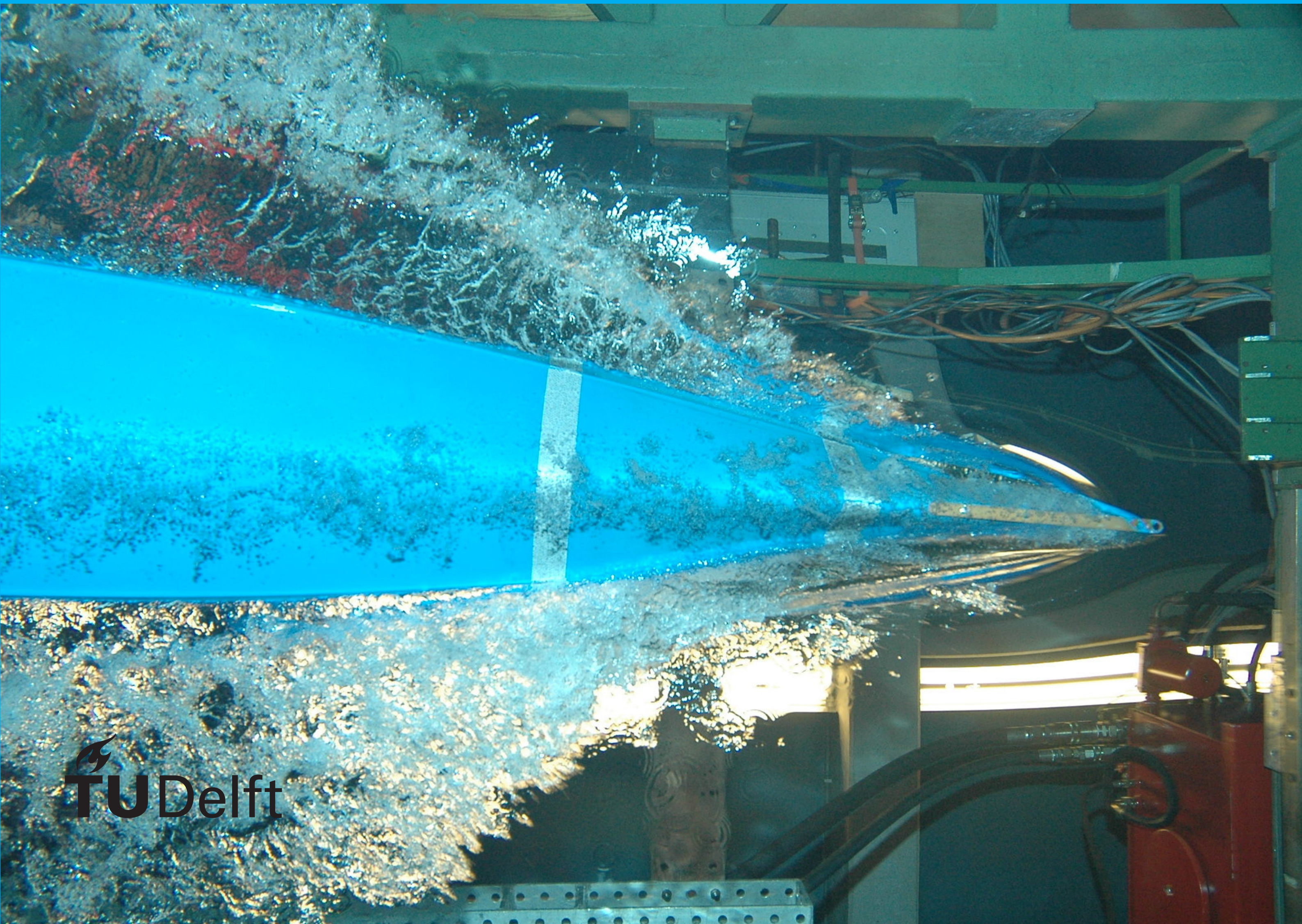
Value-sensitive Evaluation of Hybrid Human-AI Chatbots in Customer Services

Quentin Lee



Value-sensitive Evaluation of Hybrid Human-AI Chatbots in Customer Services

by

Quentin Lee

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on December 14 2022.

Student number:	4687671	
Project duration:	November 8, 2021 – December 14, 2022	
Thesis committee:	Prof. dr. ir. G.J.P.M. Houben,	TU Delft, Full Professor
	Dr. J. Yang,	TU Delft, Assistant Professor
	Dr. A. Lukina,	TU Delft, Assistant Professor
	Dr. X. Tian,	KWR Water Research Institute

Abstract

The state-of-the-art shows the potential of chatbots and other Machine Learning (ML) models to perform many tasks of high quality. Especially chatbots are already used by many companies to assist their customer service. However, chatbots will likely never be able to perform all tasks perfectly. Therefore, it is still the question whether such a chatbot is valuable for a business. Current research fails to describe how chatbots should be evaluated to compute the value of a chatbot for a business. In this research, we design an evaluation framework capturing the value of a chatbot in customer service. This framework consists of several key dimensions which should be computed in order to determine the value of the chatbot. To show that this evaluation framework captures the value of a chatbot, we perform a case study on water utility companies in The Netherlands. This case study showed that the designed evaluation framework does capture the value of a chatbot in customer service.

Preface

This research has been done in collaboration with KWR Water Research Institute. I want to thank KWR for the opportunity to do this research. We also want to thank all people from the water utility companies in The Netherlands for their collaboration in this research. Finally, we want to thank Dr. X. Tian from KWR and Dr. J. Yang from TU Delft for their supervision during the research.

