Informatics

Nutzer:innen-Zufriedenheit durch die Erkennung von Emotionen in deutschsprachigen Mensch-Roboter-Konversationen mit Q.bo One erhöhen

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Media and Human-Centered Computing

eingereicht von

Rafael Vrecar, BSc
Matrikelnummer 01627765

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Peter Purgathofer
Mitwirkung: Dr.phil. Mag.phil. Astrid Weiss

Wien, 19. April 2022

Rafael Vrecar

Peter Purgathofer

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Increasing User Satisfaction by Detecting Emotions in German-language Human-Robot Conversations with Q.bo One

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Media and Human-Centered Computing

by

Rafael Vrecar, BSc
Registration Number 01627765

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Peter Purgathofer
Assistance: Dr.phil. Mag.phil. Astrid Weiss

Vienna, 19th April, 2022

Rafael Vrecar   Peter Purgathofer

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Rafael Vrecar, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 19. April 2022

Rafael Vrecar

Danksagung

Mein herzliches Dankeschön gilt ...

- ... Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Peter Purgathofer für die Betreuung dieser Arbeit und die herzliche Begleitung meines Studienweges in vielen sehr spannenden Lehrveranstaltungen und meiner Tätigkeit als Tutor.
- ... Dr.phil. Mag.phil. Astrid Weiss für die ausgezeichnete Beratung, viele Diskussionen (und willkommene Herausforderungen), das Wecken meines Interesses an dieser Thematik und den generellen Umgang mit Studierenden, der das Lernen zu einer Freude bereitenden, aber auch fordernden (und spaßigen!) Tätigkeit gemacht hat.
- ... Salzburg Research GmbH [mbH], namentlich Mathias Schmoigl-Tonis, MSc für das Bereitstellen der Ressourcen und die konkrete Vorstellung von potenziellen Forschungszielen sowie Erklärungen bezüglich verschiedener Technologien.
- ... Prof. Dr. Kerstin Fischer und Prof. Dr. Barbara Kühnlenz für die Teilnahme an meinem Interview und die wertvollen Einsichten, die sie mir boten.
- ... allen Teilnehmer:innen meiner empirischen Studie.
- ... Aaron Wedral, BSc für die überaus hilfreichen Erklärungen und Tipps zum Thema “Statistik”.
- ... Dipl.-Ing. Janis Lena Meißner für ihre wertvollen Tipps bzgl. Studiendesign.
- ... Anna Dobrosovestnova, MA MSc für ihre Tipps die Analyse der Interviews betreffend.
- ... Glenda Hannibal, MA für ihre Tipps bezüglich Expert:innen-Interviews.
- ... Associate Prof. Dipl.-Ing. Dr.techn. Hilda Tellioğlu für die Präsentationstipps.
- ... meiner Schwester (Videobearbeitung), Dipl.-Ing. Dr.techn. Michael Urbanek, BSc BA MSc (Tipps beim Programmieren) und Michael Sigmund (Infrastruktur) für ihre Hilfe im Zusammenhang mit der Umfrage.

- ... meinem Papa, meiner Mama und meiner Schwester sowie Linda Schachenhofer, Dipl.-Ing. Christian Schöbel und Alexander Firbas, BSc für das sorgfältige Korrekturlesen dieser Arbeit bzw. der Umfrage.
- ... meinen Großeltern für die konstante Unterstützung ihrer Enkelkinder und das große Interesse an allem, was sie tun.
- ... meiner Familie, meinen Freund:innen und allen, die mich während meines ganzen Studienwegs begleitet haben (vor allem die HCI-Gruppe der TU Wien!) und immer einen Ausgleich zum Lernstress boten.
- ... der Technischen Universität Wien für die Gelegenheit mich mit einem spannenden und fordernden Studium zu bilden im engeren Sinne, sowie allen österreichischen Steuerzahler:innen, durch die das österreichische Bildungssystem finanziert wird, im weiteren Sinne.
- ... Dr.techn. Thomas Auzinger, MSc et al. deren Vorlage [Auz+] mir bereits ein zweites Mal erlaubte, mich beim Schreiben auch wirklich auf meine Arbeit und nicht das Layout zu konzentrieren.

Acknowledgements

I want to thank ...

- ... Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Peter Purgathofer for advising this thesis and his guidance on my study path in many exciting courses and while working as tutor.
- ... Dr.phil. Mag.phil. Astrid Weiss for her excellent advice, many discourses (and very welcome challenges), arousing my interest regarding this topic and her general way of interacting with students which makes learning not only a challenging but also a very enjoyable (and funny!) experience.
- ... Salzburg Research GmbH [mbH], namely Mathias Schmoigl-Tonis, MSc for providing resources and potential research goals as well as tutorials regarding different technologies.
- ... Prof. Dr. Kerstin Fischer and Prof. Dr. Barbara Kühnlenz for the participation in my interview and the valuable insights they provided.
- ... all participants involved in my empirical study.
- ... Aaron Wedral, BSc for his very valuable explanations and advice regarding the topic “statistics”.
- ... Dipl.-Ing. Janis Lena Meißner for her valuable advice regarding study design.
- ... Anna Dobrosovestnova, MA MSc for her advice regarding interview analysis.
- ... Glenda Hannibal, MA for her advice regarding expert interviews.
- ... Associate Prof. Dipl.-Ing. Dr.techn. Hilda Tellioğlu for the advice regarding presenting a thesis.
- ... my sister (video editing), Dipl.-Ing. Dr.techn. Michael Urbanek, BSc BA MSc (coding advice) and Michael Sigmund (infrastructure) for their help regarding the survey.

- ... my dad, my mom, and my sister, as well as Linda Schachenhofer, Dipl.-Ing. Christian Schöbel and Alexander Firbas, BSc for proofreading this thesis and survey prototypes.
- ... my grandparents for always being supportive and interested in everything their grandchildren are doing.
- ... my family, my friends and all, who have accompanied me (especially the HCI group of TU Wien!) during my studies and always were a welcome distraction from my learning stress.
- ... Technische Universität Wien for the opportunity to educate myself with interesting curricula and the Austrian taxpayers for making the Austrian education system possible.
- ... Dr.techn. Thomas Auzinger, MSc et al. whose template [Auz+] allowed me a second time to focus on content during writing and not on layout.

Kurzfassung

Diese Arbeit befasst sich mit der Frage, inwiefern sich das Vorhandensein von Emotionserkennung in einer Interaktion mit dem Roboter Q.bo One auf die Nutzer:innen-Zufriedenheit auswirkt. Dafür wird zunächst eine allgemeine Einführung in das Thema vorgenommen und ein Beispielszenario für eine Interaktion, in welcher eine ältere Person dem Roboter ihren Tag nacherzählt, definiert, welches stringent in der gesamten Arbeit und insbesondere der damit verbundenen Studie verwendet wird.

Für die Bearbeitung des Themas wurden verschiedene Methoden gewählt. Begonnen wird mit einem Überblick über vorangegangene, verwandte Arbeiten. Dabei werden zunächst verschiedene Definitionen des Begriffs "Emotion" sowie verschiedene Emotionsmodelle erläutert. Anschließend wird die Rolle von Emotion in Mensch-Roboter-Interaktion beleuchtet, bevor verschiedene Methoden und Technologien für Emotionserkennung diskutiert werden. Darauf folgend wird analysiert, welche Leistungsunterschiede es zwischen einer auf die deutsche Sprache optimierten Lösung zur Emotionserkennung bzw. einer englischsprachigen Lösung mit einer Übersetzungsebene gibt. Ferner wurden zwei Expertinneninterviews durchgeführt, welche mit einer thematischen Analyse bearbeitet wurden. Die Ergebnisse sind ein allgemeiner Überblick über das Forschungsgebiet der Emotionserkennung, welcher für diese Arbeit hilfreich war. Das Interview beeinflusste auch die Methodik des empirischen Teils hinsichtlich der Auswahl der Fragebogen-Serie und brachte die Adaption einer Forschungsfrage, wie in der Arbeit erläutert, mit sich.

Den größten Teil dieser Arbeit bildet die Studie mit Q.bo One, in welcher auf Basis vier verschiedener Szenarien (mit impliziter Emotionserkennung, explizites Nachfragen, Falscherkennung und keine Emotionserkennung) die Nutzer:innen-Zufriedenheit evaluiert wird. Dafür wird eine Videostudie verwendet, in welcher den Teilnehmenden die Möglichkeit geboten wird, sich in das oben erläuterte Beispielszenario hineinzuversetzen, ehe sie den Roboter hinsichtlich der Aspekte Anthropomorphismus, Belebtheit, Sympathie, wahrgenommener Intelligenz und Sicherheit (Godspeed-Fragebogenserie) bewerten sollen.

Die Ergebnisse der Studie suggerieren, dass eingebaute Emotionserkennung der expliziten Nachfrage bezüglich Emotionen nicht zwangsläufig überlegen ist. Ebenso legen sie die Notwendigkeit weiterer Arbeit in diesem Feld nahe, da breite Anwendungskontexte unter Umständen verschiedene Formen von Emotionsbehandlung benötigen – insbesondere auch hinsichtlich etwaiger Langzeiteffekte, da die Studie maximal eine Momentaufnahme eines längeren und breiteren Anwendungskontexts repräsentieren kann.

Abstract

This thesis deals with the question to what extent the existence of emotion detection within an interaction with the robot Q.bo One influences user satisfaction.

First of all, a general introduction to the topic is provided as well as an example scenario is defined which is used consistently within the whole thesis. In said scenario, an older person tells Q.bo One about their day. Based on said scenario the below explained video survey got conducted.

For working on this topic, different methods were chosen. The thesis starts with an overview of related work which begins with the discussion of the term "emotion" and different emotion models. Afterwards, the role of emotions in human-robot interactions as well as different methods and technologies for emotion detection are elaborated on.

After this section, the performance differences between a native German-language solution compared to an English-language solution with a translation layer are looked into. Moreover, expert interviews were conducted and have been analyzed using thematic analysis. The results are a general overview of the topic of "emotion handling" which was helpful in the process of writing this thesis. Moreover, the interview had an influence on the methodological work within the empirical part too, in form of the chosen questionnaire. Additionally, it resulted in the adaptation of a research question as elaborated within the thesis.

The biggest part of this thesis is a study with Q.bo One, in which, based on four different scenarios (implicit emotion detection, explicitly asking the person, malfunctioning detection, no emotion detection) user satisfaction is evaluated. For that, a video study is used, in which the participants are provided with the opportunity, to look into the above described example scenario before rating the robot regarding the aspects anthropomorphism, animacy, likability, perceived intelligence and safety (Godspeed Questionnaire Series).

The results of the study suggest that in-built emotion detection is not necessarily superior to explicitly asking a person about their emotional state. Moreover, they back up the necessity of further work within this field as broad application contexts might require different forms of emotion handling, especially regarding the long-term effects too, as the study can at most represent a moment within a longer, broader use context.

Contents

CHAPTER 1

Introduction

Interacting with other people is a basic human need [Dij+12; Kug21]. However, there are cases where human-human interaction is not possible on a regular basis. One can imagine an older person whose partner died and who is therefore living alone or, as COVID-19 has shown us, there are situations we can find ourselves in where keeping distance is necessary. To protect people in those situations from feeling lonely, social companion robots are an option increasing in popularity [PRA18].

In order to make the people who use these technologies feel comfortable, an approach one might think about is to approximate human-human interaction as this could increase the chance for user acceptance. Whether or how it does that and the topic of "social robots" and "communicative robots" in general are part of current research, e.g., the work of Hepp [Hep20] who provides an overview of said field, as well as the work of Ghazali et al. [Gha+20] who looked into user acceptance and furthermore, the work of Broadbent et al. [Bro+13] who discussed how the *human-likeness* of a robot's face is perceived.

Social companion robots can have a variety of different applications, e.g., in education as peer learners or tutors [Bel+18], in stroke rehabilitation [LWH16] or for coping with loneliness [Ode+20] as we will further look into within this thesis. To be precise, it focuses on the specific use case of a robot whose purpose is to ask its user "How was your day?" every evening. With that, we mean to recreate the experience of a partner or a friend asking this question and then giving the user the chance to recapitulate their day by telling the companion robot about it, similar to what writing a diary makes a person do. Diaries, for example in form of "expressive writing" as [GJ17] and [Bec+21] discussed, can have positive effects on people as they can improve their state of mind. This might also be an intended goal when designing companion robots.

However, in order to create an interaction between the human and the robot, rather than just letting the human talk to the agent, we have to make the agent come up with a response to the narration of the person's day. In this context, it is the goal of this thesis to find out, how the different ways of handling (or not handling) emotion influence the user satisfaction.

1.1 Envisioned Use Case

In order to situate the research performed in this thesis through a use case, we look into the example in Figure 1.1.

Imagine an older adult (A) who just came home from their groceries providing the robot (R) with a summary of their day.

- (R) "How was your day?"
- (A) "I got up later than usually, 8 a.m. instead of 7 a.m. For breakfast I had ham and eggs and afterwards, I read the newspaper. Then lunch got delivered. Today's menu was a tomato soup at first and an omelette afterwards. Then I went for a walk and bumped into Joseph. We talked about his children. Gregor is marring Claudia next week. Also, I noticed that the leaves are already falling down again. After the news, it is time for bed again."
- (R) "..."

Figure 1.1: Example – "How was your day?"

We leave the robot's response empty for now, but later on we will come back to this example and discuss different aspects and possible responses. However, at first there are some terms which have to be discussed.

1.2 Interaction, Conversation & Communication

Within the thesis at hand, the terms "interaction", "conversation", and "communication" are central. With "interaction" we mean – building on the definitions of Graumann (1972), Piontkowski (1982) or Boos (1997) who focus on the *reciprocity* of such a process as explained in [Chr99] – the actions humans and/or robots *reciprocally* set between each other; be it touching, talking, watching or similar.

Although, as discussed by, e.g., Warren in [War06], there does not seem to exist a generally valid definition of the term "conversation", we, for the context of this thesis, build on the definition proposed by Svennevig [Sve00] for whom a "conversation" is a *"[...] joint activity consisting of [...] actions predominantly [...] spoken utterances [...]"* [Sve00, p. 8]. Therefore, in this thesis, we mean by "conversation" the intentional verbal dialogue with all of its aspects which happens between humans and/or robots.

"Communication" is, as discussed in [Chr99], often used interchangeably with "interaction". Although, there have been ideas and proposals for differentiating between the two terms

[Chr99], it is not necessary for this thesis as "interaction", by the definition we motivated above, covers all explanatory needs in this context as we only operate in a setting where the messages sent by the robot are received by the user as otherwise there would be no interaction (although there very well would be a form of communication, based on the differentiation arguments in [Chr99]).

Moreover, there exist as per [Chr99] other forms of interaction/communication besides conversations, e.g., non-verbal communication. However, when getting back to said term "conversation", by the given definition, the in Figure 1.1 listed interaction already is a *conversation* as the robot gets a response from the human. Nevertheless, the human does not get any sort of feedback by now. Therefore, the conversation has an abrupt ending instead of an intentional one. With that, we already implicitly split up the conversation into different parts with the ending being one of them.

Further elaborating on that, a "conversation" can be split up into several *components*: Conversations have a *context* (e.g., time and space) and *content* (e.g., the events of a person's day), often also a *purpose* and goal (e.g., recapitulating one's day in order to improve a person's mental state). Moreover, conversations can happen in different languages and verbal as well as non-verbal, which affects their components and structure, but also their reception. The attributes of one's voice (e.g., pitch, speed, volume) can also influence how messages are received (e.g., when screaming out something to somebody, it is imaginable that it causes some sort of anxiety). E.g., Tusing and Dillard explored how vocal cues can influence the perception of dominance [TD00].

1.3 Emotions – Why & How to Detect Them?

Emotions are a further relevant component of conversations, as per Hendrix and Morrison *"[...] When emotions are expressed appropriately, senders are able to formulate a message that reflects their internal status and intentions while considering audience needs and perceptions. [...]"* [HM20, p. 2]. Therefore, for this thesis, we go by the definition of Adolphs in [Fox+18] who defines emotions as *"functional states"* [Fox+18, p. 6] [of humans, in our context] *"that cause feelings and behavior"* [Fox+18, p. 6]. Adolphs furthermore explicitly distinguishes "emotions" from "feelings" and "behavior". Therefore, only because a person *feels* a specific way (e.g., a person *feels* sad), does not mean the underlying emotion aligns with said feeling. Moreover, the observer, by this definition, cannot know the underlying emotion based on an observed behavior.

Taking this knowledge into account, when designing robots whose main focus is the interaction between human and robot itself, (correctly) detecting emotions a human explicitly or implicitly brings into the interaction would allow the robot to adapt its own *behavior* with regard to the *emotional state* of the person using it. In an imaginary optimal case, the robot could help the person using it to better understand their own, actual *emotions* (maybe even in contrast to the *feelings* a person experiences in a specific situation) and provide the human with a feeling of being understood. Note that this idea is only optimal towards what is technically possible but does not necessarily correspond to what people actually *want*. Therefore, a good implementation a priori does not necessarily imply higher user satisfaction.

Elaborating a bit further on the technical perspective, the explicit statement “I am happy” seems to be correctly parsed a lot more easily than the implicit emotion in the statement – to get back to the example in Figure 1.1 – “Gregor is marring Claudia next week”. The emotions a person encodes into this statement highly depend on the context. If said person considers Claudia to be the right partner for Gregor and they like the couple, happiness is a reasonable assumption. However, if said person does not trust Claudia – but wants all the best for Gregor – they might be worried about Gregor’s decision.

Assuming the person is happy about the planned wedding, a fitting response by the robot might be “Wow, I am so happy for them.” However, if the person is worried the mentioned answer would most likely not make them feel comfortable. Instead, an answer like “Do not worry, it will turn out fine” might be the better choice and convey some comfort. This example suggests that emotion detection with implicitly encoded emotions inherently is way more difficult than detecting explicitly stated ones.

Moreover, as humans are complex organisms with a variety of different nuanced emotions, it seems reasonable as a starting point, and therefore for the scope of a Master’s thesis, to only approximate the emotional state of the human within the human-robot interaction. For that, we decided to focus on the six “basic emotions” proposed by Ekman in, e.g., [Ekm92a]: anger, disgust, fear, happiness, sadness, and surprise. The topic of “emotions” is going to be further elaborated on in Chapter 2.

Regarding the detection of emotions, different approaches have been taken in the past. For example, sensors for detecting aspects of the human’s face and posture have been used in the work of Graterol et al. [Gra+21], and Albu [Alb15]. In contrast to that Aschenbrenner [Asc19], Agrawal and An [AA12] and Binali et al. [BWP10] focused on emotion detection from text which is also an essential focus point of this thesis.

1.4 Languages & Other Problems

An obvious prerequisite for detecting emotions is a conversation from which they can be detected, and in order to make a conversation even possible, the language of the participants has to be compatible. However, languages do not only differ in the words which are assigned to a specific concept but attributes like pitch or volume might differ in meaning too. For example, the question “How old are you?” has a constantly “flat” pitch in Danish whereas in German, the pitch usually goes up within the process of asking [see Appendix; EPI1, p. 4].

Considering said differences, it is only reasonable to assume that these have an impact on the way emotions can be detected in said languages. Even humans might have difficulties to make meaning of a scream coming to their ears in a different language. While it might very well imply joy in a for us foreign language, we could misunderstand it and assume it means that the screaming person is in danger. When even humans have difficulties in recognizing which emotion a sound might imply in a foreign language, it also seems likely that there is no general purpose solution available for automatically detecting emotions from spoken dialogues which does not heavily depend on the language it is applied on. Therefore, a focus has been set for this thesis in advance which lays on German. However, regarding the research landscape, as English is, compared to German, the demographically far bigger language [Dep19], it seems reasonable to assume that more research regarding emotion detection has already been done on English-language human-robot interactions compared to German-language ones. A quick search with *Google Scholar* for the term "automated emotion detection" delivered (Mar. 11, 2021) about 70.400 results, whereas when adding the word "german", the number of results was reduced to 27.100.

Therefore, approaches developed in and for the English language are analyzed regarding their transferability to German-language contexts in Chapter 3.

1.5 Motivation & Relevance

The variety of different application domains for social robots make this research relevant. However, the specific motivation behind this thesis is how said robots can be used in care for older adults[1].

As we look into this topic from a German-language point of view, the relevance especially in German-speaking countries is considered. E.g., in Germany, the population is aging as predicted in [Deu19]. This process results in the need for carers and supportive technology seeming to gain significance as having too few carers could be potentially alleviated – among others – by better technology.

From the perspective of just getting tasks which are getting more and more difficult with declining physical ability – like serving food or helping people to shower and dress – done, dialogue-based companion robots are not inherently better compared to others because the command for starting a procedure can very well be given by just pressing a button too instead of describing it verbally. However, when the idea of *social* interaction comes into play, just pressing a button is a completely different experience than talking to another person and explaining to them what to actually do and what one is comfortable with which is why we look into the applications of *social* robots within said field.

That said, as care is a field where highly personal interactions take place, trust and user satisfaction are important aspects of it. Whether social robots are trusted more by people compared to non-social ones or not has been explored in, e.g., [Ros+20].

For all of the stated reasons, this thesis aims to further investigate the role of emotion-detection in human-robot conversations.

[1] At this point, it also has to be made clear that the underlying normative question, if assistive robots can be a good substitution for personal care, will explicitly not be discussed in this thesis as it covers the topic of emotion detection and the mentioned question would highly exceed the scope of it.

1.6 Aim, Context & Scope of this Thesis

This thesis can be divided into *three* different focus points. As a basis, first of all, emotions in general are discussed with a special interest in different emotion models. Building on that, emotions specifically in the context of Human-Robot Interaction (HRI) are discussed. For that, we look into the concepts of companion robots, human-robot dialogue, and artificial empathy. Tracking back on these general insights, technological and automated emotion detection are discussed.

In this regard, it has to be stated that the focus of this thesis lies specifically on text-based emotion detection and is therefore the most discussed technology. Because of that, as first focus point, this thesis aims to explore how the currently available English-language technologies regarding emotion detection can be transferred to German-language use-cases and should give an overview regarding the work done in this field yet, considering performance implications.

Furthermore, the second focus point is to explore to what degree – if at all – emotion detection can increase user satisfaction. For that, a video-based online survey is conducted with Q.bo One (`https://thecorpora.com/`) with a small mock implementation created in advance. It has to be stated that the implementation is not meant to be used outside of the scope of this study. It instead shall show whether it is feasible to implicitly detect emotions in the interaction drafted in Figure 1.1, if explicitly asking leads to similar effects regarding user satisfaction or if emotions do not help at all in this context.

In this context, as the third focus point, the thesis wants to specifically explore how a wrongly detected emotion can negatively influence said user satisfaction, e.g., when the robot detects the user as "happy" whereas they are actually sad.

1.7 Research Questions

Based on the preceding explanations, the following research questions and hypotheses are derived (note that *RQ 1* is listed without a hypothesis as it is intentionally exploratory). More explanations about that are provided in Section 6.6.

- **RQ 1:** What are the performance differences (i.e., successful detection ratio) between a language-specific solution and using an already existing English-language solution with a translation layer added to it?
- **RQ 2a:** Can – and if so to what extent – user satisfaction be increased by in-built emotion detection?
  HYP 2a: *User satisfaction can be increased.*
- **RQ 2b:** How does the extent of user satisfaction change comparing "in-built emotion detection", "explicitly asking the users about their emotions" and "no emotion handling" in human-robot interactions?

HYP 2b: *User satisfaction can be increased in similar extents when having "in-built emotion-detection" or "explicitly asking the users about their emotions" compared to "no emotion handling".*

- **RQ 3:** How do wrongly detected emotions influence user satisfaction?
  HYP 3: *Wrongly detected emotions decrease user satisfaction.*

1.8 Methodology & Thesis Structure

In order to answer said research questions and meet the goals of this thesis, the following methodological approach has been chosen which also represents the thesis structure.

- **Literature Review:** As a first step, literature regarding emotions in general, different emotion models, the concept of companion robots, human-robot dialogue and artificial empathy is reviewed. Afterwards, an overview of the different methods on emotion detection is given.

- **Scoping Study:** For analyzing the performance differences between language-specific/native solutions and English-language solutions with a translation layer, a scoping study is conducted.

- **Expert Interviews:** Furthermore, two expert interviews are conducted in order to get a better overview of the research landscape, open questions around those topics and the general work in this field.

- **User Study with Q.bo One:** Because of the pandemic, a study design specifically optimized for a distance setting (asynchronous, online) has been developed. For that, we use the prior discussed (Figure 1.1) example.

  We compare three different settings: 1) no emotion detection 2) implicit, automated emotion detection (correct and wrong), 3) explicitly asking the people using the robot about their emotions. For that, we use pre-recorded videos of example conversations between a person and Q.bo One where we only bring in a variation regarding the responses from Q.bo One (first question by the robot and answer of the human conversation partner stay the same in all three settings). The participants then answer questions as outlined in Chapter 5.

  Afterwards, videos where Q.bo One provides emotionally unfitting responses are provided and the participants are again asked about their impressions on said conversations.

  Moreover, one decisive, special aspect of the study is that the study itself is also "conducted" by Q.bo One. The explanations are recorded videos and therefore, Q.bo One also "guides" the participants through the survey. After answering the questions on the depicted example conversations, the participants are also asked to reflect on their experience with Q.bo One. The details on the questions and evaluation thereof are outlined in Chapter 5.

Furthermore, the thesis also outlines limitations and abandoned ideas in Chapter 6, where the results of the conducted study are discussed too, prior providing a conclusion in Chapter 7.

CHAPTER 2

Related Work

2.1 What Are Emotions?

The term "emotions" has a variety of definitions. According to Kerkeni et al. [Ker+17] the concept of emotion is even one of the most difficult ones to define in Psychology.

As this thesis focuses on emotion detection in German-language human-robot conversations, it seems appropriate to start with the Duden definition [Dud] of emotion which is "psychische Erregung, Gemütsbewegung; Gefühl, Gefühlsregung" (Engl. "physical excitement, movement of the mood, feeling, stir of feeling").

Consulting the *Stanford Encyclopedia of Philosophy* [SS21], there are two different goals of defining emotions. On the one hand, one can aim to be compatible with linguistic use in our daily lives by providing a *descriptive definition* and on the other hand, a *prescriptive definition* which can be used along researchers in this field might necessarily violate some intuitive meanings of said term.

Whereas it can be argued that the definition in the Duden is a descriptive one as it is a broad one which can be used in daily life, scientists are not necessarily in accordance to said definition neither agreed on one in the first place, as, e.g., [Ker+17] pointed out. Adding to that, Scherer [Sch05] back in 2005 already pointed out a prominent mistake of using emotion and feeling interchangeably although they shall describe different concepts. This further backs up the argument for the Duden definition to be descriptive.

According to [Ker+17] Schacter et al.[1] consider emotions to be experiences which are positive or negative and associated with a particular pattern of physiological activity.

Kerkeni et al. [Ker+17, p. 2] consider emotion to be a *"fast process"* which consists of *"two stages"* and is *"focused on an event"*. They describe an emotion as dynamic

[1] no access to original source

phenomenon which consists of multiple components and is brief. Furthermore, they state that it is always triggered by a specific event.

Joseph LeDoux, who is a prominent researcher in the field of "emotion"[2], also considers emotions to be caused by a trigger, providing this definition in an interview with the *Brain World* Magazine [Emo, p. 1]: *"[...] emotions are conscious experiences that occur when we find our self in a situation where a challenge or opportunity exist."*

According to Adolphs and Anderson, who looked at the topic from a neuro-scientific perspective [AA18, p. 10], *"emotions are [...] internal states that afford flexible mapping to behavior"*.

Moreover, an argument that emotions can be directed towards something, e.g., an object, has also been proposed in the past [Whi11]. An example [Whi11, p. 281] is the emotion "fear": *"Jack does not just feel fear, but he feels fear-of-something".* According to Debus [Deb07], emotions can also be directed towards past situations/events.

Although the author considers the fundamental discussion, what "emotions" actually are to be extremely interesting, it would by far exceed the scope of this thesis, and more importantly, is not too relevant for the topic of this thesis either as abstractions for modeling emotions in a computational context have to be made anyway. Speaking of abstractions for modeling, the next paragraphs focus on different emotion models and argue why Ekman's model was chosen for the work within this thesis.

Regarding defining the term "emotion", however, we go, as already outlined in the introduction, by Adolph's definition [Fox+18] and view emotions as *"functional states"* [Fox+18, p. 6] [in our context of humans] *"that cause feelings and behavior"* [Fox+18, p. 6]. We therefore distinguish clearly between "emotions" and "feelings".

2.2 Emotion Models

Over the past decades, a variety of emotion models emerged. This thesis elaborates on a subset of them to provide a comprehensive overview although this list does not make any claims of being complete.

2.2.1 Ekman's Basic Emotions

Paul Ekman published different papers to argue for the existence of so-called "basic emotions" [Ekm92a; Ekm92b; Ekm99; EC11].

First of all, back in 1972, Ekman et al. [EFE13] argued that the term "emotions" has various aspects, e.g., different types of responses to a stimuli (physiological, motor, verbal, ...). However, they clarify, that they use the term "emotion" independently from the aspects in general (meaning that all of said aspects might apply).

[2] 19.400 results for the search "Joseph LeDoux emotion" on Google Scholar on Jan. 4, 2022

Ekman [Ekm92b] argued that they (Ekman, Friesen, & Ellsworth, 1972[3]) had found evidence for six emotions (happiness, surprise, fear, sadness, anger, and disgust in combination with contempt) used by observers to judge an emotion which is shown in a facial expression.

In 1992, Ekman [Ekm92a] proposed nine characteristics (see below) which are shared by "basic emotions". Said "basic emotions" are the ones from the prior paragraph although "disgust" is listed without "contempt". The term basic, according to [Ekm92a, p. 170] is used because *"(1) There are a number of separate emotions which differ one from another in important ways. (2) common features which these emotions display as well as their current function"*.

Ekman's Nine Characteristics

1. *Distinctive universal signals*
2. *Presence in other primates*
3. *Distinctive physiology*
4. *Distinctive universals in antecedent events*
5. *Coherence among emotional response*
6. *Quick onset*
7. *Brief duration*
8. *Automatic appraisal*
9. *Unbidden occurrence*

Figure 2.1: The 9 Characteristics in 1992 [Ekm92a, p. 175]

Note that the characteristic "Brief duration" also aligns with the prior discussed definition of emotions by Kerkeni et al. [Ker+17].

Contradicting Views

Ekman's work was also analyzed in other publications, e.g., by Sabini and Silver [SS05], who argued that said list does not comply with evolutionary thinking as jealousy and parental love are missing. They mention past arguments by [Bus00] and [Pin97] who backed the importance by highlighting the reproductive interests of humans. Moreover, Sabini and Silver [SS05], consider Ekman's view to be "reductionism". Furthermore, [SS05] they argue that love and jealousy did not make it to Ekman's list because they do not have unique facial expressions. From their perspective, however, they would suggest

[3] no access to original source

that emotions are *"tied to raw feels rather than action"* as *"Removing them from action undermines [...] an important aspect of the self"* [SS05, p. 710].

Turner and Ortony [TO92] questioned the concept of basic emotions in general. They built their argumentation upon the issue that supporters of the concept of basic emotions contradict each other when arguing which emotions are considered to be basic. They mentioned Ekman, Izard and Panksepp namely. Moreover, Turner and Ortony also mentioned their different approaches regarding understanding emotions. They ascribed Ekman's basis to be the face, Izard's rooting in biosocial considerations and Panksepp's to be the brain. However, they also state that the "basic emotions" view point has resulted in useful developments which use the face for tracking back to emotions – similarly with Panksepp's contributions to neurobiology.

Besides explaining the ambiguities and contradictions from above, Turner and Ortony [TO92, p. 570] argued *"[...] that a basic emotions view has difficulty in accounting for the great variety of emotions experienced [...] in our own culture, let alone [...] very different cultures"*. Furthermore, they used a similar argument as Sabini and Silver [SS05] later in 2005 as they argued that emotions which do not have a unique facial expression (their examples: pride, admiration, and envy) lose focus based on Ekman's basic emotions perspective. Nevertheless, according to them, those emotions are often important to understand experiences and behavior.

Turner and Ortony [TO92, p. 570] emphasize that even under the premise that *"[...] if such a set [of basic emotions] could be identified and agreed upon [...]"* it still would make it difficult to talk about *"[...] rich and diverse experience of emotions [...]"* [TO92, p. 570]. Therefore, they do not only criticize the shortcomings within the framework of "basic emotions" but also the conception of such in general.

Ekman Defending "Basic Emotions"

In 1999, Ekman [Ekm99] defended the "basic emotions" position and stated that instead of neglecting variety in affective phenomenas it is an attempt to organize those. Ekman also states that in said position, "non-basic" emotions are not allowed. As a clear distinction Ekman points out that this position shall clearly differentiate "basic emotions" from other affective phenomena. Ekman, moreover, describes the stated characteristics as *"challenges for more research [...] [which] [...] highlight the gaps in our knowledge"* [Ekm99, p. 57].

Furthermore, Ekman [Ekm99, p. 57] provides an outline that the utility of said approach would be evident ten years later by what research it would motivate to provide confirmation, contradictions or new perspective to the described position.

A bit over ten years later, in 2011, Ekman and Cordaro [EC11] again elaborated on the meaning of "basic emotions". The first argument for using the term "basic" from [Ekm92a] still held.

As second characteristic [EC11, p. 364] now listed *"[...] that emotions have evolved through adaptation to our surroundings"*.

Moreover, [EC11, p. 364] also explains that the term "emotion" does not describe a single affective state. It instead describes a *"family of related states"* [EC11, p. 364].

Additionally, the in 1992 proposed list (2.1) got enriched by *"9. Distinctive thoughts, memories, and images. 10. Distinctive subjective experience. 11. Refractory period filters information available to what supports the emotion. 12. Target of emotion unconstrained. 13. The emotion can be enacted in either a constructive or destructive fashion."* [EC11, p. 365].

The point "Coherence among emotional response" (2.1) got removed (and, moreover, the order changed). Moreover, points (2.1) 6. and 7. are elaborated further on.

This results in the following, updated, characteristics list [EC11, p. 365].

Ekman's 13 Characteristics

1. *Distinctive universal signals.*
2. *Distinctive physiology.*
3. *Automatic appraisal.*
4. *Distinctive universals in antecedent events.*
5. *Presence in other primates.*
6. *Capable of quick onset.*
7. *Can be of brief duration.*
8. *Unbidden occurrence.*
9. *Distinctive thoughts, memories, and images.*
10. *Distinctive subjective experience.*
11. *Refractory period filters information available to what supports the emotion.*
12. *Target of emotion unconstrained.*
13. *The emotion can be enacted in either a constructive or destructive fashion.*

Figure 2.2: The 13 Characteristics in 2011 [EC11, p. 365]

The distinction between "basic emotions" and other affective states is by the existence of universal signals or antecedent events which "basic emotions" do have in contrast to other affective states [EC11].

They [EC11] exemplify it by that a high amount of anger-related signals or joy-related signals, respectively, in a short period of time can suggest either a mood of irritation or cheerfulness, respectively. However, what both of these situations (irritation and cheerfulness) have in common, is that they both do not have distinctive signals which would differentiate them from the *anger* or *happiness* emotion families, respectively.

Ekman's "Basic Emotions" from a Physiological Perspective

In [EC11] Ekman and Cordaro state that evidence exists that the "autonomic nervous system" (ANS) provides specific responses which correspond to certain "basic emotions", respectively. They expect that specific motor activity patterns only exist if they had a survival value for the corresponding emotion.

Examples they [EC11] bring (including the references they provide in [EC11]):

- anger leads to an increased blood flow towards arms and hands in order to prepare for a fight [LEF90]
- similarly for fear, however to legs and feet for fleeing [Ekm04]
- enjoyment and happiness correlate with an increased release of serotonin, oxytocin and others which increases a body's energy to reduce the effects of negative emotions [Uvn98]
- surprise leads to a person raising their eyebrows and inhaling air quickly in order to prepare a reaction to a sudden, unknown stimulus [EF03]
- disgust leads to triggering the gag reflex [KA10]

Other "Basic Emotion" Models

Ekman's Model of "Basic Emotions" is not the only basic emotions model that exists. In [TR11], Tracy and Randles look into the different models from Izard [Iza11], Panksepp and Watt [PW11], Levenson [Lev11], as well as Ekman and Cordaro [EC11]. For that, they introduce the following comparative table.

As the table shows, there is a clear overlap in what "basic emotions" are and how they are named. While Panksepp and Watt [PW11] use different namings for conceptually similar emotions, the other three [Iza11; Lev11; EC11] have overlaps, even in the naming, with "sadness", "fear", "anger", and "disgust".

Moreover, the Tracy and Randles [TR11] conclude that the four models are comprehensive, concretely: *"[. . .] This convergence provides confirmatory support for the four models [. . .]"* [TR11, p. 404].

Recent Contradicting Views towards "Basic Emotions"

Ortony in 2021 [Ort22] revisited once again the concept of "basic emotions" and explains based on the presumed emotion "surprise" why the basic emotion concept might be flawed.

Theoretically and empirically supportet basic emotions according to each model			
IZARD	*PANKSEPP & WATT*	*LEVENSON*	*EKMAN & CORDARO*
Happiness	*PLAY*	*Enjoyment*	*Happiness*
Sadness	*PANIC/GRIEF*	*Sadness*	*Sadness*
Fear	*FEAR*	*Fear*	*Fear*
Anger	*RAGE*	*Anger*	*Anger*
Disgust		*Disgust*	*Disgust*
Interest	*SEEKING*	*Interest?*	
Contempt?		*Contempt*	
	LUST	*Love?*	
	CARE	*Relief?*	*Surprise*

Table 2.1: *"Similarities and discrepancies among the clear-cut basic emotions included in each of the four models. Note: ? = Included in this author(s)' model, but the author(s) suggested that clear-cut supporting evidence is not yet available."* [TR11, p. 399]

Ortony states that emotions shall provide three different attributes to be considered emotions: 1) Intention, 2) Valence, 3) Consciousness.

The argumentation regarding the first attribute is that surprise and intention by definition contradict each other. The same argumentation shall work for the third. However, the attribute "Valence" is not that clear as one can attribute surprise to be a positive or a negative affective state. However, as the argument that surprise can be an emotion based on the above definition is already contradicted by it not fulfilling one, let alone two of the three attributes. The discussion about the "valence", provided in [Ort22] is intentionally left out of the scope of this Master's thesis.

However, what these discourses make clear is that there, as already stated in the first paragraph, indeed exists a variety of different definitions, very well also contradicting ones, for the term "emotion".

As (also elaborated in [Ort22]) there is no generally accepted definition for the term "emotion", it is only logical, that the concept of "basic emotions" is subject of a long ongoing debate.

Why we still choose to use it, is elaborated below (2.2.8).

2.2.2 Pleasure-Arousal-Dominance (PAD)

In [Meh96] Mehrabian describes a model where emotions are described in three continuous dimensions as per Kühnlenz et al. [Küh+13][4]. Said three dimensions are *Pleasure, Arousal* and *Dominance*.

