychosomatics*, 65(1):14–23, 1996.

[75] C. D. Ryff and B. H. Singer. Know thyself and become what you are: A eudaimonic approach to psychological well-being. *Journal of Happiness Studies*, 9(1):13–39, 2008. ISSN 13894978. doi: 10.1007/s10902-006-9019-0.

[76] K. R. Scherer, A. Schorr, and T. Johnstone. Appraisal processes in emotion: Theory, methods, research. *Oxford University Press*, 2001.

[77] M. Schotanus-Dijkstra, P. M. ten Klooster, C. H. Drossaert, M. E. Pieterse, L. Bolier, J. A. Walburg, and E. T. Bohlmeijer. Validation of the Flourishing Scale in a sample of people with suboptimal levels of mental well-being. *BMC Psychology*, 4(1):1–10, 2016. ISSN 20507283. doi: 10.1186/s40359-016-0116-5. URL http://dx.doi.org/10.1186/s40359-016-0116-5.

[78] H. A. Schwartz, M. Sap, M. L. Kern, J. C. Eichstaedt, A. Kapelner, M. Agrawal, E. Blanco, L. Dziurzynski, G. Park, D. Stillwell, M. Kosinski, M. E. Seligman, and L. H. Ungar. Predicting individual well-being through the language of social media. *Pacific Symposium on Biocomputing*, pages 516–527, 2016. ISSN 23356936. doi: 10.1142/9789814749411{_}0047.

[79] S. Scott and S. Matwin. Text Classification Using WordNet Hypernyms. *Learning*, pages 45–51, 1998. URL http://acl.ldc.upenn.edu/W/W98/W98-0706.pdf.

[80] M. E. P. Seligman. Positive Psychology, Positive Prevention, and Positive Therapy. In Snyder, C. R & Lopez, S.J. *Handbook of Positive Psychology*, pages 3–9, 2002.

[81] M. E. P. Seligman. *Flourish: A visionary new understanding of happiness and well-being.* Simon and Schuster, 2012.

[82] A. J. Silva and A. Caetano. Validation of the Flourishing Scale and Scale of Positive and Negative Experience in Portugal. *Social Indicators Research*, 110(2):469–478, 2013. ISSN 03038300. doi: 10.1007/s11205-011-9938-y.

[83] Stockholm Exergi. Memory Lane (eng). URL https://www.stockholmexergi.se/memory-lane2/.

[84] A. Strauss and J. M. Corbin. *Grounded theory in practice*. Sage, 1997.

[85] B. Suhm. Towards best practices for speech user interface design. *EUROSPEECH 2003 - 8th European Conference on Speech Communication and Technology*, pages 2217–2220, 2003.

[86] P. Tanawongsuwan. Part-of-speech approach to evaluation of textbook reviews. *2nd International Conference on Computer and Network Technology, ICCNT 2010*, pages 352–356, 2010. doi: 10.1109/ICCNT.2010.31.

[87] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods, 2010. ISSN 0261927X.

[88] S. ter Stal, M. Tabak, H. op den Akker, T. Beinema, and H. Hermens. Who Do You Prefer? The Effect of Age, Gender and Role on Users' First Impressions of Embodied Conversational Agents in eHealth, 2020. ISSN 15327590.

[89] P. Touset. Natural Language Processing for a Psychometer. 2007.

[90] J. van Waterschoot, I. Hendrickx, A. Khan, E. Klabbers, M. de Korte, H. Strik, C. Cucchiarini, and M. Theune. BLISS: An agent for collecting spoken dialogue data about health and well-being. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, (May):449–458, 2020.

[91] D. Watson, L. A. Clark, and A. Tellegen. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of personality and social psychology*, 54(6):1063–1070, 1988. ISSN 00223514. doi: 10.1037/0022-3514.54.6.1063.

[92] J. Wei and K. Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 6382–6388, 2019. doi: 10.18653/v1/d19-1670.

[93] Z. Wei and J. A. Landay. Evaluating Speech-Based Smart Devices Using New Usability Heuristics. *IEEE Pervasive Computing*, 17(2):84–96, 2018. ISSN 15361268. doi: 10.1109/MPRV.2018.022511249.

[94] N. Wiratunga, A. Wijekoon, C. Palihawadana, K. Cooper, and V. Mendham. {F}it{C}hat: Conversational {AI} for Active Ageing. *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, pages 21–31, 2020. URL https://aclanthology.org/2020.intellang-1.3.

[95] World Health Organization. Constitution. URL https://www.who.int/about/governance/constitution.

[96] J. Wu. *WellBe: A Conversational Agent for Well-Being*. PhD thesis, UC Santa Cruz, 2020. URL https://escholarship.org/uc/item/9nx4s71n.