Quentin Lee
Delft, December 2022

Contents

List of Figures

List of Tables

1
Introduction

The current state-of-the-art conversational agents and chatbots are useful in many different fields such as customer service interaction [32] and education [36]. With current state-of-the-art techniques, chatbots have the potential to be of high quality and ability to perform many tasks. However, chatbots will likely never be able to perfectly perform all tasks. This has two consequences: 1) In order to mitigate the risk and resolve mistakes a chatbot can make, many chatbots are integrated in a hybrid human-AI manner, where there are humans in the loop who could act when the chatbot fails [10]. 2) When a chatbot fails to respond properly, there is some cost. This could be cost in terms of, decrease of customer satisfaction, as the problem for the customer is not solved, but also in terms of money and workload, as an employee has to jump in and resolve the mistake. This raises the question how it could be computed how valuable a chatbot actually would be for a business and other stakeholders.

Currently, chatbots are mostly evaluated on accuracy-based metrics [35]. These metrics, for example, measure how accurately a chatbot can classify the meaning of a message or how accurately it can extract information from messages. The problem with evaluating chatbots solely with such accuracy-based metrics is that it is context-dependent whether an accuracy score is sufficient [12, 40]. They do not measure the effect of the performance of the chatbot in the context it is used in. For example, an accuracy of 80 percent is of higher value when the chatbot solely serves for entertainment in comparison with a chatbot prescribing medicines to customers. In order to solve this problem, one should take the context of the real-world scenario into account when evaluating chatbots. Therefore, when evaluating a chatbot, the overall value of the chatbot should be considered [12] by evaluating many if not all of these important perspectives.

When computing the value of a Machine Learning system, we aim to go beyond accuracy metrics when evaluating the system by encompassing all value (or cost) derived from the use of it. To evaluate such a system, it is important to understand what actually leads to value, whether that be in a business, utility, or ethical context [12]. Operationalizing value for a specific task or context can be very challenging, as fully quantifying the value derived from a model is virtually impossible for any real-world application. Therefore, a number of simplifications have to be made, deciding on what requires extensive research into the specific area of application. Depending on the degree of simplification used to obtain the value, the metric of value can become meaningless, which implies that researchers must take care to strike an appropriate balance between informativeness of a value-based evaluation and the possibility of its operationalization.

In order to conform to the informativeness property, we must ensure that the information or value the evaluation framework provides, is meaningful to the stakeholders. We must figure out what stakeholders needs and values are and what actually creates value for them. Second, to conform to the operationalizability property, we need to convert the designed evaluation framework to concrete metrics. These metrics simplify the evaluation framework and make it possible to apply and process the results of the evaluation framework and finally conclude to a practical value.

This research will answer the question how such a value-based evaluation could be performed for chatbots in customer service. First, such a value-based evaluation of chatbot is not yet described in current literature.

Current literature does describe how systems could be designed based on the values of humans [10]. However, research about evaluating whether such system does conform to the determined values is still missing. Therefore, any research about computing the value of a chatbot is also still missing. Second, research describes what metrics evaluate the business perspective of a chatbot [35]. However, showing how these metrics would actually lead to value is still missing in current research. Finally, current research did show that it is important for both the research and enterprise community to consider the notion of value when evaluating any Machine Learning model [12]. Current research does however not yet show what this notion of value is and how we could actually compute this value.

This research will answer the following research questions and its sub-questions:

Main RQ. How can we determine the practical value of a chatbot for processing customer complaints automatically?

Sub RQ (a). How do we capture the needs, desires and values of the stakeholders?

Sub RQ (b). How do we design the evaluation framework?

Sub RQ (c). How do we validate that the evaluation framework conforms to both the informativeness and operationalizability property?

To conform to both the informativeness and operationalizability property, we use the Research through Design approach. Research through Design is a research methodology where one conducts research with the intent to create new knowledge [50]. We start with a formative study, where we do some preparations before designing the evaluation framework. The goal of this formative study is to consume already existing knowledge in this field and guide us to designing the actual evaluation framework. This formative study consists of two steps, a literature study and interviews. First with the literature study, we want to gain knowledge about what current literature describes about the problem of this research. Second, with the interviews, we want to analyse how this current literature matches with the given context and how stakeholders in this context think about this research problem. With the interviews, we also want to know how stakeholders would intuitively tackle the problem of this research, such that we can keep the execution of this framework operationalizable.

We will design the evaluation framework based on the literature study and interviews. We will combine the results from the literature study and interviews to a final list of dimensions and related metrics. We will analyze these dimensions and metrics and come with an approach to compute the value with the results from these metrics. As the design is based on these two sources of information, the resulting evaluation framework should conform to both the informativeness and operationalizability property.

Finally, once this framework is designed, we validate that this evaluation framework is operationalizable and conforms to the informativeness property. To validate that the evaluation framework conforms to these two properties, we will perform a case study on the designed evaluation framework. The case study will be performed at the KWR Water Research Institute (KWR). We perform the case study on water utility companies in The Netherlands and we will evaluate what value a chatbot has on customer service of water utilities in The Netherlands.

At the end of this research, we have designed an evaluation framework which captures the practical value of a chatbot in customer service and performed a case study of this evaluation framework on a chatbot in water utilities of The Netherlands. With this research, we want to show the key dimensions of a chatbot in a customer service, how we can evaluate each of these dimensions and how we finally aim to evaluate the value of the chatbot based on the evaluation of the chatbot on these dimensions. The project provides the following contributions:

1. An evaluation framework to compute the value of chatbots in customer service. The evaluation framework consists of dimensions and metrics. We show how with the results of these metrics we can compute the value of the chatbot.