[4] no access to original source

Pleasure describes the person's assessment of the situation (negative vs. positive). The higher this value, the better the person feels and therefore "happiness", e.g., is a likely emotion, whereas lower values indicate anger, for example (note that both mentioned emotions are part of Ekman's basic emotions model).

Arousal, regardless of the person experiencing the situation in a negative or positive way, describes how agitated the person is in said situation. A high value might indicate anger or surprise (both again part of Ekman's basic emotions), a low value might indicate boredom.

Dominance refers to how much control a person *feels* to have about the situation (surroundings and other people). The higher the value, the more control is felt.

For assessing the affective state of a person, Kühnlenz et al. [Küh+13], e.g., used the *Self-Assessment-Mannekin (SAM)* scale [BL94] as seen in Figure 2.3.

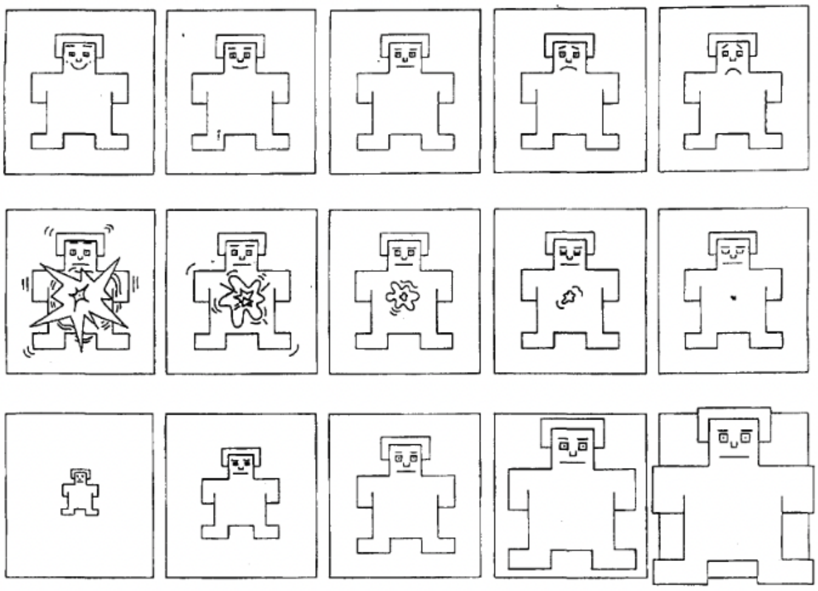


Figure 2.3: SAM Scale from [BL94]

Note that the term "Pleasure" in [BL94] is named "Valence" which aligns with the second attribute proposed by Ortony as described above [Ort22].

2.2.3 Component Process Model (CMP)

Scherer in [Sch05, p. 697] defines "emotion" as *"an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism"* (Scherer references [Sch87] and [Sch01]).

Moreover, Scherer [Sch05, p. 697] defines *"components"* of an emotion (episode) being the states of the so-called *"five subsystems"*, the *"process"* is defined by the "coordinated changes" in course of time.

Scherer provides an overview of the concept as follows [Sch05, p. 698].

Relationships between organismic subsystems and the functions and components of emotion

Emotion function	*Organismic subsystem and major substrata*	*Emotion component*
Evaluation of objects and events	*Information processing (CNS)*	*Cognitive component (appraisal)*
System regulation	*Support (CNS, NES, ANS)*	*Neurophysiological component (bodily symptoms)*
Preparation and direction of action	*Executive (CNS)*	*Motivational component (action tendencies)*
Communication of reaction and behavioral intention	*Action (SNS*	*Motor expression component (facial and vocal expression)*
Monitoring of internal state and organism–environment interaction	*Monitor (CNS)*	*Subjective feeling component (emotional experience)*

Table 2.2: *"Note: CNS 1/4 central nervous system; NES 1/4 neuro-endocrine system; ANS 1/4 auto- nomic nervous system; SNS 1/4 somatic nervous system." [Sch05, p. 698]*

2.2.4 Emotion Annotation & Representation Language (EARL)

An Emotion Annotation and Representation Language (EARL) is not an emotion model per se but rather is used to represent (different models of) emotions within a computational context, at least [SPL06] argued in that direction as they stated that there was no universally used model of emotion and that, moreover, such an EARL is difficult to create in a way that unified representation is possible.

Regardless, they proposed a first suggestion which left it to the user to *"plug in"* [SPL06, p. 88] their desired emotion model.

2.2.5 Affective Model of Interplay between Emotions & Learning by Kort et al.

Kort et al. [KRP01] looked into the interplay of emotions and learning. For that, they used an emotion model consisting of different *"sets possibly relevant to learning"* [KRP01, p. 44].

Each of them is presented on an axis ranging from −1.0 to 1.0 with 0.5-sized steps. The sets are [KRP01, p. 44] *"Anxiety-Confidence"*, *"Boredom-Fascination"*, *"Frustration-Euphoria"*, *"Dispirited-Encouraged"*, and *"Terror-Enchantment"* and have the following contents [KRP01, p. 44]:

"Anxiety ... Worry ... Discomfort ... Comfort ... Hopeful ... Confident"
"Ennui ... Boredom ... Indifference ... Interest ... Curiosity ... Intrigue"
"Frustration ... Puzzlement ... Confusion ... Insight ... Enlightenment ... Epiphany"
"Dispirited ... Disappointed ... Dissatisfied ... Satisfied ... Thrilled ... Enthusiastic"
"Terror ... Dread ... Apprehension ... Calm ... Anticipatory ... Excited"

Although this thesis does not specifically deal with learning, the usage of an axis representation was considered a good addition to this overview.

2.2.6 Plutchik's Wheel of Emotions

As per Tromp and Pechenizkiy [TP14] Plutchik in 1980 [Plu80][5] in 1980 proposed the following wheel of emotions.

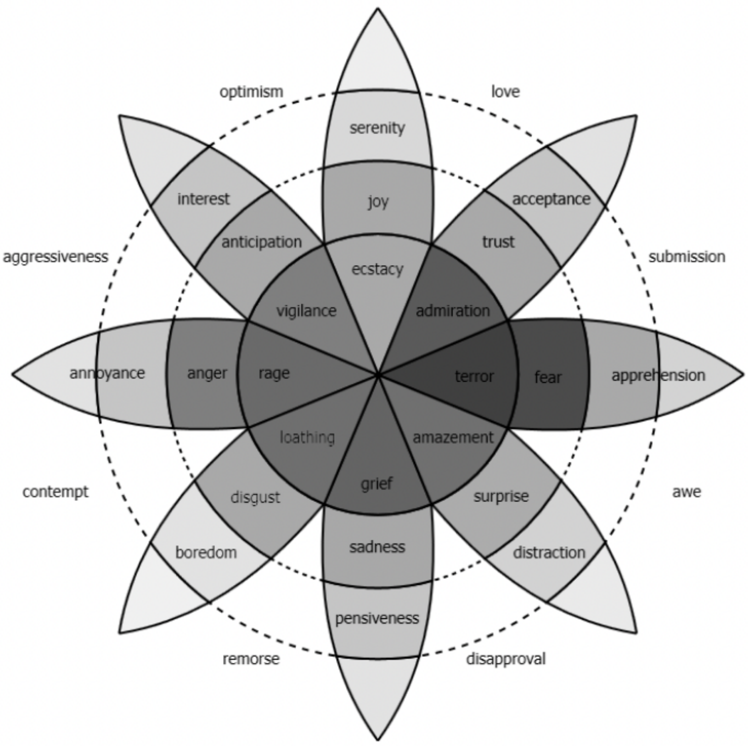


Figure 2.4: Plutchik's Wheel of Emotions from [TP14] Who Referenced [Plu80]

[5] no access to original source

As can be seen and is described in [TP14, p. 2], the wheel defines eight *"basic emotions"* in which to some extent all human emotions are to be found, therefore making the model *"complete"*, according to [TP14, p. 2], Plutchik considers these emotions furthermore to be culturally independent. Furthermore, that each emotion is the opposite of another of said eight emotions, therefore the wheel contains four axis, where each extreme is the opposite of the other. Moreover, according to [TP14, p. 2], Plutchik defines eight *"human feelings"* which are derived by combining two basic emotions. [TP14, p. 2] outline that this provides us with 16 dimensions of emotions/feelings in total.

[TP14] used Plutchik's model in emotion detection, and, moreover, Coyne et al. [CMM20] used a "wheel approach" (Geneva Emotional Wheel) to let participants of a survey rate the emotional expressions of a robot. Such an emotion model, as we can see because of this, can also be applied in the other direction (humans rate *artificial* emotional expressions of robots).

2.2.7 Additional Perspectives, Advantages & Disadvantages

At this point, it has to be made clear that the preceding analysis does not aim to be complete. There is a variety of additional perspectives and aspects which could be considered, e.g. evolution theory or genetics as well as the whole neuro-biology, etc.

However, this thesis deals with the aspect of *modeling* emotions and not with the processes happening in the human brain, e.g., when *feeling* something.

As we see based on the explanations above, scientists are not quite clear *how* to model emotions or what emotions are in the first place. However, the different models can, as we also see later on, be used to model emotions within a computational context.

This also suggests that each model has advantages and disadvantages. These, however, might very well not be absolute but relative to the use context. From a modeling perspective of a computational system, it is a reasonable aim to keep the complexity low and therefore have fewer dimensions, fewer (basic) emotions, a discrete perspective of emotions and not a continuous, as well as, e.g., not having two different emotions present in a situation at the same time. However, these abstractions lead to not being able to model the reality as exact as possible as body functions are, as we all know, not deterministic or even binary.

Choosing the "right model" therefore can highly depend on the application context too. In this context, it has to be noted that Aschenbrenner [Asc19, pp. 80–81] in 2019 proposed a framework for rating emotion models based on their validity, number of dimensions, scaling, dimension dependencies, feasibility, application, and number of emotions.

2.2.8 Why Ekman's Model Has Been Chosen as Focus Point

The in advance proposed "choice" for Ekman's model is not that strong as an argument anymore, because the implementation idea got abandoned, as explained in Section 6.6.

However, the author decided to still put his focus on Ekman's emotion model because the concept of "basic emotions" deemed itself feasible scope-wise because (in Ekman's case) as a starting point only six emotions would have to be considered. For future work, this kind of bottom-up approach, after considering the more complex models (as they simply have more emotions) within this section, might still be the way to go.

2.3 Emotion in Human-Robot Interaction

The following chapter deals with *relevance* of emotion within the field of Human-Robot Interaction. At first, the concept of companion robots is outlined, followed by elaborating on human-robot dialogue and artificial empathy in this context.

Before exploring different ideas and technologies of detecting emotions, the topic of "user satisfaction", and more specifically, how to measure it, is presented.

2.3.1 Companion Robots

Dautenhahn et al. [Dau+05] already asked back in 2005 what a companion robot actually is or, more specifically, what people expect from robots aiming to be described by the term. They found out that a vast majority saw the concept of a companion robot positively or at least neutral (less than 30 percent did not like it). Regarding the role of such a robot, around 80 percent saw it as a an assistant, a bit over 70 percent ascribed it to be a machine, a bit over 45 percent as a servant and only 20 percent as a mate or friend respectively. Tasks, people were most interested in were vacuum cleaning with over 95 percent, security and gardening with over 55 percent and entertainment with a bit over 50 percent, childcare was explicitly not desired with only a bit over 10 percent seeing it as feasible.

Jayawardena et al. [Jay+13] in 2013 designed a robotic wheel chair to also be a socially assistive companion, whereas Adam and Cavedon [AC13] looked into the application of story telling. Michaelis and Mutlu [MM18] looked into the usage of companion robots as learning-companions within the context of guided reading activities.

As companion robots in many cases might have social aspects (e.g., when considering it as mate or friend as outlined above), the definition of what a "social robot" is as per Dautenhahn [Dau21, section 38.11] appears to be relevant too: *"Social robots are a special kind of (embodied) interactive artifact (see Kahn et al., 2004[6]; Melson et al., 2009[7]) that may afford new types of interactions with people, and new roles they may adopt in society may emerge (see Dautenhahn, 2003[8])".*

This outline aims to show that, despite it far from being complete, there exists a variety of expectations and fields of usage for the concept of "companion robots".

[6] [Kah+04]
[7] [Mel+09]
[8] [Dau03]

2.3.2 Human-Robot Dialogue

Marge et al. [Mar+17] motivate the need for human-robot dialogue within the need of ensuring effective teaming between humans and robots. In this context, they outline the benefits of natural language dialogues as being familiar and flexible for humans, called *"natural communication"* within their work [Mar+17, p. 930]. Said form of communication requires several attributes. First, it has to be *"situated"*, according to them [Mar+17, p. 930], *"[...] so that it refers to the environment, build upon references to physical common ground, and must be initiated at appropriate times"* [Mar+17, p. 930].

Quoted definition is used within this thesis as the communication in our scenario also takes place in a natural way within a certain situation. Moreover, because of that, the emotions on the side of the human come into play and, although the definition above does not explicitly reference the emotional context, we consider it to be part of the dialogue's *situatedness*.

2.3.3 Artificial Empathy

Fung et al. [Fun+18, p. 177] define *empathy* as *"the recognition and sharing of the emotion of the other"*. James et al. [JWM18], referencing [Fun+18, p. 177], elaborate on the definition of *artificial empathy*. They see artificial empathy *"as the programmed affective reaction of the robot to the behavior of the human that it can sense according to the technology embedded in it"* [JWM18, p. 636].

The scenario used within this thesis aims to mock said artificial empathy within its portrayal of the interaction. To be precise, the scenario describes a person sharing their emotions to the robot. The robot reacts accordingly. However, it only applies artificial empathy according to above definition when handling the emotions via detecting it or asking explicitly about them. Artificial empathy, however, is also to be found, when the robot detects the wrong emotion(s).

2.3.4 User Satisfaction

When searching for "user satisfaction" via Google[9] within the Oxford Dictionary, the first reference one gets points to "customer satisfaction" which is defined as [Ref] *"measure of the degree to which goods and services provided by businesses and non-profit organizations satisfy the needs or expectations of those who buy or use them"*.

Transferring this definition to our use case, by going strictly by the two words building the term "user satisfaction" – user and satisfaction – the combination shall describe *how* satisfied a user is when using *something*. For this thesis, we intuitively assume the "something" to be a robot (in our case, Q.bo One) and the user to be the human interacting with said robot. Therefore, the question is not what user satisfaction is, but rather how to measure the degree of it.

[9] google.com, term: "user satisfaction oxford dictionary", Apr. 3, 2022

Lindgaard and Dudek [LD03] back in 2003 already explained that the measure for user satisfaction, e.g. being productivity for work, might differ for use cases within leisure time. They believe user satisfaction to be *"the subjective sum of the interactive experience"* [LD03, p. 430].

Fischer outlined in [Fis06, p. 122] exemplary, how different users react differently towards a computer's first question starting an interaction. For example, one might react to the question *"yes, hello, how do you do?"* with *"fine"*, whereas another might react more playfully with *"hello computer. hello. cuckoo"*. What we can see within this example is that not only users might have different tasks/intents (as explored in the paragraph above) they want to achieve when interacting with – in our case – the robot but might also react completely differently within the same situation compared to others which might therefore also influence their level of satisfaction.

As it is not the aim of this thesis to define what user satisfaction actually is, we will be satisfied with above exploration of defining the term within this thesis as the matter of what exactly the term describes is clearly of secondary nature compared to how to measure certain aspects which are likely to describe how satisfied or happy people are with a certain interaction.

Therefore, we look into the work of Kühnlenz et al. [Küh+13], Fischer [Fis06] and Dautenhahn [Dau21] in order to argue why specific measures (in our case the Godspeed Questionnaire Series) are used within the empirical part of this thesis for measuring user satisfaction.

Kühnlenz et al. [Küh+13] looked into the issue how an emotionally adapting robot to the user would increase the helpfulness by said user towards said robot. They wanted to introduce empathy within the user by triggering a feeling of similarity (note: called "Resemblance Dimension" in the conducted expert interviews in Chapter 4). Within the context of this thesis, we view "helpfulness" as part of user satisfaction as it is an effect one can also consider to be valuable within a dialogue, as a user, who "likes" the robot (or is *satisfied* with it) might be more eager to help to prevent or resolve potential misunderstandings.

Moreover, [Küh+13] also looked into how the emotional adaption influences *"HRI-key concepts anthropomorphism and animacy"* [Küh+13, p. 457] which are also relevant within the empirical part of this thesis.

Regarding the empathy towards a robot (called EDDIE), which was to guess the person a participant came up with in their had, [Küh+13, p. 468] asked the participants to rate the following statements from *"1 (not true at all) to 5 (completely) true"* [Küh+13, p. 468]:

- *"I'm happy EDDIE has guessed my person/I'm sorry that EDDIE didn't guess at my person"*

- *"I would have been sorry if EDDIE had not guessed my person/It would have been nice if EDDIE had guessed my person"*
- *"It would be a pity if somebody damaged EDDIE, and I would try to interfere"*
- *"I would have been proud if EDDIE had not guessed my person/I am proud that EDDIE did not guess my person"*

How helpful the participants actually acted towards the robot was measured via a picture labeling task, where – after labeling at least five pictures – the participant got told by the robot that they are free to leave at any time. Measurements like repeating pictures and manual typing were taken in order to make the labeling not a fun activity itself [Küh+13].

Another measure [Küh+13, p. 468] applied was a selection from the Godspeed questionnaires, where they describe the four dimensions according to Bartneck et al. [Bar+09] HRI as follows:

- *"Anthromorphism: how natural the robot appeared"*
- *"Animacy: the liveliness of the robot"*
- *"Likeability: how pleasant the robot appeared"*
- *"Perceived Intelligence: how the mental abilities of the robot were perceived"*

However, it has to be said on this occasion, that within the work of this thesis, the fifth dimension [Bar+09] "Perceived Safety" where the participants have to rate *their own* emotional state, is also considered.

Referenced within the expert interviews, Heerink et al. in 2009 [Hee+09] suggested a toolkit for measuring the acceptance of an assistive social robot. For that, they defined 13 different constructs [Hee+09, p. 529]:

1. Anxiety
2. Attitude
3. Facilitating conditions
4. Intention to use
5. Perceived adaptability
6. Perceived enjoyment
7. Perceived ease of use
8. Perceived sociability
9. Perceived usefulness

10. Social influence
11. Social presence
12. Trust
13. Use/Usage

Said constructs can be assessed via statements (two to five each) on a scale from 1 to 5 where five represents strong agreement with the corresponding statement. Note that within negative constructs (e.g., anxiety) strong agreement leads to a smaller value in satisfaction, whereas within positive constructs (e.g., intention to use) it leads to a bigger value in satisfaction [Hee+09].

This outline implies that it is not that easy to define what "user satisfaction" actually is. One might argue that a person is happy or satisfied with a robot if they are willing to help it (as explored above); however, this state might also very well be just called a *positive sentiment* or similar towards a robot. We also looked into the Heerink's toolkit [Hee+09] to measure acceptance. However, Heerink in [Hee+09, p. 528] uses the definition *"the demonstrable willingness within a user group to employ technology for the tasks it is designed to support"* for *"user acceptance"* as per [DM96, section "THE CONCEPT OF ACCEPTANCE"]. However, just because one is willing to employ technology for a specific task does not mean they are necessarily *satisfied* with it. Although this nuanced differentiation is not relevant for the work within this thesis, it has to be made clear to avoid any sort of ambiguities.

2.4 Methods for Emotion Detection

When looking into the research landscape, it quickly gets clear that there is no – at least not yet – "go to"-method for detecting emotions. Different people already came up with different approaches which this section is trying to categorize and order a bit.

First of all, within this thesis, we propose to differentiate between *content-based emotion detection* and *transmission-based emotion detection*.

First mentioned does not rely on the spoken language itself or any corresponding parameters (voice, body, etc.) but solely on the contents of the transmitted material. Therefore, the application comes into place after eventually detecting the spoken text. Considering this description, this way of detecting emotions is actually equivalent or at least very close to the concept of *sentiment analysis* which also is a key part of Chapter 3.

Second mentioned leaves the contents out of the equation and focuses on the transmission of said contents. E.g., how loud is the speaker, which parameters does the voice have, how long are different pauses, how consistent is the speaker, what is the pitch of the voice, different body parameters, etc.

2.4.1 Content-based Emotion Detection

Binali et al. [BWP10, pp. 173–174] back in 2010 proposed the following classification for text-based emotion detection approaches: 1) Keyword-based emotion detection, 2) Learning-based emotion detection, and 3) Hybrid-based emotion detection.

Keyword-based Emotion Detection

When using the *keyword-based approach*, the input is scanned for specific, predefined keywords. However, it has as disadvantages that it is specific to a certain domain and the existence of certain keywords is necessary for reliable results [BWP10]. Moreover, Canales and Martínez-Barco [CM14] – referencing Suttles and Ide [SI13] – pointed out choosing the content for an emotion lexicon is subjective. [SI13] additionally mentioned the lack of guarantee for the lexicon being comprehensive. Regarding the generation of a robust lexicon for emotion detection, Bandhakavi et al. [Ban+17] in 2017 elaborated on using blogs, news headlines and tweet as sources for learning *"word-emotion associations"* [Ban+17, p. 102] for usage in such a lexicon. As a notice for a German-language use case, the "Affective Dictionary Ulm" [HSK92] is an example for a corresponding lexicon. Additionally, e.g, Hermanns [Her95] back in 1995 already talked about *affective* words within German-language contexts.

Looking back into the example from Section 1.3, the explicit statement "I am happy" can be detected easily under the premise that "happy" is a keyword. The implicitly encoded emotion in "Gregor is marring Claudia next week" cannot detected solely based on the keyword happy as it is not in the statement. However, if the person likes Gregor and Claudia and thinks they are a good fit, said person might very well be happy. For that to be known, additional information would be required. For example, the robot would have to keep track of the relationships the person has to the mentioned two or extract information from other parameters (e.g., voice). However, that would exceed the keyword-based approach.

On an additional note, pre-processing might be helpful to increase the accuracy of the detection [BWP10].

Learning-based Emotion Detection

This approach, as the name already suggests, is based on machine learning technologies. Input is categorized into *"emotion classes"* [BWP10, p. 173] by a trained classifier. This brings the advantage of implementing adaptations to a different domain more easily because of the learning possibilities from corpora. Putting a corresponding training set into an appropriate machine learning algorithm then results in a new classification model [BWP10].

In contrast to these advantages, obtaining a large corpora might be unfeasible in some cases. Even more importantly, the distinction between emotion classes is not as clear anymore as in, e.g., the *keyword-based approach*. Additionally to these blurry boundaries, a lack of context analysis is also to be noticed within this approach. Algorithms, e.g., the mentioned support vector machine, can be used to tackle these problems [BWP10]. However, explaining them further exceeds the scope of this thesis.

Hybrid-based Emption Detection

A hybrid-based approach combines the above two in order to gain more accurate results. This can be achieved by a) training classifiers and b) adding linguistic information from, e.g., dictionaries to it [BWP10].

Binali et al. [BWP10, p. 175] in this context introduced an example for an hybrid-based emotion detection architecture as shown in their figure below.

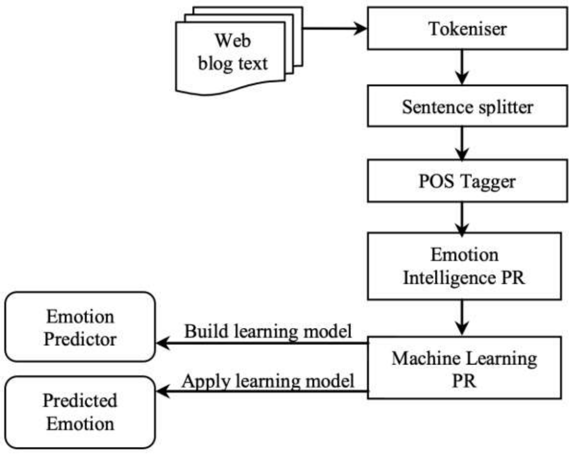


Figure 2.5: Hybrid-based Emotion Detection by Binali et al. [BWP10, p. 175]

In this case (simplified), the keyword-based component is the collected blog-data, whose text then is split into tokens followed by identifying sentences. The tokens are then tagged and afterwards, some post-processing happens. After these steps, the learning-based component takes over and the actual machine learning takes place [BWP10].

After testing their model, Binali et al. suggest – based on their data – that semantic and syntactic information improve prediction accuracy [BWP10].

Similar & Additional Models

On this occasion, it has to be mentioned that Shivhare and Khethawat [SK12] in 2012 elaborated on similar concepts. However, they differentiate between the *"Keyboard Spotting Technique"* [SK12, p. 2] and the *"Lexical Affinity Method"* [SK12, p. 3] regarding (the) keyword-based method(s). However, they consider the latter as an extension to the Keyboard Spotting Technique as it *"assigns a probabilistic «affinity» for a particular emotion to arbitrary words"* [SK12, p. 3] compared to solely considering emotional keywords.

Additionally, Agrawal and An [AA12, p. 347] see *“Linguistic Rules-based Approaches”* within these categories which exceeds relying solely on keywords by considering other aspects of the language(s). They further differentiate those rule-based approaches between the existence of *“affect lexicons”* [AA12, p. 347].

Moreover, they proposed their own framework as shown in Figure 2.6. The key concept behind said framework is that sentences are to be labeled with emotion categories. First, preprocessing helps to extract relevant words and detects *“syntactic dependencies”* [AA12, p. 347]. Afterwards, *“word-level analysis”* [AA12, p. 347] is performed which is done by computing a so-called *“emotion vector”* [AA12, p. 347] for *“affect-bearing words”* [AA12, p. 347] based on their *“semantic relatedness to emotion concepts”* [AA12, p. 347]. Following, another analysis is performed on a phrase-level which uses the newly gained context information and reiterates on the previously calculated vectors, which are in a final step then being aggregated [AA12].

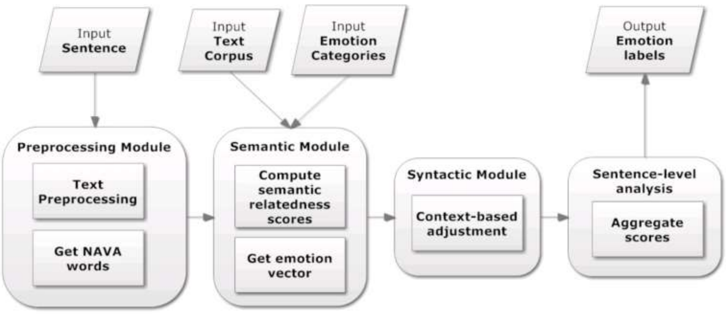


Figure 2.6: Emotion Detection Framework by Agrawal & An [AA12, p. 348]

Speech Recognition

Kunze et al. [Kun+17] specifically looked into the scenario of transferring a Wav2Letter convolutional neural network trained for automated speech recognition systems for English to a German-language context. They showed that their technique enabled faster training while not relying on resources beyond ones being fundable for consumers. When looking more detailed into their model, they concluded that adapting the weights of the network to only a small degree was already enough for achieving good performance for inner layers of the networks. The especially interesting finding in this context is that their model which was based on the English one was more performant than their German baseline model. This small notice is mentioned within this thesis as it seems to be especially useful when considering German-language contexts.

Emotion Detection vs. Sentiment Analysis

Nandwani and Verma [NV21] among others differentiate between emotion detection and sentiment analysis. According to them, emotion detection *“[. . .] determines an* *individual's emotional/mental state precisely"* [NV21, p. 80], whereas sentiment analysis assesses whether the author of a textual artifact has *"a negative, positive, or neutral attitude toward an item, administration, individual, or location"* [NV21, p. 80]. This differentiation is especially important for Chapter 3.

Similar as for emotion detection – as discussed prior – most work on sentiment analysis exists for English-language use cases is, e.g., pointed out by Dashtipour et al. [Das+16] in their review of the state-of-the-art of multilingual sentiment analysis. Lo et al. [Lo+17] review tools for a multilingual use case. However, as the topic of sentiment analysis would exceed the scope of this thesis, no more elaboration on it takes place. It, nevertheless, got introduced as it provides a good starting point for research regarding emotion detection (as sentiment analysis is a more "basic" case, as we have seen) in Chapter 3.

2.4.2 Transmission-based Emotion Detection

While the focus of this thesis lies on content-based emotion detection, there also exist methods for detecting emotions based on, e.g, detecting aspects of the human's face or posture. Considering aspects of the voice (pitch, volume, etc.) is, as already mentioned, also a possibility.

Graterol et al. [Gra+21] proposed a framework for detecting emotions and storing them in a semantic repository which is based on extensible ontology they called *"EMONTO"* [Gra+21, p. 2] (for EMotion ONTOlogy). While they also include emotion detection from text into their framework, they have a focus on emotion detection from speech too. They propose to gather data from audio, video, and images which is then to be converted into formats (feature vectors, normalized spectograms) which can be used for emotion detection. Within said framework, they define different *"modalities"* [Gra+21, p. 9] for an emotion which are *"Context, Text, Voice, Posture, Face, Gesture"* [Gra+21, p. 2, Figure 5]. Their work is an example for integrating different methods to work with each other for more robust and precise emotion detection. An image of their framework is included below.

Albu [Alb15] describe an architecture for detecting emotions from facial expressions using a Kinect camera. For the processing, Albu uses a neural network and the emotional states are described by the following expressions: *"sad, angry, neutral, happy, engaged"* [Alb15, p. 432]. Moreover, Albu proposes to add use pulse analysis for analyzing emotions too.

2.4.3 Why this Thesis Is not about Concrete Tools

The goal of this thesis is to provide an overview of concepts for emotion detection rather than specific tools. As the originally proposed idea of a concrete implementation got abandoned because of the circumstances as described in Section 6.6. However, the tools referenced in the sources discussed can be considered as a starting point for working within this field, although it has to be noted that potential updates (eventually even to the same tools) based on the time passed should be considered in advance.

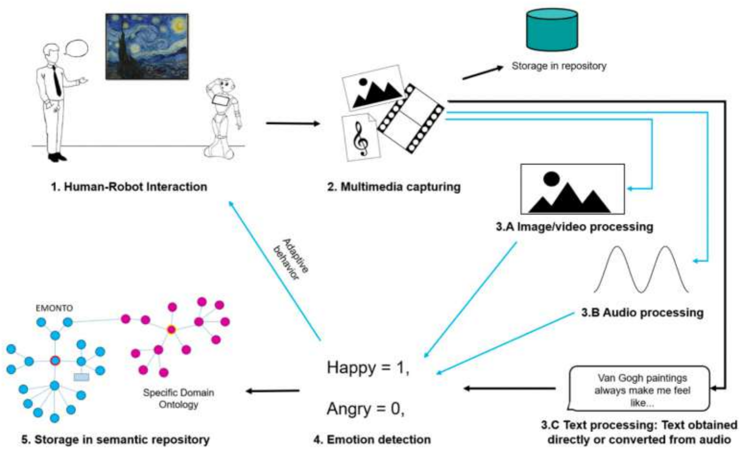


Figure 2.7: Overview of EMONTO by Graterol et al. [Gra+21, p. 6]

CHAPTER 3

Performance Differences: Native vs. Translation-layered Solutions

3.1 Preliminary Remarks

The aim of the following comparison is to answer the first research question (RQ1) which is *"What are the performance differences (i.e., successful detection ratio) between a language-specific solution and using an already existing English-language solution with a translation layer added to it?"*

For tackling this question, doing a "scoping study" as in [AO05] – considering [LCO10] – was chosen. This route was taken because of the broad scope this field encloses.

As a result, performance differences shall be summarized and presented in a concise way. It has to be stated that this analysis does not aim for completeness, as this, on the one hand, would contradict the methodological approach, and, on the other hand, would exceed the scope of a master's thesis.

3.2 Envisioned Detection Process

For better understanding, the following visualization of the envisioned detection process in the context of this thesis has been created.

Additional Remarks: Note that "recording audio" does not necessarily mean that the audio is persisted longer than the time span in which the parsing takes place. Moreover, it is not necessary that the methods are combined or that a translation layer is present. Those have only been added as they are additional scenarios this thesis looks into.

Although one can argue that the facial expression necessarily correlates with certain expressions in a specific language (for example, imagine a mouth wide open while

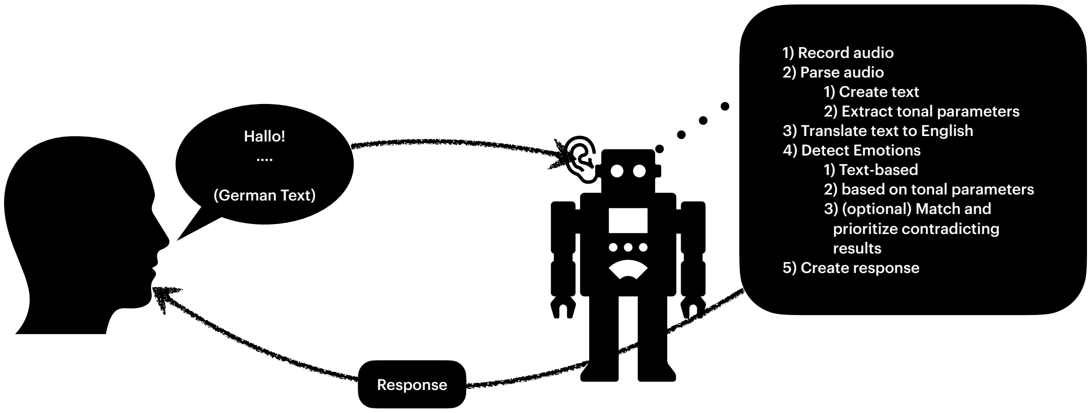


Figure 3.1: Envisioned Detection Process

screaming), these aspects are not considered within this analysis as they would increase the scope significantly. However, detecting emotions by facial expressions is briefly discussed in Chapter 2. Therefore, the only input form relevant within the envisioned detection process is audio which already got converted into text.

3.3 Scoping Study

Scoping studies are as per [AO05] divided into at least five stages. Therefore, the following paragraphs are structured accordingly.

Regarding the outlined potential purposes in [AO05], it can be said that besides "1." all three remaining apply to some extent, as the analysis aims to *"determine the value of undertaking a full systematic review"* [AO05, p. 21] for potential future work, which would go beyond the scope of this thesis, as well as *"summarize and disseminate research findings"* [AO05, p. 21], and, if possible, to *"identify research gaps in the existing literature"* [AO05, p. 21].

3.3.1 Stage 1: Identifying the Research Question

The research question (RQ1) *"What are the performance differences (i.e., successful detection ratio) between a language-specific solution and using an already existing English-language solution with a translation layer added to it?"* has already been defined in advance.

In order to find out whether the performance difference when sending the recognized text through a translation layer or not is significant, we have to define what "significant" is in our context. However, this definition is to be done after the analysis in order to not restrict research results in advance.

3.3.2 Stage 2: Identifying Relevant Studies

In order to identify relevant literature, a Google Scholar search (parameters: "Any time", "Sort by relevance", "Any type", "include citations" [but not patents]) was conducted. The following terms were used and the materials of the first two pages each were considered by titles and, if necessary, skim-reading the corresponding abstracts.

- "emotion detection performance differences english german"
- "emotion detection performance differences english german text"
- "sentiment analysis performance differences english german"
- "sentiment analysis multilingual performance differences"
- "sentiment analysis performance differences translation"

After using these five search phrases, saturation already was present as most of the articles found have already been identified prior (either as not appropriate or already noted). Because of the already big scope of this master's thesis, references within the paper have only been considered in an "unsystematic" way. This means, that the references have only been checked when (in a later step) a part of the text was quoted or referred to a specific literature entity.

The following resources have been found within this process and are considered to be relevant (first found == first listed):

- Pansy Nandwani and Rupali Verma. "A review on sentiment analysis and emotion detection from text". In: *Social Network Analysis and Mining* 11.1 (2021), pp. 1–19. ISSN: 1869-5469. DOI: 10.1007/s13278-021-00776-6
- David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. "Supervised sentiment analysis in multilingual environments". In: *Information Processing & Management* 53.3 (2017), pp. 595–607. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2017.01.004. URL: https://www.sciencedirect.com/science/article/pii/S0306457316302540
- Matheus Araujo, Julio Reis, Adriano Pereira, and Fabricio Benevenuto. "An evaluation of machine translation for multilingual sentence-level sentiment analysis". In: *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. SAC '16. Pisa, Italy: Association for Computing Machinery, 2016, pp. 1140–1145. ISBN: 978-1450337397. DOI: 10.1145/2851613.2851817
- Muhammad Haroon Shakeel, Safi Faizullah, Turki Alghamidi, and Imdadullah Khan. "Language independent sentiment analysis". In: *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*. 2020, pp. 1–5. DOI: 10.1109/AECT47998.2020.9194186
- Gayane Shalunts, Gerhard Backfried, and Nicolas Commeignes. "The impact of machine translation on sentiment analysis". In: *Data Analytics* 63 (2016), pp. 51–56. URL: https://biblio.ugent.be/publication/8116621/file/8132035#page=64

- Alexandra Balahur and Marco Turchi. "Multilingual sentiment analysis using machine translation?" In: *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Jeju, Korea: Association for Computational Linguistics, July 2012, pp. 52–60. URL: https://aclanthology.org/W12-3709

- Alexandra Balahur and Marco Turchi. "Improving sentiment analysis in Twitter using multilingual machine translated data". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, Sept. 2013, pp. 49–55. URL: https://aclanthology.org/R13-1007

- Alexandra Balahur and Marco Turchi. "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis". In: *Computer Speech & Language* 28.1 (2014), pp. 56–75. ISSN: 0885-2308. DOI: 10.1016/j.csl.2013.03.004. URL: https://www.sciencedirect.com/science/article/pii/S088523081300020X

- Tomáš Kincl, Michal Novák, and Jiří Přibil. "Improving sentiment analysis performance on morphologically rich languages: language and domain independent approach". In: *Computer Speech & Language* 56 (2019), pp. 36–51. ISSN: 0885-2308. DOI: 10.1016/j.csl.2019.01.001. URL: https://www.sciencedirect.com/science/article/pii/S0885230816300109

- Valentin Barrière and Alexandra Balahur. "Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation". In: *CoRR* abs/2010.03486 (2020). arXiv: 2010.03486. URL: https://arxiv.org/abs/2010.03486

- Alexandra Balahur and José M. Perea-Ortega. "Sentiment analysis system adaptation for multilingual processing: the case of tweets". In: *Information Processing & Management* 51.4 (2015), pp. 547–556. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2014.10.004. URL: https://www.sciencedirect.com/science/article/pii/S0306457314000934

- Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. "More than a feeling: accuracy and application of sentiment analysis". In: *SSRN Electronic Journal* (Jan. 2022). DOI: 10.2139/ssrn.3489963

- Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. "A twitter corpus and benchmark resources for german sentiment analysis". en. In: *5th International Workshop on Natural Language Processing for Social Media*. Boston MA, USA: Association for Computational Linguistics, 2017, pp. 45–51. DOI: 10.21256/zhaw-1530. URL: https://digitalcollection.zhaw.ch/handle/11475/1856

- Matheus Araújo, Adriano Pereira, and Fabrício Benevenuto. "A comparative study of machine translation for multilingual sentence-level sentiment analysis". In: *Information Sciences* 512 (2020), pp. 1078–1102. ISSN: 0020-0255. DOI: 10.1016/j.ins.2019.10.031. URL: https://www.sciencedirect.com/science/article/pii/S0020025519309879

- Alexander Hogenboom, Bas Heerschop, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. "Multi-lingual support for lexicon-based sentiment analysis guided by semantics". In: *Decision Support Systems* 62 (2014), pp. 43–53. ISSN: 0167-9236. DOI: 10.1016/j.dss.2014.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0167923614000645

- David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. "EN-ES-CS: an English-Spanish code-switching Twitter corpus for multilingual sentiment analysis". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 4149–4153. URL: https://aclanthology.org/L16-1655
- Mark Boukes, Bob Van de Velde, Theo Araujo, and Rens Vliegenthart. "What's the tone? easy doesn't do it: analyzing performance and agreement between off-the-shelf sentiment analysis tools". In: *Communication Methods and Measures* 14.2 (2020), pp. 83–104. DOI: 10.1080/19312458.2019.1671966
- Thomas Body, Xiaohui Tao, Yuefeng Li, Lin Li, and Ning Zhong. "Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models". In: *Expert Systems with Applications* 178 (2021), p. 115033. ISSN: 0957–4174. DOI: 10.1016/j.eswa.2021.115033. URL: https://www.sciencedirect.com/science/article/pii/S0957417421004747
- Mohammed Kaity and Vimala Balakrishnan. "Sentiment lexicons and non-english languages: a survey". In: *Knowledge and Information Systems* 62.12 (2020), pp. 4445–4480. DOI: 10.1007/s10115-020-01497-6
- Zilong Wang, Zhaohong Wan, and Xiaojun Wan. "Transmodality: an end2end fusion method with transformer for multimodal sentiment analysis". In: *Proceedings of The Web Conference 2020.* New York, USA: Association for Computing Machinery, 2020, pp. 2514–2520. ISBN: 978-1450370233. URL: https://doi.org/10.1145/3366423.3380000
- Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. "How translation alters sentiment". In: *Journal of Artificial Intelligence Research* 55 (2016), pp. 95–130. DOI: 10.1613/jair.4787. URL: https://jair.org/index.php/jair/article/view/10976
- Balamurali A.R., Mitesh M. Khapra, and Pushpak Bhattacharyya. "Lost in translation: viability of machine translation for cross language sentiment analysis". In: *International Conference on Intelligent Text Processing and Computational Linguistics.* Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer, 2013, pp. 38–49. ISBN: 978-3642372568. DOI: 10.1007/978-3-642-37256-8_4

It has to be stated that literature, which was found additionally to the process [i.e., when writing up the proposal or in the initial research process before defining the exact method(s) of this chapter] and has not been found again within the described process, is used regardless. However, it is explicitly stated in each of these cases.