A | Comparison of the well-being models

		Bodily functions						
		Feeling healthy	Feeling fit	No physical complaints and/or pain	Sleeping	Eating	Physical condition	Physical activity
PERMA	Positive emotion							
	Engagement							
	Relationships							
	Meaning							
	Achievement							
SDT	Autonomy							
	Relatedness							
	Competence	?	?	?	?	?	?	?
SWB	Affect	x	x					
	Life satisfaction							
PWB	Self-acceptance							
	Positive relations with others							
	Autonomy							
	Environmental mastery							
	Purpose in life							
	Personal growth							

		Mental well-being						
		Being able to remember things	Being able to concentrate	Being able to communicate	Being cheerful	Accepting yourself	Being able to handle change	Feeling in control
PERMA	Positive emotion				x			
	Engagement							
	Relationships							
	Meaning							
	Achievement							
SDT	Autonomy	x	x	x				x
	Relatedness							
	Competence	?	?			?	?	?
SWB	Affect							
	Life satisfaction					x		
PWB	Self-acceptance					x		
	Positive relations with others							
	Autonomy	x	x	x				x
	Environmental mastery	x	x	x			x	x
	Purpose in life							
	Personal growth							

		Meaningfullness						
		Having a meaningful life	Having a zest for life	Persuing ideals	Feeling confident	Accepting life	Being grateful	Lifelong learning
	Positive emotion							
	Engagement							
	Relationships							
	Meaning	x	x	x				
PERMA	Achievement							?
	Autonomy							
	Relatedness							
SDT	Competence							
	Affect							
SWB	Life satisfaction					x		
	Self-acceptance					x		
	Positive relations with others							
	Autonomy				x			
	Environmental mastery							
	Purpose in life	x	x	x				
PWB	Personal growth							x

		Quality of life						
		Enjoyment	Being happy	Feeling good	Feeling well-balanced	Feeling safe	Housing circumstances	Having enough money
PERMA	Positive emotion	x	x	x	x	x	x	x
	Engagement							
	Relationships							
	Meaning							
	Achievement						?	?
SDT	Autonomy							
	Relatedness							
	Competence							x
SWB	Affect	x	x	x	x	x		
	Life satisfaction							
PWB	Self-acceptance							
	Positive relations with others							
	Autonomy							
	Environmental mastery							
	Purpose in life							
	Personal growth							

		Participation						
		Social contacts	Being taken seriously	Doing fun things together	Having the support from others	Sense of belonging	Doing meaningful things	Being interested in society
	Positive emotion							
	Engagement					x	x	x
	Relationships	x	x	x	x	x		x
	Meaning							
PERMA	Achievement							
	Autonomy							
	Relatedness	x	x	x	x	x	x	x
SDT	Competence							
	Affect							
SWB	Life satisfaction							
	Self-acceptance							
	Positive relations with others	x	x	x	x	x	x	
	Autonomy							
	Environmental mastery							
	Purpose in life							
PWB	Personal growth							

		Daily functioning						
		Taking care of yourself	Knowing your limitations	Knowledge of health	Managing time	Managing money	Being able to work	Being able to ask for help
PERMA	Positive emotion							
	Engagement							
	Relationships							
	Meaning							
	Achievement	?			?	?	?	
SDT	Autonomy							
	Relatedness							
	Competence		x	x	x	x		?
SWB	Affect							
	Life satisfaction							
PWB	Self-acceptance		x					
	Positive relations with others							
	Autonomy	x						
	Environmental mastery				x	x		
	Purpose in life							
	Personal growth							

B | The Flourishing Scale in English and Dutch

B.1 | English version

FLOURISHING SCALE

Below are 8 statements with which you may agree or disagree. Using the 1–7 scale below, indicate your agreement with each item by indicating that response for each statement.

- 7 - Strongly agree
- 6 - Agree
- 5 - Slightly agree
- 4 - Neither agree nor disagree
- 3 - Slightly disagree
- 2 - Disagree
- 1 - Strongly disagree

____ I lead a purposeful and meaningful life

____ My social relationships are supportive and rewarding

____ I am engaged and interested in my daily activities

____ I actively contribute to the happiness and well-being of others

____ I am competent and capable in the activities that are important to me

____ I am a good person and live a good life

____ I am optimistic about my future

____ People respect me

Scoring:

Add the responses, varying from 1 to 7, for all eight items. The possible range of scores is from 8 (lowest possible) to 56 (highest PWB possible). A high score represents a person with many psychological resources and strengths

B.2 | Dutch version

FLOURISHING SCALE

Geef aan in hoeverre je het eens of oneens bent met de volgende stellingen.

1 Helemaal niet mee eens	2 Mee oneens	3 Een beetje mee oneens	4 Noch mee oneens, noch mee eens	5 Een beetje mee eens	6 Mee eens	7 Heel erg mee eens

____ Ik leid een doelgericht en zinvol leven

____ Mijn sociale relaties zijn ondersteunend en belonend.

____ Ik ben geëngageerd met en geïnteresseerd in mijn dagelijkse activiteiten.

____ Ik draag actief bij aan het geluk en welzijn van anderen.

____ Ik ben capabel en in staat om de activiteiten uit te voeren die belangrijk voor mij zijn.

____ Ik ben een goed persoon en leid een goed leven.

____ Ik ben optimistisch over mijn toekomst.

____ Mensen respecteren mij.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi. D., Oishi, S., & Biswas-Diener, R. (2009). New measures of well-being: Flourishing and positive and negative feelings. *Social Indicators Research, 39*, 247-266.