2. Show how this evaluation framework could actually be applied in a real-world scenario with the use of a case study.

3. Show how the results from the evaluation framework can result in useful insights of the chatbot for businesses. These insights assist businesses in deciding whether the chatbot would be valuable for their business to use.

2 Related work

In this chapter, work related to this research is presented. We will discuss what type of chatbots there are and what techniques are generally being used. Then, some state-of-the-art Natural Language Processing (NLP) chatbots will be discussed. After that, we will discuss what Machine Learning Decision Rejection algorithms exist. Machine Learning Decision Rejection is important to create hybrid human AI systems with Machine Learning models as we must determine when to redirect a task to a human. We will also take a closer look at automation and what we should take into account when deciding what to automate when automating (business) processes, such as automating resolving customer complaints. Finally, we will discuss how and from what different perspectives a chatbot can be evaluated on its practical value.

2.1. Chatbots

2.1.1. Chatbots in general

Chatbots are a new kind of conversational agent that mimic human conversations using different techniques [37]. Chatbots can be designed for many different use cases. It could be a used for entertainment purposes or assistance in businesses. These days, many companies also use chatbots for their customer service, since manually answering every questions from customers is very time-consuming [49].

When developing a chatbot, there are 2 two major types of chatbots. These are **social** and **task-oriented** chatbots [31]. A social chatbot is a chatbot that is designed to interact with a human and have more of an unstructured conversation. A task-oriented chatbot is a chatbot that is designed to perform a specific task.

2.1.2. Types of chatbots

When building a chatbot, there are a few different approaches one can take. The first approach is a **rule-based** chatbot. A rule-based chatbot is a chatbot that uses a set of rules to determine what to say. The other approach is the data-driven approach [31].

A **data-driven** chatbot is a chatbot that uses a set of data to determine what to say. This is also one of the more recent approaches. In order to build a data-driven chatbot, one could use different approaches. The first approach is the **Information Retrieval** approach. With such chatbots, we query the answer using a search engine. The second approach is the **Machine Learning** approach, where we use machine learning models to determine the answer. The two most popular Machine Learning techniques are Reinforcement Learning and Sequence-to-Sequence Learning.

2.2. NLP chatbots

This section will discuss what techniques are being used in state-of-the-art NLP chatbots.

If we look at the system proposed by Lalwani[27], it shows that they split their chatbot system in different parts. First they try to find the context of the message. During this stage,vs it will also be classified as a specific question or a message for a "normal" conversation. If the message is for a normal conversation, Artificial

Intelligence Markup Language (AIML) is used to match the message to a pattern using pattern matching. If the message is a specific question, the system will try to find the answer to the question in the questions set. To extract information from the message, they use NLP methods such as lemmatization and POS tagging with WordNet. Since every question can be asked in many different ways, the system tries to find the question with the highest similarity to the message.

When analyzing other systems, such as the one described by Handoyo [19], they often have at least two components. These components are **intent classification** and **entity recognition**. The intent classification component is used to determine the intent of the message. The entity recognition is used to extract the entities in the message. With the intent and extracted entities, an answer can be generated.

Another necessary component found in literature, is a so-called conversational flow or **dialogue management**. This is used to determine what type of message the chatbot should send back. To implement the conversational flow or dialogue management, we can see a few different approaches. As described by Handoyo [19], we could simply create a flow in the form of a diagram. Two other approaches are described by Ayanouz [4]. The first approach uses a feedback algorithm where we learn from mistakes for future conversations. The second approach is the so-called policy learning approach. In this approach, we try to learn the happy paths of the conversation and during the conversation we will try to follow one of the happy paths.

The final important component missing in this architecture is some sort of **answer generation** where based on the intent and extracted information, we generate a response for the user. This can be done in a few different ways. We could simply use a set of predefined question-answer pairs. Another approach is to use a machine learning model to generate the answer.

The remainder of this section will more in-depth discuss how each component can be implemented with the current state-of-the-art.

2.2.1. Intent classification

If we look at different methods used for intent classification, we can see that there are many different methods which could be used. Jiao [24] describes two different methods. First, they used a the RASA NLU pipeline to train to classify the intents. This pipeline uses transformers. The other method described is to use neural networks to classify the intents. It showed that the RASA NLU method is more accurate than the neural network method. Handoyo described a method to use similary methods to classify the intents of the message.

Kulkarni [26] describes a method to use vectorization where we convert the message into a bag-of-words (BOW) model which is then used to classify the intents using a trained classifier.

Finally, Mathew [29] describes a way to use K-Nearest Neighbors (KNN) to classify the intents of the message. KNN also uses similarities to classify the intent of the message.

2.2.2. Named entity recoginition (NER)

For entitiy recoginition, Jiao [24] also uses the transformer from the RASA NLU pipeline. Outside of that, not many papers about chatbots explicitly described methods for named entity recognition. However, if we dive a bit further in the research, we can see that there are multiple different methods for named entity recognition.

Wu [48] shows that you could also use CRF (Conditional Random Fields) or DNN (deep neural networks) for NER. It showed that DNN outperforms CRF on the task of NER. Both methods use word embeddings to represent the words in the message. Many other papers use some sort of neural network to perform NER. Hofer [20] uses word embeddings, long short-term memory (LSTM) and dense neural network layer to perform NER. They improved their model by pretraining the model, such that few-shot learning can be applied. Pretraining the neural network on a large labeled corpus or tuning the hyperparameters allowed us to still have a decent score for NER. It however was not able to outperform a model trained on a large corpus.