The following resources additionally emerged of this process:

- Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. "Multilingual sentiment analysis: state of the art and independent comparison of techniques". In: *Cognitive Computation* 8.4 (2016), pp. 757–771. ISSN: 1866-9964. DOI: 10.1007/s12559-016-9415-7
- Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. "Multilingual sentiment analysis: from formal to informal and scarce resource languages". In: *Artificial Intelligence Review* 48.4 (2017), pp. 499–527. ISSN: 1573-7462. DOI: 10.1007/s10462-016-9508-4

- Karsten Tymann, Matthias Lutz, Patrick Palsbröker, and Carsten Gips. "Gervader-a german adaptation of the vader sentiment analysis tool for social media texts". In: *LWDA*. 2019, pp. 178–189
- Kerstin Denecke. "Using sentiwordnet for multilingual sentiment analysis". In: *2008 IEEE 24th International Conference on Data Engineering Workshop*. 2008, pp. 507–512. DOI: 10.1109/ICDEW.2008.4498370

Furthermore note that the following resources were not accessible for the author although having been considered relevant at first glance:

- Gayane Shalunts and Gerhard Backfried. "SentiSAIL: sentiment analysis in english, german and russian". In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Ed. by Petra Perner. Springer. 2015, pp. 87–97. ISBN: 978-3319210247. DOI: 10.1007/978-3-319-21024-7_6
- Maritza Bustos-López, Nicandro Cruz-Ramírez, Alejandro Guerra-Hernández, Laura Nely Sánchez-Morales, and Giner Alor-Hernández. "Emotion detection from text in learning environments: a review". In: *New Perspectives on Enterprise Decision-Making Applying Artificial Intelligence Techniques*. Ed. by Julian Andres Zapata-Cortes, Giner Alor-Hernández, Cuauhtémoc Sánchez-Ramírez, and Jorge Luis García-Alcaraz. Cham: Springer International Publishing, 2021, pp. 483–508. ISBN: 978-3030711153. DOI: 10.1007/978-3-030-71115-3_21
- Ashima Yadav and Dinesh Kumar Vishwakarma. "Sentiment analysis using deep learning architectures: a review". In: *Artificial Intelligence Review* 53.6 (2020), pp. 4335–4385. DOI: 10.1007/s10462-019-09794-5

3.3.3 Stage 3: Study Selection

After the second stage, the resources found were studied in more detail by reading the abstracts and when the judgement was not clear afterwards, reading introductions, discussions, and conclusions first, or – if applicable – tables of contents, too.

Based on that, the following resources have been selected for further review (alphabetically per first autor):

- Balamurali A.R., Mitesh M. Khapra, and Pushpak Bhattacharyya. "Lost in translation: viability of machine translation for cross language sentiment analysis". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer, 2013, pp. 38–49. ISBN: 978-3642372568. DOI: 10.1007/978-3-642-37256-8_4
- Matheus Araujo, Julio Reis, Adriano Pereira, and Fabricio Benevenuto. "An evaluation of machine translation for multilingual sentence-level sentiment analysis". In: *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. SAC '16. Pisa, Italy: Association for Computing Machinery, 2016, pp. 1140–1145. ISBN: 978-1450337397. DOI: 10.1145/2851613.2851817
- Matheus Araújo, Adriano Pereira, and Fabrício Benevenuto. "A comparative study of machine translation for multilingual sentence-level sentiment analysis". In: *Information Sciences* 512 (2020), pp. 1078–1102. ISSN: 0020-0255. DOI: 10.1016/j.ins.2019.10.031. URL: https://www.sciencedirect.com/science/article/pii/S0020025519309879

- Alexandra Balahur and Marco Turchi. “Multilingual sentiment analysis using machine translation?” In: *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Jeju, Korea: Association for Computational Linguistics, July 2012, pp. 52–60. URL: https://aclanthology.org/W12-3709

- Alexandra Balahur and Marco Turchi. “Improving sentiment analysis in Twitter using multilingual machine translated data”. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, Sept. 2013, pp. 49–55. URL: https://aclanthology.org/R13-1007

- Alexandra Balahur and Marco Turchi. “Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis”. In: *Computer Speech & Language* 28.1 (2014), pp. 56–75. ISSN: 0885-2308. DOI: 10.1016/j.csl.2013.03.004. URL: https://www.sciencedirect.com/science/article/pii/S088523081300020X

- Alexandra Balahur and José M. Perea-Ortega. “Sentiment analysis system adaptation for multilingual processing: the case of tweets”. In: *Information Processing & Management* 51.4 (2015), pp. 547–556. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2014.10.004. URL: https://www.sciencedirect.com/science/article/pii/S0306457314000934

- Valentin Barrière and Alexandra Balahur. “Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation”. In: *CoRR* abs/2010.03486 (2020). arXiv: 2010.03486. URL: https://arxiv.org/abs/2010.03486

- Thomas Body, Xiaohui Tao, Yuefeng Li, Lin Li, and Ning Zhong. “Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models”. In: *Expert Systems with Applications* 178 (2021), p. 115033. ISSN: 0957–4174. DOI: 10.1016/j.eswa.2021.115033. URL: https://www.sciencedirect.com/science/article/pii/S0957417421004747

- Alexander Hogenboom, Bas Heerschop, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. “Multi-lingual support for lexicon-based sentiment analysis guided by semantics”. In: *Decision Support Systems* 62 (2014), pp. 43–53. ISSN: 0167-9236. DOI: 10.1016/j.dss.2014.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0167923614000645

- Mohammed Kaity and Vimala Balakrishnan. “Sentiment lexicons and non-english languages: a survey”. In: *Knowledge and Information Systems* 62.12 (2020), pp. 4445–4480. DOI: 10.1007/s10115-020-01497-6

- Tomáš Kincl, Michal Novák, and Jiří Přibil. “Improving sentiment analysis performance on morphologically rich languages: language and domain independent approach”. In: *Computer Speech & Language* 56 (2019), pp. 36–51. ISSN: 0885-2308. DOI: 10.1016/j.csl.2019.01.001. URL: https://www.sciencedirect.com/science/article/pii/S0885230816300109

- Pansy Nandwani and Rupali Verma. “A review on sentiment analysis and emotion detection from text”. In: *Social Network Analysis and Mining* 11.1 (2021), pp. 1–19. ISSN: 1869-5469. DOI: 10.1007/s13278-021-00776-6

- Gayane Shalunts, Gerhard Backfried, and Nicolas Commeignes. “The impact of machine translation on sentiment analysis”. In: *Data Analytics* 63 (2016), pp. 51–56. URL: https://biblio.ugent.be/publication/8116621/file/8132035#page=64

- Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. "How translation alters sentiment". In: *Journal of Artificial Intelligence Research* 55 (2016), pp. 95–130. DOI: 10.1613/jair.4787. URL: https://jair.org/index.php/jair/article/view/10976

3.3.4 Stage 4: Charting the Data

For selecting which resources are to what extent relevant for the final report, a similar scheme of recording information from the resources as proposed in [AO05, p. 27] was used, with, however, removing *"Intervention type, and comparator (if any); duration of the intervention"* and *"Study populations (carer group; care recipient group)"* as these two were not considered applicable for this setting. Moreover, the study location was not considered relevant either. These edits result in the following format as per [AO05, p. 27]:

- *"Author(s), year of publication [...]"*
- *"Aims of the study"*
- *"Methodology"*
- *"Outcome measures"*
- *"Important results"*

Note that we remove the first item in the following sections as we list a full bibliography entry as identifier for each of the resources.

Balamurali A.R., Mitesh M. Khapra, and Pushpak Bhattacharyya. "Lost in translation: viability of machine translation for cross language sentiment analysis". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer, 2013, pp. 38–49. ISBN: 978-3642372568. DOI: 10.1007/978-3-642-37256-8_4

- **Aims of the study:** AR et al. [AKB13] aim to show that using Machine Translation systems for Cross Language Sentiment Analysis does not perform better than a system trained using only *"a few polarity annotated documents"* [AKB13, p. 38] for a specific language.
- **Methodology:** They used English, French, German, and Russian within their testing and downloaded movie reviews randomly from https://imdb.com. They only used reviews with two or lower or eight or higher (out of ten) as rating. For Russian, they had based on the data to fall back to book reviews.
- **Outcome measures:** Accuracy of detection (*"sentiment classification accuracy"* [AKB13, p. 39])

- **Important results:** *"In-language sentiment analysis clearly outperforms cross language sentiment analysis"* [AKB13, p. 45]

Matheus Araujo, Julio Reis, Adriano Pereira, and Fabricio Benevenuto. "An evaluation of machine translation for multilingual sentence-level sentiment analysis". In: *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. SAC '16. Pisa, Italy: Association for Computing Machinery, 2016, pp. 1140–1145. ISBN: 978-1450337397. DOI: 10.1145/2851613.2851817

- **Aims of the study:** Araujo et al. [Ara+16] want to do an evaluation of existing techniques to conduct language specific sentiment analyisis.
- **Methodology:** They did a quantitative analysis of a variety of multi-language approaches which they base on nine different datasets.
- **Outcome measures:** Harmonic mean between precision and recall, as well as coverage.
- **Important results:** Based on their results, translating the input might work better than using a language specific solution.

Matheus Araújo, Adriano Pereira, and Fabrício Benevenuto. "A comparative study of machine translation for multilingual sentence-level sentiment analysis". In: *Information Sciences* 512 (2020), pp. 1078–1102. ISSN: 0020-0255. DOI: 10.1016/j.ins.2019.10.031. URL: https://www.sciencedirect.com/science/article/pii/S0020025519309879

This work builds on [Ara+16] and is enriched by more datasets (14 instead of 9). Because of said work backs up the prior findings, it has been decided to spare the summary on this occasion.

Alexandra Balahur and Marco Turchi. "Multilingual sentiment analysis using machine translation?" In: *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Jeju, Korea: Association for Computational Linguistics, July 2012, pp. 52–60. URL: https://aclanthology.org/W12-3709

- **Aims of the study:** Examining how sentiment analysis performs when being provided with machine translated input.

- **Methodology:** Using Bing, Google, and Moses to perform the (statistical) machine translation (French, German, and Spanish as languages) part and then comparing the results.
- **Outcome measures:** *"classifying sentiment from the translated data"* [BT12, p. 54]
- **Important results:** If the translation has low quality, the performance decreases. Moreover, Statistical Machine Translation (SMT) systems are *"mature enough"* [BT12, p. 58] to generate training data for foreign languages. In the worst case scenario, a performance decrease of eight percent at max was found.

Alexandra Balahur and Marco Turchi. "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis". In: *Computer Speech & Language* 28.1 (2014), pp. 56–75. ISSN: 0885-2308. DOI: 10.1016/j.csl.2013.03.004. URL: https://www.sciencedirect.com/science/article/pii/S088523081300020X

- **Aims of the study:** Analyzing how machine translation and supervised methods for building models for sentiment analysis for languages with less resources.
- **Methodology:** comparing results with supervised learning using four different features.
- **Outcome measures:** sentiment classification performance
- **Important results:** gap in classification is minimal, max. 12 percent less performance compared to "native" training sets, *"incorrect translations imply an increment of the features, sparseness and more difficulties in identifying a hyperplane which separates the positive and negative examples in the training phase."* [BT14, p. 64]

Alexandra Balahur and Marco Turchi. "Improving sentiment analysis in Twitter using multilingual machine translated data". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, Sept. 2013, pp. 49–55. URL: https://aclanthology.org/R13-1007

Alexandra Balahur and José M. Perea-Ortega. "Sentiment analysis system adaptation for multilingual processing: the case of tweets". In: *Information Processing & Management* 51.4 (2015), pp. 547–556. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2014.10.004. URL: https://www.sciencedirect.com/science/article/pii/S0306457314000934

As this work focuses on the specific input format "tweet" and on the Spanish language, it got decided to not further elaborate on it as it would exceed the targeted scope of this thesis.

Valentin Barrière and Alexandra Balahur. "Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation". In: *CoRR* abs/2010.03486 (2020). arXiv: 2010.03486. URL: https://arxiv.org/abs/2010.03486

This work focuses on tweets too and was not considered to be an enhancement for this analysis after skim reading the paper. It still is listed here for completeness as it might be relevant for future work elaborating further on this topic.

Thomas Body, Xiaohui Tao, Yuefeng Li, Lin Li, and Ning Zhong. "Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models". In: *Expert Systems with Applications* 178 (2021), p. 115033. ISSN: 0957–4174. DOI: 10.1016/j.eswa.2021.115033. URL: https://www.sciencedirect.com/science/article/pii/S0957417421004747

This work deals with a specific technique (back-and-forth-translation for generating more robust datasets). As translation in this context is used as a tool for improving the quality of a dataset in the target language, it is only mentioned in a short notice, but not elaborated further on as it does exceed the scope of the research question. However, it might still be a valuable reference for future work. Hence, it remains listed here.

Alexander Hogenboom, Bas Heerschop, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. "Multilingual support for lexicon-based sentiment analysis guided by semantics". In: *Decision Support Systems* 62 (2014), pp. 43–53. ISSN: 0167-9236. DOI: 10.1016/j.dss.2014.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0167923614000645

- **Aims of the study:** comparing performance of different sentiment analysis methods (machine translation vs. mapping of sentiment scores), lexicon-based context
- **Methodology:** translating text from a target language to a reference language (baseline performance), mapping of sentiment scores from the reference language lexicon to a new one in the target language
- **Outcome measures:** accuracy of sentiment classification
- **Important results:** semantic relations within as well as between languages have to be considered carefully, as simply translating the text from a target language to a reference language might not be performant enough

Mohammed Kaity and Vimala Balakrishnan. "Sentiment lexicons and non-english languages: a survey". In: *Knowledge and Information Systems* 62.12 (2020), pp. 4445–4480. DOI: 10.1007/s10115-020-01497-6

- **Aims of the study:** analyze available resources regarding building sentiment lexicons for languages other than English, reviewing existing tools
- **Methodology:** literature review
- **Outcome measures:** qualitative
- **Important results:** Most researchers use a variety of translation techniques to build lexicons for languages other than English, transferring learning, graph-based approaches, and merge-based approaches are also utilized.

Tomáš Kincl, Michal Novák, and Jiří Přibil. "Improving sentiment analysis performance on morphologically rich languages: language and domain independent approach". In: *Computer Speech & Language* 56 (2019), pp. 36–51. ISSN: 0885-2308. DOI: 10.1016/j.csl.2019.01.001. URL: https://www.sciencedirect.com/science/article/pii/S0885230816300109

- **Aims of the study:** Investigating the possibilities of language- as well as domain-independent techniques for sentiment analysis
- **Methodology:** improving *"the classifiers performance by utilizing surrounding context."* [KNP19, p. 36]
- **Outcome measures:** classification accuracy
- **Important results:** using character n-gram based models performs well for multi-language and multi-domain datasets

Pansy Nandwani and Rupali Verma. "A review on sentiment analysis and emotion detection from text". In: *Social Network Analysis and Mining* 11.1 (2021), pp. 1–19. ISSN: 1869-5469. DOI: 10.1007/s13278-021-00776-6

This paper does not feature a performance comparison. However, it provides a general overview regarding the topic of emotion detection and also differentiates between sentiment analysis and emotion detection. Therefore, it was used to enhance the related work section instead of further discussing it here (as it does not exactly fit the purpose of this scoping study).

Gayane Shalunts, Gerhard Backfried, and Nicolas Commeignes. "The impact of machine translation on sentiment analysis". In: *Data Analytics* 63 (2016), pp. 51–56. URL: https://biblio.ugent.be/publication/8116621/file/8132035#page=64

- **Aims of the study:** Find out how automatic machine translation impacts sentiment analysis
- **Methodology:** Using two state-of-the art technologies (SentiSAIL, SDL Language Weaver) for assessment
- **Outcome measures:** *"average inter-annotator agreement rate to the average SentiSAIL-annotator agreement rate"* [SBC16, p. 54]
- **Important results:** Performance *"rates [. . .] on the original and translated corpora are comparable"* [SBC16, p. 51], worst-case performance decrease within 5 percent

---

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. "How translation alters sentiment". In: *Journal of Artificial Intelligence Research* 55 (2016), pp. 95–130. DOI: 10.1613/jair.4787. URL: https://jair.org/index.php/jair/article/view/10976

- **Aims of the study:** Systematic analysis of different methods [1) translating text, 2) translating resources for analysis, e.g., lexicon] of sentiment analysis for non-English language
- **Methodology:** Analysis of the above described methods, usage of social media posts as input
- **Outcome measures:** accuracy of sentiment analysis
- **Important results:** translation to English leads to competitive results, automatically translated sentiment lexicons (English) enrich Arabic sentiment analysis tools

As the additionally emerged resources, besides [Tym+19] and [Den08] did not feature a performance comparison but were topically interesting, they were also used to enhance the Related Work Section (2).

Karsten Tymann, Matthias Lutz, Patrick Palsbröker, and Carsten Gips. "Gervader-a german adaptation of the vader sentiment analysis tool for social media texts". In: *LWDA*. 2019, pp. 178–189

- **Aims of the study:** Applying the creation process of a sentiment analysis tool for English (i.e., VADER) to create a German-language version (GerVADER)
- **Methodology:** adapting a reference process
- **Outcome measures:** performance criteria
- **Important results:** adaptation was successful, performance optimizations may be used for enhancements

---

Kerstin Denecke. "Using sentiwordnet for multilingual sentiment analysis". In: *2008 IEEE 24th International Conference on Data Engineering Workshop*. 2008, pp. 507–512. DOI: 10.1109/ICDEW.2008.4498370

- **Aims of the study:** Introducing a methodology for determining text polarity (multilingual setting)
- **Methodology:** lexical resources for English (SentiWordNet), translating document from non-English to English
- **Outcome measures:** accuracy in polarity classification
- **Important results:** SentiWordNet delivers decent results in a multilingual setting

3.3.5 Stage 5: Collating, Summarizing & Reporting the Results

Based on the described process, the following findings have been collated.

Moreover, based on [LCO10], the report consists of themes and a table of strengths and gaps in the findings, before answering the research question within the scope of this thesis.

3.4 Report of the Results

The following report at first lists the results which emerged from the conducted study. Afterwards it compares, based on said results, the strengths and gaps in the collated literature and closes with a conclusion.

Note that the findings mostly deal with sentiment analysis and not the more sophisticated concepts of emotion detection (as compared in Section 2, referencing [NV21]). This came due to the fact that the above referenced first search term provided no relevant findings, and the second term only uncovered [NV21] and a non-accessible source ([Bus+21]). However, when introducing the term *sentiment analysis* into the search field, useful results were found.

Themes

Performance, mostly in form of accuracy, is a theme every found resource tackles. Despite, accuracy, e.g., [Ara+16] also tackles coverage (messages which can be classified as either positive or negative as a percentage of the whole dataset) as a measure.

Different languages are also part of (nearly) every resource which was found within the process (namely Arabic, English, French, German, Russian, and Spanish).

Approaches to sentiment analysis are also discussed. [MSK16] differentiates between either translating the input (i.e., text to analyze) or the resources used (e.g., the lexicon with annotations).

Different translation technologies (e.g., Bing, Google) are discussed in [BT12] or [SBC16] (SDL Language Weaver). However, *different technologies for sentiment analysis* are also discussed, e.g., in [SBC16] (SentiSAIL), [Tym+19] (VADER and GerVADER), and [Den08] (SentiWordNet).

Strengths & Gaps

The following table (Table 3.1) compares the strengths and gaps the found literature provides. Thematically related strengths and gaps are listed in the same line.

Conclusion

First of all, it has to be stated that this conclusion is specifically based on this scoping analysis, whereas a broader *discussion* is to be found in Section 6.2.

With regard to the exploratory first research question *"What are the performance differences (i.e., successful detection ratio) between a language-specific solution and using an already existing English-language solution with a translation layer added to it?"*.

It can be said that the performance differences indeed exist, however, they are rather small in many cases: 5 percent [SBC16], 7 percent [BT12], 12 percent [BT13] (worst case each). Based on the findings, using a translation layer might be a feasible solution compared to implementing a language-specific solution in many cases. However, e.g., [AKB13] rather argued towards a language-specfic solution.

Strengths	Gaps
many resources successfully applied translation in sentiment analysis	sentiment analysis is less sophisticated than emotion detection, therefore no direct conclusions may be drawn
all use sentiment analysis based on text-input, therefore comparable	as speech detection is considered within the resources (although it would be necessary for a spoken interaction with a robot), the error rate could very well increase because speech detection is neither error free nor noise free (e.g., filling words like "erm" might very well influence sentiment, as they could indicate confusion, stress, etc. as negative examples, or screams of joy as a positive example)
different translator software was used which loosens the connection to a specific translator technology	the language space consisted of English, French, German, Russian, and Spanish mainly (Arabic as an exception) and is therefore not easily transferable to other languages with maybe less sophisticated translation technologies.

Table 3.1: Strengths & Gaps of the Collated Literature

CHAPTER 4

Expert Interviews

As part of this thesis, two expert interviews have been conducted. The goal behind this addition was to get initially a comprehensive overview of the research landscape within the field of *social robots* and *emotion detection*, as well as open questions surrounding said topics.

For that, two experts on the mentioned topics have been interviewed, namely Prof. Dr. Kerstin Fischer and Prof. Dr. Barbara Kühnlenz.

Kerstin Fischer currently is *Professor WSR* at the Department of *Design og Kommunikation*, as well as *Associate Professor* at the research areas *Communication* and *Knowledge, Learning and IT* at the University of Southern Denmark [Soua]. Kerstin Fischer's research interests include but are not limited to the fields of *Robots*, *Human-Robot Interaction*, *Social Robots*, *Robotics*, and *Grammer* [Soub] and are relevant for this thesis as can be seen in the Bibliography.

Barbara Kühnlenz currently is *Professor of Business Psychology with focus on Technical Innovation* at the Coburg University of applied sciences and arts having *Human-Machine Interaction*, *Acceptance and User Experience*, *Interactive and Social Robotics*, and *Persuasive Technology* as relevant research fields for this thesis [Uni] (see Bibliography for cited work).

The two experts have been chosen as they are very suitable for the topic of this thesis with their work and said work came up many times in the literature research process. The contact has been established by Astrid Weiss.

4.1 Preparation

The interviews have been prepared by first developing an interview guide which contained the following questions for both interviewees:

1. To what extent did emotion detection play a role in your research activities until now? (original: "Inwiefern hat Emotionserkennung in Ihrer bisherigen Forschungstätigkeit eine Rolle gespielt?") – estimated time for this question: 4 to 6 minutes.

2. What is the most important output you take away from your past work with robots and especially in the communication between human and robot? (original: "Was ist der wichtigste Output, den Sie aus Ihrer Arbeit mit Robotern und speziell in der Kommunikation zwischen Mensch und Roboter aus Ihrer bisherigen Tätigkeit mitnehmen?") – estimated time for this question: 3 to 5 minutes.

3. As follow-up: Which advice can you provide within this context in order to avoid beginner's mistakes. Especially: What are, in your opinion, the open research questions which are worth looking into in this context? (original: "In weiterer Folge: Welche Tipps können Sie in diesem Zusammenhang geben, um klassische Anfänger:innen-Fehler zu vermeiden? Insbesondere: Was sind, Ihrer Meinung nach, die offenen Forschungsfragen, die in diesem Zusammenhang untersuchenswert sind?") – estimated time for this question: 2 to 4 minutes.

4. What are, following your experience, the biggest obstacles in creating a satisfying communication flow between human and robot? To what extent do emotions play a role in this regard? (original: "Was sind Ihrer Erfahrung zu Folge die größten Hürden, einen zufriedenstellenden Kommunikationsfluss zwischen Mensch und Roboter zu ermöglichen? Inwiefern spielen Emotionen hierbei eine Rolle?") – estimated time for this question: 3 to 5 minutes.

As per request, the guide was sent to Barbara Kühnlenz in advance (see Appendix; the introduction does also apply to the interview with Kerstin Fischer). Kerstin Fischer did not request it and therefore did not receive it in advance.

Additionally, both experts got specific questions targeted at their respective research fields:

- Kerstin Fischer:
    1. You looked into different languages and dialects and focused on to what extent speakers are perceived differently when their language melody gets transferred to other languages. To what extent can this play a decisive role in emotion detection too? Is it even possible to create a "general" emotion detection or does it seem more likely that each language or even each dialect require manual and detailed adaptation thereof? (original: "Sie haben sich mit verschiedenen Sprachen bzw. Dialekten auseinandergesetzt[1] und sich dabei darauf fokussiert, inwiefern die Sprecher:innen anders wahrgenommen werden,

[1]https://portal.findresearcher.sdu.dk/en/publications/studying-language-attitudes-using-robots, last accessed (post interview): Nov. 27, 2021

wenn ihre Sprachmelodie auf andere Sprachen übertragen wird. Inwiefern kann das auch in der Emotionserkennung eine entscheidende Rolle spielen? Ist es überhaupt möglich, eine "generelle" Emotionserkennung anzufertigen oder scheint es wahrscheinlicher, dass diese für jede Sprache bzw. gar jeden Dialekt manuell und detailliert angepasst werden muss?") – estimated time for this question: 7 to 14 minutes.

2. Where are, in your opinion, the "structural" differences I should pay attention to or use (i.e. how the language is built) between English and German relevant for my thesis? (original: "Wo liegen Ihrer Meinung nach die für meine Arbeit relevanten "strukturellen" Unterschiede (d.h. wie die Sprache aufgebaut ist) zwischen Englisch und Deutsch, die ich beachten sollte bzw. mir zunutze machen könnte?) – estimated time for this question: 2 to 4 minutes.

- Barbara Kühnlenz:

  1. You already in the past looked into the question to what extent the perceived "helpfulness" of robots increases when they adapt to the emotions of their human counterpart. Following that, it seems reasonable to assume that emotion detection – when reliable – positively influences "helpfulness" too. For which other parameters besides "helpfulness" can emotion detection be of relevance too and, especially, to what extent are negative effects imaginable if emotions are detected wrongly/not exactly and, as a result, the "mood adaptation" fails? (original: "Sie haben sich in der Vergangenheit bereits damit auseinandergesetzt, inwiefern die wahrgenommene "helpfulness" von Robotern dadurch erhöht wird, dass sie sich an die Emotionen des menschlichen Gegenübers anpassen[2]. Dementsprechend liegt es nahe, dass sich die automatisierte Emotionserkennung, sofern zuverlässig, ebenfalls positiv auf die "helpfulness" auswirkt. Für welche anderen Parameter neben "helpfulness" kann sie aber noch von Bedeutung sein und vor allem, inwiefern sind negative Effekte denkbar, wenn Emotionen falsch/ungenau erkannt werden und dementsprechend auch die "Mood-Adaption" fehlschlägt?") – estimated time for this question: 5 to 12 minutes.

Before conducting the interviews, the participants received a consent form (to be found in the Appendix) via e-mail which they were asked to fill in and sign.

4.2 Conduct

The interviews have been conducted in German via the video conference software *Zoom* and got recorded and stored on the author's local drive. They were not uploaded to a cloud service of any kind and were not shared with anyone else either.

[2]https://mediatum.ub.tum.de/doc/1216553/file.pdf, last accessed (post interview): Nov. 27, 2021

4.3 Analysis

The interviews have been analyzed via a *thematic analysis*, building on the work of Braun and Clarke [BC06].

Thematic analysis provides a variety of positive attributes, as outlined in [BC06, p. 97]:

- flexible
- quick and easy compared to other methods
- enables social interpretation of data
- comprehensive description of data
- unanticipated insights may occur
- ability to highlight similarities as well as differences

In the specific context of this thesis, it has been chosen because it offered the author the possibility to engage in-depth with the gathered data and also because a too rigid interpretation and data analysis was not deemed necessary as the interviews themselves were not intended to generate new insights, but rather to structure the research landscape and engage with it for this thesis.

Braun and Clarke outline different *phases* of a thematic analysis [BC06, p. 87] (phase names rephrased in nominalized form for more conciseness, descriptions condensed focusing on relevant aspects):

	Phase	Description
1.	Data Familiarization	Transcription of the data, reading said transcript multiple times, as well as noting initially arising ideas
2.	Initial Coding	Systematically creating codes for interesting data features, structure data by code
3.	Theme Search	Find corresponding themes for different sets of codes
4.	Theme Review	Check whether themes fit a) coded extracts (Level 1) and b) whole data set (Level 2), generation of a “thematic map”
5.	Theme Definition & Theme Naming	Refining aspects of themes as well as the “overall story” of the analysis; creating names and definitions for the found themes
6.	Report Production	Selecting examples, relating back to initial questions, producing “scholary report”

Table 4.1: Phases of a Thematic Analysis [BC06, p. 87]

The in Table 4.1 listed phases were considered and followed within the process of analyzing the two expert interviews.

4.3.1 Schedule

It has to be stated that Phases 1 to 5 have been carried out consecutively for each interview. Therefore the schedule was (the smaller the number the earlier in the process):

1. Phase 1 to 5 – K. Fischer
2. Phase 1 to 5 – B. Kühnlenz
3. Phase 6 – K. Fischer & B. Kühnlenz at the same time

This order has been chosen not to mix up the statements of the separate interviews. However, in the following paragraphs the results are listed per phases. That means that Phase n is in the same section for both interviews. This decision has been made to give the reader a better comparability.

4.3.2 Phase 1: Data Familiarization

Both interviews were transcribed. For the transcription, the recorded video was played at 0.5 times the original speed. It was paused when necessary and passages were watched repeatedly if something was not clearly understood. This procedure ensured that the transcriptions stayed rather close to the original wording. However, filling words as, e.g., “erm” or similar were removed. Moreover, sentences have in some cases been slightly restructured in order to ensure a more coherent reading flow. Nevertheless, there was a specific focus on not changing the content and not leaving out any details. The transcriptions were written in gender-sensitive language even if the interview or parts of it were not.

After transcribing, the transcripts were read one time to ensure a relatively “obstacle-free” reading flow and small changes were done if necessary.

During a third read, initial thoughts have been noted. Also, some potential code words have already been added in an unstructured form.

4.3.3 Phase 2: Initial Coding

Phase 2 was done iteratively. During the first iteration, the initially found potential code words from Phase 1 got approved or declined. In a second iteration, the approved words were refined to have consistent wording for the same concepts.

The following code words were chosen (alphabetically listed per interview, crossed out words show evolution in second iteration):

Interview with K. Fischer: Arousal Level (of People) • Collaborative Work (Study re: Creativity) • Computer Voice • Dialects & Languages (also to be found as “Languages & Dialects”) • Differences between English & German (e.g., different words) • Emotion Agreement • Emotion Detection • ~~Emotion Modeling~~ ⇒ Emotion Models • Emotion Models • Emotion Production • ~~Emotion Recognition~~ ⇒ Emotion Detection • Emotional

Expression (of Robot) • Emotional Mismatch • Excuses (Differences between Danish & German) • Expectation (re: Emotional Expression) • Eye-Tracking • Filler Words (add.: not per se "imperfect" as they have function) • Foreign Language • Hints & Tricks (for Work in this Field) • Intonation • Language Signals (original: "Sprachsignale") • Languages & Dialects (also to be found as "Dialects & Languages") • Modification of written Language to be more speech-like (e.g., Google Duplex) • Pitch (original: "Tonhöhe"; Example Difference Danish & German) • Prosody (re: Linguistics) • Responsiveness (of Robot) • Robot (Definition): What is a Robot? – Social Agent or Machine? • Robot Language • Societal Conventions • Spoken Language vs. Written Language (also to be found as "Written Language vs. Spoken Language") • Stiff-Upper-Lip • User Satisfaction • Voice Range (original: "Stimmlage") • Written Language vs. Spoken Language (also to be found as "Spoken Language vs. Written Language")

Interview with B. Kühnlenz: Adaptation Effect • Adaptation of the emotional Expression • Animacy • Anthropomorphism • back up detected Emotion in a Dialogue • behavioristic Measure of Helpfulness • Benefit vs. Effort (e.g., Programming Effort; also listed as "Effort vs. Benefit") • Borders (of emotional Interaction) • Bridge(-Function of the Robot) • Chances vs. Dangers (also listed as "Dangers vs. Chances") • Choice of Words • cognitive (strictly contentwise) Goals • cognitive Goals • cognitive Task Context • collaborative Setting • Computing Resources needed (for Emotion Detection) • Confounding Variables (e.g., Background Noise) • Contexts • Countenance (original: "Mimik") • Course of Time • Dangers vs. Chances (also listed as "Chances vs. Dangers") • defective Emotion Detection • ~~defective Speech Recognition~~ ⇒ defective & deficient Speech Recognition • ~~deficient Speech Recognition~~ ⇒ defective & deficient Speech Recognition • Delivery & Pick-Up Services (also listed as "Pickup & Delivery Services") • Design too anthropomorph • Detection Performance • Dissolution of Boundaries between Human and Robot (Danger) • Effects • Effort vs. Benefit (e.g., Programming Effort; also listed as "Benefit vs. Effort") • Emotion • emotional Adaptation • Empathy-Altruism-Hypothesis (original: "Empathie-Altruismus-Hypothese") • Emphasis (original: "Betonung") • Enquiring (re: Misunderstanding) • Error Tolerance • Errors (vs. Success) • Expectations • Face Recognition • Facial-Action-Units • False Positive (original: "Falscherkennung") • Frequencies • Goals • Habituation Effect • Heerink Dimensions (e.g., Trust, Sociability, social Presence, Enjoyment) • Helpfulness • Hints & Tricks • Human and Robot automatically adapt to each other • Human simply is an emotional Being • Industry 4.0 • Influence (of Human-Robot Interaction) on Individual & Society • incongruent Task Context • Lab- vs. Real-World-Settings (also listed as "Real-World- vs. Lab-Settings") • Language • Measures – objective vs. subjective • mental Strain • Methods • Miscommunication Handling • Models • Naturalness • Novum of an Experiment (as added Value) • One cannot not communicate (Watzlawick) • Over-Usage • Parameters (of UX etc.) • Patterns • physiological Measures (Skin Conductance Level, Heart Rate Variability) • Pick-Up & Delivery Services (also listed as "Delivery & Pick-Up Services") • Pondering (Trade-Offs) • pro-social Effects • pro-social Sense (use Emotions) • Prosody • Real-World- vs. Lab-Settings (also listed as "Lab- vs. Real-World-Settings") • reduced Context • Resemblance Dimension (original: "Ähnlichkeitsdimension") •

semantic Emotion Detection • Sentiment Analysis • situative Empathy • social Design • social Dimension • Social effects do not always lead to desired Results • social Goal • social Interaction • social-psychological Approach • Speech Recognition • subjective Questionnaires • Support Dialogue • task-based Human-Robot Interaction • Uncanny-Valley-Effect • User Experience (Dimensions) • Users' Assessment • Users' Background • verbal social Interaction • What can a Robot do? • What is minimally necessary for the desired Effects? • "Züricher Lächelmodell" (not found on the Internet)

4.3.4 Phase 3: Theme Search

Building on the codes, themes were created which consist of different codes (initial map) and connections for the interview with Kerstin Fischer between the themes (refined map) and codes were noted in two "maps" for Phase 3. For the interview with Barbara Kühnlenz, instead of connections, more refined meta terms have been chosen to further order the elements as the amount of data was considered too much to be ordered by connecting items (however, later on they were structured based on the meta themes, more below). The maps of this phase evolve into the "thematic map" of Phase 4. See below.

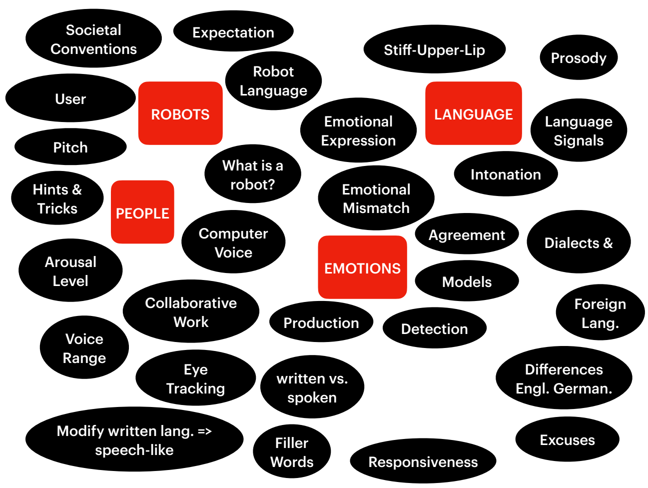


Figure 4.1: Initial Map of Phase 3 (K. Fischer)

As we can see, the map of the interview with Kerstin Fischer has a far more concise list of topics and terms, whereas the initial sorting process of the interview with Barbara Kühnlenz was way more intense as there were more different terms to be categorized. The initial categorization, however, did not deem sufficient at all. For the interview with Barbara Kühnlenz, the red elements got initially selected as meta themes and the black elements as sub themes. However, this changed over the course of the analysis (see below).