Dutch translation:
Dr. Marieke van Egmond & Dr. Katja Hanke
Jacobs University Bremen
Bremen International Graduate School of Social Sciences (BIGSSS)
Germany

C | Final conversation in Dutch

Agent	***User***
Introduction	
Hallo! Ik ben Marli. Ik zou graag in gesprek gaan om je beter te leren kennen, en te leren hoe jij in het leven staat. Hoe mag ik jou noemen?	
	Text input
Hallo <naam>, wat leuk je te spreken! Nog even wat praktische dingen. Wees niet bang om te veel te praten, hoe meer jij praat hoe meer ik kan leren! Omdat ik maar een chatbot ben heb ik soms moeite om mensen te begrijpen. Daarom wil ik je vragen om langzaam en duidelijk te praten. Soms zal ik je vragen om een vraag te beantwoorden door een van de meerkeuze opties te kiezen. Dit helpt mij om jouw antwoord te dubbelchecken. Als je een vraag niet snapt of te moeilijk vindt dan ben je vrij om de vraag niet te beantwoorden door niks op te nemen. En natuurlijk ben je altijd vrij om het gesprek te stoppen. Ik zal niet boos worden. Is alles zo duidelijk?	
	Ja/nee knoppen
Mooi! Dan kunnen we beginnen.	
Affect	
Laten we beginnen zoals mensen altijd beginnen... Hoe gaat het met je?	
	Gesproken antwoord
In een week komen veel emoties voorbij en veel situaties die voor bepaalde emoties zorgen. Sommige emoties zijn simpel te benoemen zoals blij, boos, of bang. Anderen kunnen lastiger te identificeren zijn, bijvoorbeeld als je je jaloers, beschaamd of geliefd voelt. Als je kijkt naar de afgelopen week, welke emoties heb jij vooral gevoeld?	
	Gesproken antwoord
Op een schaal van 1 tot 7, van altijd negatief naar altijd positief, hoe zou je deze zin afmaken? Mijn emoties waren de afgelopen week...	
	Knoppen voor verificatie
Sommige mensen zijn heel blij met hoe hun leven is, anderen zien dingen in hun leven liever anders. Hoe tevreden ben jij met je leven zoals het nu is?	
	Gesproken antwoord
Op een schaal van 1 tot 7, hoe erg ben je het eens met de stelling: ik ben tevreden met hoe mijn leven nu is?	
	Knoppen voor verificatie

Activities	
De volgende vragen gaan over activiteiten die jij doet. Kun je beschrijven welke dingen een vast onderdeel van je week zijn?	Gesproken antwoord
Dankjewel voor je antwoord! Sommige mensen geven aan dat ze hun dagelijkse activiteiten een sleur vinden, maar anderen zijn juist geïnteresseerd en gemotiveerd voor de dingen die ze doen. Hoe gemotiveerd ben jij voor jouw dagelijkse activiteiten?	Gesproken antwoord
Op een schaal van 1 tot 7, hoe erg ben je het eens met deze stelling: ik ben gemotiveerd en geïnteresseerd in mijn dagelijkse activiteiten?	Knoppen voor verificatie
Oke. En heb je voldoende activiteiten die je leuk vind om te doen? Welke activiteiten zijn dat?	Gesproken antwoord
Wat fijn om zo meer te leren over wie je bent! Misschien ga ik die activiteiten ook wel eens proberen, wie weet is het wat voor mij. In mijn vrije tijd ga ik graag door de collectie kattenfilmpjes op het internet heen. Zo'n lief zacht beestje maakt mijn dag altijd beter. Maar terug naar jou. Ik heb ondertussen al een goed beeld van hoe jouw dagen eruit zien. Ik vraag me nog wel af of het je lukt je dagen in te vullen zoals je zou willen. In welke mate ben je in staat je dag te vullen met activiteiten die je wil doen?	Gesproken antwoord
Op een schaal van 1 tot 7, hoe erg ben je het eens met de stelling: ik ben capabel en in staat om activiteiten uit te voeren die belangrijk voor mij zijn?	Knoppen voor verificatie
Social	
Naast wat je op een dag zoal doet, ben ik ook wel benieuwd naar met wie je deze activiteiten doet. Voor iedereen is dit anders. Sommigen hebben vooral veel contact met hun ouders, kinderen of andere familie. Anderen zoeken vooral hun vrienden of collega's op om contact mee te hebben. Met welke mensen heb jij in een normale week contact?	Gesproken antwoord
Het is altijd fijn om contact met anderen te hebben. Hoewel ik als chatbot mensen soms ook wel heel uitputtend kan vinden. Dan ben ik blij als ze weer weg zijn. Kijkend naar de mensen die je net hebt genoemd, hoeveel plezier haal je uit het contact met hen?	Gesproken antwoord
En bij wie van deze mensen kun je terecht als je hulp nodig hebt?	Gesproken antwoord
Op een schaal van 1 tot 7, hoe erg ben je het eens met de stelling: mijn sociale relaties zijn ondersteunend en belonend?	Knoppen voor verificatie