However, if we look at NER approaches, we can see that there are many different approaches. Gong [18] describes how to use BERT for NER. In their approach they simply put the output of BERT into a neural network which will then apply NER. Since we want to label every word of the message, there was experimented with both Gated recurrent Unit (GRU) and LSTM and because of faster training times, GRU was preferred in this approach.

2.2.3. Answer generation

When analysing answer generation methods specifically used in chatbots, we can see that many chatbots do not use very complex methods for answer generation. Some chatbots do not even have a specific algorithm, because the goal is to retrieve data from the user. Therefore, simple responses such as, "data incorrect" or "data received" are sufficient [19].

Other chatbots use intent classification and NER to retrieve the relevant data and this will be used to generate an answer using predefined question-answer pairs [24]. Another way of storing question-answer pairs is with the use of AIML. With AIML we store patterns of questions and their corresponding answers in XML files. If a pattern matches with the message, the answer will be generated based on the predefined answer.

Ayanouz [4] describes a method where we let the machine learn question-answer pairs using a neural network and based on the given question, the corresponding answer will be returned by the algorithm.

2.2.4. Dialogue and context management

Dialogue and context management are two complex systems that should be implemented in chatbots. First, dialogue management determines what to answer given the current state of the conversation. Next, context management stores and manages information from previous messages so they can be used in future messages.

When designing a dialogue management system, there are two different approaches we can take. The first approach is the **Strong System Initiative Interactions**. In this approach, the chatbot takes the initiative in the conversation. This is the easier approach to implement as the chatbot has the control over the conversation [16]. The second approach is the **Weak System Initiative Interactions**. In this approach the chatbot does not necessarily take the initiative and often just replies with opinions not necessarily asking for a response from the user.

When analyzing the literature of dialogue management, many different approaches can be found to implement the dialogue management system. Bocklisch [8] describes a method to use the RASA pipeline for Natural Language Understanding (NLU) for dialogue management. In this approach, a vector is being created which contains the last action and its intent and a vector of slots with their corresponding value. Then it learns what the next best action would be given the last action and intent using examples and active learning. The advantage of this is that you do not need a lot of training data. However, this also means that the accuracy of the chatbot will not be very good in the beginning, but it will improve increasingly once the chatbot starts to learn from its mistakes.

Another approach described by Finch [16] uses predicates and inference for their dialogue management system. When the user sends a message, a list of predicates will be defined using NLP algorithms. These defined predicates will be matched with defined logical conditions to find the right corresponding answer. The advantage of this approach is that it is easy to implement. However, the conversation feels less natural since the chatbot will give the same response every time for the same message.

2.3. Machine Learning Decision Rejection

When using NLP and Machine Learning algorithms to classify data, it is sometimes difficult to classify messages. Many classification algorithms force the classifier to make a classification. However, in these cases where the classifier does not actually know the answer, the algorithm should not be forced to make a classification [13]. Instead, an algorithm should be implemented which decides whether the input should be accepted or rejected. When designing such an algorithm, there are two values which are important to consider; the cost of rejecting an answer and the cost of accepting an incorrect answer.

To implement a decision rejection algorithm, there are two main approaches one can consider. The first one is a *confidence-based* approach. In this approach, a confidence score is set for the output of the classifier and this will be compared with a threshold. If the confidence is lower than the threshold, the classification will be rejected. There are multiple approaches to decide on the threshold. One could use a probabilistic approach in which we calculate a probabilistic model to decide on the optimal threshold [47]. Another option is to decide on a rejection cost and surrogate risk loss function to train the model on and decide on a threshold [33] or rejection condition. An example of a rejection condition is that the answer is rejected if the confidence scores of all answers are negative [13]. What type of rejection condition is chosen depends on the problem and the chosen loss function as not all loss functions will have similar values.

In the other approach, called *separation-based* approach, a rejection classifier is trained simultanuously with the classifier to learn when to reject the answer of the classifier. For this approach, it is very important to know the cost of rejecting [3] and the cost of giving an incorrect answer [13], so the rejection classifier can be trained to minimize the total cost. However, this is difficult, especially in the case of multi-class classification [33]. Multi-class classification is a classification problem where the algorithm must classify the data into three or more classes instead of two.

Finally, there is a final approach worth mentioning. This approach looks at the problem differently. All answers from the algorithm can be assigned to the accept set and reject set. Given the two sets from the training phase, the cosine similarity of each answer will be calculate with regards to these two sets. Given the cosine similarity, the answer will be either accepted or rejected [38]. So in this approach, it will be calculated if the answer looks more similar to a correct or rejected answer.

2.4. Automating (business) processes

When taking a broader look at this problem of automating business processes, some interesting insights could be found. Schumann [42] describes that improvements can take place on individual level, department level and enterprise level when specifically focusing on office automation. Some examples of how office automation could possibly result in improvements are:

1. individual level: reduced workload or improved quality of the delivered work.
2. Department level: better coordination within department or better access of information.
3. Enterprise level: more flexibility in the way the business processes are executed

2.4.1. Different levels and types of automation

Parasuram et al. [34] created a model for different types and levels of automation. This model could help deciding what type of automation should be used and how one should reason whether it would be beneficial. According to this model by Parasuram et al., there are ten different levels of automation. These are:

1. The computer offers no assistance: human must take all decision and actions.
2. The computer offers a complete set of decision/action alternatives, or
3. narrows the selection down to a few, or
4. suggests one alternative, and
5. executes that suggestion if the human approves, or
6. allows the human a restricted time to veto before automatic execution, or
7. Executes automatically, then necessarily informs humans, and
8. Informs the human only if asked, or
9. Informs the human only if it, the computer, decides to.
10. The computer decides everything and acts autonomously, ignoring the human.