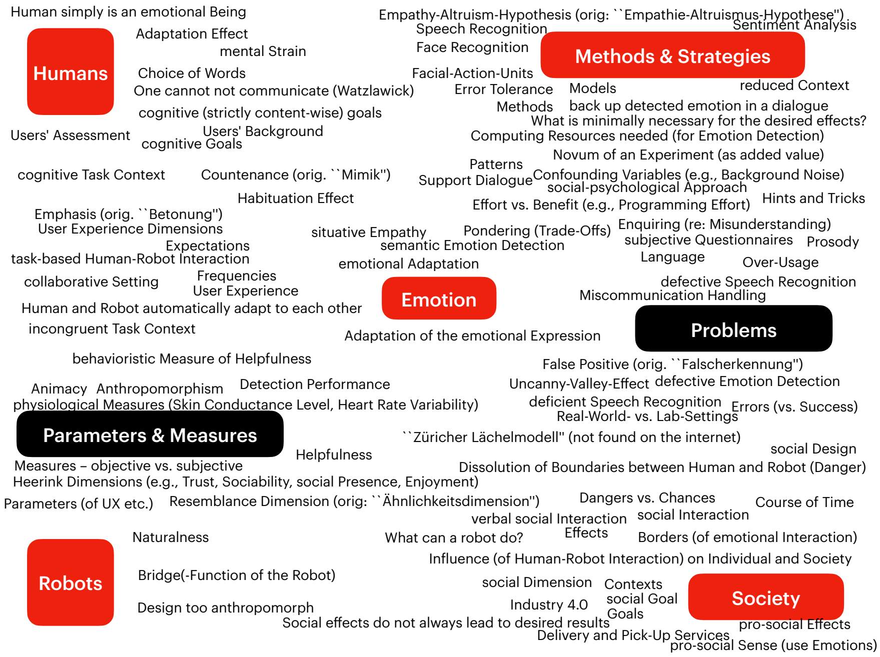


Figure 4.2: Initial Map of Phase 3 (B. Kühnlenz)

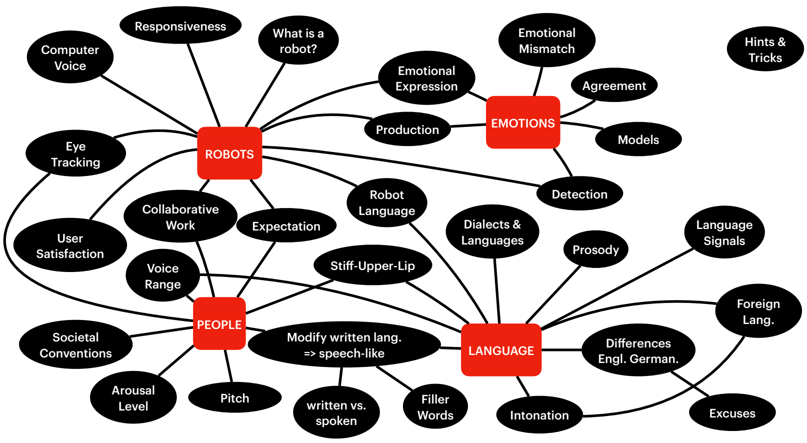


Figure 4.3: Refined Map with Connections of Phase 3 (K. Fischer)

4.3.5 Phase 4: Theme Review

The found themes were reviewed and, if necessary, further divided or put together to one bigger theme. The graphic below shall illustrate this evolution as it shows the resulting "thematic map".

As we can see here, another theme "methods" got added and, for better visibility, the new resulting connections have been colored blue for the interview with Kerstin Fischer.

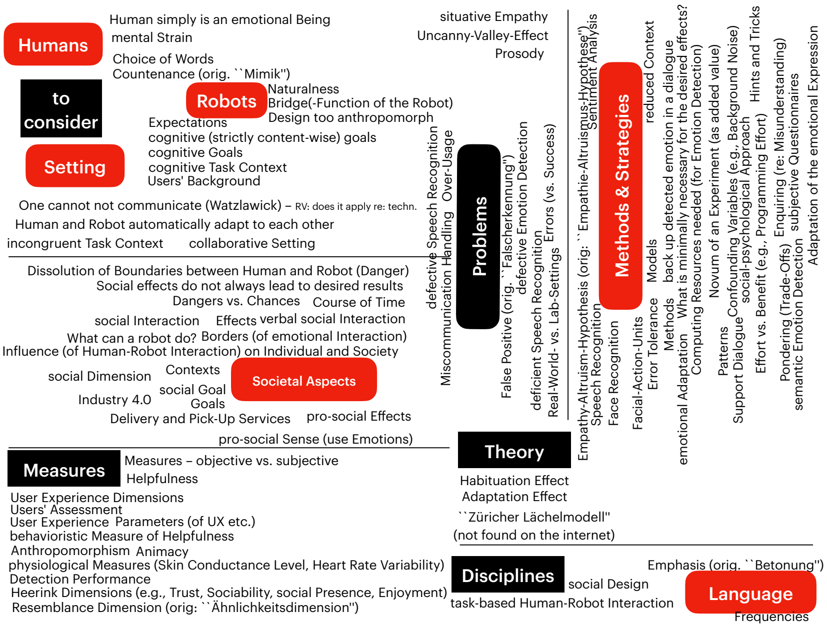


Figure 4.4: Refined Map with Stricter Categorization of Phase 3 (B. Kühnlenz)

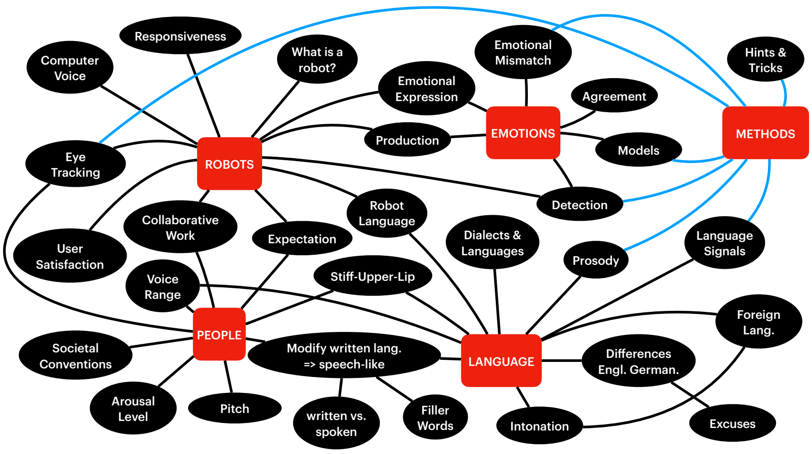


Figure 4.5: Initial Map of Phase 4 (K. Fischer)

For the interview with Barbara Kühnlenz, the themes got refined as "Disciplines", "Language" and "Theory" got removed and included into the other themes.

Building on the created "thematic maps", refinements have been made which lead to the following final maps. For the interview with Barbara Kühnlenz it has to be said that the red elements symbolize practical aspects, whereas the black elements symbolize theoretical aspects. Both of them are on "equal" thematic layers.

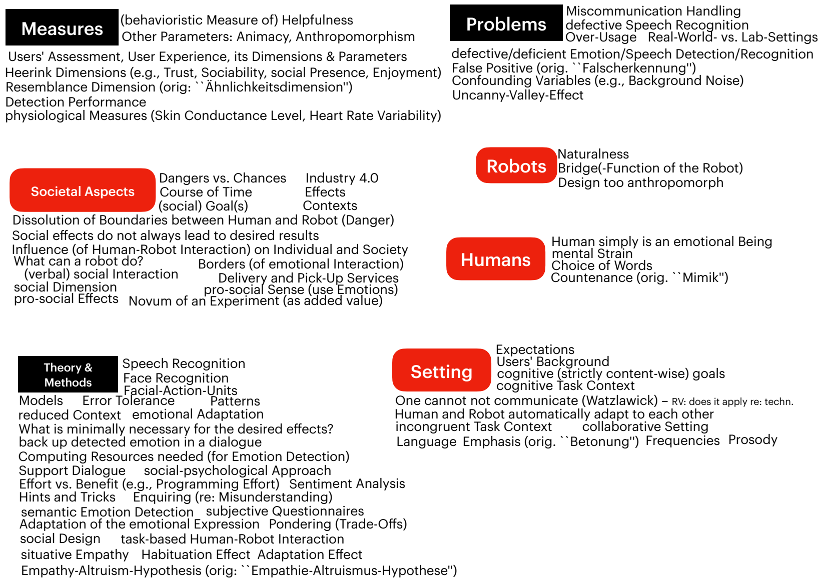


Figure 4.6: Initial Map of Phase 4 (B. Kühnlenz)

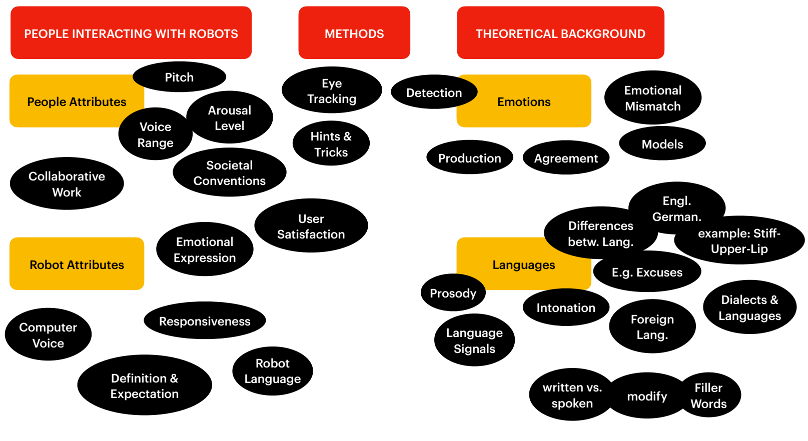


Figure 4.7: Refined Map of Phase 4 (K. Fischer)

We can see that – compared to the initial map of Phase 4 – the themes have evolved into sub-themes. To be precise: We now have 3 (meta-)themes which are "People interacting with Robots", "Methods", and "Theoretical Background". "Methods" remained a (meta-)theme whereas the other themes got changed into sub-themes and rather focus on the "attributes" now (people + robots) as this is more precise. "Emotions" and "Language" together joined the (meta-)theme "Theoretical Background" for the interview with Kerstin Fischer.

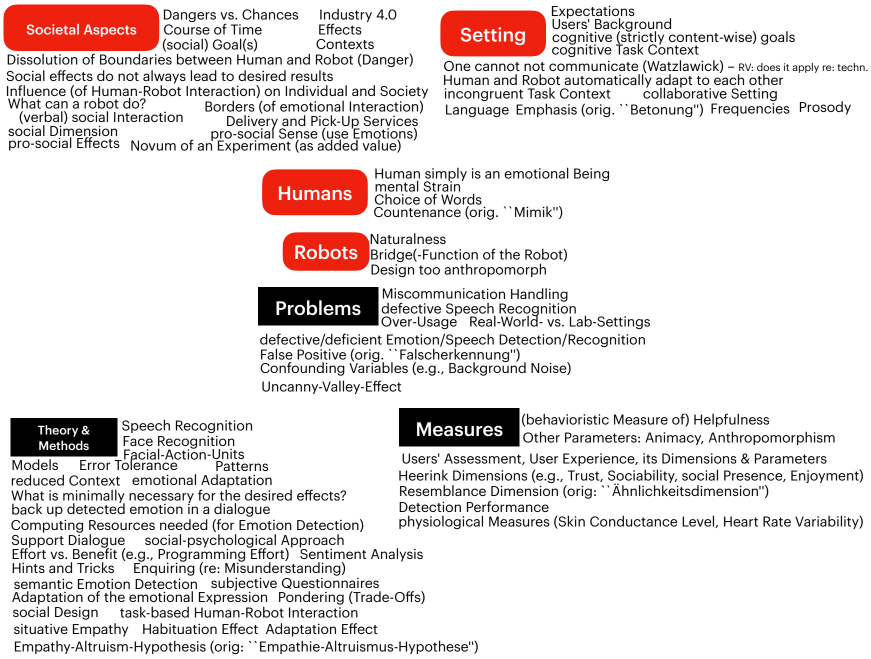


Figure 4.8: Refined Map of Phase 4 (B. Kühnlenz)

Moreover, the aspects (black ellipses) of different themes got refined as (for better visibility) the connections have been removed and instead aspects fitting together are now symbolized via the shorter distance between them for the interview with Kerstin Fischer.

For the interview with Barbara Kühnlenz, the topics got further structured as their position changed. On top the "Societal Aspects" and "Setting" are positioned to symbolize that each interaction is influenced by societal aspects and also the general evolvement of the topic within society is influenced by different interactions (e.g., people being happy or unhappy with said technology). Each interaction includes (a) human(s) and (a) robot(s) and can, based on all of the mentioned themes, have different problems. User satisfaction, etc. can be evaluated by different "Measures" and "Theory & Methods" back our work in those fields up.

4.3.6 Phase 5: Theme Definition & Theme Naming

The found themes have already been defined in Phase 4 and did not change afterwards as a quite significant restructuring process already happened in said phase.

However, an order for the report has been chosen and first initial definitions have been attempted. These steps shall help to produce a concise and coherent report in the next step.

A (meta-)theme is "People interacting with Robots". For that we have aspects, which do not necessarily have to do with our specific context “Emotions” and detection of them in specific “Languages”. However, emotions and languages have aspects which might have come up within the interview which do not necessarily only have to with Human-Robot Interaction for the interview with Kerstin Fischer.

Regarding the interview with Barbara Kühnlenz, we already mentioned the structural division into the “practical” and “theoretical” aspects and that narrative got selected for this phase too.

To detect emotions and to make human-robot interaction even possible, we have to use a variety of different methods (also to measure the quality/success of said interactions) of which a few have come up within the interview with Kerstin Fischer.

These considerations lead us to the following theme definitions within the context of the interview with Kerstin Fischer:

- **People interacting with Robots:** As name for this theme the term “Human-Robot Interaction” has intentionally not been chosen as this should not be mixed up neither with the research area nor with a specific interaction. It covers the aspects of the interview where attributes of said interactions but also of “people” and “robots” separately have been discussed, e.g., the emotional expression a robot can offer or what the term “robot” means for different people. Moreover, collaborative work between those two entities has also come up.
- **Methods:** Making mentioned interactions between people and robots possible requires certain methods. In the context of the interview, emotion detection was the specific focus regarding methods. However, methods have also been discussed in a broader context as measuring the pitch of a human’s voice, for example, may deem relevant in other contexts too.
- **Theoretical Background:** Within this theme, there is a clear distinction between the sub-theme of “Emotions” and “Languages”. Within the context of the interview, because of the background of the expert, only Danish, English and German have been discussed. The theme “Emotions” focused rather on what emotions mean within an interaction than what emotions actually are.

. . . as well as to the following theme definitions within the context of the interview with Barbara Kühnlenz:

- **Practical Aspects:** By this, all aspects which are directly relevant within an interaction are meant, whereas with theoretical aspects (below) a rather non-directly-participating but analyzing view is taken.
- **Theoretical Aspects:** These include all analyzing aspects as well as theoretical background and methods. However, the problems within an interaction are bound

to the specific interaction and therefore have been changed to a practical rather than theoretical aspect within this analysis.

More important than the dual differentiation between theory and practice seem to be the themes in layer two:

- **Societal Aspects:** Each interaction happens within a societal context. How robots are perceived, what we can and want to do with them, etc. is influenced by the society we find ourselves in.
- **Setting:** Moreover, each specific interaction happens within a certain setting. Users have different expectations, a different background, and also different goals – or more simply, they might just use a different robot.
- **Humans** & **Robots** are integral parts of each interaction as they are the "participants" so to speak.
- Interactions can be compromised/interrupted by different kinds of **problems** as, e.g., defective speech recognition.
- Different theoretical aspects ("**Theory**") are relevant or valuable to further look into. Moreover, there are also different "**Methods**" for detecting emotions or analyzing the performance of a tool in this context.
- Performance, user satisfaction, etc. can be evaluated by different "**Measures**".

At this point, it has to be stated that the definition of the themes are rather broad and short (compared to what [BC06, p. 92f.] mention in their description). This can be tracked back to the fact that the interviews themselves were rather short and have been conducted during an early stage of this thesis and therefore provide a broad overview rather than create new insights or data. However, this does not change the fact that their [BC06, p. 92f.] methodology has been followed to the extent applicable in this context.

4.3.7 Phase 6: Report Production

The Final Phase produced the report presented in the next section. This has been achieved by revisiting the results of the phases prior and writing up the results in a structured way where themes are described instead of just named and connections and comparisons between the two interviews are drawn.

4.4 Report of the Results

Although the interview guide (despite the person-specific question at the end) was basically the same, the results turned out quite differently. Whereas the interview with

Kerstin Fischer was more of a discussion or conversation, the interview with Barbara Kühnlenz was more of a master-apprentice style with a higher information density. Both are considered equally valuable for the work within this thesis.

The topics were, as we can see in the maps above, similar, but with different focus points. On the one hand, the interview with Kerstin Fischer had a high focus on language, whereas this topic in contrast during the interview with Barbara Kühnlenz did not come up that much. On the other hand, the interview with Barbara Kühnlenz delivered a big amount of measures and methods regarding emotion detection.

Both interviews also discussed different roles and attributes of humans and robots in a human-robot interaction. However, the interview with Barbara Kühnlenz focused also on the societal implications and situations, whereas the interview with Kerstin Fischer rather discussed emotions as a general topic.

4.4.1 Key Focus Points

There are a few key focus points of the arisen themes which are especially interesting for this thesis as they provided the author with new insights for the subsequent research work. Those are discussed in the following paragraphs.

Languages (K. Fischer)

The topic of languages was particularly interesting for Chapter 3. Differences between languages were briefly discussed as well as the general difference between spoken and written text in any language.

First, languages vary regarding intonation and this "has a huge influence" (K. Fischer) on human-robot interactions, specifically on "how the robot is perceived" (K. Fischer). This concerns but is not limited to the way of speaking. It also influences the "emotional expression" (K. Fischer).

A key learning in this regard was also that the "intonation contours" (K. Fischer) are transferred from the mother tongue to speaking a foreign language. E.g., a German-speaking person might (subconsciously) transfer the German-specific "intonation contours" to their English. Therefore, there are not only lexical and grammatical differences between languages but also practical differences in intonation. This might therefore be something which has to be considered in emotion detection.

An example regarding intonation differences was also given. "Hvor gammel er du?" (Danish for "How old are you?") is pronounced "flat" (K. Fischer), whereas the same phrase in German "Wie alt bist du?" is usually combined with providing a higher pitch while speaking ("One goes up with their voice" – K. Fischer). Therefore, we can see, that the same phrase in different languages might have completely different intonations. Adding dialects and non-mother-tongue intonations to this equation, we see that the complexity of modeling it within a technological system might increase.

Another example for the differences between English and German in that case is that as per Kerstin Fischer the "prosodic range" is bigger in English (e.g., for expressing pleasure one might say "Oh, dear, that is lovely" which has a more intense intonation than in German. As an example for intercultural differences between English and German the English "Stiff upper lip" (e.g., discussed in [CC13]) has been mentioned by Kerstin Fischer.

Reflecting on the focus point of text-based emotion detection in this thesis, we can state two thoughts: 1) intonation is not relevant as it is not encoded within written text (at least not by default), however 2) leaving out intonation also removes a lot of information which could be very useful regarding emotions encoded within said intonation.

However, there are also semantic differences which might be relevant for text-based emotion detection. An example from the interview is that in German one might say "Tut mir leid, der Bus ist jetzt weg." ("I am sorry, the bus already left") although the person saying this is not the bus driver. In contrast, in Danish one usually does not say this (K. Fischer).

Regarding the differences between written and spoken text in general, it was stated that "filling words" are deliberately encoded into a robots speaking as experiments have shown a positive impact of this practice (K. Fischer). Google Duplex [LM18] was mentioned as a side note because said technology already makes use of encoding filling words into text, Kerstin Fischer mentioned.

To conclude this focus point with, the expectation of the author that text-based emotion detection with semantically analyzing the content currently is more advanced than other techniques was contradicted. Namely, Kerstin Fischer explained emotion detection based on language signals is advanced compared to semantic ones.

Theory & Methods (B. Kühnlenz)

Another key focus point are the discussed theoretical aspects and methods from the interview with Barbara Kühnlenz.

Speech recognition is a key aspect of text-based emotion detection as it enables the option of transferring text to a processable form. Face recognition is another method for emotion detection, in this context Facial-Action units were brought up by Barbara Kühnlenz [Küh+13] as they were used in a study to adapt the emotional expression to the user's emotional state [Küh+13]. This approach is explicitly not the semantic one this thesis focuses on.

Moreover, it was also spoken about the aspect that, which Barbara Kühnlenz found in one study, successful emotion detection lets users tend to have fault tolerance with the robot (B. Kühnlenz).

Two aspects which are also considered for the study design in Chapter 5 were brought up in the interview: 1) What is minimally necessary for the desired effects (Effort vs.

Benefit)? 2) Back up detected emotion in dialogue (also to avoid misunderstandings). The first aspect is important in order to use resources (development, computing resources, etc.) where they are necessary because if emotion detection (if relevant) by explicitly asking the person interacting with a robot is equally satisfying for them as implicitly detected ones, the resources needed for such a complex project might be better used elsewhere. For deciding which method for coping with emotions is used, the task context as per Barbara Kühnlenz is important and has to be looked into, especially thinking about the findings that a "reduced context" with just enquiring about emotions was sufficient in some cases.

Therefore, the conducted study (Chapter 5) has a scenario where the person interacting with the robot is explicitly asked about their emotions.

Furthermore, Barbara Kühnlenz explained that emotional adaptation, e.g., give back a smile in the right moment, increases the situative empathy and also provides better values in a variety of user experience dimensions.

Taking the approach of emotional adaptation further, as Barbara Kühnlenz explained, based on a social psychological approach, humans tend to provide more helpfulness towards another person if the person in need of help has similarity to the person on the verge of helping ("empathy-altruism hypothesis", e.g., in [BLS15]). This hypothesis was transferred to the human-robot interaction, providing the author with the reference to the paper [Küh+13] as also discussed in Chapter 2.

Other aspects to consider, mentioned in said interview, were "Habituation Effect(s)" and "Adaptation Effect(s)". More precisely, how any form of habituation affects the interactions after a (longer) period. Moreover, the question where adaptation is desired and where not also has to be answered – and arguably before designing an artifact. This all comes down to the "setting" a human-robot interaction takes place in (see 4.4.2).

A fitting addition to these paragraphs from the interview with Kerstin Fischer is that eye-tracking has also been mentioned as method.

Emotions (Theoretical Background, K. Fischer)

On a short notice, the topic of emotions was also discussed briefly. Important keywords in this aspect where the "production" of emotions, "agreement" on emotions, "models" of emotions (discussed in more detail in Chapter 2) and "emotional mismatch" which leads the author to deciding on using wrongly detected emotions as a scenario within Chapter 5.

Measures (B. Kühnlenz)

Regarding measuring user satisfaction, the interview with Barbara Kühnlenz played a key role in deciding on the study design (Chapter 5) where the mentioned paper [Küh+13] inspired to use the Godspeed Questionnaire Series [Bar+09] to measure user satisfaction.

Moreover, other parameters than helpfulness (also in a behavioristic sense), animacy, anthropomorphism and the "Resemblance Dimension" as well as the "Heerink Dimensions" [Hee+09] proved to be a good starting point for further looking into and are discussed in Chapter 2.

Physiological Measures (e.g., Skin Conductance Level, Heart Rate Variability) were also discussed briefly but do not directly influence text-based emotion detection. However, the idea of combining methods to further back up what was detected with one method is an obvious one in this context.

4.4.2 Valuable Basic Aspects

Additional aspects which are not as relevant for this thesis or are more "common sense" are discussed below as they still provide value to the overall story this thesis is telling.

Problems (B. Kühnlenz)

When designing an emotion detection system, depending on the technological basis and use context, different problems might arise. Some of them have been discussed within the interview with Barbara Kühnlenz.

Speech recognition, e.g., can be disturbed because of background noise. Those discrepancies are specifically important if there are different settings in lab vs. real-world experiments. This is an advice Barbara Kühnlenz specifically provided when asked about hints and tricks – to not forget that things in labs might work differently compared to real-world experiments.

Another problem mentioned was "over usage" when asking people about their emotions, e.g., there are "critical points which could be surpassed" (B. Kühnlenz) and therefore provide a negative effect. Barbara Kühnlenz also mentioned "Miscommunication Handling" as field of interest.

Wrongly detected emotions were also briefly mentioned and are therefore explicitly part of the study (Chapter 5).

Moreover, the "Uncanny-Valley-Effect" (e.g., in [SN07]) was mentioned in context of expectations which could occur based on a social design and are then not fulfilled (B. Kühnlenz).

People Interacting with Robots (both) & Setting (B. Kühnlenz)

Although this topic has only been named in the final theme map for the interview with Kerstin Fischer, it includes the themes "humans" and "robots" from the interview with Barbara Kühnlenz as well as the themes "people attributes" and "robot attributes" from the interview with Kerstin Fischer.

Regarding the attributes, pitch (of the voice), the voice range and the arousal level were named. Note that, although mentioned in the context of human speakers, robots' voices can obviously also have different pitches and ranges.

Moreover, robots can also be used in a collaborative work setting, where different attributes (e.g., emotional adaption might be even counterproductive) are desired. For example, Barbara Kühnlenz stated that a design might also very well be too anthropomorphic in some use cases. For example, in a said collaborative setting, a too anthropomorphic design might even lead to losing focus (e.g., a robot telling a joke).

However, as stated by Barbara Kühnlenz, humans are social beings and therefore, there do exist certain patterns regarding emotions we intuitively also apply in human-robot interactions which we cannot deny as such. Moreover, mental strain, the choice of words and facial expressions (dt. "Mimik") came also up on short notes.

Attributes a robot might have are naturalness, a computer voice and different levels of responsiveness (derived from both interviews). Moreover, the emotional expression of a robot and the robot's language came also up in the interview with Kerstin Fischer. Additionally, it was discussed that different people might have different definitions and expectations for what a "robot" actually is – whether they rather see it as technology or as social agent. In this context societal aspects also have to be mentioned are discussed in the next section.

Another theme which is derived from the interview with Barbara Kühnlenz is the setting in which an interaction happens. There are always expectations which a person brings into an interaction as well as their background. Interactions can have cognitive goals (probably even strictly content-wise) and there is also the cognitive task context which might be part of the interaction. However, Barbara Kühnlenz also explained the assumption that there might always be societal goals within an interaction (e.g., invoking helpfulness as discussed in the interview).

In this context, the famous quote by Watzlawick that one cannot not communicate has also been mentioned by Barbara Kühnlenz. However, one might ask if this does also apply when interacting with a robot because the robot has to be turned on to actually recognize an interaction partner.

The collaborative setting as well as incongruent task contexts have been mentioned as well. Moreover, the adaptation by person and robot towards each other was also noted.

Regarding language, emphasis, frequencies and prosody were mentioned briefly.

Societal Aspects (B. Kühnlenz)

The interview with Barbara Kühnlenzalso had a focus on societal aspects which emerged from the process of the interview.

It was mentioned that an eye has to be kept on the dangers which might very well come with the chances different aspects of human-robot interaction provide. A concrete example was the dissolution of boundaries between human and robot. Moreover, social effects do not necessarily lead to the desired results. This statement is explicitly relevant to the thesis at hand as the scenario discussed in Figure 1.1 describes a social interaction. Therefore, it has to be reflected on whether we actually want to achieve a successful interaction as outlined or if we rather do not want to replace human interactions of that kind at all.

Furthermore, human-robot interaction of course might influence individuals and society as a whole. Therefore, Barbara Kühnlenz emphasized that we also shall reflect which boundaries we might want to set regarding an emotional dimension in human-robot interaction.

Industry 4.0, as well as delivery- and pickup-services have been mentioned as short notes too.

To conclude with, emotions can be useful in a pro-social sense and therefore lead to pro-social effects as per Barbara Kühnlenz.

4.5 Discussing the Quality Aspect

To ensure the quality of this analysis, Braun's and Clarke's "15-Point Checklist of Criteria for Good Thematic Analysis" [BC06, p. 96] as listed in Table 4.2 has been considered.

The **Transcription (1)** happened as described in the first Phase and therefore rather rigidly.

Coding – (2): Although equal attention is difficult to measure (as time stopping would have been disturbing) it was tried to equally consider all items as they were all placed in the first theme maps. **(3)** The themes have been generated by reviewing every single coded element. Themes got modified within the iterations when other aspects arose and other themes were considered more fitting. **(4)** The full transcript has been reviewed again for writing this analysis. **(5)** The themes have been checked although different data points might fit multiple themes as the themes are all part of the first layer theme "Human-Robot Interaction" which also leads to **(6)** as a clear distinction simply does not seem possible and, moreover, not appropriate because the aspects all interfere with each other, e.g., problems highly depend on the used method (text-based emotion detection does not seem to interfere with mimics or similar, for example).

Analysis – (7): This phase was not too much of work as the interview data itself was very structured. **(8)** can also be backed up with the transcripts in the Appendix. However, as most of the spoken text was relevant, it is rather a summary and linking the two interviews than having to start with making sense and extracting only the important information. **(9)** also was a key consideration as the story told shall help the author to further elaborate on the different topics and after collating Chapter 2, there were no contradictions to the analysis found which should further back the structuring and story up, as it was implicitly checked against the research landscape. **(10)** the analysis

Process	*No.*	*Criteria*
Transcription	*1*	*The data have been transcribed to an appropriate level of detail, and the transcripts have been checked against the tapes for 'accuracy'.*
Coding	*2*	*Each data item has been given equal attention in the coding process.*
	3	*Themes have not been generated from a few vivid examples (an anecdotal approach), but instead the coding process has been thorough, inclusive and comprehensive.*
	4	*All relevant extracts for all each theme have been collated.*
	5	*Themes have been checked against each other and back to the original data set.*
	6	*Themes are internally coherent, consistent, and distinctive.*
Analysis	*7*	*Data have been analysed – interpreted, made sense of – rather than just paraphrased or described.*
	8	*Analysis and data match each other – the extracts illustrate the analytic claims.*
	9	*Analysis tells a convincing and well-organized story about the data and topic.*
	10	*A good balance between analytic narrative and illustrative extracts is provided.*
Overall	*11*	*Enough time has been allocated to complete all phases of the analysis adequately, without rushing a phase or giving it a once-over-lightly.*
Written report	*12*	*The assumptions about, and specific approach to, thematic analysis are clearly explicated.*
	13	*There is a good fit between what you claim you do, and what you show you have done – ie, described method and reported analysis are consistent.*
	14	*The language and concepts used in the report are consistent with the epistemological position of the analysis.*
	15	*The researcher is positioned as active in the research process; themes do not just 'emerge'.*

Table 4.2: 15-point checklist of criteria for good thematic analysis [BC06, p. 96]

consists very much of extracts as the information density, as we can see above, was very high. Therefore, it can be argued that not much of a rigid "data polishing" process was necessary as the interview direction was rather defined in advance as Barbara Kühnlenz even prepared to answer the very specific questions provided in advance.

Overall – (11): How much time is enough, is a non-concrete measure; however, the analysis made up a huge part of this thesis and the excessively described approach might imply that the taken time was enough.

Written report – (12): can be considered fulfilled with the same argument as **(11)**. **(13)** can be checked based on the extensive description. However, after reading this chapter multiple times, no inconsistencies were found by the author; the same goes for **(14)**. **(15)** the position is as active as it can be as the topics of interest were very much asked after by the interviewer. However, within the process, the experts were not interrupted as the connections they draw based on the questions asked were considered of high value by the interviewer/author.

4.6 Reflection, Discussion & Conclusion

When reflecting on the conducted interviews, it has to be stated that, as one can see with the extensive reporting above, that they have been very valuable to the thematic depth of this thesis.

Regarding the analysis method it can be said that it worked well in this context although the topics/themes were already predefined in advance because of the interview guide. However, the structuring and sorting could completely build on this method which justifies its usage. The method, in general, was rigidly followed. There were only small adaptations in the reporting phase as new aspects came up. Additionally, the in [BC06] mentioned analytical aspects had to come short as the interviews were to get an overview of the research landscape rather than to answer a specific research question. Therefore, the report is more descriptive than suggested in [BC06]. This backs up the suspicion that the data generated from the interviews does significantly differ from the kind of data used to exemplify said method. It was, regarding the report, more loosely followed – caused by the data – compared to the other phases. The resulting maps of the phases, however, still look a bit different compared to the example pictures provided in [BC06].

In contrast to other, more data of the interview might have been used as there not really was something left out because the questions where asked strategically, based on the interview guide; therefore there are no patterns which are irrelevant.

To discuss the method further, the guidelines are rather "vague" which might be inherently rooted in the method itself as it depends highly on topic *and* setting (or, more specifically, also on how the interview turns out). Therefore, e.g., the number of iterations in our very specific context might be even a better indicator regarding quality as, because of the setting, it was rather a "How to structure the data?" and not a "What data to actually get out of this?".

Regardless of the arguable points, the interviews deemed itself already relevant and helpful as references to work of the interviewees have been given which were of use in Chapter 2, as mentioned prior.

A doubt the author got after a deeper look into related work also manifested itself within the interviews as text-based emotion detection *might* be less reliable/resource-effective than other methods based on the pitch of the voice, facial expressions or similar.

The idea to look into the influence of wrongly detected emotions got reassured as Kerstin Fischer mentioned that this might be a new aspect to the field which can be of interest.

To conclude with, the "thematic analysis" was useful as it helped to structure the output of the interviews. Moreover, the final maps deemed very helpful when wanting to actually go back to the transcript for finding more details to write up the final report.

CHAPTER 5

Study with Q.bo One

Whereas *RQ1* was answered through a systematic literature review, *RQ2a, RQ2b,* and *RQ3* are addressed through an empirical approach. To answer whether user satisfaction can be increased by in-built emotion detection or not (RQ2a) and if so to what extent, as well as how said satisfaction changes in different settings (RQ2b) and, moreover, how wrongly detected emotions have an influence in this regard (RQ3), a video-based study with Q.bo One was set-up.

5.1 Setting

The setting focuses on two key aspects: 1) Video clips of Q.bo One (`https://thecorpora.com/)`) guide participants through the study, and 2) the study is conducted in German. Moreover, it is completely asynchronous as there is no interaction with a human or a robot necessary. It is an online, video-based (transcripts of the spoken texts are provided to the participants as well) survey conducted with (a self-hosted) LimeSurvey (instance) (`https://www.limesurvey.org/`). The participants watch videos of different scenarios after they have been provided with a specific setting they should imagine themselves to be in.

It is worth mentioning that the setting is designed in a way the participant should gain the impression that the survey itself is conducted by Q.bo One. Therefore, there are two layers: the emotion detection layer (1) with scenarios for emotion detection with Q.bo One as well as the "meta-layer" (2) which is in the end, asking questions about the survey conduct by Q.bo One. The goal of the collected data with this meta scenario is to gain a bit more information about the general sentiment towards Q.bo One, independently of the impact of the tested conversation scenario. E.g., maybe Q.bo One "itself" is no robot people want to share their emotions with after all.

5.2 Scenarios

There exist four different conversation scenarios in the study. Each scenario addresses one of the in RQ2b mentioned settings:

- **Scenario 1:** implicit emotion detection – Q.bo One correctly detects the emotion a participant has, based on the information given
- **Scenario 2:** explicitly asking the participant about their emotion – “How do you feel about ...?”
- **Scenario 3:** wrongly detecting an emotion (based on Ekman – see Chapter 2) – e.g., enjoyment = detected, fear = actual emotion
- **Scenario 4:** no emotion detection – Q.bo One only thanks the conversation partner without addressing their emotional state

Which scenario a participant has to assess during the study is decided automatically by a modulo operation the tool calculates based on the incremental ID of the survey participant. That said, it is possible that because of participants not finishing the full survey different scenarios have a different amount of people taking part in. However, this downside was accepted based on the limitations of the tool and it was imagined in advance by the researcher that the numbers should not differ too much.

5.3 Expected Findings

The expected findings, based on the hypotheses set in advance, are as follows: 1) User satisfaction can be increased with implicit emotion detection as well as explicitly asking the users compared to having no emotion detection at all and 2) wrongly detected emotions negatively influence and therefore decrease user satisfaction.

5.4 Measures

Regarding the quantification of said user satisfaction, the Godspeed Questionnaire Series ([Bar+09]) was chosen. This choice is based on the results of the expert interviews, namely the one with Prof. Dr. Barbara Kühnlenz. Said questionnaire [Bar+09] provides the user with the opportunity to rate the interaction with a robot regarding the aspects anthropomorphism, animacy, likability, perceived intelligence, and perceived safety. This is done by offering different keywords within each aspect and providing a discrete scale from 1 (least) to 5 (most). The measures were chosen as they offer a standardized scale, which can be assumed to measure the underlying concepts.

Besides the sole Godspeed Questionnaire, a few demographic attributes are collected, namely *age, gender*[1], *highest successfully completed education,* and *prior experience with robots*[2].

Regarding the above mentioned "meta-layer", it is asked whether the formal German form of addressing someone ("Sie") or the informal form ("Du") is preferred (the study is conducted with the formal form). Furthermore, the participants are asked whether they think they assessed the robot differently because of the fact that Q.bo One "itself" conducted the survey.

Additionally, each step of the study offers the possibility to add comments and remarks in a text field. At the end participants can offer their participation in an interview and win a prize[3]. The whole survey is to be found in the Appendix (it also includes the transcripts of what the robot actually says).

There are no mandatory open questions, but all questions with a scale are mandatory, as otherwise, e.g., the Godspeed scales could not have been calculated reliably.

5.5 The Robot – Q.bo One

The participants saw videos of Q.bo One in the following perspective.

Figure 5.1: Subset of Q.bo One's States – Vary in Number of Points the Mouth Consists of

The robot was voiced by Google Cloud (`https://cloud.google.com/text-to-speech?hl=de`, last accessed on Jan. 6, 2022) with the following settings.

[1]For asking about gender, the suggestions made by Spiel et al. have been considered [SHL19].
[2]The scale has been based on the prior work of Hannibal et al. [HWC21].
[3]Winners are decided by draw.

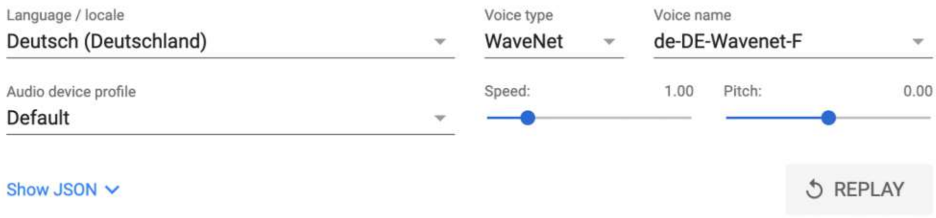


Figure 5.2: Screenshot of the Google Cloud Settings

5.6 Reasoning behind & Limitations of this Setting

The COVID-19 pandemic (as of December 2021) still has a huge impact on life in Austria. Therefore, to not endanger any participants, a setting making it possible to conduct the survey fully online was chosen.