Naast dat de contacten voor jou plezier en hulp kunnen geven, hoe denk je dat de ander het contact met jou ervaart?	Gesproken antwoord
Waarom denk je dat anderen het contact met jou zo ervaren?	Gesproken antwoord
Op een schaal van 1 tot 7, hoe erg ben je het eens met de stelling: ik draag actief bij aan het geluk en welzijn van anderen?	Knoppen voor verificatie
Niet al het contact is altijd even makkelijk. Soms heb ik momenten dat ik iets duidelijk probeer te maken, maar dat die ander telkens door mij heen blijft praten. Of niet eens moeite lijkt te doen om mij te begrijpen. Ik voel me dan niet zo gerespecteerd. In hoeverre voel jij je gerespecteerd in je sociale relaties?	Gesproken antwoord
Op een schaal van 1 tot 7, hoe erg ben je het eens met de stelling: mensen respecteren mij?	Knoppen voor verificatie
Meaning	
We hebben nu veel dingen besproken die van buiten heel belangrijk lijken. Maar het is ook belangrijk hoe dat van binnen voelt. De sociale contacten waar we het net over hadden, in hoeverre geven die voor jou het leven zin?	Gesproken antwoord
Kun je nog andere dingen noemen die voor jou het leven de moeite waard maken?	Gesproken antwoord
Dat is mooi geformuleerd. Als virtueel wezen hebben mijn makers mij een duidelijk doel gegeven. Daar kan ik niks aan veranderen. Gelukkig kunnen mensen wel zelf beslissen wat ze willen doen en bereiken, en waarom ze dat willen. Kan je een voorbeeld geven van een doel dat je hebt met je leven?	Gesproken antwoord
Op een schaal van 1 tot 7, hoe erg ben je het eens met de stelling: ik leid een doelgericht en zinvol leven?	Knoppen voor verificatie
Ik vind mijzelf vaak heel grappig, maar anderen zijn het daar niet altijd mee eens. Wat vind je goede eigenschappen van jezelf?	Gesproken antwoord
Eigenschappen zoals aardig, zorgzaam of vrijgevig worden vaak toegeschreven aan een goed persoon. Dingen als egoïstisch, roekeloos of gesloten worden vaak genoemd over een slecht persoon. Als je naar jezelf kijkt, in hoeverre vind je jezelf een goed persoon?	Gesproken antwoord
Bedankt voor je openheid.	

Jullie mensen zijn sociale wezens en beïnvloeden elkaar en jullie omgeving continu. Soms op een positieve manier en soms ook in de negatieve zin. Denk terug aan die eigenschappen die je net hebt genoemd, op wat voor manier beïnvloeden die de mensen om je heen?	Gesproken antwoord
Kun je een voorbeeld geven van hoe je mensen om je heen op een positieve of negatieve manier beïnvloedt?	Gesproken antwoord
Mooi verwoord. Hoe tevreden ben je over de invloed die je hebt?	Gesproken antwoord
Op een schaal van 1 tot 7, hoe erg ben je het eens met de stelling: ik ben een goed persoon en leid een goed leven?	Knoppen voor verificatie
En als laatste onderwerp een paar vragen over de toekomst. Als je aan je toekomst denkt, wat voor beeld zie je dan voor je?	Gesproken antwoord
Welke wensen heb je voor je toekomst?	Gesproken antwoord
Hoeveel vertrouwen heb je dat die wensen echt uit zullen komen?	Gesproken antwoord
Op een schaal van 1 tot 7, hoe erg ben je het eens met de stelling: ik ben optimistisch over mijn toekomst?	Knoppen voor verificatie
Wrapping up	
Dat was het alweer. Bedankt voor al je antwoorden en je openheid! Als je een van deze vragen lastig vond en je graag door zou praten met iemand, klik dan op de volgende link. https://www.rijksoverheid.nl/onderwerpen/geestelijke-gezondheidszorg/vraag-en-antwoord/waar-vind-ik-hulp-bij-psychische-problemen. Leuk om je te ontmoeten! Geniet van de rest van je dag! Het gesprek is nu afgelopen. Je kunt het scherm sluiten.	

D | Personas used in the experiment

Persona 1 – Drukke moeder

Naam

Wilke Hendriks

Leeftijd

43

Gezinssituatie

Gescheiden met 3 kinderen (7, 10 en 13 jaar oud)

Werk

Schadebehandelaar bij een verzekeringsmaatschappij

Beschrijving

Wilke is een vrouw die haar vrienden en familie zorgzaam en gezellig zouden noemen. Toch is het ook duidelijk dat ze erg veel op haar bordje heeft. Na de scheiding 4 jaar geleden is de zorg van de 3 kinderen volledig op haar aangekomen. Hoewel ze dit met alle liefde doet is het lastig om dit te combineren met haar fulltimebaan bij een verzekeringsmaatschappij. Omdat het al zo druk is voelt ze weinig ruimte voor haar eigen behoeften en heeft ze haar eigen wensen uit het oog verloren. Ook haar hobby's en vrienden hebben vaak geen plek in haar agenda.