These levels of automation go from no automation at all to completely automating the process. The levels in between each remove the human from some extend from the process. The more human interference is removed from the process however, the higher the chance that a mistake or failure from the automation will go undetected and possibly cause additional costs.

When automating processes, there are four different types or classes in which the automation can be classified [34]. These are:

1. Information acquisition (e.g. use software and/or sensors to gather data and register/process it in the system)
2. Information analysis (e.g. use software to analyze data and show it to the user)
3. Decision and action selection (automate the decision process of what action should be executed by the system)
4. Action implementation (execute the action automatically)

2.4.2. Risks in automation

When applying automation, there are risks which can occur. While some simple risks are errors made by the automated process, there are also other categories which might suffer from automation. Such a risk is for example reduced situational awareness or skill degradation of employees[34]. Therefore, when deciding to automate a process, it is important to consider the risks and to reason whether the benefits of automating the process would outweight the risks and costs.

2.5. Evaluating and measuring practical value of a chatbot

According to a model for chatbot evaluation created by Peras [35], chatbots can be evaluated in five different perspectives. These are:

1. The **user experience** perspective looks mostly at the usability of the chatbot. It will be analysed how easy it is for a user to use the chatbot and whether the user expectations of using a chatbot are satisfied.
2. The **linguistic** perspective analyses whether the chatbot can give appropriate responses in correct grammar and spelling. It will also look at the quality of the responses and whether the responses fit the current conversation.
3. The **technology** perspective analyses to what extend the chatbot express human like behaviour.
4. The **information retrieval** perspective looks at how well the chatbot can meet the information requirements of the user. Therefore, it will be analysed how precise the chatbot is in its response to the user and whether the information it returns is correct.
5. The **business** perspective analyses how the chatbot impacts businesses. It will be analysed for example how many conversations can be automated and how many employees are still required per x conversations.

With each of these perspectives, chatbots can be evaluated both in a qualitative and quantitative manner. Each perspective has its own advantages and disadvantages. For example, evaluating the user experience is very time consuming and expensive [35] as it often requires people to evaluate the chatbot and fill in surveys. Other perspectives such as the information retrieval perspective are easier to evaluate. However, it does not provide any qualitative results. Therefore, when evaluating a chatbot, it is important to consider which perspectives should be evaluated based on what is important for the business using the chatbot.

2.6. Cost-benefit analysis of Information Systems

When analyzing the current state-of-the-art research of computing the value of a ML model, the field closest to this is the field of cost-benefit analysis of Information Systems. In this field, different approaches to cost-benefit analysis are described for information systems. An analysis of this field would give a good starting point computing the value of ML models and chatbots in particular.

One common problem in doing such analysis, is that some benefits and costs are intangible, and that these intangible measures cannot easily be assigned a value to. King[25] proposes several approaches to still use these intangible costs and benefit in your analysis. First, they describe to set lower or upper bounds of these intangible benefits and that the value should be in between them. Next, also trade-offs can be described where for example a lower intangible benefit should also have a different benefit of a higher value.

Other aspects that should be taken account of in cost-benefit analysis are risks and uncertainty. When deciding on the formula for the cost-benefit analysis, these risks and uncertainties could be taken account of with the use of discount factors and the probability of the risk happening [15].

A cost-benefit analysis would be very useful when implementing and incorporating an information system, as it will help considering alternatives and making decisions for the information system [15]. There are many different approaches to take when performing a cost-benefit analysis on an information system as described by Sassone [39]. First, there is the **break-even analysis**. In this analysis, it is being compared whether the costs and benefits are equal. Disadvantage here is that it does not really take into account uncertainty and risks. However, it is very cheap do to such analysis. Next, there is the **subjective analysis**. When performing a subjective analysis, managers or stakeholders make decisions based on the different calculated costs and benefits. One disadvantage here is that the result of the cost benefit analysis is based on the knowledge of the ones computing the result. If those people have good knowledge, then the result of the analysis will be useful, otherwise the result might not be a good estimate of the value information system. Another approach is the **cost effectiveness analysis**. In such analysis we try to find the best option among multiple options. After that, there is **the time savings times salary (TSTS)** approach. In this approach, it will be computed how much in terms of workers' time is saved. Finally, there is the **work value model**. In this model, instead only computing how much time is saved, it also calculates how time and resources are allocated within a company. When assuming that the changed allocation of resources are an effect of optimizing behaviour, the values of the workers can be inferred.

2.7. The value of Machine Learning models

Originally derived from value-sensitive design methodologies [17], in the context of AI, value can capture and quantify the entirety of the advantages that users and other stakeholders experience due to the deployment of the chosen ML model and provides a clear target to optimize for.

Value is especially critical to properly evaluating systems where humans interact with AI, as it allows for the formulation of a metric that encompasses the benefits and costs of multiple involved parties, i.e. stakeholders. This is especially true considering value can also be an effective way of implementing certain social principles or moral requirements into the building and evaluation of ML models [44]. When considering the context of chatbots, value can be a useful metric to gauge the benefit derived from employing the ML model in this context from the perspective of the stakeholders interacting with it on either side.