Moreover, the online setting – as it only takes about ten minutes and does not require any form of commuting – might allow to make participation easier and therefore accessible for more diverse people which would increase the ecological validity of the results. Additionally, this setting makes it easier to reach a bigger number of people in a shorter period of time (e.g., parallel study conduction without a researcher actively interfering is possible). Furthermore, the higher number of participants seems to be a reasonable argument to make more general claims about human behavior.

In advance of the study, the option of filming the whole human-robot conversation with the human counterpart visible was considered. However, it was decided against it, as participants should not feel like mere observers of "others interacting", but rather immersed into the setting. Instead, the opportunity for people to imagine to directly interact with the robot was chosen. Moreover, considering the "Theory of Mind" (see, e.g. [Bar99]), in order to enable participants to see the scenario from the protagonist's perspective they might have needed to see a human counterpart to whom they can closely relate. However, this would have arisen the necessity to create a variety of different videos with a diverse set of protagonists and, therefore, would have exceeded the scope of a Master's thesis.

It has to be stated that the implementation is only a mock implementation with using Google Cloud (`https://cloud.google.com/`) and just filming the robot face (Setup: iPhone 13 Pro, portrait mode). This was done for two reasons: 1) programming errors etc. do not confound the study, 2) the time invested was focused on the actual findings and not on creating a "production-ready artifact".

Regarding the limitations, it has to be stated that there might very well be a mismatch between participants (younger people, using the internet, smartphones, etc. very often) and the targeted user group (older people, not as used to said technologies). However, the author assumes that the effects occurring are age-independent and therefore designed the study in a way to be able to look for correlations regarding age in the data a posteriori.

Moreover, the video setting obviously is no direct interaction in a physical space and there also is a lack of interactivity as people do not actually talk to the robot, but think about a predefined scenario.

Note that all of the discussed limitations were a priori noted. In Chapter 6, a posteriori limitations which arose when conducting the study and reviewing the results are discussed.

5.7 Data Protection & Ethical Considerations

As can be seen in the Appendix, the LimeSurvey instance was hosted on servers within the TU Wien facilities. Moreover, the data got stored anonymously and besides time-stamps, no information is connected to the answers, not even IP addresses get stored.

If and only if a participants decides that they are available for a post-survey interview, they provide an e-mail address, phone number or similar which is then connected to their data but is removed before reviewing the data for analysis purposes.

The participants, moreover, were informed that they can stop the survey at any time by closing their browser window if they decide they do not want to participate anymore. As the only personal data linkable to them is only stored if they decide to do so, a stopped survey ensures that they remain anonymous anyway.

To conclude with, they have been informed that they can contact the author or his advisor (Astrid Weiss) at any time via email (any time is emphasized).

Moreover, the TU Wien Research Ethics Department (`https://www.tuwien.at/en/research/rti-support/research-ethics`) has been contacted regarding the study and approved its conduction.

5.8 Report of Results

5.8.1 Scale & Duration

Within one month, 100 people completed the survey (24 with correct emotion detection, 25 with the robot explicitly asking about the emotional state, 22 with wrong emotion detection, and 29 with no emotion handling). However, more people started but did not submit the survey and where therefore not included into the following calculations. The reason behind that was that we could not differentiate between people who completed only a part of the survey and never came back and others who just restarted again and complete it on the second occasion.

The average duration for completing the survey was 12 minutes and 24 seconds. The lowest value was 3 minutes and 36 seconds and the maximum value 2 hours, 45 minutes and 52 seconds. The median in duration was 9 minutes and 13 seconds.

5.8.2 Participants Demography

The average age was 28.19 years with the oldest person being 84 years old and the youngest person 18 years old. The median, however was 23 years and the mode 21 years.

Regarding gender of the participants, one person did not want to provide information, one person chose to describe it, 41 people identified themselves as female and 57 as male.

Looking at the education of the participants 39 completed "AHS with Matura" (Allgemeinbildende Höhere Schule mit Matura, engl. "General Educational Higher School with school leaving examination"), 10 "BHS with Matura" (Berufsbildende Höhere Schule mit Matura, engl. "Vocational Higher School with school leaving examination"), 27 bachelor studies, 16 master or diploma studies, 3 PhD or higher, 2 Apprenticeship with Matura ("school leaving exam") and 2 without, and one person completed compulsory school.

5.8.3 Prior Knowledge

Prior knowledge regarding robots was distributed as follows (multiple selections were possible):

- 81 from culture (literature, movies, radio, magazines, television)
- 33 from their free time (crafting, projects, scientific magazines, family, friends)
- 29 from education (lectures, thesis, internship)
- 29 randomly (shopping, participation in a study, events)
- 20 from work (building, programming, research projects)

However, note that one person was unclear what exactly was meant with the question and noted that "gotten to know" and "getting in contact" are two different things.

5.8.4 Ability to Change Perspectives

Rating, how well they think they could see the scenario from the perspective of the person being part of it, the participants answered as follows:

- 40 good
- 24 rather good
- 19 medium
- 9 very good
- 4 rather bad

- 3 not at all or very bad
- 1 bad

They provided some comments which were (complete list but shortened answers):

- "The robot voice telling the own narration from one's day too makes the situation *weird*".
- "I do not think that everyone would tell that much at the beginning, maybe after a learning phase".
- "After the first sentences I rather read the transcription instead of listening actively".
- "Far away life-style".
- "That is a very sad vision. From today's perspective, I know that the robot only collects data without understanding, showing empathy or concluding any sort of consequence from it. Perhaps, as an older person, I would lose this analytic view and find myself within the illusion that the robot actually is interested in me. I consider that to be a certain way of misleading and therefore as a sad vision. It can work, but it is dishonest".

5.8.5 Way of Addressing

Regarding the formal way of addressing the participants, 27 did not consider it to be weird, 23 a bit weird, 23 did not actively notice it, 15 not really weird, 8 neither weird nor normal, and 4 very weird. However, 42 would have preferred the informal wording "Du", "33" did not know their perference, and 25 would not have preferred "Du".

When asked whether they think that they would have assessed the robot differently if it did not ask the questions itself but instead a not participating third person, 26 think that it would have not made any difference, 25 think that there is a difference but they do not know the extent, 23 they have assessed the robot more positively because of that, 15 do not think they can assess that, 7 think it made a difference that they graded the robot a bit more negatively, 3 think more positively, and 1 thinks more negatively.

In that context, addressing the perspective and form of survey explanations, one person stated that the preference between informal and formal might depend on the social environment of the participants and that they consider it a bit irritating when a stranger addresses them informally and therefore also a computer.

One did not understand the question and asked whether it is with regard to a human counterpart or just because of the grammatical construction. They added that informal vs. formal implies a measurement of human closeness or distance (e.g., consider supervisors, older people, strangers, parents, friends, . . .). But because of a machine talking, the formal vs. informal does not apply and is "generic".

The different layers were considered to be confusing by one person.

Another adds that they noticed a habituation effect regarding the unnatural and mechanical way of speaking which influences the general impression in a positive way.

One would have expected an answer which relates more to the narration and the conversation finished abruptly and, therefore, they did not feel that they were listened to.

Another participant noted that the formal addressing increased “the distance and coldness of the interaction”.

Moreover, one participant considered the introduction of the scenario at the beginning to be “really good and important” and therefore, they think that they would not have acted differently regarding formal vs. informal addressing.

Another person noted that the repetition of the exact wording let the scenario and questions “coalesce” and that therefore a not participating third person would have been helpful.

Finally, one participant added that the voices of robots have to sound more realistically to be convincing would be the first step and only the second would be that the “Robot/AI or however one want to call the thing” listens and answers in a sense-making way. They added that it would be interesting if it meant something to the person that one is agreeing with them and whether that is enough. For them personally, it would not be very much if one is understanding and agreeing all the time. Leaving aside that it is not that easy that the “robot” creates this connection, how often would it be the “wrong” one?

5.8.6 Other Comments

Other remarks were that the transcript led to skipping most parts of the video, that the topic is an exciting one, that the study is interesting although the conduct is highly based on assumptions which might influence the results, that there was no neutral answer regarding the knowledge about robots.

One considered the evaluation based on one answer of the robot to be restricting and their assessments to be rather unspecific.

Another one considered it to be exciting receiving the survey by the robot itself and that the survey was not too long.

Note that there were other positive comments which are left out here intentionally as they did not provide concrete feedback or remarks.

5.8.7 Godspeed Values of Different Scenarios

The following section deals with the different values of the Godspeed Questionnaire used within the work for this thesis. For that the results have been curated and the median and mean have been calculated for each scale for each scenario.

Within the graphs, the following notation is used:

- Asking about a person's emotional state
- No emotion handling
- Emotion detection
- Wrong emotion detection

For easier recognizing, consider the following thought:

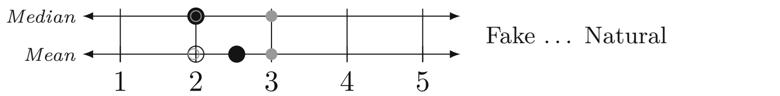
Filled means that emotion handling exists, the bigger the circle the "smarter" the handling, i.e. bigger circle means detection. (Different shades only to make overlays visible.)

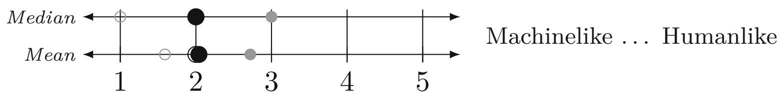
Note that all values got rounded to two decimals (round up from 5 onwards).

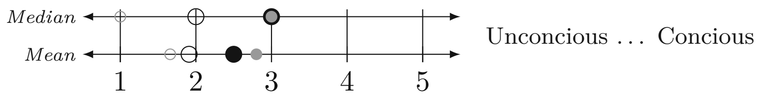
Block 1: Anthropomorphism

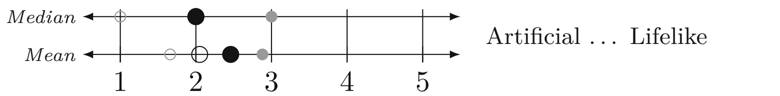
We can clearly see that in the scenario, in which the robot asks about the emotional state of the user (grey-filled, smaller circle), the scores are the highest. The robot is perceived as more natural, human- as well as lifelike and with a higher state of "consciousness". Note that all values are nowhere near close to four, let alone five and therefore in all scenarios Q.bo One was not perceived as very "anthropomorphic".

Asking a person about their emotions was, as we can see here, perceived as more natural as well as more humanlike compared to implicit detection.







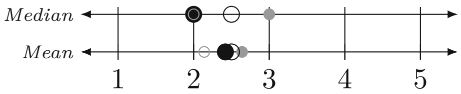


Also worth mentioning is the fact that wrongly detected emotions in all cases scored higher "anthropomorphic" values than no emotion detection at all.

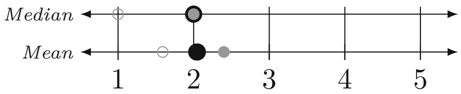
Block 2: Animacy

The second block draws a similar picture in the sense that the explicit asking about a person's emotional state leads to higher values on the Godspeed scales, with the exception that the median is equivalent in more cases and even higher regarding the responsiveness of the robot.

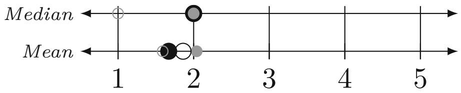
Interestingly, the robot is perceived as more lifelike when it explicitly asks about one's feelings instead of solely detecting them and using the results within the answer. Moreover, emotion handling leads to a slightly higher perception regarding interactivity.



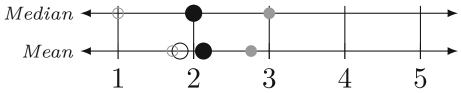
Dead ... Alive



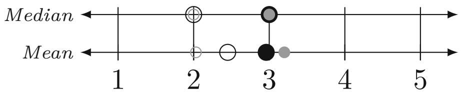
Stagnant ... Lively



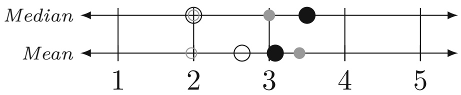
Mechanical ... Organical



Artificial ... Lifelike



Inert ... Interactive

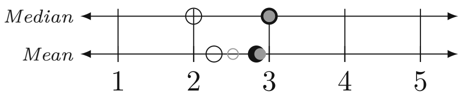


Apathetic ... Responsive

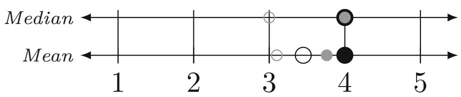
A person, however, noted that the given answer is a "classic psychotherapeutic" one and therefore "seems to be rather analyzing than expressing sympathy".

Block 3: Likeability

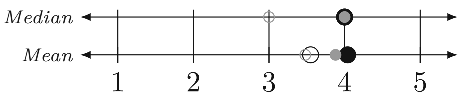
We can see that the robot was perceived as rather friendly but that the participants did not really like it per se. They saw the experience rather neutral and neither really unpleasant nor pleasant, as well as neither awful nor nice.



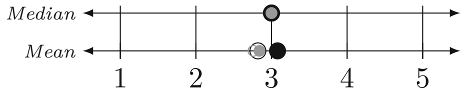
Dislike ... Like



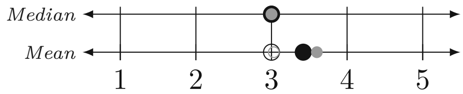
Unfriendly ... Friendly



Unkind ... Kind



Unpleasant . . . Pleasant



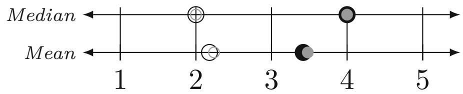
Awful . . . Nice

However, the robot was perceived more kind than unkind.

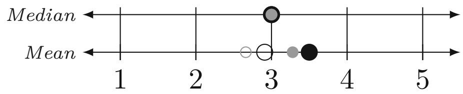
Block 4: Perceived Intelligence

This Godspeed block shows that the scenarios with emotion handling lead to the robot being perceived as more competent as can be seen especially within the median scale. However, the difference on the mean scale is also bigger compared to most other scales within these results.

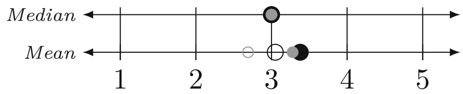
Emotion handling also leads to higher values regarding perceiving the robot as intelligent. The robot is considered neither foolish nor sensible but neutral on this scale. However, the implicit emotion detection leads to it being seen as more sensible.



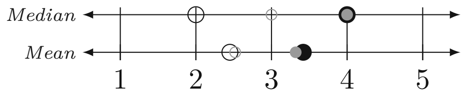
Incompetent . . . Competent



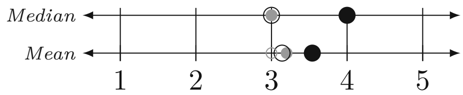
Ignorant . . . Knowledgeable



Irresponsible . . . Responsible



Unintelligent . . . Intelligent



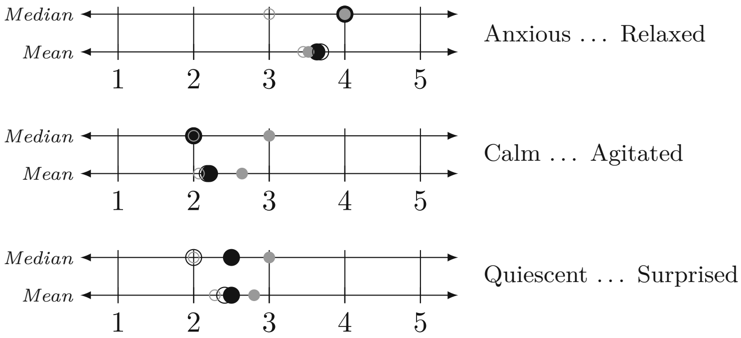
Foolish . . . Sensible

In general, the scales of this block suggest besides competence and intelligence a rather neutral stance of the participants.

A participant stated the following (scenario: wrong detection): "As I know that this is a robot, I do not expect 'intelligent' (in this case: connections detecting) answers. If the robot had, e.g., an AI, then I would be a lot more surprised about that 'mistake' (i.e., the assumption that I am happy about the wedding), respectively, perceiving the robot as highly incompetent or manipulative.

Block 5: Perceived Safety

Regarding their own emotional state, participants were rather relaxed than anxious and also calm rather than agitated. Interestingly, the values of surprise and agitation were the highest when dealing with the scenario of the robot asking explicitly about the emotional state of the participant.



5.8.8 Analysis of Variances & Kruskal-Wallis Test

Regarding the following analysis, the work of Kühnlenz (then: Gonsior) et al. [Gon+11] has been considered to be a reference regarding structure and analysis methodology. We apply the same reasoning for the assumption reliability and internal consistency, namely that the questionnaires are previously evaluated.

Subject Groups & Key Concepts

The participants got divided into four different groups as explained above based on the scenario that got selected because of the survey ID. The four different scenarios are abbreviated as follows when being referenced in the following sections:

1. Scenario – **C**orrect Emotion **D**etection: CD
2. Scenario – **E**xplicitly **A**sking about one's emotional state: EA
3. Scenario – **W**rong Emotion **D**etection: WD
4. Scenario – **N**o Emotion **H**andling: NH

Whether the different ways of (not) handling emotion have improved the human-robot interaction regarding the five concepts tested via the Godspeed Questionnaire Series [Bar+09], namely anthropomorphism (abbr. ANTHR), animacy (ANIMA), likability (LIKAB), perceived intelligence (PEINT), and perceived safety (PESAF), or not is evaluated via analysis of variance (ANOVA) or a Kruskal-Wallis test, respectively, based on the corresponding distributions.

The analysis has been conducted using Python[4].

The average values of the key concepts and the corresponding standard deviations are as listed in Table 5.1 and shown graphically in Figure 5.3.

	ANTHR	**ANIMA**	**LIKAB**	**PEINT**	**PESAF**
CD ($n = 24$)	2.39 (0.94)	2.38 (0.79)	3.48 (0.89)	3.45 (0.86)	$2.\dot{7}$ (0.45)
EA ($n = 25$)	2.85 (0.84)	2.74 (0.71)	3.39 (0.62)	3.31 (0.61)	$2.98\dot{6}$ (0.46)
WD ($n = 22$)	1.99 (0.81)	2.22 (0.55)	$3.0\overline{18}$ (0.90)	2.75 (0.79)	$2.\overline{75}$ (0.52)
NH ($n = 29$)	1.72 (0.76)	1.84 (0.60)	2.98 (0.77)	2.62 (0.55)	2.60 (0.62)

Table 5.1: Mean (Standard Deviation)

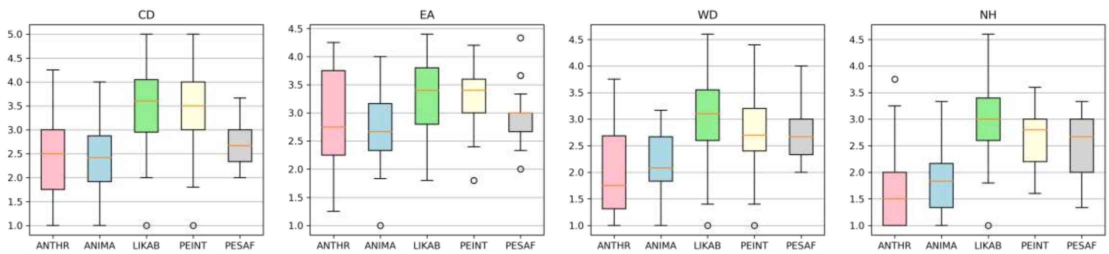


Figure 5.3: Mean Values of the 5 Godspeed Concepts in the 4 Different Scenarios

Distributions of Results

First, the distributions resulting from the questionnaire answers have been tested whether they are normal distributions or not ($alpha = 0.05$). The results of said tests are shown below (Table 5.2) as well as the distributions are displayed as violin plots (Figure 5.4).

	ANTHR	**ANIMA**	**LIKAB**	**PEINT**	**PESAF**
CD	✓	✓	✓		✓
EA	✓	✓	✓		✓
WD	✓	✓	✓	✓	✓
NH		✓	✓	✓	✓

Table 5.2: Distribution Overview

Because of these distributions, we can apply an ANOVA for all key concepts except anthropomorphism and perceived intelligence which are analyzed with a Kruskal-Wallis test. The p-values adjusted with the Holm–Bonferroni method in both cases.

[4]using `matplotlib`, `numpy`, `pandas`, `scikit_posthocs`, `sciipy`, and `seaborn`

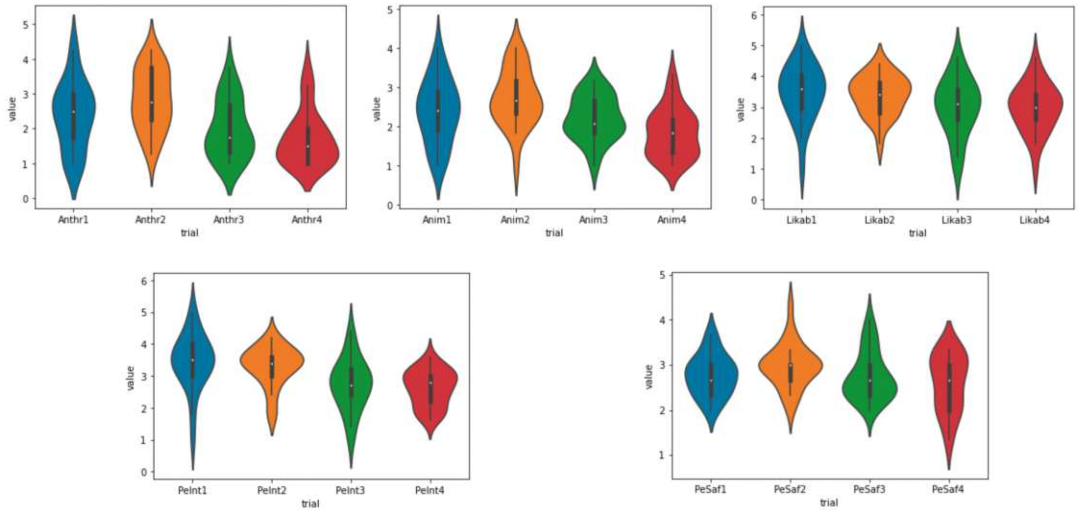


Figure 5.4: Distributions as Violin Plots

Looking into the results of the ANOVA for likability ($F = 2.453$, $p = 0.068$) and perceived safety ($F = 2.354$, $p = 0.077$) we see that the different scenarios had no(t enough) impact as the results lie within the same distribution.

However, when reviewing the results for animacy, we see that the results are (likely) not of the same distribution ($F = 8.021$, $p = 7.975 * 10^{-5}$), as there are significant differences between the following scenarios shown by a post hoc t-test.

Animacy:
CD vs. NH: 0.039
EA vs. WD: 0.039
EA vs. NH: $4.8 * 10^{-5}$

The conducted post hoc analyses indicates that the participants rated *Animacy* in the CD condition significantly higher ($M = 2.38$, $SD = 0.79$) compared to the NH condition ($M = 1.84$, $SD = 0.60$). In the EA condition, animacy was also significantly higher rated ($M = 2.74$, $SD = 0.71$) than in the WD ($M = 2.22$, $SD = 0.55$) and NH conditions ($M = 1.84$, $SD = 0.60$).

The Kruskal-Wallis test for anthropomorphism and perceived intelligence also revealed that the results are (likely) not of the same distribution, with p-values of 0.0001 and $1.057 * 10^{-5}$ and corresponding H-values of 21.025 and 25.786, respectively. Further looking into these results by applying a post hoc Conover's test.

Anthropomorphism:
CD vs. NH: 0.022
EA vs. WD: 0.005
EA vs. NH: $2.8 * 10^{-5}$

For *Anthropomorphism* the same statements apply as for Animacy, except for differences regarding mean and standard deviation (CD: $M = 2.39$, $SD = 0.94$ and EA: $M = 2.85$, $SD = 0.84$ and WD: $M = 1.99$, $SD = 0.81$ and NH $M = 1.72$, $SD = 0.76$)

Perceived Intelligence
CD vs. WD: 0.002
CD vs. NH: $3.5 * 10^{-5}$
EA vs. WD: 0.006
EA vs. NH: $1.65 * 10^{-4}$

For *Perceived Intelligence* the same statements apply as for Animacy and Anthropomorphism, again except for differences regarding mean and standard deviation. However, in this case, there were significant differences between the conditions CD and NH to be found too (CD: $M = 3.45$, $SD = 0.86$ and EA: $M = 3.31$, $SD = 0.61$ and WD: $M = 2.75$, $SD = 0.79$ and NH $M = 2.62$, $SD = 0.55$).

5.8.9 Age Correlations

When looking into the coefficients of correlation between age and the average values of the Godspeed results, no significant correlations could be found as the biggest coefficient was 0.283 (likability when no emotion handling is conducted) and the smallest coefficient was -0.380 (perceived intelligence when wrong emotion detection is conducted). However, based on the distribution of the age values of the sample (see Figure 5.5), this was not a surprising result.

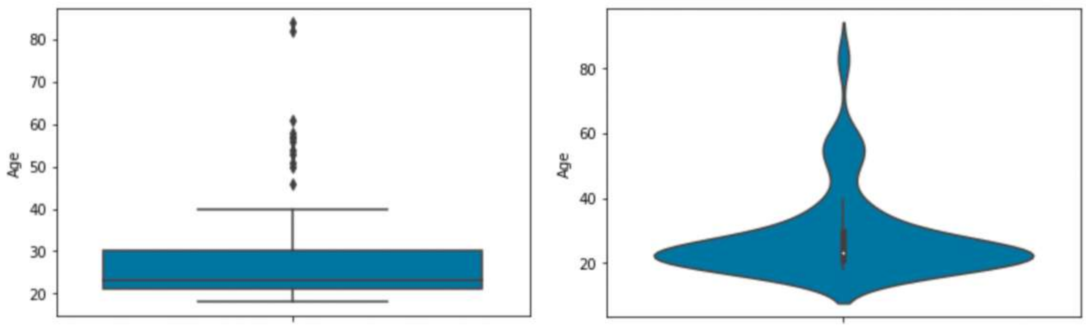


Figure 5.5: Distribution of Age

5.8.10 Summary

Within the results of this survey, the scenarios with emotion handling scored higher Godspeed values than the ones without. Further differentiating between the two different ways of emotion handling, explicitly asking the participants about their emotional state led to higher values than implicit emotion detection, in general.

In contrast to that, wrong emotion detection in most cases still led to higher values than no emotion detection at all.

However, the blocks anthropomorphism and animacy had a tendency towards lower values compared to likability or perceived intelligence.

When looking for significant differences between the four scenarios, we found that there exist ones between 1) correct emotion detection and no emotion handling, 2) explicitly asking about one's emotional state and wrong detection, and 3) explicitly asking about one's emotional state and no emotion handling for anthropomorphism, animacy and perceived intelligence. When considering perceived intelligence, there are also significant differences between correct emotion detection and no emotion handling.

Significant correlations concerning the age of the participants and the respective Godspeed values could not be found.

73 percent of the participants considered themselves to be able to see the scenario from the perspective of the person being part of it *better than medium*. 92 percent considered themselves to have this ability to *medium or better* extent.

CHAPTER 6

Discussion

The following chapter deals with discussing, first of all, the results found in the different steps of the thesis. This especially includes the findings regarding the study as they shall form the empirical, entirely new part of this thesis, whereas the other chapters are mainly based on collating and discussing literature.

6.1 Related Work

A considerable part of the Related Work Section (2) of this thesis deals with the concept of emotion models and what emotion actually is according to various ideas. Whereas the first is directly relevant for actually modeling emotions within a computational context, second has indirect implications in that case as it defines what one models, and, moreover, might be introducing bias regarding what one wants to find out with a specific study design. Therefore, it was considered to be vital as well as it shall help with the implementation of different emotion detection solutions when one knows about the concept they are trying to model on a computational, artificial level.

Moreover, the originally proposed focus on Ekman's basic emotion model lost its justification a bit the moment the implementation got removed of the scope of this thesis. However, as also shown based on the vibrant discussions outlined within the Related Work Section (2), its relevance is clearly given within this field. Therefore, the focus was not out of place after all.

Outlining the concepts of companion robots, artificial empathy and different measures for user satisfaction was relevant because it provides the basis for the empirical part of the thesis at hand.

To conclude with, the overview of the methods for emotion detection does not aim to achieve completeness. It rather shall provide a comprehensive collection of different ways one could approach the idea of detecting emotion in human-robot conversations when planning on doing an actual implementation or another study. As the text-focus was a priori defined, other ways were outlined rather briefly.

6.2 Performance Differences: Native vs. Translation-layered Solutions

As concluded, most of the sources suggested that translation might very well be good enough to provide reliable sentiment analysis with tools developed for English-language use cases. However, it was already briefly discussed that sentiment analysis is a way simpler use case than (sophisticated) emotion detection. Positive, negative, or neutral attitude towards something is a broader classification than, e.g. anger, or fear (which are both negative in this context).

Moreover, the input also has to be discussed. Tools that work reliable on a curated text from a news outlet might deliver less optimal results when being provided with input which went through a speech recognition layer and contains, e.g., filling words and other noise. Moreover, filling words could indicate something like boredom and sentiment/emotions would be altered when removing them.

Regarding the translation technology, we have seen that different providers (Google, Bing, e.g.) delivered all promising results. Therefore, different translation technologies might be mature enough overall to be used as an additional layer for English-focused tools and therefore emerge as an alternative to language-specific solutions. However, one must not forget that the process of translation also costs resources, as pointed out by [AKB13].

Furthermore, using translation does not always mean, translating the input to English, e.g., [MSK16] pointed out that a lexicon or similar can also be translated to enable a sentiment analysis tool to deal with input in a language different from English.

The focus on the languages English, French, German, and Spanish – and on rare occasions Arabic and Russian – has to be considered as limitation as it does not allow for general conclusions as different languages might have different translation technologies regarding, e.g., reliability.

Moreover, it only was a scoping study [AO05; LCO10] and therefore not a full review. This has also to be considered when putting the results into perspective.

To conclude this section of the discussion, sentiment analysis might be a good starting point when searching for emotion detection from text. However, that does not imply at all that text-based emotion detection is the "go to" emotion detection after all.

6.3 Expert Interviews

The results and limitations of the expert interviews are discussed in Sections 4.5 and 4.6 as the interviews form a distinct entity and are therefore discussed separately. Said interviews, however, might have had an influence on Chapter 2, as well as Chapter 5, as they have been conducted in advance and therefore the "exploration space" was defined, at least partially, by the interviews (e.g., suggested papers by experts).

6.4 Study Results

The study results suggest that people slightly prefer being explicitly asked about their emotions over the robot implicitly detecting them. However, this statement has to consider that the setting is narrow in the sense that only one robot was used, the demography of the participants was not as diverse as it could be doing this study at a larger scale, and that the shown scenario only provides one very specific interaction with a person being asked about their feelings regarding a wedding.

Statistically significant differences, however, were to be found between the correct emotion detection, explicitly asking about one's emotional state and not handling the emotional state of the user. Animacy, anthropomorphism, and perceived intelligence were all rated significantly higher when correctly detecting one's emotions or explicitly asking about them compared to not handling the emotional state at all. The same applies on all three occasions too when comparing explicitly asking about one's emotion's to wrong detection. However, there are no significant differences between correct detection and wrong detection regarding any concept despite "Perceived Intelligence". The fact that, wrong detection does not lead to as many statistically significantly worse results compared to no emotion handling at all might come due to the fact that users tend to forgive robots errors as discussed by Mirnig et al. [Mir+17] where an "errant" robot was liked better by the participants.

Therefore, we cannot argue that the statement holds in general. However, it provides us with an interesting aspect we might want to look into further as the technical complexity of just asking a person about their emotional state in a human-robot interaction scenario is way smaller than detecting it under different circumstances. Therefore, we could save a lot of time and computing resources in that context.

On a side-note, the expert interviews also suggested that implicit emotion detection might not be necessary in all cases regarding user satisfaction. The results back this assumption up – at least to some degree.

A comment from the survey sees explicitly asking about the emotions and the corresponding answer as "psychotherapeutic" answer which is rather analyzing than showing sympathy (which might remind one about Weizenbaum's ELIZA [Wei66]. It would be interesting to follow up whether the robot is considered as companion or as some sort of psychotherapeutic agent/professional and if there are differences regarding the various scenarios.

Considering *RQ 2a*, it cannot be necessarily answered that the user satisfaction can be increased (which was the corresponding hypothesis) by in-built emotion detection. In the context of this scenario, explicitly asking the user about their emotions performed better in almost all parameters. However, emotion handling in general performed better compared to no or wrong emotion handling which leads to *RQ 2b* which asked about the extent of user satisfaction comparing “in-built emotion detection”, “explicitly asking the users about their emotions“ and “no emotion handling”. In this context, the results suggest that compared to “no emotion handling” the first two performed better. Therefore, emotions are to be considered in related use cases, based on the results of the survey.

Wrongly detected emotions (*RQ3*) have a negative influence on user satisfaction compared to “correct” emotion handling, however, compared to no emotion detection at all the participants were actually liking the “wrongly detecting” robot better. However, based on the one scenario, there cannot be any predictions made what happens with repeatedly failing emotion detection. Maybe the robot was liked better because it at least reacted to the scenario instead of providing a general “thank you for sharing your thoughts” answer. Therefore, it would have to be investigated what happens when the answer does indeed reference to the scenario, but ignores emotions. It is entirely possible that a more specific but still not emotion-considering answer would lead to better results. Wrongly detected emotions therefore likely decrease user satisfaction compared to emotion handling answers but one mistake does not necessarily push the users away from the robot entirely, at least based on the survey results.

In this context, a comment made by one of the participants that they would be more surprised about the mistake if the robot had some sort of AI suggests that the assumptions brought into the scenario might be quite different as well, because it is not stated that the robot does not have an AI in a corresponding real-world scenario.

However, further following the above thought, the setting of the study might very well also have an influence on the perception of the robot as the thought “if the robot does not interact directly with me, it cannot be this smart” or similar could arise.

The limitations are further discussed in the section below (Section 6.5).

6.5 Limitations

When talking about the limitations of the video-based online study, the following only include the a posteriori limitations which arose during conducting the study or when reviewing the results. The a priori noted limitations of the setup are to be found in Section 5.6.

The diversity of the participants is for sure a limitation of the conducted study. As the envisioned demography for future users (older people) differs from the participants (avg. age about 28 years).

In the end, this thesis turned out to be rather a “where” to start or “that is a good method” suggestion than providing definitive answers on “how to do emotion detection right”. This is not entirely unexpected as the study was of course limited to one robot and the distance scenario also might have interfered. However, this should not discourage one to further look into the questions of “When is emotion detection useful?” and “Which kind of emotion detection is appropriate for which use context?". Because, as – a bit surprisingly – (mocked) emotion detection was not as appealing to the participants as simply asking a question about the emotional state.

6.6 Abandoned Ideas

In order, to put the work in this thesis into perspective, the author decided to also reflect on abandoned ideas which might be useful to set Section 7.2 into perspective.

First (before removing it entirely), HYP1 was slightly modified (compared to the one in the propsal) in order to make it more concrete. This means, "..., without impacting detection performance significantly" got added. However, within the process it came clear that the question was actually exploring the wrong issue.

RQ 1 and HYP 1 after the adjustment looked as follows.

RQ 1: Which structural differences between English and German are relevant for speech-to-text recognition and emotion detection (e.g., vocabulary, sentence structure, or grammar) and which methods built for English-language conversations can be applied?
HYP 1: *There are differences regarding pitch and volume of the spoken language as words describing the same emotional state can be built of completely different letters and therefore have a completely different "sound-character". Methods building on the voice attributes can be applied, and textual methods can be sent through a translation layer, without impacting detection performance significantly.*

However, there were two major issues with this research question: 1) In order to challenge it in a valid an scientifically upright way, a lot of linguistic domain knowledge is needed which was not part of the author's curriculum, and 2) before actually looking into the structural differences relevant for transferring solutions from one language to another, thinking about the actual need of said transfers might be the correct first step. Therefore, this thesis now deals with the concept of putting a translation layer between the English-language emotion detection and the German-language text and looking into the performance implications rather than working on a language level per se.

Moreover, originally proposed as RQ2 was "Are there conflicts between certain methods which outcomes probably contradict each other?" (corresponding HYP: "The different methods can be combined by, e.g., evaluating the input in parallel and possible contradictions can be resolved by prioritizing methods which are more likely to give a sufficient answer"). The reasoning behind abandoning this question was its triviality.

It seems very reasonable to assume that different methods, even operating with different types of input, do indeed lead to different results (e.g., recognizing facial parameters compared to text-based emotion detection). Therefore, in order to not interrupt an interaction, some type of error handling has to be proposed. Intuitively, this can either be to neglect emotion detection in general in contradicting cases, or to implement prioritization in some way (even if the prioritization is chosen "randomly" from case to case).

Moreover, as said question is a technological detail rather than a fundamental question, which backs up the decision to abandon it even further.

Furthermore, regarding the proposed RQ3 (which got split up into RQ2a and RQ2b), the question changed from a "yes/no" type to "to what extent" and therefore offers the possibility to answer in a more nuanced way.

Additionally, the idea of exploring a possible implementation of such an emotion detection has been completely abandoned as it turned out to be not entirely feasible in order to answer the research questions. First of all, it would only have been used for one scenario (implicit emotion detection); secondly, it could, as already mentioned, introduce errors which influence the study results, and, thirdly, the balance between necessary time and benefits in order to answer the research questions was not given as per the author.

Based on abandoning the idea of an actual implementation, the scenario(s) got also narrowed down to one specific.

Regarding the "meta-learnings", it has to be said, that the scope envisioned at first was way to ambitious for a Master's thesis. Therefore, the idea of an implementation simply had to be abandoned.

CHAPTER 7

Conclusion & Future Work

7.1 Conclusion

Emotion handling is already a well-studied topic within the field of Human-Robot interaction. This work focused on three different areas: 1) How to detect emotions (text-based)? 2) What do different languages imply for different technologies? 3) How does emotion detection influence user satisfaction? Within the work of this thesis, three differentiations for dealing with emotions have been defined: 1) emotion detection, 2) explicitly asking the user about their emotions, 3) no emotion handling. The scenario of detection has been differentiated in a) correct and b) wrong detection.

Before touching the actual models for emotions applied in Computer Science, it has been seen that scientists are yet to agree upon what an "emotion" is. However, the exact definition of what an emotion actually is might not be necessary in order to "work" with emotions from a Human-Robot Interaction perspective, as people only have to agree upon a model which works for a specific context. For detecting emotions, this thesis differentiated between *content-based emotion detection* and *transmission-based emotion detection*. When deciding on implementing emotion detection for a specific use case and following deciding on a method for emotion detection, or more broadly speaking, emotion handling, it might be a favorable idea to consider the complexity, reliability and potential user experience enrichments carefully in advance to work resource-efficient.