Persona 2 – Man in de sleur

Naam

Wout Blankema

Leeftijd

39

Gezinssituatie

Getrouwd met Suzanne, 2 kinderen (15 en 19)

Werk

Servicemonteur werktuigbouwkunde

Beschrijving

Wout heeft zijn leven al een tijdje op de rit. De kinderen zijn bijna volwassen, zijn baan is steady en zijn huwelijk loopt prima. Eigenlijk is elke dag hetzelfde: vroeg opstaan, krant lezen, naar werk, keuvelen met collega's, om 5 uur terug naar huis waar het eten klaar staat, en 's avonds tv kijken op de bank. De laatste paar maanden begint Wout steeds meer na te denken over zijn leven. Toen hij 20 was werd hij al vader en in de jaren daarna voelt het alsof zijn leven geleefd werd. Nu zit hij op de bank en vraagt hij zich af of dit het is. Alles in zijn leven loopt prima, maar wil hij dit wel? Waar is de hoop en energie gebleven waar hij als jonge jongen van overstroomde?

Persona 3 – Thuis door chronische ziekte

Naam

Steffan Molenaar

Leeftijd

34

Gezinssituatie

Single

Werk

Ziektewet

Beschrijving

Steffan werkte als verpleger in het ziekenhuis tot hij in de eerste weken dat Corona bij ons in het land was ziek werd. Nu, 2 jaar later, worden zijn klachten *long covid* genoemd. Bij Steffan komt dit vooral uit in vermoeidheid en kortademigheid. Waar hij voor de ziekte genoot van hele middagen wandelen of sporten is hij nu na 20 minuten dood op. Door deze vermoeidheidsklachten houdt hij werken niet vol. Omdat Steffan alleen woont heeft hij ook niet veel contact met anderen en voelt het alsof het leven op pauze is gezet. Hoelang dit nog gaat duren kunnen de artsen niet zeggen.

Persona 4 – Depressie

Naam

Nicole van der Velden

Leeftijd

31

Gezinssituatie

Single

Werk

Administratief medewerker in de zorg

Beschrijving

Voor buitenstaanders lijkt Nicole een heel gewone vrouw. Ze werkt al enkele jaren als administratief medewerker, kletst met collega's en spreekt soms af met vrienden. Maar wat niet veel mensen weten is dat ze ook al een aantal jaar aan een depressie lijdt. Hierdoor haalt Nicole geen vreugde of energie uit de dingen die ze doet. 's Ochtends sleept ze zichzelf met moeite uit bed. Vaak zonder de energie om te ontbijten gaat ze op een lege maag naar haar werk. Als ze thuiskomt ploft ze weer op de bank, waar ze haar tijd vult met gamen en series kijken. Pas diep in de nacht valt ze in slaap, want dan is ze moe genoeg om alle gedachtes te verdrijven die haar de hele dag door tergen.

Persona 5 – Gepensioneerde

Naam

Brechje Maas

Leeftijd

68

Gezinssituatie

Weduwe

Werk

Gepensioneerd docent Frans

Beschrijving

Brechje heeft jarenlang voor de klas gestaan op de middelbare school niet ver van haar huis. Omgeven door collega's en leerlingen genoot ze van haar dagen. Sinds ze een jaar geleden met pensioen is gegaan is ze in een gat gevallen. Waar haar dagen altijd goed gevuld waren, zijn ze nu bijna helemaal leeg. Met wat hobby's en een beetje rommelen in huis kan ze wel een paar uur per dag vullen, maar daarna is het meestal wachten op de avond zodat ze weer kan gaan slapen. Ook het sociale contact dat ze op haar werk had heeft ze nu niet meer. Steeds meer mist ze haar man die alweer 20 jaar geleden is overleden. Pas als je het niet meer hebt besef je hoe vol je leven was, en hoe leeg het kan zijn.

Persona 6 – Perfectionisme en angst

Naam

Tom van Dijk

Leeftijd

20

Gezinssituatie

Single, woont in een studentenhuis

Werk

Student psychologie

Beschrijving

Tom is tweedejaars student psychologie en heeft het erg naar zijn zin op zijn studie. Alleen is het voor Tom een stuk lastiger om dit leven te navigeren dan voor de meeste mensen. Tom heeft last van angsten en perfectionisme, waardoor de kleinste dingen al een strijd in zijn hoofd veroorzaken. Iets inleveren voor een cijfer, mensen om hulp vragen, naar een onbekende situatie… voor al deze situaties heeft Tom uren met stress. Het is alsof hij een groot betonblok achter zich aan sleept en het super veel moeite kost voordat iets lukt. Al deze angsten hebben ervoor gezorgd dat Tom zich vaak uitgeput voelt en daar lijden zijn sociale contacten en zijn schoolresultaten onder.