To get a better notion of how one should calculate the practical value of a chatbot, it should first be researched how the value of a Machine Learning model can be calculated as the chatbot to be used is based on a Machine Learning model.

When calculating the value of a Machine Learning model, current literature shows that it is important to take into account three metrics [41]. These metrics are:

1. Cost of the default flow,
2. Cost of wrong prediction,
3. Benefit of correct prediction.

When incorporating a ML model in a business, there are typically three flows to be implemented in the business process. These flows are the flow without ML model, the flow where the ML model makes a wrong prediction and the flow where the ML model makes a correct prediction. Each of these flows corresponds respectively with the three described metrics.

The reason for using value to build and evaluate ML models is to achieve improved outcomes, i.e. higher value for stakeholders. Thus, it is key to establish a baseline to which we can compare a successful model, resulting in a value higher than the one of the baseline, and an unsuccessful model to make value-sensitive judgements on model performance.

3 Method

This chapter describes the method of this research. It describes which steps are taken in this research and how each of these steps answers each of the different research questions. Broadly, this research consists of four steps.

These fours steps are:

1. Formative study: Literature review
2. Formative study: Capture stakeholder needs
3. Design evaluation framework
4. Validate evaluation framework

The figure below describes more in depth how each of these steps are merged together into one research project.

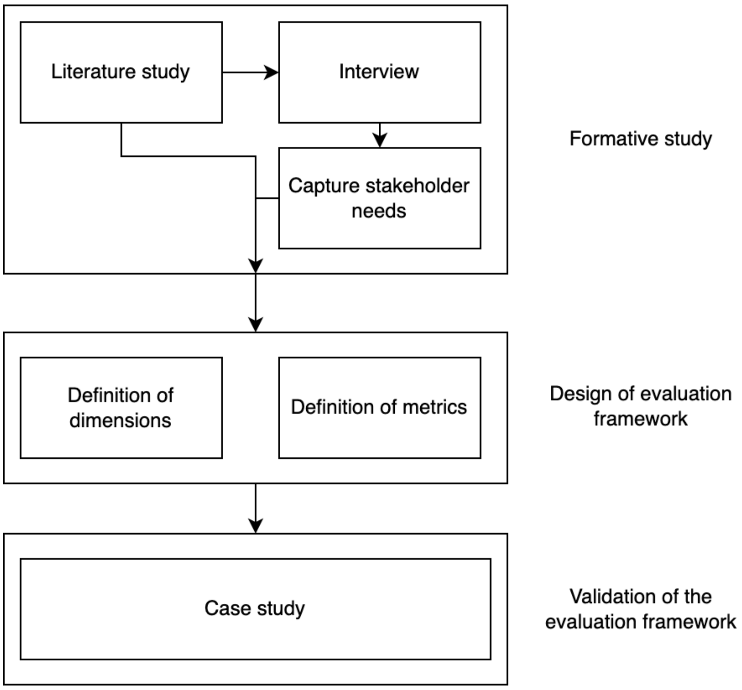


Figure 3.1: High level design of this research

3.1. Formative study: Literature study

The first step in this research is to perform a literature study. This literature study has two goals. First, we want to analyse what has already been researched in this research context. Second, we want to research what the best approaches are to tackle certain steps in this research such as the interviews and the case study.

3.1.1. Literature survey: researching the state-of-the-art

In this literature study about the state-of-the-art of evaluating chatbots in customer service, we study different aspects of evaluating chatbots in customer service. We describe which fields we study and how we come to the final analytical dimensions with these fields. Figure 3.2 shows how this literature study is structured.

When performing the literature study on state-of-the-art chatbots in customer service, there are different topics to research. First, research should be conducted in the field of **online systems**. Chatbots are part of an online system. A chatbot should adhere to the qualities of an online system. An online system refers to any system available by means of accessing it through the internet. Such systems have different requirements such as being responsible, available, and hardware agnostic. Therefore, it should be analyzed whether such requirements can be affected by a Machine Learning model like a chatbot. Research in online systems showed that three aspects of online systems should be analyzed. These are Quality of Service (QoS), Quality of Experience (QoE) and Quality of Business (QoBiz) [45]. Every dimension of the evaluation framework should fit in one of these three aspects. Online systems are the baseline from which we will continue our literature study

During the literature study, in addition to the field of online systems, there are three fields of research that should be studied. The first field is **Robotic Process Automation (RPA)**. This is the most important field that should be researched. RPA is the application of technology and methodologies aiming to automate repetitive human tasks [23]. As a chatbot in customer service fits the description of RPA, it should be evaluated as a RPA system. Research in how RPA systems are evaluated should also describe how we can evaluate chatbots on their degree of automation.

Second, we should study **SERVQUAL**. SERVQUAL is a scale to measure service quality [6]. This shows how we should measure a service to determine its service quality. Conducting a literature review in this field shows what importances there are in services to measure.

Finally, a literature study in **customer service** itself is necessary. Studying customer service aspects shows what aspects are most important in customer service to measure. Analyzing those aspects and determining which aspects could actually be influenced by a chatbot in customer service should show what dimensions in a chatbot should be computed in order to find the practical value of a chatbot in customer service of water utility.