In order to ground these considerations on data, after sharpening the research focus with expert interviews, a scoping study on the performance implications when using English-language tool with a translation layer for German-language use cases compared to native solutions has been conducted. Moreover, this thesis looked into four different scenarios and how they influence the perception of the Godspeed concepts *Anthropomorphism, Animacy, Likability, Perceived Intelligence,* and *Perceived Safety* with a video-based online survey. Said scenarios were correct emotion detection (CD), explicitly asking the user about their emotions (EA), wrong emotion detection (WD), and no emotion handling at all (NH).

Resulting data suggests that it is possible that in-built emotion detection does not lead to better results than, e.g., explicitly asking the user about their emotions as the results of our study suggest and the expert interviews already implied. However, for that there has not been found statistical significance. In contrast to that, *Animacy*, *Anthropomorphism*, and *Perceived Intelligence* are significantly higher within the CD condition compared to the NH condition. Moreover, said concepts are significantly higher rated when the EA condition is present compared to the NH as well as the WD condition. Interestingly, only for the concept of *Perceived Intelligence*, there are significant differences between CD and WD, but not for the other concepts. As pointed out in the discussion, when referencing the work by Mirnig et al. [Mir+17], users tend to forgive robots when they make a mistake.

To conclude with, it might not be necessary, to "reinvent the wheel" when implementing emotion detection. Translation layers might provide acceptable performance compared to native solutions. However, that does not imply one should use *content-based emotion detection* at all. *Transmission-based emotion detection* or a combination of both might be a more feasible way. Regardless, emotion detection might not be necessary for the targeted user satisfaction, explicitly asking might very well be enough. When deciding on doing emotion detection, one should not forget that it still might be a good idea to explicitly ask the human interaction counterparts (maybe even based on the detected emotion) about their emotional state anyway, as this could be perceived even more satisfactory.

7.2 Future Work

Based on the outcomes of this thesis, following ideas for future work arise:

- The scenario within this thesis only included one "basic emotion". However, it would be interesting to look into the outcomes (maybe even with the same setup regarding comparability) with different depicted emotions and whether a specific kind of emotion handling is most suitable for a specific emotional state. This was not possible within this Master's thesis because of the limited amount of resources.
- Moreover, the setting only looked into one specific robot. However, the robot can also influence the user satisfaction as it is not detached from the concept of emotion detection in our context because it has a very specific appearance and voice etc. Doing a study with the same setup but a set of different robots might also lead to discrepancies within the results.
- The study could be made more sophisticated by removing the video setting and enabling the participants to actually interact with a robot (distance or presence

format). However, for this an implementation or a more time-consuming "Wizard-of-Oz" approach (e.g., [Mar+17]) is necessary.

- As this thesis only considered the German language landscape specifically, in-depth comparisons between a variety of languages might also be interesting to further elaborate on in the context of emotion detection.
- Moreover, not only the scenario can be varied but also the technologies used for the actual in-built emotion detection.

This list does not aim to be complete and only reflects the author's ideas regarding future work based on the outcomes of this thesis.

APPENDIX A

Appendix

A.1 Re: Expert Interviews

The following pages include the authorized PDF documents transcribing the conducted expert interviews. As the inclusion only works on full pages, the transcriptions start on the next page.

The following quotation has been used within this thesis:

[EPI1] = Expert Interview (no. 1) with Kerstin Fischer

[EPI2] = Expert Interview (no. 2) with Barbara Kühnlenz

Moreover, the interview guide (sent to Barbara Kühnlenz) and consent form are provided after the transcriptions.

Interviewprotokoll

Prof. Kerstin Fischer

Erklärungen

VRE: … Rafael Vrecar

FIS: … Prof. Kerstin Fischer

Das Interview-Protokoll wurde in zwecks Einheitlichkeit durchgängig geschlechtergerechter Sprache abgefasst, auch, falls einzelne Passagen im Interview selbst nicht in einer solchen Sprache gesagt wurden.

Einleitung

[Inhaltlich nicht relevanter Smalltalk]

VRE: Vielen Dank für die Teilnahme, ich bin mir sicher, dass das hilfreich sein wird. Mein Name ist Rafael Vrecar und mir fehlt jetzt "nur noch" die Diplomarbeit zum Abschluss meines Studiums (Media and Human Centered Computing an der TU Wien) und im Rahmen dieser Diplomarbeit möchte ich untersuchen, inwiefern sich Emotionserkennung in Konversationen zwischen Mensch und Roboter eben auch auf die Nutzer:innenzufriedenheit auswirkt. Dabei liegt mein Fokus auf der deutschen Sprache und deshalb ist es für mich insbesondere interessant, inwiefern sich die deutsche Sprache hier in der Methodik zur Emotionserkennung evtl. von der englischen Sprache unterscheidet. Wenig überraschend gibt es natürlich zur englischen Sprache die meiste Literatur, als "lingua franca" in der Wissenschaft; und ob das überhaupt möglich ist, dass man sagt: Ok, man hat jetzt generelle Methoden, die man eigentlich in verschiedenen Sprachen anwenden kann – oder ob im Prinzip jede Art von Emotionserkennung von Sprache zu Sprache so unterschiedlich ist, dass man hier hochspezifische Lösungen braucht. Das wäre das generelle Abstract, sozusagen, von dem, was wir hier machen.

[Inhaltlich nicht relevante Fragen bzgl. Einverständniserklärung, Interview etc.]

Emotionen in bisheriger Forschungstätigkeit

VRE: Gleich zur ersten Frage: Inwieweit [hat] Emotionserkennung bis jetzt bei dir in der Forschungsarbeit & -tätigkeit eine Rolle gespielt … also ob du damit jetzt schon viel zu tun hattest, und wenn ja, wie und was?

FIS: Ich habe mich eher damit beschäftigt, welchen emotionalen Ausdruck Roboter haben sollten und nicht so sehr zu erkennen, welche[n] der Mensch hat bzw. benutzt – aber in der Interaktion geht es natürlich immer darum, die Emotion zu teilen. Wenn man zum Beispiel eine Geschichte erzählt, dann versucht der:die Sprecher:in, dem:der Höher:in eine Möglichkeit zu geben, vorherzusehen, wie er:sie auf die Geschichte zu reagieren hat – also ob es eine schreckliche Geschichte wird, oder eine gute Geschichte. Man kann auch zeigen, solange die Leute nicht wissen, worauf das hinausläuft, sie einen extrem hohen Arousal-Level haben. Also da ist es auch für die Höher:innen total wichtig, zu wissen, welche Emotionen sind mit der Geschichte verbunden um quasi die erwarteten Reaktionen kommunizieren zu können. Das ist jetzt nur ein Beispiel dafür wie die Emotionen quasi interaktiv, behandelt bzw. ausgehandelt werden und eben sowohl für das Erkennen durch die Roboter als auch die Produktion von Emotion, von den richtigen Emotionen zur rechten Zeit – wie wichtig das ist.

VRE: Danke erstmal, das ist schon ein guter Einstieg, dass ich ungefähr einen Überblick habe, worum es bei deiner Arbeit geht. Jetzt die direkt darauf aufbauende Frage: Du hast ja schon relativ viel mit Robotern und ich nehme auch an, durch deine Arbeit mit der Emotionserzeugung, in der Kommunikation zwischen Mensch und Roboter, gemacht. Und wenn du das jetzt herunterbrechen müsstest, auf einen Output, wenn du einer laien Person erzählst, das ist der wichtigste Output, den ich bis jetzt aus meiner Arbeit mitnehme … Kann man das irgendwie so herunterbrechen oder gibt es da mehrere Sachen, wo du sagst – in a nutshell – das habe ich jetzt in meiner Erfahrung gelernt bzw. mitgenommen?

*[**FIS:** Rückfrage auf Bezug]*

VRE: … in Bezug auf Mensch und Roboter, also die Interaktion, also de facto Human-Robot-Interaction.

FIS: … Also es muss responsiv sein, also man muss sich aufeinander beziehen – das ist eigentlich das Wichtigste in der Interaktion.

VRE: ... Also quasi immer wenn ich als Mensch eine Anfrage stelle bzw. erzähle, dann will ich und dann ist es wichtig, dass der Roboter schnell und inhaltlich korrekt im Zusammenhang antwortet …

FIS: Genau, das ist genau das, was ich unter Responsivität verstehe, dass es erstens ein Timing-Issue ist, also du kannst das nicht fünf Sekunden später machen und, dass es eben angemessen sein muss für die jeweiligen kommunikativen Zwecke …

VRE: … Also im Prinzip kann die Antwort noch so perfekt sein, wenn es 30 Sekunden braucht, bis mir der Roboter antwortet, dann ist die Kommunikation völlig unintuitiv und man wird nicht zufrieden sein …

FIS: … Genauso wenn der emotionale Ausdruck fünf Minuten später kommt. Stell dir vor, du erzählst irgendwie aufgeregt, was irgendein A****loch zu dir gesagt hat – und der Roboter sagt erstmal: Das ist eine sehr interessante Geschichte; und dann: Oh nein, das ist ja fürchterlich; das ist dann zu spät, das muss direkt in der Situation kommen … du teilst ja quasi die Emotion in der Situation und deswegen muss das einfach schnell und angemessen kommen.

VRE: … Auf die Erfahrung aufbauend, wenn man jetzt eine derartige Interaktion designed, auch hinsichtlich Emotionserkennung oder generell, was sind da so die klassischen Anfänger:innenfehler, die man macht oder hast du da in deiner Erfahrung irgendwelche Sachen, wo du sagst: Ok, das war jetzt der Learning-Output von irgendwelchen Projekten, Studien, … da sind wir das falsch angegangen, das sollte man in Zukunft auf keinen Fall so angehen?

FIS: Naja, in dem 2019er Paper, von dem ich jetzt gesprochen habe; da haben wir ja die anderen Arbeiten zu Emotion in Human-Robot-Interaction reviewed und festgestellt, dass die fast alle irgendwelche emotionalen States vorher festlegen und dann sagen: The Robot is happy; und so funktioniert das in der Interaktion aber nicht … in der Interaktion werden viele Emotionen hintereinander benutzt und nicht ein State und die werden zu interaktiven Zwecken eingesetzt … an manchen Stellen ist man auch einfach sozial gezwungen, Emotion zu zeigen. Also, wenn irgendetwas schief geht, dann muss man sagen: Oh, das tut mir leid; egal ob es einer:m jetzt wirklich leid tut.

VRE: … Also die gesellschaftlichen Konventionen müssen vom Roboter auch erfüllt werden …

FIS: … Genau. Und da gibt es auch Konventionen hinsichtlich des emotionalen Ausdrucks und wenn der Roboter die nicht "matched", dann wird er einfach als unangenehm wahrgenommen werden.

VRE: Wenn man jetzt die Kommunikation als Ganzes betrachtet; es gibt ja einige Sachen, wo wir vermutlich schon relativ gut sind mittlerweile, bspw. gesprochenen Text dann wirklich zu erkennen und zu parsen, was der Inhalt ist, wird vielleicht leichter sein als die implizit encodierten Emotionen darin festzustellen, da braucht man dann vielleicht Zusatz-Features wie Eye-Tracking und Stimmlage und all die möglichen Attribute, wie der Mensch Emotionen ausdrücken kann. Meine Frage: Was sind da deiner Meinung nach noch die größten Hürden und offenen Fragen, bis wir wirklich sagen können: Ok, jetzt haben wir Assistive Roboter, die massentauglich sind, jetzt sind die Leute im Großen und Ganzen so zufrieden, dass wir sagen: Jetzt ist das in unserer Gesellschaft angekommen?

FIS: Tatsächlich ist es umgekehrt. Was es schon relativ gut gibt, ist eine Emotionserkennung aus den Sprachsignalen, weil das einfach generell ist und man das anhand von tausenden Daten testen kann. Also z. B. an einer Stimme zu erkennen, ob jemand ärgerlich ist, also [an] den prosodischen Parametern, das gibt es schon viele Jahre. Das ist weniger das Problem als tatsächlich die Semantik herauszufinden, weil wir Menschen einfach unglaublich flexibel sind, was das Verstehen [betrifft] – also sprachliche Ressourcen sind einfach begrenzt, und wir schaffen es trotzdem, mit diesen sprachlich begrenzten Ressourcen durch Kontextabhängigkeit, durch unsere Kreativität, immer neue Inhalte auszudrücken. Also gerade das [computationale] Sprachverstehen, das tut's überhaupt noch nicht … und deswegen ist es sehr, sehr schwer für Computer zu wissen, an welchen Stellen sie welche Emotion vorbringen müssen, weil sie einfach nicht verstehen, was jetzt gerade gesagt wird.

VRE: Generell ist es wahrscheinlich so, also das ist jetzt meine naive Annahme, da ist es wahrscheinlich besser, wenn der Roboter als völlig emotionsloses Wesen etwas antwortet, als wenn er dann mit einer falschen Emotion etwas antwortet … also wenn ich jetzt erzähle: Oh, ich bin traurig … und er dann mit einem Smiley-Gesicht antwortet: Ja, ich bin auch glücklich oder irgend so etwas … Ich weiß nicht, hast du da irgendwelche Daten oder Erfahrungen aus deiner Forschungsarbeit bis jetzt? Welchen Negativeffekt falsch erkannte Emotionen haben können auf die generelle Interaktion?

FIS: Ich glaube nicht, dass das große interaktionale Probleme auslöst, wenn der Roboter falsch liegt, aber was die Leute da vorziehen, also gar keine Emotion oder falsche Emotion, oder schlecht getimte Emotion … das wäre tatsächlich 1. eine empirische Frage und 2. würde ich davon ausgehen, dass da sehr viel interpersonale Variation besteht, also dass es wirklich darauf ankommt, was die Leute von so einem Roboter erwarten oder in welchem Ausmaß die den als sozialen Agenten betrachten … denn es gibt ja viele Leute, immer noch um die 25%, die den Roboter eben nicht als sozialen Agenten ansehen, sondern als Maschine und für die ist das wahrscheinlich angenehm, wenn der Roboter nicht vorgibt, irgendwelche Emotionen zu haben, während die, die den Roboter behandeln als ob es ein Mensch wäre, für die ist es wahrscheinlich seltsam, wenn der Roboter mit einem versteinerten Plastikgesicht durch die Gegend fährt, ohne irgendetwas zu tun.

VRE: Das war jetzt der allgemeine Teil, Dankeschön schon einmal. Jetzt habe ich mir angesehen, du hast dich ja mit Sprachen und Dialekten auseinandergesetzt und dich, wenn ich das richtig verstanden habe, auch darauf fokussiert, inwieweit die Sprecher:innen anders wahrgenommen werden, wenn die Sprachmelodie auf andere Sprachen übertragen wird. Habe ich das jetzt ungefähr …

FIS: Nein, also wenn man jetzt in einer Fremdsprache spricht, dann passiert es sehr häufig, dass man die Intonationskonturen der Muttersprache in die andere Sprache mit-hinein nimmt, also, wenn man z. B. eine Frage stellt, wie es im Deutschen üblich ist, und dann aber Dänisch versucht zu sprechen, also so eine Frage stellt: Wie alt bist du? – auf Dänisch. Da sagt man “Hvor gammel er du?” und das ist flach wie Holland … und auf Deutsch fragt man ja [andere Betonung]: Wie alt bist du? und geht mit der Stimme hoch und wenn man jetzt sagt “Hvor gammel er du?”, dann hat das einfach interaktionale Konsequenzen.

VRE: Inwiefern bzw. kann man daraus etwas für die Mensch-Roboter-Interaktion spezifisch mitnehmen … ?

FIS: Auf jeden Fall, dass Intonation einen riesigen Einfluss hat, das haben wir jetzt in mehreren Studien gezeigt, dass die Art und Weise wie gesprochen wird – und das betrifft dann auch den emotionalen Ausdruck – einen extremen Einfluss darauf hat, wie der Roboter wahrgenommen wird. In einem unserer letzten Experimente haben wir gezeigt, dass die Sprechweise des Roboters dazu führt, dass Teams besser zusammenarbeiten und kreativer sind, also wirklich handfest bessere Resultate liefern und dabei hatte der Roboter wirklich nur 2x2-minütige Statements abgeliefert in der einen oder anderen Sprechweise; und in der anderen Studie haben wir jetzt gerade gezeigt, dass so eine Instruktion des Roboters, die auch irgendwie nur zwei Minuten lang war, [in] signifikant besseren Ergebnissen in einer komplett unabhängigen Aufgabe resultiert hat; also das hat einen massiven Einfluss darauf, wie der Roboter wahrgenommen wird, aber auch wie die Menschen nachher funktionieren, die in solche Situationen gebracht werden.

VRE: Weil wir ja jetzt bei Stimmlage und Tonhöhe sind … Wie weit sind wir denn da deiner Meinung nach schon, dass wir eine zufriedenstellende computergenerierte Stimme haben, also wenn der Computer einen Text vorliest, der nicht vorher von einer:m Sprecher:in aufgenommen wurde. Ist das schon wirklich nah daran, deiner Erfahrung nach … ?

FIS: Die Systeme haben sich die letzten Jahre massiv verbessert, das muss man einfach sagen. Die klingen ziemlich natürlich, also nicht schrecklich wie … *[Referenz auf Hinweisstimme, dass das Meeting nun aufzeichnet wird]* … aber die sind halt alle trainiert oder ausgerichtet auf geschriebene Sprache und … also auf das Vorlesen geschriebener Sprache und nicht auf Interaktion und das merkt man prosodisch einfach schon massiv, dass man einfach viele Sachen ganz anders machen würde, wenn man sie eben jetzt spontan aussprechen würde.

VRE: Vor allem, weil ich nehme mal an, Roboter werden jetzt nicht irgendwelche Füllwörter wie ähm, äh, was im Sprechen einfach ganz normal ist, weil wenn jemand mit mir Schriftsprache redet und dann auch keine Pausen macht oder so, dann ist das ja schon einmal etwas ganz anderes, das ist dann ja so, als ob eine Geschichte vorgelesen würde …

FIS: Genau, und das ist halt auch ein anderer, weiterer Faktor …

VRE: Dass, es da vielleicht sogar Sinn machen würde, intelligent Pausen und Füllwörter einzubauen, wenn man näher an die Mensch-Mensch-Interaktion herankommen wollte …

FIS: Tun wir jetzt auch, da haben wir schon ein Experiment gemacht und gezeigt, dass das einen positiven Einfluss hat. Google Duplex macht das zum Beispiel auch, also damit die sich wirklich natürlichsprachlich anhören, bauen die wirklich Füllwörter ein – beeindruckend.

VRE: Ja, das finde ich auch spannend, weil dann denkt man sich: Okay, man will eigentlich … normalerweise ist es ja so, dass man mit der Technik die Vollkommenheit anstrebt, die der Mensch gar nicht erreichen kann … und hier baut man dann wieder bewusst die Unvollkommenheit des Menschen ein, damit man sich bei der Interaktion dann besser fühlt …

FIS: Ja, aber das hat ja nichts mit Unvollkommenheit zu tun, das sind ja auch alles Strategien aus der Interaktion, die dann auch ihre Funktion haben.

VRE: … Was jetzt auch eigentlich schon meine letzte Frage, von denen, die ich vorbereitet habe, [ist] und zwar: Die Astrid hat mir gesagt, du solltest bzw. könntest dich damit auskennen, so – in a nutshell – heruntergebrochen, was sind denn so diese strukturellen Unterschiede, die man zwischen Englisch, und gerade auch Deutsch, berücksichtigen muss, wenn man jetzt Emotionen erkennen will oder semantisch arbeiten will, mit Texten?

FIS: Also semantisch; also lexikalische natürlich, man benutzt einfach andere Worte im Englischen als im Deutschen; semantisch glaube ich eher nicht, es gibt vielleicht unterschiedliche pragmatische Strategien, das stiff-upper-lip, dass das schon so [einen] Namen hat, zeigt ja, dass es da interkulturelle Unterschiede in der emotionalen bzw. der Erwartung hinsichtlich des emotionalen Ausdrucks gibt. Das haben wir tatsächlich in unseren Studien auch gefunden. Also jetzt als Beispiel eine Studie von 2019, dass die Dän:inn:en z. B. nicht … wenn irgendwas schief geht, wofür sie nichts können, da sagt man auf Deutsch, z. B., trotzdem: Das tut mir leid. Also: Tut mir leid, der Bus ist jetzt weg; auch wenn man selber nicht der:die Busfahrer:in war, der:die jetzt weggefahren ist, und das macht man auf Dänisch einfach nicht. Also da gibt es bspw. unterschiedliche Erwartungen an den emotionalen Ausdruck; und jetzt zu dem sprachlichen Verfahren das zu tun, da gibt es auch unterschiedliche Erwartungen. Also grundsätzlich sind zum Beispiel die Konturen im Englisch, also der range, der prosodische, ist viel größer … Also z. B. um jetzt Pleasure auszudrücken, dass einem etwas gut gefällt: Oh, dear, that is lovely. … Man sagt auf Deutsch ja nicht (stimmlich wie Englisch): Oh, das ist ja wundervoll … Das macht man nicht. Also die grundsätzlichen Erwartungen an den emotionalen Ausdruck sind andere …

[Dank und Debriefing, inhaltlich nicht relevant]

Interviewprotokoll

Prof. Barbara Kühnlenz

Erklärungen

VRE: … Rafael Vrecar

KUE: … Prof. Barbara Kühnlenz

Das Interview-Protokoll wurde in zwecks Einheitlichkeit durchgängig geschlechtergerechter Sprache abgefasst, auch, falls einzelne Passagen im Interview selbst nicht in einer solchen Sprache gesagt wurden.

Einleitung

VRE: Mein Name ist Rafael Vrecar und ich studiere an der TU Wien "Media and Human Centered Computing" (Master). Mir fehlt nun noch die Diplomarbeit – gleich noch einmal vielen Dank, ich bin mir sicher, dass die Einsichten, die ich durch das Interview gewinne, einen großen Beitrag zu meiner Diplomarbeit leisten werden. Im Rahmen dieser möchte ich untersuchen, inwiefern sich Emotionserkennung in Mensch-Roboter-Konversationen auf die Nutzer:innen-Zufriedenheit auswirkt. Dabei möchte ich meinen Fokus stark auf die deutsche Sprache legen, denn zur englischen findet man – wenig überraschend – deutlich mehr [Anm. Material]. Der Fokus ist auf die semantische Emotionserkennung – also auch in Richtung "Sentiment Analyse" gelegt – ob da eine einfache Implementierung, die auch relativ wenig Rechenaufwand auf dem Q.bo One notwendig macht, möglich ist.

KUE: Es geht als auch viel um Wortwahl? [VRE bejaht]. Das ist ein wenig anders, als mein Fokus bisher war, aber ich hoffe, es ist trotzdem spannend. [VRE erklärt, dass das Abstecken des Research-Landscapes ohnehin ein Ziel dieses Interviews ist und vergewissert sich, ob es keine Rückfragen mehr gibt. KUE verneint dies und erklärt, die Wichtigkeit, das Protokoll/Transkript zu erhalten, um dieses auch zu autorisieren. Nach kurzer Rücksprache wird festgelegt, dass das gesamte Transkript und nicht nur einzelne Passagen übermittelt werden.]

Emotionen in bisheriger Forschungstätigkeit

VRE: Inwiefern – in welcher Form auch immer – hat Emotion in Ihrer bisherigen Forschungstätigkeit eine Rolle gespielt?

KUE: Durchaus keine nicht ganz unwesentliche. Ich habe vor allem mit Emotionen in der Mimik – also Gesichtserkennung – gearbeitet, mit Facial-Action-Units. Dann haben wir auch die Sprache zur Emotionserkennung und zwar nicht im semantischen Sinne [Anm. wie Diplomarbeit von VRE], sondern in der Prosodie. [Rückfrage seitens VRE zum Verständnis von Prosodie] – Betonung, Pitches, auf den entsprechenden Frequenzen. Dabei haben wir den Ansatz der emotionalen Anpassung verfolgt, in den Mensch-Roboter-Interaktionen. Wir haben Emotionserkennung in

Gesichtsausdrücken mittels dieser Facial-Action-Units verwendet, zur emotionalen Anpassung an den:die User:in, vor allem in der aufgabenbezogenen Mensch-Roboter-Interaktion, da wir die These hatten, dass wir immer einen kognitiven Aufgabenkontext haben, mit kognitiven Zielen, die zu erreichen sind, aber gleichzeitig läuft parallel auch immer die soziale Ebene mit. Da gibt es soziale Unterziele, die man erreichen muss, um in manchen Kontexten beispielsweise etwas Hilfsbereitschaft erzeugen zu können. Der Roboter kann zum Beispiel eine Brücke für eine:n andere:n User:in sein, indem Bring- und Holdienste erledigt werden, was zum Beispiel unser Kontext im EU-Projekt damals war und damit wirkt der:die User:in quasi als Brücke zu einem:einer anderen User:in und damit braucht es eben das soziale Ziel der Hilfsbereitschaft. Damals haben wir in einem ersten Schritt mit einem Züricher Lächelmodell, wie das damals hieß, eine signifikante höhere situative Empathie dadurch erzeugt, dass wir aus diesen Facial-Action-Units die Emotionen erkannt haben und der Roboter sich darauf eingestellt hat, also beispielsweise ein Lächeln im richtigen Moment erwidert hat. Dadurch kam es zu einer höheren situativen Empathie und auch auf anderen User Experience Dimensionen zu höheren Werten in der User:innen-Bewertung, in den subjektiven Fragebögen, die dann ausgefüllt wurden. Dann sind wir noch einen Schritt weitergegangen, weil wir einen sozialpsychologischen Ansatz entwickelt haben, der nach der Empathie-Altruismus-Hypothese besagt, dass bei Menschen mehr hilfreiches Verhalten auftritt, wenn wir eine Ähnlichkeit empfinden, zur hilfsbedürftigen Person. Wir haben das dann auf den Roboter übertragen, um festzustellen, ob es dort auch funktioniert, und haben dabei die emotionale Anpassung als Ähnlichkeitsdimension benutzt. Das haben wir auf zweierlei Art gemacht: einmal implizit durch Anpassung des Emotionsausdrucks, einmal wieder durch die Gesichtsmimik, wie vorher schon erprobt, und dann eben auch noch durch die Erkennung der Prosodie, wie wir gerade eben schon besprochen haben. Dann hat der Roboter das in einem ersten Smalltalk so erkannt und sich dann entsprechend des Erfolgs in der gemeinsamen Aufgabe aber eben auch entsprechend der gezeigten Emotion der:des User:in in Prosodie und Gesichtsausdruck an diese:n angepasst. Die zweite Variable war dann explizite Anpassung [KUE weist auf mögliche Relevanz hin], also die erkannte Emotion in einem Dialog absichern, sodass man trotzdem im Smalltalk fragt "Wie fühlen Sie sich gerade?" und die Ähnlichkeit dann dadurch erzeugt, dass man dann sagt "Mensch, mir geht es heute genauso!", "Ich fühle mich auch so lala", bzw. "ziemlich gut" [Anm. Beispielantworten]. Dadurch konnten wir signifikant erhöhte Hilfsbereitschaft erzielen und auch erhöhte Werte in Animacy und Anthropomorphismus.

[Es folgt eine Besprechung bzgl. Literaturempfehlungen, die für das Transkript nicht relevant ist.]

Erkenntnisse & Tipps

VRE: Wenn Sie sich Ihre ganze Forschungstätigkeit bzgl. Mensch-Roboter-Interaktion ansehen. Was wäre der wichtigste Output, den Sie einer laien Person mitgeben würden (in ein paar Sätzen oder als einzelne Phrase o. Ä.)?

KUE: Der wichtigste Output ist die Erkenntnis, dass es in jeder Interaktion nicht nur die kognitiven streng inhaltlichen Ziele gibt, sondern auch soziale Ziele, die es zu erreichen gibt, und dass wir uns in diesem Bereich, wenn es um soziale Kommunikationsziele geht, Emotionen durchaus zu nutze machen können – im pro-sozialen Sinne natürlich, in der Mensch-Roboter-Interaktion und hier sind zusammengefasst die wichtigsten Erkenntnisse, dass eine erfolgreiche Emotionserkennung vor allem dafür sorgt, dass wir diese pro-sozialen Effekte erzielen können, dadurch dass wir uns auf den:die User:in einstellen können – aber nicht nur das. Wenn diese Emotionserkennung erfolgreich ist – so hat sich bei unseren Experimenten gezeigt – dann führt das zu einer gewissen Fehlertoleranz. Wir haben es bei der Technik immer mit Fehlern zu tun – fehlerhafte Spracherkennung, Hintergrundlärm, sonstige Störvariablen. Wenn wir eine erfolgreiche emotionale Anpassung durchführen können, durch eine gute Emotionserkennung, dann macht das den:die User:in fehlertoleranter gegenüber dem Roboter.

VRE: Das heißt, wenn der Roboter dann ein paar kleine "Aussetzer" hat …

KUE: Als Beispiel bei uns war das dann in diesen Real-World-Experimenten in der Münchner Innenstadt, wo wir wahnsinnig viel Hintergrundlärm hatten, da hat natürlich die Spracherkennung nicht so gut funktioniert wie im Labor. Hier hatten wir dann nicht nur diese erhöhte situative Empathie und Hilfsbereitschaft, oder eben auch auf andere Dimensionen der User Experience, sondern dass es diese schlechte Spracherkennung in diesen "noisy environments" kompensieren kann und zwar wird das in der User Experience sichtbar.

VRE: Sie haben vorher schon quasi auf meine nächste Frage übergeleitet. Was sind so die klassischen Tipps bzw. die klassischen Anfänger:innen-Fehler, die man vermeiden sollte, wenn man sich mit diesem Forschungsfeld beschäftigt, oder haben Sie da auch Erkenntnisse wie: Oh, das würde ich jetzt im Nachhinein anders machen, verglichen mit vor ein paar Jahren?

KUE: Auf jeden Fall Störungen in der Erkennungsleistung einkalkulieren. Man probiert ja meistens die Sachen vorher im Labor aus. Man muss sich einfach darüber bewusst sein, dass das draußen in der Welt alles nicht ganz so gut funktionieren kann – oder eben auch, dass es einen gewissen Tageseinfluss gibt. Man probiert es an einem ruhigen Tag am Lehrstuhl aus und führt dann die Experimente einfach durch, wenn ein bisschen mehr los ist und dann hat man mehr Hintergrund-Geräusche und die Modelle sind dann nicht ganz so robust in der Erkennungsleistung. Also eben wirklich die Störungen einkalkulieren – das gilt für alle Modalitäten, sei es eben für die Sprache, oder auch für die Mimik. Immer darauf achten, dass das ja auch in der realen Welt optimalerweise funktionieren soll und darauf aufbauend eben auch die Erkenntnis, dass man sich nie auf die Emotionserkennung alleine verlassen sollte und das Ganze auch immer noch durch einen expliziten Dialog unterstützen oder absichern, durch eine Nachfrage des Roboters …

VRE: … also als Faustregel: Lieber einmal zu oft nachfragen als einmal zu wenig? …

KUE: Da gibt es natürlich auch gewisse kritische Punkte. Ich habe mich eine Weile auch mit "Miscommunication Handling" beschäftigt und da hat man natürlich auch wieder gesehen, dass ab einem gewissen Punkt beim Nachfragen, um eben diese potenziellen Missverständnisse zu behandeln, auch kritische Punkte überschritten werden können: "over-usage". Auch das musste man natürlich ein bisschen mit-evaluieren; dann natürlich auch immer ein gewisses Abwägen von Aufwand und Nutzen. Wir dürfen nicht vergessen, dass die ganzen Erkennungsleistungen auch ein hoher Programmieraufwand sind. Da hat sich aus meiner Erfahrung gezeigt, dass es nicht immer notwendig ist, man muss sich da immer den Aufgabenkontext ein bisschen anschauen: Welche Effekte will ich für welchen Aufgabenkontext wirklich erzielen? Wir müssen nicht immer alles anwenden, was wir in-petto haben. Es lohnt sich auch in den Studien zu schauen: Was ist minimal notwendig für unsere Effekte, die wir brauchen? … um dann eben Aufwand und Nutzen ein wenig abzuwägen. Als Beispiel haben wir einmal festgestellt, dass es zu unglaublich guten Ergebnissen geführt hat, wenn wir bzgl. Prosodie und Mimik alles erkannt und angepasst haben, manchmal hat aber auch der reduzierte Kontext durchaus gereicht, dass man das Ganze nur durch die verbale Interaktion löst, durch die Frage: Wie geht es Ihnen? Mir geht es auch so und so – ohne, dass die absolut robuste Emotionserkennung da jetzt dahinter laufen musste. Wenn es um soziale Effekte geht, die man evozieren möchte, dann reicht manchmal auch die soziale Interaktion mit dem Roboter, das fand ich auch ganz interessant und sich eben dann auch immer den Aufgabenkontext anzusehen, aber auch die Benutzer:innen-Hintergründe und welche Erwartungen die mitbringen, das kann nämlich auch durchaus unterschiedlich sein, je nach Benutzer:innen-Hintergrund – und dann, um auch noch einmal auf den Aufgabenkontext einzugehen: Je nachdem, in welchem Kontext wir die HRI [Anm. Human-Robot Interaction] haben, gibt es andere soziale Effekte oder sind vielleicht auch nicht immer soziale Effekte zielführend. Ein Beispiel wären hier die kollaborativen Industrie 4.0 Anwendungen. Da hat sich in unseren Forschungen auch gezeigt, dass es unterschiedliche Bedarfe geben kann. Zum Beispiel eben bei der wirklich kollaborativen Interaktion, wo man manchmal auch inkongruente Aufgaben zueinander erfüllen muss, könnte ein zu anthropomorphes Design oder zu viele soziale Hinweisreize manchmal auch störend sein, weil sich ja Mensch und Roboter automatisch aneinander anpassen und nicht bei jeder Anwendung – gerade im industriellen Kontext – ist das auch erwünscht. So eben eher bei kongruenten oder anleitenden Anwendungen, wo wirklich die Anpassungseffekte, soziale Effekte, erwünscht sind, da muss das dann eher ein Ziel für uns sein.

VRE: Wenn ich das jetzt richtig verstehe, um es salopp zu formulieren: Wenn man mit dem T-Mobile/Telekom/… Chatbot chattet, dann will man nicht zuerst hundertmal gefragt werden, wie es so geht … und mir geht es auch so … sondern da will man einfach die Seite finden, wo man den Handy-Vertrag ändern kann … das ist ja auch diese Industrie-Anwendung …

KUE: Zum Beispiel, ja. Das ist auch ein gutes Beispiel.

Offene Fragen des Forschungsfelds

VRE: Gut, dann habe ich das richtig verstanden. Generell bzgl. Emotionserkennung: Was sind Ihrer Meinung nach – auch vielleicht breiter im Forschungsfeld – die offenen Forschungsfragen oder Gebiete, wo es noch nicht so weit ist mit den Untersuchungen, also wo man viel Bedarf hätte?

KUE: Ich denke, dass in Experimenten viele Effekte, auch durch das Novum des Experiments selbst, entstehen und man durchaus auch genauer hinschauen kann: Was machen Gewöhnungseffekte über einen längeren Zeitraum mit den Interaktionen? Wo sind die Grenzen dieser emotionalen Interaktion? Was macht das auch mit dem Individuum und der Gesellschaft? Also ich finde das sehr wichtig, das in den breiteren Kontext einzuordnen. Also das immer mehr noch einzuordnen, was es wirklich für Effekte gibt – positive wie negative. Man muss das Ganze natürlich auch kritisch sehen und schauen, wo wollen wir auch Grenzen sehen in der emotionalen Dimension der Interaktion und diese vielleicht auch bewusst immer wieder setzen. Da hat dann auch wieder ein Abwägen von Nutzen und aber auch vielleicht Gefahren/Chancen, also die Gefahr der Entgrenzung zwischen Mensch und Roboter, zum Beispiel. Da kann man gar nicht genug forschen.

VRE: … ja, weil im Prinzip auch im Wandel der Zeit werden sich immer wieder ganz neue gesellschaftliche Situationen ergeben. Wenn man denkt, die Human-Robot Interaction vor der Pandemie und jetzt während der Pandemie, das sind ja auch wieder ganz neue Anforderungen, das ist mir zum Beispiel gerade eingefallen …

[Kurzes Gespräch zur INTERACT 2021, das inhaltlich für diese Arbeit nicht relevant ist]

KUE: … absolut, die gesellschaftlichen Anforderungen verändern sich tatsächlich und besonders ist die Pandemie eben ein gutes Beispiel dafür.

[Kurzes Gespräch zur Pandemie, das inhaltlich für diese Arbeit nicht relevant ist]

Hürden für Kommunikation & Wichtigkeit von Emotion

VRE: … Was sind denn Ihrer Meinung nach die größten Hürden, um einen erfolgreichen Kommunikationsfluss zwischen Roboter und Mensch sicherzustellen? Ich nehme an, es kommt auch hier wieder auf den Kontext der Interaktion an und die Hürden sind natürlich auch wieder verschieden, aber [wie lauten] diese generell – und auch im emotionalen Kontext, bei doch sozialer Interaktion und nicht nur bei rein "professioneller" Maschineninteraktion. Wie wichtig sind denn hier Emotionen überhaupt?

KUE: Ja, sie sind nicht unwichtig, weil der Mensch einfach ein emotionales Wesen ist und vor allem können wir es, glaube ich, auch gar nicht ausblenden. Jede Kommunikation – man sagt ja so schön: Man kann nicht nicht kommunizieren; nach Watzlawick – und ich glaube, über all dort, wo eine vor allem eine verbale soziale Interaktion stattfindet, stecken ja die Emotionen schon mit drinnen. Das heißt, man kann sie gar nicht negieren – sie werden entstehen, in einem:einer menschlichen User:in; einfach weil wir genauso unsere Muster auf die Mensch-Roboter Interaktion anwenden – und das Ganze bringt uns natürlich auch eine gewisse Natürlichkeit – also insofern würde ich die Wichtigkeit als sehr hoch einschätzen; zumindest die Berücksichtigung dieser Emotionen; ob und inwieweit man sie verwendet und was man mit ihnen macht, ist wieder ganz etwas anderes. Die Wichtigkeit würde ich sehr hoch einschätzen.

VRE: … und generell, die Hürden – wo stolpert man denn am ehesten drüber?

KUE: Ja – wie auch schon vorgegriffen, man stolpert über mangelhafte Spracherkennung, einfach, dass die Technik manchmal noch nicht so funktioniert, oder eben auch fehlerhafte Emotionserkennung ist auch etwas, das passieren kann und womit man umgehen muss. Vor allem in Szenarien, die sich in der realen Welt bewegen gibt es viele dieser Störungen und damit ist die Robustheit dieser Emotionserkennung eine der großen Hürden, die sich aber technisch auch immer mehr verbessern wird, denke ich. Hierbei sind auch die Erwartungen nicht zu vernachlässigen – das hängt natürlich wieder vom Benutzer:innen-Hintergrund und vom ganzen Kontext der Interaktion ab, aber die User:innen-Erwartungen sind auch etwas, das man nicht vergessen darf, die womöglich zur Hürde werden, wenn Uncanny-Valley-Effekte dadurch entstehen, dass ein Roboter vielleicht durch sein Design oder durch den Kontext gewisse Erwartungen bei dem:der User:in hervorruft, die dann nicht erfüllt werden können.