E | Information brochures and consent forms

E.1 | Original experiment

Informatiebrochure over het onderzoek

Naam onderzoek: Gesprekken over welzijn met een spraak bot

Verantwoordelijke onderzoeker: Mariët Theune

Uitvoerder van het onderzoek: D.J. Kwakkel, Interaction Technology, Universiteit Twente

Inleiding

Wij vragen u om mee te doen aan een wetenschappelijk onderzoek voor het afstudeerproject van Daniëlle Kwakkel. Meedoen is vrijwillig, maar uw schriftelijke toestemming is nodig. Voordat u beslist of u wilt meedoen aan dit onderzoek, krijgt u uitleg over wat het onderzoek inhoudt. Lees deze informatie rustig door en neem contact op met de onderzoeker als u vragen heeft.

Beschrijving en doel van het onderzoek

In dit onderzoek willen we een computersysteem ontwikkelen dat met mensen in het Nederlands kan praten en hen kan helpen hun welzijn en geluk te verhogen. Hiervoor wordt technologie ontwikkeld die belangrijke informatie over geluk en welzijn uit Nederlandse teksten en audio-opnames kan halen zodat de computer leert waar mensen gelukkig van worden en wanneer zij zich goed voelen. Hiervoor vragen we deelnemers om gesprekken te voeren via de computer. We hebben voorbeelden van vragen en antwoorden nodig die over welzijn en geluk gaan om ons computersysteem te kunnen verbeteren en om te onderzoeken welke vragen wel of niet nuttig zijn.

Wat wordt er van u verwacht?

In dit onderzoek gaat u een gesprek voeren via de computer. Dit gesprek voert u in de rol van één van de persona's die u aan het begin van het gesprek krijgt. De computer stelt u vragen over uw dagelijks leven en over wat u gelukkig maakt. Dit gesprek duurt tussen de 15 en 30 minuten Hier is een voorbeeld van het type vragen in het gesprek: 'Hoe geïnteresseerd bent u in uw dagelijkse activiteiten?', 'Hoe optimistisch bent u over uw toekomst?' of 'Hoe zinvol vind u de dingen waarmee u uw tijd vult?'

Wij maken geluidsopnamen van dit gesprek en verwerken uw antwoorden met een computersysteem tot een tekstbestand. Deze tekst zal gebruikt worden voor het bouwen van een classificatiesysteem. We vragen geen persoonlijke gegevens zoals uw geslacht of leeftijd. De enige persoonlijke identificeerbare informatie die wordt gevraagd is uw naam. Deze zal echter niet in het tekstbestand terecht komen. Afgezien van uw stem zijn de opnames dus anoniem.

Risico's en ongemakken

De rol waaruit u dit gesprek voert zit niet lekker in zijn vel en heeft een laag welzijn. Voor sommige mensen kan het lastig zijn of dichtbij komen om zich in te leven in zo'n rol. Als dit voor u geldt dan vragen we u niet deel te nemen aan dit onderzoek. Wilt u wel meedoen maar komt u vragen tegen waar u liever geen antwoord op wilt geven, dan kunt u deze vragen gewoon overslaan.

Wat doen wij met uw gegevens?

Wij bewaren en gebruiken de geluidsopnamen voor onderzoeksdoeleinden voor de duur van het onderzoek, daarna worden ze vernietigd. De geluidsopnamen worden alleen gebruikt voor het automatisch genereren en verifiëren van de transcriptie. Voor de rest van het onderzoek zullen de geanonimiseerde transcripties worden gebruikt. Wij analyseren met behulp van het computersysteem uw antwoorden en proberen daar informatie uit te halen om de computer te leren automatisch iemands welzijn in te schatten.

We maken u attent op de volgende twee aspecten van ons onderzoek: geluidsopnamen zijn nooit volledig anoniem omdat wij uw stemgeluid opnemen. Om deze redenen vragen we uw toestemming om de geluidsopnamen te mogen maken, en deze opnamen en de daaruit voortvloeiende onderzoeksgegevens te mogen bewaren en te gebruiken voor onderzoeksdoeleinden. Uw geluidsopnamen zullen niet bij andere partijen terecht komen.

Vrijwilligheid

U kunt op elk moment tijdens het onderzoek uw deelname stopzetten en uw toestemming intrekken. U kunt het gesprek stoppen door het internet venster te sluiten. U hoeft niet aan te geven waarom u stopt. De data van het gesprek wordt per vraag opgeslagen. Na het sturen van audio kan uw data niet meer worden verwijderd omdat er geen persoonlijke informatie van u wordt bewaard.

Wat gebeurt er met mijn gegevens?

De geluidsopnamen en onderzoeksgegevens die we in dit onderzoek verzamelen, zullen door mij gebruikt worden voor het maken van mijn afstudeerverslag en voor het ontwikkelen van het sprekende computersysteem.