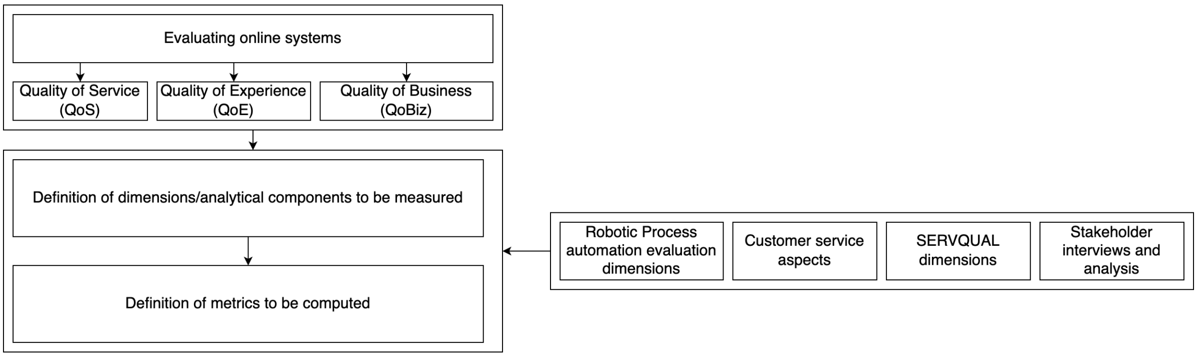

Evaluating online systems
Quality of Service (QoS)
Quality of Experience (QoE)
Quality of Business (QoBiz)
Definition of dimensions/analytical components to be measured
Definition of metrics to be computed
Robotic Process automation evaluation dimensions
Customer service aspects
SERVQUAL dimensions
Stakeholder interviews and analysis

Figure 3.2: Design of the literature study

3.2. Formative study: Capturing stakeholder needs, desires and values

The second step for this research is to capture the stakeholder needs, desires and values. The goal of this study is to compare what we found in the literature study with the thoughts and opinions of the stakeholder. The literature study contains knowledge about the general use case. However, the context of this research are water utility companies in The Netherlands and therefore a study should be conducted with stakeholders of water utility companies in The Netherlands to determine what knowledge found in the literature study also holds for customer service of water utility companies in The Netherlands.

In order to capture stakeholder needs, desires and values, stakeholder interviews are conducted. With this stakeholder interview, we can collect the knowledge and opinions from the stakeholders.

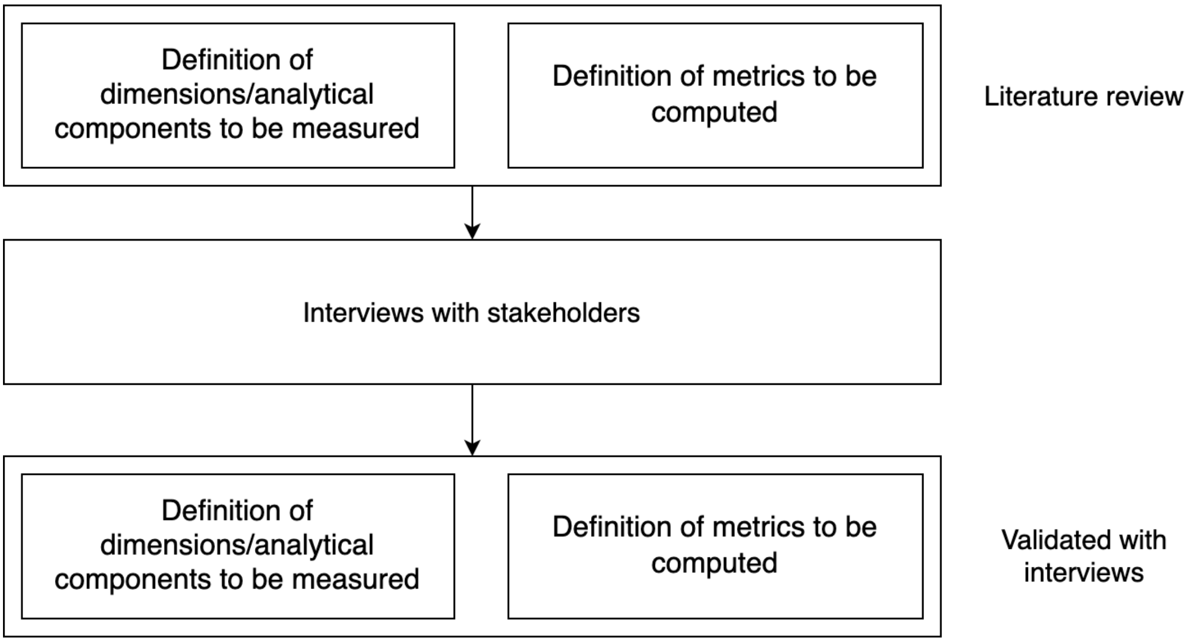

Definition of dimensions/analytical components to be measured
Definition of metrics to be computed
Literature review
Interviews with stakeholders
Definition of dimensions/analytical components to be measured
Definition of metrics to be computed
Validated with interviews

Figure 3.3: How the interviews will affect the results from the literature study

3.2.1. Interview design

One important part of this research is to interview stakeholders about chatbots in customer service. This interview has two goals. The first goal is to determine what aspects stakeholders find important in their customer service. This will help us achieve the informativeness property. We will analyse what aspects stakeholders find important in their customer service. It is important to understand what aspects in customer service are important for the stakeholders and what they find important for chatbots in customer service as those aspects need to be computed for the practical value to be meaningful. The second goal is to understand how employees in customer service would compute the practical value and corresponding metrics of their customer service with a chatbot. This knowledge would help us to conform to the operationalizability property. A more discussion-based approach will be used where we discuss with the interviewees what metrics they would use to compute the practical value of a chatbot in their customers service and how they would compute each of these metrics.