Helpfulness, andere Parameter & Zusammenfassung

VRE: Damit sind wir im Prinzip bei der letzten Frage. Ich habe mir für jedes Gegenüber eine eigene Frage überlegt, wobei das in diesem Fall ohnehin nur mehr ein Rekapitulieren und Zusammenfassen wird, weil wir viele Aspekte davon schon besprochen haben: Wenn ich das richtig verstanden habe, haben Sie sich in der Vergangenheit schon sehr intensiv damit auseinandergesetzt, wie die emotionale Anpassung des Roboters die wahrgenommene Helpfulness erhöht – insofern, wenn eine automatisierte Emotionserkennung, sofern zuverlässig, vorliegt, wird sich das – ich meine, in Sozialwissenschaften ist es generell eine schlechte Idee, Deduktionen anzustellen, aber wenn die emotionale Anpassung dadurch unterstützt wird, legt es nahe, dass dadurch die Helpfulness noch größer wird. Jetzt würde mich aber auch noch interessieren, erstens, welche anderen Parameter neben der Helpfulness es noch gibt – denn das ist ja nicht der einzige Parameter, den man bei so einer Interaktion analysieren kann – die dadurch erhöht oder vielleicht sogar verringert werden, je nachdem; und vor allem, wie es sich auswirkt – vorher haben wir schon über Fehlertoleranz gesprochen – wenn jetzt in irgendeinem Kontext eine Emotion völlig falsch erkannt wird; inwiefern es also besser wäre, den Roboter sozusagen "emotionslos" agieren zu lassen, also neutral, als das Risiko einer Falscherkennung einzugehen, wo dann vielleicht der:die User:in sich vor den Kopf gestoßen fühlt?

KUE: Tatsächlich haben wir ja nicht nur die Helpfulness untersucht, sondern auch viele andere Dimensionen der User-Experience. An dieser Stelle auch noch ein kleiner Nachtrag an die Tipps: Es ist immer gut auch zu mischen, also zwischen objektiven und subjektiven Maßen. Man kann physiologische Messungen anstellen, z. B. Hautleitwert, Herzratenvariabilität, wie sich das Ganze auch auf die mentale Belastung auswirkt von dem:der User:in. Hier haben wir z. B. ein behavioristisches Maß der Helpfulness gehabt, denn was macht der:die User:in wirklich. Denn man kann zum einen die Hilfsbereitschaft abfragen: Wie groß ist die Hilfsbereitschaft? – das heißt aber noch nicht, dass es der:die User:in auch wirklich macht, denn da kommen dann wieder auch andere Variablen auch noch zum Zug, inwieweit das umgesetzt wird. Beides konnten wir aber positiv bewerten und hatten dabei signifikante Erhöhungen, auch im Verhalten der User:innen. Wenn wir jetzt noch einmal auf die Fragebögen schauen, auf anderen User-Experience-Dimensionen gab es auch signifikante Erhöhungen, wie laut Godspeed die Dimensionen Animacy, Anthropomorphismus – die Heerink Dimensionen haben wir auch abgefragt, z. B. Trust, Sociability, soziale Präsenz, oder eben auch einfach Enjoyment während der Aktion und womöglich auch eine erhöhte Intention to use, so heißen diese Dimensionen laut Heerink; dann eben auch die Fehlertoleranz, die wir schon angesprochen haben. Das leitet auch schon dazu über: Was ist eben, wenn das Ganze einmal nicht so gut funktioniert? Das war der zweite Teil Ihrer Frage. Dann sieht man schon, dass diese sozialpsychologischen Effekte, die man da anvisiert hat, womöglich nicht funktionieren. Gerade, weil die Interaktion dann ja vielleicht auch darauf ausgelegt ist, der:die User:in merkt, in welche Richtung das Ganze geht, man hat die soziale Dimension schon eingeführt, und wenn die dann nicht funktioniert, haben wir einen gewissen Uncanny-Valley-Effekt: Wir haben gewisse Erwartungen hervorgerufen, vielleicht auch durch ein soziales Design des Roboters, die wir dann eben nicht komplett erfüllen können; und dann stürzen natürlich auch entsprechend diese User-Experience-Dimensionen ab, ganz zu schweigen davon, dass die sozialen Effekte, die erwünscht waren, vielleicht nicht funktionieren, und zusätzlich gibt es vielleicht auch die Gefahr von

Missverständnissen, weil auch die finden oft schon auf einer sozialen Dimension ihre Wurzeln; oder wenn eben auch eine Industrie 4.0 Anwendung, als Beispiel, mit sozialen Effekten angereichert wird, könnte sich das in so einem Fall sogar auf geringere Effektivität oder geringere Produktivität in der Kollaboration von Mensch und Roboter auswirken. Das hat, muss ich sagen, auch insgesamt immer gut funktioniert, weil wir uns eben nicht nur alleine auf die Emotionserkennung verlassen haben, sondern das immer auch mit expliziten Komponenten abgesichert haben. So hat man auch die soziale Dimension, wenn sie erwünscht ist, auch ein bisschen unabhängig von solchen Störvariablen.

VRE: Das heißt also, salopp gesagt, wenn z. B. so ein Roboter-Arm bei einer Autokonstruktion auf einmal anfängt, seinem:seiner Steuerer:in einen Witz zu erzählen, wird die Produktivität vielleicht nicht so hoch bleiben, aber wenn das in irgendeinem sozialen Kontext ist, wird das vielleicht sogar die Stimmung heben können, wenn der Witz gut ist?

KUE: Ja, das stimmt, oder wenn wir jetzt anleitende Bewegungen in der Kollaboration durch soziale Hinweisreize der Anpassung unterstützen wollen, dann funktioniert das nicht und der:die User:in ist irritiert und muss gegen diese Störvariablen vielleicht sogar ankämpfen, dann ist das ja eine höhere mentale Belastung, das haben unsere Experimente z. B. auch gezeigt, wenn die Erwartungen nicht damit übereinstimmen – gerade im industriellen Kontext hat man an Roboter nicht diese Erwartungen. Wenn der Roboter jetzt einen Witz erzählt, dann stört das einfach, weil das nicht zu den Erwartungen passt. Gerade in solchen Kontexten müssten dann auch die Erwartungen der User:innen entsprechend angepasst bzw. aufgebrochen werden, durch zum Beispiel flankierende Maßnahmen.

[Debriefing und Dank, welche thematisch für diese Arbeit nicht relevant sind]

Expert:innen-Interviews
Sommer 2021
Version 1.1.2

Master's Thesis:
Interviewleitfaden

Rafael Vrecar
Matr. Nr.: 01267765
Studium: 066 935

Increasing User Satisfaction by detecting Emotions in German-language Human-Robot Conversations with Q.bo One

Einleitung

Zunächst danke, dass Sie an meinem Interview teilnehmen. Sie leisten damit einen großen Beitrag, dass meine Diplomarbeit die im Voraus gesetzten Anforderungen erfüllt.

Kurz zu meiner Person: Mein Name ist Rafael Vrecar und mir fehlt de facto "nur noch" die Diplomarbeit zum Abschluss meines Studiums "MSc Media and Human-Centered Computing" an der TU Wien.

Im Rahmen ebendieser Diplomarbeit möchte ich untersuchen, inwiefern sich Emotionserkennung in Mensch-Roboter-Konversationen auf die Nutzer:innen-Zufriedenheit auswirkt. Mein Fokus liegt dabei auf der deutschen Sprache. Interessant ist für mich also insbesondere, inwiefern sich die deutsche Sprache hier von den Methoden für englische Sprache unterscheidet (wenig überraschend fand ich zur englischen Sprache die meiste Literatur) und inwiefern es überhaupt möglich ist, eine "generische" Lösung zu finden.

Haben Sie vorab irgendwelche Fragen zum Interview oder zur Einverständniserklärung?

Fragen an alle

1. Inwiefern hat Emotionserkennung in Ihrer bisherigen Forschungstätigkeit eine Rolle gespielt?
(4-6 Minuten)
2. Was ist der wichtigste Output, den Sie aus Ihrer Arbeit mit Robotern und speziell in der Kommunikation zwischen Mensch und Roboter aus Ihrer bisherigen Tätigkeit mitnehmen?
(3-5 Minuten)
3. In weiterer Folge: Welche Tipps können Sie in diesem Zusammenhang geben, um klassische Anfänger:innen-Fehler zu vermeiden? Insbesondere: Was sind, Ihrer Meinung nach, die offenen Forschungsfragen, die in diesem Zusammenhang untersuchenswert sind?
(2-4 Minuten)
4. Was sind Ihrer Erfahrung zu Folge die größten Hürden, einen zufriedenstellenden Kommunikationsfluss zwischen Mensch und Roboter zu ermöglichen? Inwiefern spielen Emotionen hierbei eine Rolle?
(3-5 Minuten)

Datum: 29. Juli 2021 Seite: 1/2

Personenspezifische Fragen

Barbara Kühnlenz

Sie haben sich in der Vergangenheit bereits damit auseinandergesetzt, inwiefern die wahrgenommene "helpfulness" von Robotern dadurch erhöht wird, dass sie sich an die Emotionen des menschlichen Gegenübers anpassen[1]. Dementsprechend liegt es nahe, dass sich die automatisierte Emotionserkennung, sofern zuverlässig, ebenfalls positiv auf die "helpfulness" auswirkt. Für welche anderen Parameter neben "helpfulness" kann sie aber noch von Bedeutung sein und vor allem, inwiefern sind negative Effekte denkbar, wenn Emotionen falsch/ungenau erkannt werden und dementsprechend auch die "Mood-Adaption" fehlschlägt?
(5-12 Minuten)

[1] https://mediatum.ub.tum.de/doc/1216553/file.pdf

Wenn Sie ein Übersetzung für dieses Formular wünschen, bitte kontaktieren Sie mich (siehe unten angeführte E-Mail-Adresse).
(above: details for request for translation into German)

TU WIEN Informatics

Declaration of Consent

The following declaration is valid for the interview regarding my (Rafael Vrecar's) diploma thesis. Please read this carefully before eventually declaring your consent and contact me at any time if you have any sort of questions. Your trust is very important to me.

In the context of my diploma thesis with the title "Increasing User Satisfaction by detecting Emotions in German-language Human-Robot Conversations with Q.bo One" at TU Wien, I conduct several expert interviews for getting a better overview of the topic emotion detection in dialogue-based Human-Robot Interaction. Within the course of the interview, it is also possible that the focus might change a bit as I plan on rather loosely structuring it. Regardless of this, you will not be required to share any sensitive data within the course of the whole interview.

The interview will take about 25 minutes but 40 minutes at most and will be held via Zoom.
(If you prefer another platform, please let me know.)

I would like to record the interview with video and sound by using the Zoom functionality, as I plan to transcribe the interview afterwards. Moreover, I plan on taking notes during the interview.

The recording will only be stored locally on my device (operating system: macOS) and will not be published. However, the storage of the recording by Zoom cannot be controlled by me but I will not trigger any options to store it via Zoom. Moreover, it is intended that your name, professional title and affiliation are to be used for direct quotes in the thesis and associated academic publications. However, you below can request to receive a preview (which you can then reject) before declaring your consent for publishing them.

Of course you can stop the interview at any point without giving any reasons. Moreover, you can retract your declaration of consent partially or completely within five Austrian business days after the interview.

You may contact me via e-mail or by phone at any time:
Rafael Vrecar (rafael.vrecar@tuwien.ac.at, +43 664 912 73 24)

My thesis is supervised by *Prof. Peter Purgathofer (peter.purgathofer@tuwien.ac.at)* and *Dr. Astrid Weiss (astrid.weiss@tuwien.ac.at)*.

Research Institution:
TU Wien: Human-Computer Interaction (E193/5)
sekretariat@igw.tuwien.ac.at
+43 (01) 58801-18703
Argentinierstraße 8
AT-1040 Vienna

Interviewer: *Rafael Vrecar, BSc*
(signed in Vienna on 26th July 2021)

....

Declaration of consent: With the following signature, I, ____ (full name), declare my consent to participate in the interview under the mentioned conditions and agree with the described usage of my data.

Preferred language for the Interview:
[ ] I do not mind being interviewed in German (default case).
[ ] I prefer English over German.

Handling of quotes:
[ ] I want to review the quotes being published.
[ ] Reviewing is not necessary for me.

Location, Date: ______________________________

Signature: ______________________________

Page 1 of 1

A.2 Re: Study with Q.bo One

The survey has been conducted in German. The translation from [Bar] was used.

See the following pages for a printable version of the online(!) conducted survey.

The start page is provided in web view as it was not included in the standardized exported document.

Emotionserkennung in Mensch-Roboter-Konversationen (Jänner 2022)

Liebe:r Teilnehmer:in!

Vielen Dank, dass Sie sich Zeit nehmen beim empirischen Teil der Diplomarbeit "Nutzer:innen-Zufriedenheit durch die Erkennung von Emotionen in deutschsprachigen Mensch-Roboter-Konversationen mit Q.bo One erhöhen" einen Beitrag zu leisten! Die Wissenschaft lebt von Freiwilligen wie Ihnen und ist, wie wir die letzten Monate gesehen haben, ein sehr wichtiges Gut unserer Gesellschaft, das es zu beschützen gilt!

Als kleines Dankeschön werden 3x jeweils 30 Euro verlost – mehr dazu am Ende dieser Befragung.

Im Folgenden finden Sie die Datenschutzerklärung inkl. Teilnahmebedingungen, welche Ihrer Zustimmung bedarf. Bitte lesen Sie sich diese sorgfältig durch. Ohne Zustimmung ist eine Teilnahme leider nicht möglich.

Dauer: ca. 10 Minuten

Anforderung(en): keine vorangegangene akademische Arbeit zum Thema

Zustimmungserklärung

Diese Studie soll über eine online Umfrage untersuchen, inwiefern sich Emotionserkennung in deutschsprachigen Mensch-Roboter-Konversationen auf die Nutzer:innen-Zufriedenheit auswirkt.

Sie müssen in dieser Studie nicht(!) direkt mit einem Roboter interagieren, sondern sehen sich lediglich Videos an, auf deren Basis Sie dann einige Fragen Ihre Eindrücke betreffend beantworten. Dafür wird die "Godspeed Questionnaire Series" (https://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/, abgerufen am 20. Dezember 2021) verwendet (ergänzt um ein paar weitere, individuell festgelegte Fragen + "The robot may not notice my discomfort" von Hannibal et al. bzgl. Vorfahrung mit Robotern https://ieeexplore.ieee.org/document/9515513, abgerufen am 27. Dezember 2021).

Sie können die Studie jederzeit abbrechen, indem Sie einfach das Browserfenster bzw. den Tab schließen. Ihre Daten werden dann auch nur anonym ohne zuordnende Informationen gespeichert.

Bei Fragen, Problemen, Wünschen, Beschwerden etc. können Sie sich ***JEDERZEIT(!)*** an mich (Rafael Vrecar, BSc; rafael.vrecar@tuwien.ac.at) bzw. meine Betreuerin (Dr. Astrid Weiss, astrid.weiss@tuwien.ac.at) wenden.

Datenschutzerklärung

Die Lime-Survey-Instanz wird lokal an der TU Wien (Arbeitsgruppe Human-Computer Interaction) betrieben (der Server steht bei uns im Institutsgebäude – Argentinierstraße 8, 2. Stock, 1040 Wien). Es werden keinerlei Daten außer denjenigen, die Sie im Rahmen der Umfrage (+ Zeitstempel der Beantwortung) bei uns dauerhaft, also über die Dauer, die technisch gesehen notwendig ist, hinaus gespeichert.

Bitte beachten Sie ferner, dass wir Ihre Daten anonym behandeln. Die (freiwillige!) Angabe einer E-Mail-Adresse (letztes Fenster) wird zwar mit Ihren Daten verknüpft, allerdings vor der Auswertung explizit entfernt und somit nicht mit Ihren Antworten in Verbindung gebracht.

Zugriff auf besagte Daten hat nur Rafael Vrecar, BSc (Studienleiter) und Dr. Astrid Weiss (Betreuerin) sowie die LimeSurvey-Nutzer:innen der Arbeitsgruppe mit entsprechenden administrativen Zugriffsberechtigungen.

Ich habe die beiliegenden Erklärungen (vmtl. Scrollen notwendig!) gelesen, akzeptiere sie und möchte an der Umfrage teilnehmen. ☐

Weiter

LimeSurvey

Dauer: ca. 10 Minuten

Anforderung(en): keine vorangegangene akademische Arbeit zum Thema

Teil A: Einführung

Ihr Browser unterstützt leider den HTML Video Tag nicht. Bitte rufen Sie folgenden Link auf: https://www.vrecar.eu/robot/qbo-one-video-material/qbo-one-intro.mp4 Bitte sehen Sie sich das Video an. Das Transkript zum Video lautet wie folgt:

Die folgende Studie beschäftigt sich mit der Fragestellung, inwiefern es die Zufriedenheit der Nutzer:innen mit mir beeinflusst, wenn ich die Emotionen in einer Konversation mit Ihnen beachte.

Dazu werde ich Ihnen zunächst ein Szenario vorstellen, in welches Sie sich bitte "hineindenken". Danach wird Ihnen meine Antwort präsentiert und darauf folgend bitte ich um die Beantwortung einiger Fragen zu dem gezeigten Szenario.

Nach diesen folgen noch einige abschließende Fragen.

Zu Beginn möchte ich gleich einige Informationen zu Ihrer Person erheben.

A1. **Wie alt sind Sie (in Jahren)?**

☐☐☐☐☐☐☐☐☐☐

A2. **Teilen Sie mir bitte, wenn Sie möchten, Ihr Geschlecht mit.**

- weiblich ☐
- männlich ☐
- nicht-binär ☐
- Ich möchte lieber keine Angabe machen. ☐
- Sonstiges ☐

Sonstiges

☐

LimeSurvey

A3. Wie lautet Ihre höchste abgeschlossene Ausbildung?

- [ ] Pflichtschule
- [ ] Lehre (ohne Matura)
- [ ] AHS mit Matura
- [ ] BHS mit Matura
- [ ] Lehre mit Matura
- [ ] Bachelorstudium (bzw. 1. Studienabschnitt)
- [ ] Masterstudium bzw. Diplomstudium
- [ ] Doktorat oder höher
- [ ] keine(s) der genannten

A4. Welches Vorwissen (und woher) bzgl. Robotern haben Sie?

- [ ] Ich kenne Roboter aus dem Arbeitsumfeld (Bauen, Programmieren, Forschungsprojekte).
- [ ] Ich kenne Roboter aus der Kultur (Literatur, Filme, Radio, Zeitschriften, TV).
- [ ] Ich kenne Roboter aus meiner Freizeit (Bastelprojekte, wissenschaftliche Zeitschriften, Familie, Freund:innen).
- [ ] Ich kenne Roboter aus der Bildung (Lehrveranstaltungen, Abschlussarbeiten, Praktika).
- [ ] Ich kenne Roboter zufällig (in einem Geschäft, Teilnahme an einer Studie, Veranstaltungen).

A5. Kommentare, Anmerkungen (optional!):

LimeSurvey

Teil B: Vorgestelltes Szenario

Ihr Browser unterstützt leider den HTML Video Tag nicht. Bitte rufen Sie folgenden Link auf: https://www.vrecar.eu/robot/qbo-one-video-material/qbo-one-vorgestelltes_szenario.mp4 Bitte sehen Sie sich das Video an. Das Transkript zum Video lautet wie folgt:

Stellen Sie sich bitte vor, Sie sind bereits etwas älter und leben allein. Ich bin dafür zuständig, dass Sie trotzdem Ansprache haben und jeden Abend vom jeweiligen Tag erzählen, um diesen noch einmal zu rekapitulieren. Dies soll Ihrem Gedächtnis (ein bisschen wie das Schreiben eines Tagebuchs) sowie gegen Einsamkeit helfen.

Ich frage Sie also:

"Wie war Ihr Tag?"

Sie antworten mir:

"Ich bin später aufgestanden als normalerweise – 8 Uhr anstelle von 7 Uhr. Zum Frühstück hatte ich Schinken und Eier und danach habe ich die Zeitung gelesen. Danach wurde das Mittagessen geliefert. Das heutige Menü war Tomatensuppe als Vorspeise und danach ein Omelett. Danach bin ich spazieren gegangen und habe zufällig Josef getroffen. Wir haben über seine Kinder geredet. Gregor heiratet nächste Woche Claudia. Außerdem ist mir aufgefallen, dass die Blätter schon wieder von den Bäumen fallen. Nach den Nachrichten ist jetzt dann wieder Zeit für's Bett."

Was Sie außerdem wissen müssen: Kürzlich haben Sie mir erzählt, dass Sie Claudia nicht mögen, weil sie so "berechnend und falsch" sei und dass Sie sich um Gregor sorgen, weil der immer so "ein lieber Bub" gewesen sei.

Versuchen Sie nun wirklich, sich in die Erzählung der älteren Person hineinzuversetzen und bewerten Sie anschließend, wie gut Sie meinen, dass Ihnen das gelungen ist.

Wenn Sie dies gemacht haben und bereit für meine Antwort auf die gelistete Erzählung sind, drücken Sie bitte auf "Weiter".

B1. Wie gut meinen Sie, sich in das vorgestellte Szenario hinein versetzen zu können?

- gar nicht bis sehr schlecht ☐
- schlecht ☐
- eher schlecht ☐
- mittelmäßig ☐
- eher gut ☐
- gut ☐
- sehr gut ☐
- Ich kann das leider überhaupt nicht beurteilen. ☐

B2. Kommentare, Anmerkungen (optional!):

Teil C: Fragebogen

Ihr Browser unterstützt leider den HTML Video Tag nicht. Bitte rufen Sie folgenden Link auf: https://www.vrecar.eu/robot/qbo-one-video-material/qbo-one-detection.mp4 Bitte sehen Sie sich das Video an. Das Transkript zum Video lautet wie folgt:

Zu aller erst: Sie können sich dieses Video gerne beliebig oft ansehen.

Ich erinnere Sie kurz, dass Sie mir von Ihrer Morgenroutine und dem Mittagessen erzählt haben. Danach haben Sie beim Spazieren Josef getroffen; sein Sohn Gregor, der immer ein lieber Bub gewesen sei, heiratet nächste Woche Claudia, die Sie nicht mögen, weil Sie berechnend und falsch sei. Sie erklärten mir, dass Sie sich deshalb um Gregor sorgen.

Das war also Ihre Erzählung.

Nun erzähle ich Ihnen, was ich auf diese Erzählung antworten würde.

Meine Antwort lautet:

Oje. Ich hoffe, das geht gut. Sie haben mir ja erzählt, dass Sie Claudia nicht mögen, weil sie "berechnend und falsch" ist. Ich kann verstehen, dass Sie Angst um Gregor haben.

Nun, da Sie meine Antwort kennen, beurteilen Sie bitte im Folgenden Ihren Eindruck von mir.

C1. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Unecht ... 1 ... 2 ... 3 ... 4 ... 5 ... Natürlich	☐	☐	☐	☐	☐
Wie eine Maschine ... 1 ... 2 ... 3 ... 4 ... 5 ... Wie ein Mensch	☐	☐	☐	☐	☐
Hat kein Bewusstsein ... 1 ... 2 ... 3 ... 4 ... 5 ... Hat ein Bewusstsein	☐	☐	☐	☐	☐
Künstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Realistisch	☐	☐	☐	☐	☐

C2. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Tot ... 1 ... 2 ... 3 ... 4 ... 5 ... Lebendig	☐	☐	☐	☐	☐
Unbewegt ... 1 ... 2 ... 3 ... 4 ... 5 ... Lebendig	☐	☐	☐	☐	☐
Mechanisch ... 1 ... 2 ... 3 ... 4 ... 5 ... Organisch	☐	☐	☐	☐	☐
Künstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Realistisch	☐	☐	☐	☐	☐
Träge ... 1 ... 2 ... 3 ... 4 ... 5 ... Interaktiv	☐	☐	☐	☐	☐
Apathisch ... 1 ... 2 ... 3 ... 4 ... 5 ... Reagierend	☐	☐	☐	☐	☐

C3. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Nicht mögen ... 1 ... 2 ... 3 ... 4 ... 5 ... Mögen	☐	☐	☐	☐	☐

LimeSurvey

	1	2	3	4	5
Unfreundlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Freundlich	☐	☐	☐	☐	☐
Unhöflich ... 1 ... 2 ... 3 ... 4 ... 5 ... Höflich	☐	☐	☐	☐	☐
Unangenehm ... 1 ... 2 ... 3 ... 4 ... 5 ... Angenehm	☐	☐	☐	☐	☐
Furchtbar ... 1 ... 2 ... 3 ... 4 ... 5 ... Nett	☐	☐	☐	☐	☐

C4. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Inkompetent ... 1 ... 2 ... 3 ... 4 ... 5 ... Kompetent	☐	☐	☐	☐	☐
Ungebildet ... 1 ... 2 ... 3 ... 4 ... 5 ... Unterrichtet	☐	☐	☐	☐	☐
Verantwortungslos ... 1 ... 2 ... 3 ... 4 ... 5 ... Verantwortungsbewusst	☐	☐	☐	☐	☐
Unintelligent ... 1 ... 2 ... 3 ... 4 ... 5 ... Intelligent	☐	☐	☐	☐	☐
Unvernünftig ... 1 ... 2 ... 3 ... 4 ... 5 ... Vernünftig	☐	☐	☐	☐	☐

C5. Bitte bewerten Sie Ihren emotionalen Zustand auf diesen Skalen:

	1	2	3	4	5
Ängstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Entspannt	☐	☐	☐	☐	☐
Ruhig ... 1 ... 2 ... 3 ... 4 ... 5 ... Aufgewühlt	☐	☐	☐	☐	☐
Still ... 1 ... 2 ... 3 ... 4 ... 5 ... Überrascht	☐	☐	☐	☐	☐

C6. Kommentare, Anmerkungen (optional!):

LimeSurvey

Teil D: Fragebogen

Ihr Browser unterstützt leider den HTML Video Tag nicht. Bitte rufen Sie folgenden Link auf: https://www.vrecar.eu/robot/qbo-one-video-material/qbo-one-frage.mp4 Bitte sehen Sie sich das Video an. Das Transkript zum Video lautet wie folgt:

Zu aller erst: Sie können sich dieses Video gerne beliebig oft ansehen.

Ich erinnere Sie kurz, dass Sie mir von Ihrer Morgenroutine und dem Mittagessen erzählt haben. Danach haben Sie beim Spazieren Josef getroffen; sein Sohn Gregor, der immer ein lieber Bub gewesen sei, heiratet nächste Woche Claudia, die Sie nicht mögen, weil Sie berechnend und falsch sei. Sie erklärten mir, dass Sie sich deshalb um Gregor sorgen.

Das war also Ihre Erzählung.

Nun erzähle ich Ihnen, was ich auf diese Erzählung antworten würde.

Meine Antwort lautet:

Wie geht es Ihnen mit der Hochzeit von Gregor und Claudia?

Sie können Ihre Emotionen gerne unten im Feld für Kommentare und Anmerkungen kurz beschreiben. Bedenken Sie, dass dies in dem gegenwärtigen Setup die verbale Antwort ersetzen soll, die ich sonst von Ihnen auf die Nachfrage erhalten würde.

Nun, da Sie meine Antwort kennen, beurteilen Sie bitte im Folgenden Ihren Eindruck von mir.

D1. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Unecht ... 1 ... 2 ... 3 ... 4 ... 5 ... Natürlich	☐	☐	☐	☐	☐
Wie eine Maschine ... 1 ... 2 ... 3 ... 4 ... 5 ... Wie ein Mensch	☐	☐	☐	☐	☐
Hat kein Bewusstsein ... 1 ... 2 ... 3 ... 4 ... 5 ... Hat ein Bewusstsein	☐	☐	☐	☐	☐
Künstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Realistisch	☐	☐	☐	☐	☐

D2. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Tot ... 1 ... 2 ... 3 ... 4 ... 5 ... Lebendig	☐	☐	☐	☐	☐
Unbewegt ... 1 ... 2 ... 3 ... 4 ... 5 ... Lebendig	☐	☐	☐	☐	☐
Mechanisch ... 1 ... 2 ... 3 ... 4 ... 5 ... Organisch	☐	☐	☐	☐	☐
Künstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Realistisch	☐	☐	☐	☐	☐
Träge ... 1 ... 2 ... 3 ... 4 ... 5 ... Interaktiv	☐	☐	☐	☐	☐
Apathisch ... 1 ... 2 ... 3 ... 4 ... 5 ... Reagierend	☐	☐	☐	☐	☐

LimeSurvey

D3. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Nicht mögen ... 1 ... 2 ... 3 ... 4 ... 5 ... Mögen	☐	☐	☐	☐	☐
Unfreundlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Freundlich	☐	☐	☐	☐	☐
Unhöflich ... 1 ... 2 ... 3 ... 4 ... 5 ... Höflich	☐	☐	☐	☐	☐
Unangenehm ... 1 ... 2 ... 3 ... 4 ... 5 ... Angenehm	☐	☐	☐	☐	☐
Furchtbar ... 1 ... 2 ... 3 ... 4 ... 5 ... Nett	☐	☐	☐	☐	☐

D4. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Inkompetent ... 1 ... 2 ... 3 ... 4 ... 5 ... Kompetent	☐	☐	☐	☐	☐
Ungebildet ... 1 ... 2 ... 3 ... 4 ... 5 ... Unterrichtet	☐	☐	☐	☐	☐
Verantwortungslos ... 1 ... 2 ... 3 ... 4 ... 5 ... Verantwortungsbewusst	☐	☐	☐	☐	☐
Unintelligent ... 1 ... 2 ... 3 ... 4 ... 5 ... Intelligent	☐	☐	☐	☐	☐
Unvernünftig ... 1 ... 2 ... 3 ... 4 ... 5 ... Vernünftig	☐	☐	☐	☐	☐

D5. Bitte bewerten Sie Ihren emotionalen Zustand auf diesen Skalen:

	1	2	3	4	5
Ängstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Entspannt	☐	☐	☐	☐	☐
Ruhig ... 1 ... 2 ... 3 ... 4 ... 5 ... Aufgewühlt	☐	☐	☐	☐	☐
Still ... 1 ... 2 ... 3 ... 4 ... 5 ... Überrascht	☐	☐	☐	☐	☐

D6. Kommentare, Anmerkungen (optional!):

LimeSurvey

Teil E: Fragebogen

Ihr Browser unterstützt leider den HTML Video Tag nicht. Bitte rufen Sie folgenden Link auf: https://www.vrecar.eu/robot/qbo-one-video-material/qbo-one-falsch.mp4 Bitte sehen Sie sich das Video an. Das Transkript zum Video lautet wie folgt:

Zu aller erst: Sie können sich dieses Video gerne beliebig oft ansehen.

Ich erinnere Sie kurz, dass Sie mir von Ihrer Morgenroutine und dem Mittagessen erzählt haben. Danach haben Sie beim Spazieren Josef getroffen; sein Sohn Gregor, der immer ein lieber Bub gewesen sei, heiratet nächste Woche Claudia, die Sie nicht mögen, weil Sie berechnend und falsch sei. Sie erklärten mir, dass Sie sich deshalb um Gregor sorgen.

Das war also Ihre Erzählung.

Nun erzähle ich Ihnen, was ich auf diese Erzählung antworten würde.

Meine Antwort lautet:

Schön, dass Sie sich freuen, dass die beiden heiraten.

Nun, da Sie meine Antwort kennen, beurteilen Sie bitte im Folgenden Ihren Eindruck von mir.

E1. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Unecht ... 1 ... 2 ... 3 ... 4 ... 5 ... Natürlich	☐	☐	☐	☐	☐
Wie eine Maschine ... 1 ... 2 ... 3 ... 4 ... 5 ... Wie ein Mensch	☐	☐	☐	☐	☐
Hat kein Bewusstsein ... 1 ... 2 ... 3 ... 4 ... 5 ... Hat ein Bewusstsein	☐	☐	☐	☐	☐
Künstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Realistisch	☐	☐	☐	☐	☐

E2. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Tot ... 1 ... 2 ... 3 ... 4 ... 5 ... Lebendig	☐	☐	☐	☐	☐
Unbewegt ... 1 ... 2 ... 3 ... 4 ... 5 ... Lebendig	☐	☐	☐	☐	☐
Mechanisch ... 1 ... 2 ... 3 ... 4 ... 5 ... Organisch	☐	☐	☐	☐	☐
Künstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Realistisch	☐	☐	☐	☐	☐
Träge ... 1 ... 2 ... 3 ... 4 ... 5 ... Interaktiv	☐	☐	☐	☐	☐
Apathisch ... 1 ... 2 ... 3 ... 4 ... 5 ... Reagierend	☐	☐	☐	☐	☐

E3. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Nicht mögen ... 1 ... 2 ... 3 ... 4 ... 5 ... Mögen	☐	☐	☐	☐	☐

LimeSurvey

	1	2	3	4	5
Unfreundlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Freundlich	☐	☐	☐	☐	☐
Unhöflich ... 1 ... 2 ... 3 ... 4 ... 5 ... Höflich	☐	☐	☐	☐	☐
Unangenehm ... 1 ... 2 ... 3 ... 4 ... 5 ... Angenehm	☐	☐	☐	☐	☐
Furchtbar ... 1 ... 2 ... 3 ... 4 ... 5 ... Nett	☐	☐	☐	☐	☐

E4. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Inkompetent ... 1 ... 2 ... 3 ... 4 ... 5 ... Kompetent	☐	☐	☐	☐	☐
Ungebildet ... 1 ... 2 ... 3 ... 4 ... 5 ... Unterrichtet	☐	☐	☐	☐	☐
Verantwortungslos ... 1 ... 2 ... 3 ... 4 ... 5 ... Verantwortungsbewusst	☐	☐	☐	☐	☐
Unintelligent ... 1 ... 2 ... 3 ... 4 ... 5 ... Intelligent	☐	☐	☐	☐	☐
Unvernünftig ... 1 ... 2 ... 3 ... 4 ... 5 ... Vernünftig	☐	☐	☐	☐	☐

E5. Bitte bewerten Sie Ihren emotionalen Zustand auf diesen Skalen:

	1	2	3	4	5
Ängstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Entspannt	☐	☐	☐	☐	☐
Ruhig ... 1 ... 2 ... 3 ... 4 ... 5 ... Aufgewühlt	☐	☐	☐	☐	☐
Still ... 1 ... 2 ... 3 ... 4 ... 5 ... Überrascht	☐	☐	☐	☐	☐

E6. Kommentare, Anmerkungen (optional!):

LimeSurvey

Teil F: Fragebogen

Ihr Browser unterstützt leider den HTML Video Tag nicht. Bitte rufen Sie folgenden Link auf: https://www.vrecar.eu/robot/qbo-one-video-material/qbo-one-ohne.mp4 Bitte sehen Sie sich das Video an. Das Transkript zum Video lautet wie folgt:

Zu aller erst: Sie können sich dieses Video gerne beliebig oft ansehen.

Ich erinnere Sie kurz, dass Sie mir von Ihrer Morgenroutine und dem Mittagessen erzählt haben. Danach haben Sie beim Spazieren Josef getroffen; sein Sohn Gregor, der immer ein lieber Bub gewesen sei, heiratet nächste Woche Claudia, die Sie nicht mögen, weil Sie berechnend und falsch sei. Sie erklärten mir, dass Sie sich deshalb um Gregor sorgen.

Das war also Ihre Erzählung.

Nun erzähle ich Ihnen, was ich auf diese Erzählung antworten würde.

Meine Antwort lautet:

Danke für Ihre Erzählung.

Nun, da Sie meine Antwort kennen, beurteilen Sie bitte im Folgenden Ihren Eindruck von mir.

F1. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Unecht ... 1 ... 2 ... 3 ... 4 ... 5 ... Natürlich	☐	☐	☐	☐	☐
Wie eine Maschine ... 1 ... 2 ... 3 ... 4 ... 5 ... Wie ein Mensch	☐	☐	☐	☐	☐
Hat kein Bewusstsein ... 1 ... 2 ... 3 ... 4 ... 5 ... Hat ein Bewusstsein	☐	☐	☐	☐	☐
Künstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Realistisch	☐	☐	☐	☐	☐

F2. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Tot ... 1 ... 2 ... 3 ... 4 ... 5 ... Lebendig	☐	☐	☐	☐	☐
Unbewegt ... 1 ... 2 ... 3 ... 4 ... 5 ... Lebendig	☐	☐	☐	☐	☐
Mechanisch ... 1 ... 2 ... 3 ... 4 ... 5 ... Organisch	☐	☐	☐	☐	☐
Künstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Realistisch	☐	☐	☐	☐	☐
Träge ... 1 ... 2 ... 3 ... 4 ... 5 ... Interaktiv	☐	☐	☐	☐	☐
Apathisch ... 1 ... 2 ... 3 ... 4 ... 5 ... Reagierend	☐	☐	☐	☐	☐

F3. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Nicht mögen ... 1 ... 2 ... 3 ... 4 ... 5 ... Mögen	☐	☐	☐	☐	☐

LimeSurvey

	1	2	3	4	5
Unfreundlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Freundlich	☐	☐	☐	☐	☐
Unhöflich ... 1 ... 2 ... 3 ... 4 ... 5 ... Höflich	☐	☐	☐	☐	☐
Unangenehm ... 1 ... 2 ... 3 ... 4 ... 5 ... Angenehm	☐	☐	☐	☐	☐
Furchtbar ... 1 ... 2 ... 3 ... 4 ... 5 ... Nett	☐	☐	☐	☐	☐

F4. Bitte beurteilen Sie Ihren Eindruck von mir auf diesen Skalen:

	1	2	3	4	5
Inkompetent ... 1 ... 2 ... 3 ... 4 ... 5 ... Kompetent	☐	☐	☐	☐	☐
Ungebildet ... 1 ... 2 ... 3 ... 4 ... 5 ... Unterrichtet	☐	☐	☐	☐	☐
Verantwortungslos ... 1 ... 2 ... 3 ... 4 ... 5 ... Verantwortungsbewusst	☐	☐	☐	☐	☐
Unintelligent ... 1 ... 2 ... 3 ... 4 ... 5 ... Intelligent	☐	☐	☐	☐	☐
Unvernünftig ... 1 ... 2 ... 3 ... 4 ... 5 ... Vernünftig	☐	☐	☐	☐	☐

F5. Bitte bewerten Sie Ihren emotionalen Zustand auf diesen Skalen:

	1	2	3	4	5
Ängstlich ... 1 ... 2 ... 3 ... 4 ... 5 ... Entspannt	☐	☐	☐	☐	☐
Ruhig ... 1 ... 2 ... 3 ... 4 ... 5 ... Aufgewühlt	☐	☐	☐	☐	☐
Still ... 1 ... 2 ... 3 ... 4 ... 5 ... Überrascht	☐	☐	☐	☐	☐

F6. Kommentare, Anmerkungen (optional!):

LimeSurvey

Teil G: Fragen zum Ablauf der Umfrage

Ihr Browser unterstützt leider den HTML Video Tag nicht. Bitte rufen Sie folgenden Link auf: https://www.vrecar.eu/robot/qbo-one-video-material/qbo-one-meta.mp4 Bitte sehen Sie sich das Video an. Das Transkript zum Video lautet wie folgt:

Danke an dieser Stelle schon einmal für Ihren wertvollen Input. Ich würde Sie nun noch bitten, einige allgemeine Fragen zur Umfragedurchführung durch mich zu beantworten.