Uw geluidsopnames worden alleen gebruikt voor het maken van de transcripties voor mijn afstudeerproject en zullen daarna verwijderd worden.

We bewaren alle onderzoeksgegevens op beveiligde wijze volgens de richtlijnen van Universiteit Twente en de wet.

Heeft u vragen over het onderzoek?

Heeft u vragen, aanmerkingen of klachten over dit onderzoek? Of wilt u meer informatie hebben? Dan kunt u contact op nemen met de uitvoerder van dit onderzoek: Daniëlle Kwakkel, bereikbaar via [redacted]. Ook kunt u contact op nemen met de verantwoordelijke onderzoeker: Mariët Theune, bereikbaar op [redacted]. Adres: Universiteit Twente, Drienerlolaan 5, 7522 NB, Enschede. Gebouw: [redacted], kamer: [redacted].

Ethische toetsing en klachten

Als u niet tevreden bent met hoe dit onderzoek wordt uitgevoerd, of als u zorgen, klachten of algemene vragen heeft over het onderzoek of uw rechten als deelnemer, neem dan contact op met de Ethische Commissie, faculteit EEMCS, Universiteit Twente, tel: 053 -489 6719, email ethicscommittee-cis@utwente.nl.

Toestemmingsverklaring

Als u aan dit onderzoek mee wilt doen, vragen we u een toestemmingsverklaring te ondertekenen. Door uw schriftelijke toestemming geeft u aan dat u de informatie heeft begrepen en instemt met deelname aan het onderzoek.

Toestemmingsverklaring

Verklaring deelnemer

Bent u ouder dan 18 jaar?

☐ ja

☐ nee

Ik heb uitleg gekregen over het doel van het onderzoek. Ik heb vragen mogen stellen over het onderzoek. Ik neem vrijwillig deel aan het onderzoek. Ik begrijp dat ik op elk moment tijdens het onderzoek mag stoppen als ik dat wil. Ik begrijp hoe de gegevens van het onderzoek bewaard zullen worden en waarvoor ze gebruikt zullen worden. Ik stem in met deelname aan het onderzoek zoals beschreven in het informatiedocument.

☐ ja

☐ nee

Ik geef toestemming om audio-opnamen van mij te maken voor dit onderzoek en deze opnames op te slaan volgens de geldende regels van Universiteit Twente.

☐ ja

☐ nee

Ik geef toestemming om de gemaakte opnamen te gebruiken voor wetenschappelijk onderzoek zoals beschreven in de bijgevoegde informatie brochure.

☐ ja

☐ nee

E.2 | Experiment with personas

Informatiebrochure over het onderzoek

Naam onderzoek: Gesprekken over welzijn met een spraak bot

Verantwoordelijke onderzoeker: Mariët Theune

Uitvoerder van het onderzoek: D.J. Kwakkel, Interaction Technology, Universiteit Twente

Inleiding

Wij vragen u om mee te doen aan een wetenschappelijk onderzoek voor het afstudeerproject van Daniëlle Kwakkel. Meedoen is vrijwillig, maar uw schriftelijke toestemming is nodig. Voordat u beslist of u wilt meedoen aan dit onderzoek, krijgt u uitleg over wat het onderzoek inhoudt. Lees deze informatie rustig door en neem contact op met de onderzoeker als u vragen heeft.

Beschrijving en doel van het onderzoek

In dit onderzoek willen we een computersysteem ontwikkelen dat met mensen in het Nederlands kan praten en hen kan helpen hun welzijn en geluk te verhogen. Hiervoor wordt technologie ontwikkeld die belangrijke informatie over geluk en welzijn uit Nederlandse teksten en audio-opnames kan halen zodat de computer leert waar mensen gelukkig van worden en wanneer zij zich goed voelen. Hiervoor vragen we deelnemers om gesprekken te voeren via de computer. We hebben voorbeelden van vragen en antwoorden nodig die over welzijn en geluk gaan om ons computersysteem te kunnen verbeteren en om te onderzoeken welke vragen wel of niet nuttig zijn.

Wat wordt er van u verwacht?

In dit onderzoek gaat u een gesprek voeren via de computer. Dit gesprek voert u in de rol van één van de persona's die u aan het begin van het gesprek krijgt. De computer stelt u vragen over uw dagelijks leven en over wat u gelukkig maakt. Dit gesprek duurt tussen de 15 en 30 minuten Hier is een voorbeeld van het type vragen in het gesprek: 'Hoe geinteresseerd bent u in uw dagelijkse activiteiten?', 'Hoe optimistisch bent u over uw toekomst?' of 'Hoe zinvol vind u de dingen waarmee u uw tijd vult?' U beantwoord deze vragen zoals de persoon waarin jij je inleeft ze zou beantwoorden.

Wij maken geluidsopnamen van dit gesprek en verwerken uw antwoorden met een computersysteem tot een tekstbestand. Deze tekst zal gebruikt worden voor het bouwen van een classificatiesysteem. Er worden geen persoonlijke gegevens zoals je geslacht of leeftijd gebruikt in het classificatiesysteem. De enige persoonlijke identificeerbare informatie die wordt gevraagd zijn je naam en SONA identiteitsnummer. Deze zullen echter niet in het tekstbestand terecht komen, en alleen gebruikt worden voor het toekennen van de studiepunten in het SONA systeem.

Risico's en ongemakken

De rol waaruit u dit gesprek voert zit niet lekker in zijn vel en heeft een laag welzijn. Voor sommige mensen kan het lastig zijn of dichtbij komen om zich in te leven in zo'n rol. Als dit voor u geldt dan vragen we u niet deel te nemen aan dit onderzoek. Wilt u wel meedoen maar komt u vragen tegen waar u liever geen antwoord op wilt geven, dan kunt u deze vragen gewoon overslaan.

Wat doen wij met uw gegevens?

Wij bewaren en gebruiken de geluidsopnamen voor onderzoeksdoeleinden voor de duur van het onderzoek, daarna worden ze vernietigd. De geluidsopnamen worden alleen gebruikt voor het automatisch genereren en verifiëren van de transcriptie. Voor de rest van het onderzoek zullen de geanonimiseerde transcripties worden gebruikt. Wij analyseren met behulp van het computersysteem uw antwoorden en proberen daar informatie uit te halen om de computer te leren automatisch iemands welzijn in te schatten.

We maken u attent op de volgende twee aspecten van ons onderzoek: geluidsopnamen zijn nooit volledig anoniem omdat wij uw stemgeluid opnemen. Om deze redenen vragen we uw toestemming om de geluidsopnamen te mogen maken, en deze opnamen en de daaruit voortvloeiende onderzoeksgegevens te mogen bewaren en te gebruiken voor onderzoeksdoeleinden. Uw geluidsopnamen zullen niet bij andere partijen terecht komen.

Vrijwilligheid

U kunt op elk moment tijdens het onderzoek uw deelname stopzetten en uw toestemming intrekken. U kunt het gesprek stoppen door het internet venster te sluiten. U hoeft niet aan te geven waarom u stopt. De data van het gesprek wordt per vraag opgeslagen. Na het sturen van audio kan uw data niet meer worden verwijderd omdat er geen persoonlijke informatie van u wordt bewaard.

Wat gebeurt er met mijn gegevens?

De geluidsopnamen en onderzoeksgegevens die we in dit onderzoek verzamelen, zullen door mij gebruikt worden voor het maken van mijn afstudeerverslag en voor het ontwikkelen van het sprekende computersysteem.

Uw geluidsopnames worden alleen gebruikt voor het maken van de transcripties voor mijn afstudeerproject en zullen daarna verwijderd worden.

We bewaren alle onderzoeksgegevens op beveiligde wijze volgens de richtlijnen van Universiteit Twente en de wet.

Heeft u vragen over het onderzoek?

Heeft u vragen, aanmerkingen of klachten over dit onderzoek? Of wilt u meer informatie hebben? Dan kunt u contact op nemen met de uitvoerder van dit onderzoek: Daniëlle Kwakkel, bereikbaar via [redacted]. Ook kunt u contact op nemen met de verantwoordelijke onderzoeker: Mariët Theune, bereikbaar [redacted]. Adres: Universiteit Twente, Drienerlolaan 5, 7522 NB, Enschede. Gebouw: [redacted], kamer: [redacted].

Ethische toetsing en klachten

Als u niet tevreden bent met hoe dit onderzoek wordt uitgevoerd, of als u zorgen, klachten of algemene vragen heeft over het onderzoek of uw rechten als deelnemer, neem dan contact op met de Ethische Commissie, faculteit EEMCS, Universiteit Twente, tel: 053 -489 6719, email ethicscommittee-cis@utwente.nl.

Toestemmingsverklaring

Als u aan dit onderzoek mee wilt doen, vragen we u een toestemmingsverklaring te ondertekenen. Door uw schriftelijke toestemming geeft u aan dat u de informatie heeft begrepen en instemt met deelname aan het onderzoek.

Toestemmingsverklaring

Verklaring deelnemer

Bent u ouder dan 18 jaar?

☐ ja

☐ nee

Ik heb uitleg gekregen over het doel van het onderzoek. Ik heb vragen mogen stellen over het onderzoek. Ik neem vrijwillig deel aan het onderzoek. Ik begrijp dat ik op elk moment tijdens het onderzoek mag stoppen als ik dat wil. Ik begrijp hoe de gegevens van het onderzoek bewaard zullen worden en waarvoor ze gebruikt zullen worden. Ik stem in met deelname aan het onderzoek zoals beschreven in het informatiedocument.

☐ ja

☐ nee

Ik geef toestemming om audio-opnamen van mij te maken voor dit onderzoek en deze opnames op te slaan volgens de geldende regels van Universiteit Twente.

☐ ja

☐ nee

Ik geef toestemming om de gemaakte opnamen te gebruiken voor wetenschappelijk onderzoek zoals beschreven in de bijgevoegde informatie brochure.

☐ ja

☐ nee