G1. Wie haben Sie es empfunden, dass ich Sie gesiezt habe?

- nicht komisch ☐
- wenig komisch ☐
- weder noch ☐
- etwas komisch ☐
- sehr komisch ☐
- Das kann ich leider nicht beurteilen. ☐
- Ich habe es nicht bemerkt bzw. es ist mir nicht bewusst aufgefallen. ☐

G2. Wäre es Ihnen lieber gewesen, wenn ich Sie "geduzt" hätte?

- Ja ☐
- Nein ☐
- Weiß nicht ☐

G3. Glauben Sie, Sie haben mich anders bewertet, weil ich Ihnen die Fragen selbst gestellt habe und sie nicht aus der Sicht einer unbeteiligten dritten Person gestellt wurden?

- Ja, positiver. ☐
- Ja, etwas positiver. ☐
- Ja, aber ich weiß nicht inwiefern. ☐
- Ja, etwas negativer ☐
- Ja, negativer. ☐
- Nein, kein Unterschied. ☐
- Das kann ich leider nicht beurteilen. ☐

G4. Kommentare, Anmerkungen (optional!):

LimeSurvey

Teil H: Abschluss

Danke für Ihre Teilnahme!

Um am Gewinnspiel (3x je 30 Euro) teilzunehmen, tragen Sie bitte Ihre E-Mail-Adresse (wahlweise auch gerne Telefonnummer) unten ein.

Falls Sie außerdem bereit wären, im Rahmen eines kleinen Online-Interviews noch ein paar kurze Fragen zu dieser Interaktion zu beantworten, markieren Sie bitte dieses Feld.

Beachten Sie, dass die Bereitschaft für ein Interview Ihre Gewinnchancen verdreifacht.

Auf dieser Seite haben Sie außerdem noch einmal die Möglichkeit, Kommentare zur Umfrage abzugeben.

Dankeschön!

Bitte drücken Sie am Ende noch einmal auf [Absenden], um Ihre Angaben zu speichern.

H1. Bitte geben Sie an, ob Sie an einem Interview teilnehmen würden (sofern Sie keine Angabe machen, wird davon ausgegangen, dass Sie NICHT(!) teilnehmen möchten.)

Ich möchte an einem Interview teilnehmen. ☐

Ich möchte NICHT an einem Interview teilnehmen ☐

H2. Kontaktinformation (E-Mail oder Telefonnummer)

H3. Kommentare, Anmerkungen, Feedback (optional!):

Erneut herzlichen Dank für Ihre Teilnahme!

Die Umfrage ist abgeschlossen, sie können dieses Fenster bzw. diesen Tab nun schließen!

List of Figures

List of Tables

Bibliography

[AA12] Ameeta Agrawal and Aijun An. "Unsupervised emotion detection from text using semantic and syntactic relations". In: *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. Vol. 1. 2012, pp. 346–353. DOI: 10.1109/WI-IAT.2012.170.

[AA18] Ralph Adolphs and David J Anderson. *The neuroscience of emotion: a new synthesis*. Princeton University Press, 2018. ISBN: 978-0691174082. URL: https://books.google.at/books?id=tXCYDwAAQBAJ.

[AC13] Carole Adam and Lawrence Cavedon. "Once upon a time... a companion robot that can tell stories". In: *Proceedings of Intelligent Virtual Agents*. Ed. by Ruth Aylett, Brigitte Krenn, Catherine Pelachaud, and Hiroshi Shimodaira. Vol. 8108. Lecture Notes in Computer Science. Springer. Berlin, Heidelberg, 2013, pp. 434–435. DOI: 10.1007/978-3-642-40415-3.

[AKB13] Balamurali A.R., Mitesh M. Khapra, and Pushpak Bhattacharyya. "Lost in translation: viability of machine translation for cross language sentiment analysis". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer, 2013, pp. 38–49. ISBN: 978-3642372568. DOI: 10.1007/978-3-642-37256-8_4.

[Alb15] Veaceslav Albu. "Neural network based model for emotion recognition". In: *Workshop on Foundations of Informatics*. 2015, p. 423.

[AO05] Hilary Arksey and Lisa O'Malley. "Scoping studies: towards a methodological framework". In: *International Journal of Social Research Methodology* 8.1 (2005), pp. 19–32. DOI: 10.1080/1364557032000119616.

[APB20] Matheus Araújo, Adriano Pereira, and Fabrício Benevenuto. "A comparative study of machine translation for multilingual sentence-level sentiment analysis". In: *Information Sciences* 512 (2020), pp. 1078–1102. ISSN: 0020-0255. DOI: 10.1016/j.ins.2019.10.031. URL: https://www.sciencedirect.com/science/article/pii/S0020025519309879.

[Ara+16] Matheus Araujo, Julio Reis, Adriano Pereira, and Fabricio Benevenuto. "An evaluation of machine translation for multilingual sentence-level sentiment analysis". In: *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. SAC '16. Pisa, Italy: Association for Computing Machinery, 2016, pp. 1140–1145. ISBN: 978-1450337397. DOI: 10.1145/2851613.2851817.

[Asc19] Andreas Aschenbrenner. "Emotionserkennung bei Nachrichtenkommentaren mittels Convolutional Neural Networks und Label Propagationsverfahren". Apr. 2019. URL: http://nbn-resolving.de/urn:nbn:de:bvb:19-252950.

[Auz+] Thomas Auzinger et al. *Thesis Template - TU Wien - Faculty of Informatics*. https://gitlab.com/ThomasAUZINGER/vutinfth. Accessed on Sep. 25, 2020.

[Ban+17] Anil Bandhakavi, Nirmalie Wiratunga, Stewart Massie, and Deepak Padmanabhan. "Lexicon generation for emotion detection from text". In: *IEEE Intelligent Systems* 32.1 (2017), pp. 102–108. DOI: 10.1109/MIS.2017.22.

[Bar] Christoph Bartneck. *The Godspeed Questionnaire Series*. https://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/. Accessed on Jan. 6, 2022, saved to https://web.archive.org/web/20220418180403/https://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/ on Apr. 18, 2022.

[Bar+09] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots". In: *International Journal of Social Robotics* 1.1 (2009), pp. 71–81. DOI: 10.1007/s12369-008-0001-3.

[Bar99] Simon Baron-Cohen. *Evolution of a theory of mind?* Oxford University Press, 1999. ISBN: 978-0192632593. DOI: 10.1093/acprof:oso/9780192632593.003.0013.

[BB20] Valentin Barrière and Alexandra Balahur. "Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation". In: *CoRR* abs/2010.03486 (2020). arXiv: 2010.03486. URL: https://arxiv.org/abs/2010.03486.

[BC06] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa. URL: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

[Bec+21] Elizabeth Bechard, John Evans, Eunji Cho, Yufen Lin, Arthi Kozhumam, Jill Jones, Sydney Grob, and Oliver Glass. "Feasibility, acceptability, and potential effectiveness of an online expressive writing intervention for covid-19 resilience". In: *Complementary Therapies in Clinical Practice* 45 (2021), 101460: 1–8. ISSN: 1744-3881. DOI: 10.1016/j.ctcp.2021.101460. URL: https://www.sciencedirect.com/science/article/pii/S1744388121001596.

[Bel+18] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. "Social robots for education: a review". In: *Science Robotics* 3.21 (Aug. 2018), eaat5954: 1–9. DOI: 10.1126/scirobotics.aat5954. URL: https://www.science.org/doi/abs/10.1126/scirobotics.aat5954.

[BL94] Margaret M. Bradley and Peter J. Lang. "Measuring emotion: the self-assessment manikin and the semantic differential". In: *Journal of Behavior Therapy and Experimental Psychiatry* 25.1 (1994), pp. 49–59. ISSN: 0005-7916. DOI: 10.1016/0005-7916(94)90063-9. URL: https://www.sciencedirect.com/science/article/pii/0005791694900639.

[BLS15] C Daniel Batson, David A Lishner, and Eric L Stocks. "The empathy-altruism hypothesis". In: *The Oxford Handbook of Prosocial Behavior* (Jan. 2015), pp. 259–281. DOI: 10.1093/oxfordhb/9780195399813.013.023. URL: https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195399813.001.0001/oxfordhb-9780195399813-e-023.

[Bod+21] Thomas Body, Xiaohui Tao, Yuefeng Li, Lin Li, and Ning Zhong. "Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models". In: *Expert Systems with Applications* 178 (2021), p. 115033. ISSN: 0957–4174. DOI: 10.1016/j.eswa.2021.115033. URL: https://www.sciencedirect.com/science/article/pii/S0957417421004747.

[Bou+20] Mark Boukes, Bob Van de Velde, Theo Araujo, and Rens Vliegenthart. "What's the tone? easy doesn't do it: analyzing performance and agreement between off-the-shelf sentiment analysis tools". In: *Communication Methods and Measures* 14.2 (2020), pp. 83–104. DOI: 10.1080/19312458.2019.1671966.

[BP15] Alexandra Balahur and José M. Perea-Ortega. "Sentiment analysis system adaptation for multilingual processing: the case of tweets". In: *Information Processing & Management* 51.4 (2015), pp. 547–556. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2014.10.004. URL: https://www.sciencedirect.com/science/article/pii/S0306457314000934.

[Bro+13] Elizabeth Broadbent, Vinayak Kumar, Xingyan Li, John Sollers 3rd, Rebecca Q. Stafford, Bruce A. MacDonald, and Daniel M. Wegner. "Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality". In: *PLOS ONE* 8.8 (Aug. 2013), pp. 1–9. DOI: 10.1371/journal.pone.0072589.

[BT12] Alexandra Balahur and Marco Turchi. "Multilingual sentiment analysis using machine translation?" In: *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Jeju, Korea: Association for Computational Linguistics, July 2012, pp. 52–60. URL: https://aclanthology.org/W12-3709.

[BT13] Alexandra Balahur and Marco Turchi. "Improving sentiment analysis in Twitter using multilingual machine translated data". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, Sept. 2013, pp. 49–55. URL: https://aclanthology.org/R13-1007.

[BT14] Alexandra Balahur and Marco Turchi. "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis". In: *Computer Speech & Language* 28.1 (2014), pp. 56–75. ISSN: 0885-2308. DOI: 10.1016/j.csl.2013.03.004. URL: https://www.sciencedirect.com/science/article/pii/S088523081300020X.

[Bus+21] Maritza Bustos-López, Nicandro Cruz-Ramírez, Alejandro Guerra-Hernández, Laura Nely Sánchez-Morales, and Giner Alor-Hernández. "Emotion detection from text in learning environments: a review". In: *New Perspectives on Enterprise Decision-Making Applying Artificial Intelligence Techniques*. Ed. by Julian Andres Zapata-Cortes, Giner Alor-Hernández, Cuauhtémoc Sánchez-Ramírez, and Jorge Luis García-Alcaraz. Cham: Springer International Publishing, 2021, pp. 483–508. ISBN: 978-3030711153. DOI: 10.1007/978-3-030-71115-3_21.

[Bus00] David M Buss. *The dangerous passion: why jealousy is as necessary as love and sex*. The Free Press, Simon and Schuster, 2000. ISBN: 0684867869. URL: https://books.google.at/books?id=M5EzujGkAn4C.

[BWP10] Haji Binali, Chen Wu, and Vidyasagar Potdar. "Computational approaches for emotion detection in text". In: *4th IEEE International Conference on Digital Ecosystems and Technologies*. IEEE. 2010, pp. 172–177. DOI: 10.1109/DEST.2010.5610650.

[CC13] Andrea Capstick and David Clegg. "Behind the stiff upper lip: war narratives of older men with dementia". In: *Journal of War & Culture Studies* 6.3 (2013), pp. 239–254. DOI: 10.1179/1752627213Z.00000000021.

[Chr99] Ursula Christmann. "Kommunikation, Interaktion". In: *Kölner Psychologische Studien* 4 (Jan. 1999), pp. 22–26.

[Cie+17] Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. "A twitter corpus and benchmark resources for german sentiment analysis". en. In: *5th International Workshop on Natural Language Processing for Social Media*. Boston MA, USA: Association for Computational Linguistics, 2017, pp. 45–51. DOI: 10.21256/zhaw-1530. URL: https://digitalcollection.zhaw.ch/handle/11475/1856.

[CM14] Lea Canales and Patricio Martínez-Barco. “Emotion detection from text: a survey”. In: *Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC)*. Quito, Ecuador: Association for Computational Linguistics, Oct. 2014, pp. 37–43. DOI: 10.3115/v1/W14-6905.

[CMM20] Adam K. Coyne, Andrew Murtagh, and Conor McGinn. “Using the geneva emotion wheel to measure perceived affect in human-robot interaction”. In: *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2020, pp. 491–498.

[Das+16] Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. “Multilingual sentiment analysis: state of the art and independent comparison of techniques”. In: *Cognitive Computation* 8.4 (2016), pp. 757–771. ISSN: 1866-9964. DOI: 10.1007/s12559-016-9415-7.

[Dau+05] Kerstin Dautenhahn, Sarah Woods, Christina Kaouri, Michael L Walters, Kheng Lee Koay, and Iain Werry. “What is a robot companion - friend, assistant or butler?” In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2005, pp. 1192–1197. DOI: 10.1109/IROS.2005.1545189.

[Dau03] Kerstin Dautenhahn. “Roles and functions of robots in human society: implications from research in autism therapy”. In: *Robotica* 21.4 (2003), pp. 443–452. DOI: 10.1017/S0263574703004922.

[Dau21] Kerstin Dautenhahn. *The encyclopedia of human-computer interaction, 2nd edition (38. human-robot interaction)*. Accessed on Mar. 11, 2021. 2021. URL: https://web.archive.org/web/20210116002048/https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/human-robot-interaction.

[Deb07] Dorothea Debus. “Being emotional about the past: on the nature and role of past-directed emotions”. In: *Noûs* 41.4 (2007), pp. 758–779. DOI: 10.1111/j.1468-0068.2007.00669.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0068.2007.00669.x.

[Den08] Kerstin Denecke. “Using sentiwordnet for multilingual sentiment analysis”. In: *2008 IEEE 24th International Conference on Data Engineering Workshop*. 2008, pp. 507–512. DOI: 10.1109/ICDEW.2008.4498370.

[Dep19] Statista Research Department. *The most spoken languages worldwide in 2019*. Accessed on Mar. 11, 2021. 2019. URL: https://web.archive.org/web/20210202200226/https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/.

[Deu19] Statistisches Bundesamt Deutschland. *Bevölkerung in Deutschland.* Accessed on Mar. 11, 2021. 2019. URL: https://web.archive.org/web/20210228065111/https://service.destatis.de/bevoelkerungspyramide/.

[Dij+12] Ate Dijkstra, Gülendam Hakverdioğlu Yönt, Esra Akin Korhan, Marta Muszalik, Kornelia Kędziora-Kornatowska, and Mizue Suzuki. "The care dependency scale for measuring basic human needs: an international comparison". In: *Journal of Advanced Nursing* 68.10 (2012), pp. 2341–2348. DOI: 10.1111/j.1365-2648.2011.05939.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2648.2011.05939.x.

[DM96] Andrew Dillon and Michael G. Morris. "User acceptance of information technology: theories and models". In: *Annual Review of Information Science and Technology (ARIST)* 31 (1996), pp. 3–32. ISSN: 0066-4200. URL: https://www.learntechlib.org/p/82513.

[Dud] Duden. *Website des Duden.* https://www.duden.de/rechtschreibung/Emotion. Accessed on Dec. 31, 2021.

[EC11] Paul Ekman and Daniel Cordaro. "What is meant by calling emotions basic". In: *Emotion Review* 3.4 (2011), pp. 364–370. DOI: 10.1177/1754073911410740.

[EF03] Paul Ekman and Wallace V Friesen. *Unmasking the face: a guide to recognizing emotions from facial clues.* Vol. 10. Malor Books, 2003. ISBN: 978-1883536367. URL: https://books.google.at/books?id=TukNoJDgMTUC.

[EFE13] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: guidelines for research and an integration of findings.* Vol. 11. Pergamon general psychology series. Elsevier Science, 2013. ISBN: 978-1483147635. URL: https://books.google.at/books?id=dOFFBQAAQBAJ.

[Ekm04] Paul Ekman. "Emotions revealed". In: *BMJ* 328.Suppl S5 (2004). DOI: 10.1136/sbmj.0405184. URL: https://www.bmj.com/content/328/Suppl_S5/0405184.

[Ekm92a] Paul Ekman. "An argument for basic emotions". In: *Cognition and Emotion* 6.3-4 (1992), pp. 169–200. DOI: 10.1080/02699939208411068.

[Ekm92b] Paul Ekman. "Are there basic emotions?" In: *Psychological Review* 99.3 (1992), pp. 550–553. DOI: 10.1037/0033-295X.99.3.550. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2F0033-295X.99.3.550.

[Ekm99] Paul Ekman. "Basic emotions". In: *Handbook of Cognition and Emotion* 98.16 (1999), pp. 45–60.

[Emo] Margaret Emory. *On Fear, Emotions, and Memory: An Interview with Dr. Joseph LeDoux*. https://brainworldmagazine.com/on-fear-emotions-and-memory-an-interview-with-dr-joseph-ledoux/. Accessed on Jan. 4, 2022, saved to https://web.archive.org/web/20220418181225/https://brainworldmagazine.com/on-fear-emotions-and-memory-an-interview-with-dr-joseph-ledoux/ on Apr. 18, 2022.

[Fis06] Kerstin Fischer. "What computer talk is and isn't: human-computer conversation as intercultural communication". In: *Linguistics-Computational Linguistics* 17 (2006). URL: http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/repository/downloads/fischer-computertalk.pdf.

[Fox+18] Andrew S Fox, Regina C Lapate, Alexander J Shackman, and Richard J Davidson. *The nature of emotion: fundamental questions*. Series in Affective Science. Oxford University Press, 2018. ISBN: 978-0190873127. URL: https://books.google.at/books?id=fh5pDwAAQBAJ.

[Fun+18] Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. "Towards empathetic human-robot interactions". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Ed. by Alexander Gelbukh. Springer, 2018, pp. 173–193. ISBN: 978-3319754871. DOI: 10.1007/978-3-319-75487-1_14.

[Gha+20] Aimi Shazwani Ghazali, Jaap Ham, Emilia Barakova, and Panos Markopoulos. "Persuasive robots acceptance model (pram): roles of social responses within the acceptance model of persuasive robots". In: *International Journal of Social Robotics* 12.5 (2020), pp. 1075–1092. ISSN: 1875-4805. DOI: 10.1007/s12369-019-00611-1.

[GJ17] Chloe A. Greenbaum and Shabnam Javdani. "Expressive writing intervention promotes resilience among juvenile justice-involved youth". In: *Children and Youth Services Review* 73 (2017), pp. 220–229. ISSN: 0190-7409. DOI: 10.1016/j.childyouth.2016.11.034. URL: https://www.sciencedirect.com/science/article/pii/S0190740916304996.

[Gon+11] Barbara Gonsior, Stefan Sosnowski, Christoph Mayer, Jürgen Blume, Bernd Radig, Dirk Wollherr, and Kolja Kühnlenz. "Improving aspects of empathy and subjective performance for hri through mirroring facial expressions". In: *2011 RO-MAN*. IEEE. 2011, pp. 350–356. DOI: 10.1109/ROMAN.2011.6005294.

[Gra+21] Wilfredo Graterol, Jose Diaz-Amado, Yudith Cardinale, Irvin Dongo, Edmundo Lopes-Silva, and Cleia Santos-Libarino. "Emotion detection for social robots based on nlp transformers and an emotion ontology". In: *Sensors*

21.4 (2021), 1322: 1–19. ISSN: 1424-8220. DOI: 10.3390/s21041322. URL: https://www.mdpi.com/1424-8220/21/4/1322.

[Har+22] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. "More than a feeling: accuracy and application of sentiment analysis". In: *SSRN Electronic Journal* (Jan. 2022). DOI: 10.2139/ssrn.3489963.

[Hee+09] Marcel Heerink, Ben Krose, Vanessa Evers, and Bob Wielinga. "Measuring acceptance of an assistive social robot: a suggested toolkit". In: *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*. 2009, pp. 528–533. DOI: 10.1109/ROMAN.2009.5326320.

[Hep20] Andreas Hepp. "Artificial companions, social bots and work bots: communicative robots as research objects of media and communication studies". In: *Media, Culture & Society* 42.7-8 (2020), pp. 1410–1426. DOI: 10.1177/0163443720916412.

[Her95] Fritz Hermanns. "Kognition, Emotion, Intention. Dimensionen lexikalischer Semantik". In: *Die Ordnung der Wörter: Kognitive und lexikalische Strukturen*. Ed. by Gisela Harras. Berlin/New York: De Gruyter, 1995, pp. 138–178. ISBN: 978-3110144383. DOI: 10.1515/9783110622485. URL: https://www.degruyter.com/document/doi/10.1515/9783110622485/pdf#page=144.

[HM20] Rachel E Hendrix and Carley C Morrison. "Student emotional responses to different communication situations." In: *Journal of Applied Communications* 104.3 (2020), pp. 1–20. URL: https://link.gale.com/apps/doc/A643734348/AONE?u=anon~2e37f748&sid=googleScholar&xid=967a4ac8.

[Hog+14] Alexander Hogenboom, Bas Heerschop, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. "Multi-lingual support for lexicon-based sentiment analysis guided by semantics". In: *Decision Support Systems* 62 (2014), pp. 43–53. ISSN: 0167-9236. DOI: 10.1016/j.dss.2014.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0167923614000645.

[HSK92] Michael Hölzer, Nicola Scheytt, and Horst Kächele. "Das „Affektive Diktionär Ulm“ als eine Methode der quantitativen Vokabularbestimmung". In: *Textanalyse: Anwendungen der computerunterstüzten Inhaltsanalyse. Beiträge zur 1. TEXTPACK-Anwenderkonferenz*. Ed. by Cornelia Züll and Peter Ph. Mohler. Wiesbaden: VS Verlag für Sozialwissenschaften, 1992, pp. 131–154. ISBN: 978-3322942296. DOI: 10.1007/978-3-322-94229-6_7.

[HWC21] Glenda Hannibal, Astrid Weiss, and Vicky Charisi. "'The robot may not notice my discomfort' – examining the experience of vulnerability for trust in human-robot interaction". In: *2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2021, pp. 704–711. DOI: 10.1109/RO-MAN50785.2021.9515513.

[Iza11] Carroll E Izard. "Forms and functions of emotions: matters of emotion–cognition interactions". In: *Emotion Review* 3.4 (2011), pp. 371–378. DOI: 10.1177/1754073911410737.

[Jay+13] Chandimal Jayawardena, Nilufar Baghaei, Kathiravelu Ganeshan, and Abdolhossein Sarrafzadeh. "Designing a socially assistive companion robotic wheel chair: robochair". In: *2013 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM)*. 2013, pp. 231–236. DOI: 10.1109/RAM.2013.6758589.

[JWM18] Jesin James, Catherine Inez Watson, and Bruce MacDonald. "Artificial empathy in social robots: an analysis of emotions in speech". In: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 2018, pp. 632–637. DOI: 10.1109/ROMAN.2018.8525652.

[KA10] Naomi Koerner and Martin M Antony. "Special series on disgust and phobic avoidance: a commentary". In: *International Journal of Cognitive Therapy* 3.1 (2010), pp. 52–63. DOI: 10.1521/ijct.2010.3.1.52.

[Kah+04] Peter H. Kahn, Batya Friedman, Deanne R. Perez-Granados, and Nathan G. Freier. "Robotic pets in the lives of preschool children". In: *CHI '04 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '04. Vienna, Austria: Association for Computing Machinery, 2004, pp. 1449–1452. ISBN: 1581137036. DOI: 10.1145/985921.986087.

[KB20] Mohammed Kaity and Vimala Balakrishnan. "Sentiment lexicons and non-english languages: a survey". In: *Knowledge and Information Systems* 62.12 (2020), pp. 4445–4480. DOI: 10.1007/s10115-020-01497-6.

[Ker+17] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, and Mohamed Ali Mahjoub. "A review on speech emotion recognition: case of pedagogical interaction in classroom". In: *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. 2017, pp. 1–7. DOI: 10.1109/ATSIP.2017.8075575.

[KNP19] Tomáš Kincl, Michal Novák, and Jiří Přibil. "Improving sentiment analysis performance on morphologically rich languages: language and domain independent approach". In: *Computer Speech & Language* 56 (2019), pp. 36–51. ISSN: 0885-2308. DOI: 10.1016/j.csl.2019.01.001. URL: https://www.sciencedirect.com/science/article/pii/S0885230816300109.

[KRP01] Barry Kort, Rob Reilly, and Rosalind W Picard. "An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion". In: *Proceedings IEEE International Conference on Advanced Learning Technologies*. 2001, pp. 43–46. DOI: 10.1109/ICALT.2001.943850.

[Kug21] Kseniia Kugai. "Communication. its basic concepts". In: *The driving force of science and trends in its development*. European Scientific Platform. 2021. URL: https://er.knutd.edu.ua/bitstream/123456789/17892/1/Kugai_Communication_Basic_concepts.pdf.

[Küh+13] Barbara Kühnlenz, Stefan Sosnowski, Malte Buß, Dirk Wollherr, Kolja Kühnlenz, and Martin Buss. "Increasing helpfulness towards a robot by emotional adaption to the user". In: *International Journal of Social Robotics* 5.4 (2013), pp. 457–476. ISSN: 1875-4805. DOI: 10.1007/s12369-013-0182-2.

[Kun+17] Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. *Transfer learning for speech recognition on a budget*. 2017. DOI: 10.48550/ARXIV.1706.00290. arXiv: 1706.00290 [cs.LG]. URL: https://arxiv.org/abs/1706.00290.

[LCO10] Danielle Levac, Heather Colquhoun, and Kelly K O'Brien. "Scoping studies: advancing the methodology". In: *Implementation Science* 5.1 (2010), pp. 1–9. ISSN: 1748-5908. DOI: 10.1186/1748-5908-5-69.

[LD03] Gitte Lindgaard and Cathy Dudek. "What is this evasive beast we call user satisfaction?" In: *Interacting with Computers* 15.3 (2003), pp. 429–452. DOI: 10.1016/S0953-5438(02)00063-2.

[LEF90] Robert W Levenson, Paul Ekman, and Wallace V Friesen. "Voluntary facial action generates emotion-specific autonomic nervous system activity". In: *Psychophysiology* 27.4 (1990), pp. 363–384. DOI: 10.1111/j.1469-8986.1990.tb02330.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.1990.tb02330.x.

[Lev11] Robert W Levenson. "Basic emotion questions". In: *Emotion Review* 3.4 (2011), pp. 379–386. DOI: 10.1177/1754073911410743.

[LM18] Yaniv Leviathan and Yossi Matias. *Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone*. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html. Accessed on Jan. 5, 2022. 2018.

[Lo+17] Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. "Multilingual sentiment analysis: from formal to informal and scarce resource languages". In: *Artificial Intelligence Review* 48.4 (2017), pp. 499–527. ISSN: 1573-7462. DOI: 10.1007/s10462-016-9508-4.

[LWH16] Tamara Lorenz, Astrid Weiss, and Sandra Hirche. “Synchrony and reciprocity: key mechanisms for social companion robots in therapy and care”. In: *International Journal of Social Robotics* 8.1 (2016), pp. 125–143. ISSN: 1875-4805. DOI: 10.1007/s12369-015-0325-8.

[Mar+17] Matthew Marge, Claire Bonial, Brendan Byrne, Taylor Cassidy, A. William Evans, Susan G. Hill, and Clare Voss. *Applying the wizard-of-oz technique to multimodal human-robot dialogue*. 2017. DOI: 10.48550/ARXIV.1703.03714. URL: https://arxiv.org/abs/1703.03714.

[mbH] Salzburg Research Forschungsgesellschaft m.b.H. *Website der Salzburg Research Forschungsgesellschaft m.b.H.* https://www.salzburgresearch.at. Accessed on Feb. 19, 2021.

[Meh96] Albert Mehrabian. “Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament”. In: *Current Psychology* 14.4 (1996), pp. 261–292. ISSN: 1936-4733. DOI: 10.1007/BF02686918.

[Mel+09] Gail F. Melson, Peter H. Kahn, Alan Beck, Batya Friedman, Trace Roberts, Erik Garrett, and Brian T. Gill. “Children’s behavior toward and understanding of robotic and living dogs”. In: *Journal of Applied Developmental Psychology* 30.2 (2009), pp. 92–102. ISSN: 0193-3973. DOI: 10.1016/j.appdev.2008.10.011. URL: https://www.sciencedirect.com/science/article/pii/S0193397308001329.

[Mir+17] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. “To err is robot: how humans assess and act toward an erroneous social robot”. In: *Frontiers in Robotics and AI* 4 (2017). ISSN: 2296-9144. DOI: 10.3389/frobt.2017.00021. URL: https://www.frontiersin.org/article/10.3389/frobt.2017.00021.

[MM18] Joseph E. Michaelis and Bilge Mutlu. “Reading socially: transforming the in-home reading experience with a learning-companion robot”. In: *Science Robotics* 3.21 (2018), eaat5999: 1–11. DOI: 10.1126/scirobotics.aat5999. URL: https://www.science.org/doi/abs/10.1126/scirobotics.aat5999.

[MSK16] Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. “How translation alters sentiment”. In: *Journal of Artificial Intelligence Research* 55 (2016), pp. 95–130. DOI: 10.1613/jair.4787. URL: https://jair.org/index.php/jair/article/view/10976.

[NV21] Pansy Nandwani and Rupali Verma. “A review on sentiment analysis and emotion detection from text”. In: *Social Network Analysis and Mining* 11.1 (2021), pp. 1–19. ISSN: 1869-5469. DOI: 10.1007/s13278-021-00776-6.

[Ode+20] Gaby Odekerken-Schröder, Cristina Mele, Tiziana Russo-Spena, Dominik Mahr, and Andrea Ruggiero. "Mitigating loneliness with companion robots in the covid-19 pandemic and beyond: an integrative framework and research agenda". In: *Journal of Service Management* (2020). ISSN: 1757-5818. URL: https://www.emerald.com/insight/content/doi/10.1108/JOSM-05-2020-0148/full/html.

[Ort22] Andrew Ortony. "Are all "basic emotions" emotions? a problem for the (basic) emotions construct". In: *Perspectives on Psychological Science* 17.1 (2022), pp. 41–61. DOI: 10.1177/1745691620985415.

[Pin97] Steven Pinker. *How the mind works*. New York: Norton, 1997.

[Plu80] Robert Plutchik. "A general psychoevolutionary theory of emotion". In: *Theories of Emotion*. Ed. by Robert Plutchik and Henry Kellerman. Academic Press, 1980, pp. 3–33. ISBN: 978-0125587013. DOI: 10.1016/B978-0-12-558701-3.50007-7. URL: https://www.sciencedirect.com/science/article/pii/B9780125587013500077.

[PRA18] Isabel Pedersen, Samantha Reid, and Kristen Aspevig. "Developing social robots for aging populations: a literature review of recent academic sources". In: *Sociology Compass* 12.6 (2018), e12585. DOI: 10.1111/soc4.12585. URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/soc4.12585.

[PW11] Jaak Panksepp and Douglas Watt. "What is basic about basic emotions? lasting lessons from affective neuroscience". In: *Emotion Review* 3.4 (2011), pp. 387–396. DOI: 10.1177/1754073911410741.

[Ref] Oxford Reference. *Website of the Oxford Reference*. https://www.oxfordreference.com/view/10.1093/acref/9780191792236.001.0001/acref-9780191792236-e-146. Accessed on Apr. 3, 2022, saved to https://web.archive.org/web/20220418181703/https://www.oxfordreference.com/view/10.1093/acref/9780191792236.001.0001/acref-9780191792236-e-146 on Apr. 18, 2022.

[Ros+20] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. "How social robots influence people's trust in critical situations". In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2020, pp. 1020–1025. DOI: 10.1109/RO-MAN47096.2020.9223471.

[SB15] Gayane Shalunts and Gerhard Backfried. "SentiSAIL: sentiment analysis in english, german and russian". In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Ed. by Petra Perner. Springer. 2015, pp. 87–97. ISBN: 978-3319210247. DOI: 10.1007/978-3-319-21024-7_6.

[SBC16] Gayane Shalunts, Gerhard Backfried, and Nicolas Commeignes. "The impact of machine translation on sentiment analysis". In: *Data Analytics* 63 (2016), pp. 51–56. URL: https://biblio.ugent.be/publication/8116621/file/8132035#page=64.

[Sch01] Klaus R Scherer. *Appraisal considered as a process of multi-level sequential checking*. Ed. by Klaus R Scherer, Angela Schorr, and Tom Johnstone. New York and Oxford: Oxford University Press, 2001, pp. 92–120.

[Sch05] Klaus R Scherer. "What are emotions? and how can they be measured?" In: *Social Science Information* 44.4 (2005), pp. 695–729. DOI: 10.1177/0539018405058216.

[Sch87] Klaus R Scherer. "Toward a dynamic theory of emotion: the component process model of affective states". In: *Geneva Studies in Emotion and Communication* 1 (1987), pp. 1–98. URL: https://ci.nii.ac.jp/naid/10027433990/en/.

[Sha+20] Muhammad Haroon Shakeel, Safi Faizullah, Turki Alghamidi, and Imdadullah Khan. "Language independent sentiment analysis". In: *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*. 2020, pp. 1–5. DOI: 10.1109/AECT47998.2020.9194186.

[SHL19] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. "How to do better with gender on surveys: a guide for hci researchers". In: *Interactions* 26.4 (June 2019), pp. 62–65. ISSN: 1072-5520. DOI: 10.1145/3338283.

[SI13] Jared Suttles and Nancy Ide. "Distant supervision for emotion classification with discrete binary values". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer, 2013, pp. 121–136. ISBN: 978-3642372568.

[SK12] Shiv Naresh Shivhare and Saritha Khethawat. *Emotion detection from text*. 2012. DOI: 10.48550/ARXIV.1205.4944. arXiv: 1205.4944 [cs.HC]. URL: https://arxiv.org/abs/1205.4944.

[SN07] Jun'ichiro Seyama and Ruth S Nagayama. "The uncanny valley: effect of realism on the impression of artificial human faces". In: *Presence* 16.4 (2007), pp. 337–351. DOI: 10.1162/pres.16.4.337.

[Soua] University of Southern Denmark. *Kerstin Fischer – CV*. https://findresearcher.sdu.dk:8443/ws/portalfiles/portal/cv/e213ebee-0ec6-4b9f-ba09-d7bfe5b33599?locale=en_GB. Accessed on Nov. 27, 2021.

[Soub] University of Southern Denmark. *Kerstin Fischer – Fingerprint*. https://portal.findresearcher.sdu.dk/en/persons/kerstin/fingerprints/. Accessed on Nov. 27, 2021.

[SPL06] Marc Schröder, Hannes Pirker, and Myriam Lamolle. “First suggestions for an emotion annotation and representation language”. In: *Proceedings of LREC*. Vol. 6. 2006, pp. 88–92. URL: http://lrec.elra.info/proceedings/lrec2006/workshops/W09/Emotion-proceeding.pdf#page=97.

[SS05] John Sabini and Maury Silver. “Ekman’s basic emotions: why not love and jealousy?” In: *Cognition & Emotion* 19.5 (2005), pp. 693–712. DOI: 10.1080/02699930441000481.

[SS21] Andrea Scarantino and Ronald de Sousa. “Emotion”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University, 2021.

[Sve00] Jan Svennevig. *Getting acquainted in conversation: a study of initial interactions*. Vol. 64. Pragmatics & Beyond New Series. John Benjamins Publishing, 2000. ISBN: 978-9027299055. URL: https://books.google.at/books?id=uZE9AAAAQBAJ.

[TD00] Kyle James Tusing and James Price Dillard. “The sounds of dominance. vocal precursors of perceived dominance during interpersonal influence”. In: *Human Communication Research* 26.1 (2000), pp. 148–171. DOI: 10.1111/j.1468-2958.2000.tb00754.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-2958.2000.tb00754.x.

[TO92] Terence J Turner and Andrew Ortony. “Basic emotions: can conflicting criteria converge?” In: *Psychological Review* 99.3 (1992), pp. 566–571. URL: https://psycnet.apa.org/buy/1992-41841-001.

[TP14] Erik Tromp and Mykola Pechenizkiy. “Rule-based emotion detection on social media: putting tweets on plutchik’s wheel”. In: *arXiv preprint arXiv:1412.4682* (2014). DOI: 10.48550/ARXIV.1412.4682. URL: https://arxiv.org/abs/1412.4682.

[TR11] Jessica L. Tracy and Daniel Randles. “Four models of basic emotions: a review of ekman and cordaro, izard, levenson, and panksepp and watt”. In: *Emotion Review* 3.4 (2011), pp. 397–405. DOI: 10.1177/1754073911410747.

[Tym+19] Karsten Tymann, Matthias Lutz, Patrick Palsbröker, and Carsten Gips. “Gervader-a german adaptation of the vader sentiment analysis tool for social media texts”. In: *LWDA*. 2019, pp. 178–189.

[Uni] Coburg University. *Barbara Kühnlenz – CV*. https://www.coburg-university.de/about-us/academic-center-for-sciences-and-humanities/prof-dr-barbara-kuehnlenz.html. Accessed on Nov. 27, 2021.

[Uvn98] Kerstin Uvnäs-Moberg. "Oxytocin may mediate the benefits of positive social interaction and emotions". In: *Psychoneuroendocrinology* 23.8 (1998), pp. 819–835. ISSN: 0306-4530. DOI: 10.1016/S0306-4530(98)00056-0. URL: https://www.sciencedirect.com/science/article/pii/S0306453098000560.

[VAG16] David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. "EN-ES-CS: an English-Spanish code-switching Twitter corpus for multilingual sentiment analysis". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 4149–4153. URL: https://aclanthology.org/L16-1655.

[VAG17] David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. "Supervised sentiment analysis in multilingual environments". In: *Information Processing & Management* 53.3 (2017), pp. 595–607. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2017.01.004. URL: https://www.sciencedirect.com/science/article/pii/S0306457316302540.

[War06] Martin Warren. *Features of naturalness in conversation*. Vol. 152. John Benjamins Publishing, 2006.

[Wei66] Joseph Weizenbaum. "Eliza–a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1 (1966), pp. 36–45.

[Whi11] Demian Whiting. "The feeling theory of emotion and the object-directed emotions". In: *European Journal of Philosophy* 19.2 (2011), pp. 281–303. DOI: 10.1111/j.1468-0378.2009.00384.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0378.2009.00384.x.

[WWW20] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. "Transmodality: an end2end fusion method with transformer for multimodal sentiment analysis". In: *Proceedings of The Web Conference 2020*. New York, USA: Association for Computing Machinery, 2020, pp. 2514–2520. ISBN: 978-1450370233. URL: https://doi.org/10.1145/3366423.3380000.

[YV20] Ashima Yadav and Dinesh Kumar Vishwakarma. "Sentiment analysis using deep learning architectures: a review". In: *Artificial Intelligence Review* 53.6 (2020), pp. 4335–4385. DOI: 10.1007/s10462-019-09794-5.