The following approach is used when designing the interview. We design a semi-structured interview as semi-structured interviews have the advantage of achieving the objective of the interview while allowing for a better understanding of the interviewees opinion or perspective as well [11]. First, the research question and its sub-questions of the interview are defined. This is important as we need to know what information we want to gather from these interviews. These are however not the exact questions being asked during the interview. We do not want to ask the exact research questions as this might not lead to optimal answers. Next, the interview questions will be defined. When defining these questions, it should clearly be stated to which research sub-question(s) this question belongs. In order to get most out the answers from the participants, it is important that the questions are open-ended and to not assume anything during the questions in order to give participants as much freedom to answer as possible [43]. As we have existing research in the field of chatbots in customer service and ask open-ended questions to the interviewees, we use a mixed deductive-inductive approach in this interview. We want to match gained knowledge from the interview with knowledge from the literature study. We also want to create new knowledge if possible from the interview. Therefore a mixed deductive-inductive approach is used for this interview

In a deductive approach in qualitative research, analysis is based on pre-existing theories [5]. As there is a lot of research about the importances of customer service, the key of these interviews is to find relations between the pre-existing theory and this specific use case. With the pre-existing research, pre-defined categories can be generated and it will be analyzed what pre-existing research still holds in this specific use case of customer service in water utilities. In deductive research, one tests existing theories and tests in the theory if it applies to a specific use case [21]. In this study, theory about customer service, customer satisfaction and chatbots are applied to the use case of customer service in water utility companies and it is being tested which theories also applies to this use case. Not all theory applies to this use case as a water utility company is different from other companies. One reason why water utility companies are different than other companies is that customers do not have a choice whether they want to join a water utility company or not. In an inductive approach, we want to generate knowledge purely based on statements made by the interviewees [5].

Finally, for the interview design, a pilot interview is conducted such that flaws in the interview design can be detected and that the interview can be improved based on the results from the pilot interviews [43]. These interviews are conducted to participants who could also be actual participants for the real interviews.

The design of the interview is as follows. First, the main question of the interview is determined, which is then divided in several sub-questions. Then finally, for each of these sub-questions, concrete interview questions are designed to be asked to the interviewees.

Interview main question: What do customer service employees in water utilities find important when incorporating a chatbot in a customer service of water utility companies in The Netherlands.

Sub-questions:

1. What are the most important aspects in customer service?
2. What problems do they want to solve with chatbots?
3. What (aspects in customer service) do they want to improve with chatbots?
4. How do they want to improve their customer service with chatbots?
5. What difficulties are expected when incorporating a chatbot?

Interview questions:

1. What aspects do you find most important in customer service? (sub-question 1)
2. What current issues do you find in your customer service workflow? (sub-question 2)
3. Do you see any issues in your employee satisfaction working in the customer service? (sub-question 2)
4. How do you expect a chatbot to affect your customer service (positively or negatively)? (sub-questions 3, 4)
5. What aspects in customer service should not get worse because of a chatbot? (sub-question 5)
6. What difficulties would you expect when incorporating/implementing a chatbot in the company? (sub-question 5)

For the second part of the interview, we will discuss with the interviewee what aspects they would actually measure in order to find out if the chatbot would be valuable for their water utility. Then it will be discussed how they would actually measure these aspects. It is important to know how people in the domain of customer service in water utilities would measure these aspects and whether it would be possible to measure these aspects with the available resources. The goal of this second part is to determine how employees of customer service would evaluate their customer service. As their method would be based on their values and their insights in how they could compute these metrics in their customer service, this discussion would help us to conform to both the informativeness and operationalizability property.

3.2.2. Participant selection

Besides the interview itself, it is also important to select the appropriate participants [43]. Therefore, the stakeholders are first defined. As the scope of this research is the customer service of water utilities, the stakeholders are employees of customer service in water utility companies and preferably employees in a managing function. Participants from this stakeholder group work a lot with customer service and know about their problems, their needs, and how they would value chatbots in their customer service. When determining how many participants are required for this research, the number depends on the context [7]. As the pool of participants and time are limited for this use case, the following approach is used. First, a group of five participants is interviewed. In order to determine whether this group of participants is sufficient, results from these interviews will be analysed whether they are similar or completely different. If in general, the answers to the interview are in the same scope, it could be determined that the interviews are sufficient as there is not much more to explore as all answers are in the same scope. We could reason that the whole solution space has been explored. In the interview analysis, it is described how it will be determined whether the answers between the different participants are similar.

3.2.2.1. Interview data analysis strategy

After the interviews, the data is analyzed. In order to analyze the data, the thematic content analysis method will be utilized. Thematic content analysis is used to create a descriptive presentation of data from qualitative interviews [2]. After the interview, data from the interview is extracted and filtered. Keywords and important sentences are extracted from the interviews and then labeled. When labeling these keywords, we use both the deductive and inductive approach. First, we use the results from the literature study to create those labels. Using this approach, we can easily check whether the information matches the results from literature study. Second, as the questions are open-ended, we also perform an inductive analysis where we create the labels solely based on the results from the interview.

More concretely, in this research we use a variation of the approach described by Braun and Clarke [9]. Originally, this approach consisted of 6 phases. However, given the small size of interviewees and the design of the interview, the approach described by Braun and Clarke has been adapted to fit the design of this interview. This adapted approach has the following steps:

1. Gather all data from interviews.
2. Convert raw data into codes. These codes describe the main topics of the interview.
3. Group codes into overarching themes. Themes are the final results and are most relevant to be used to answer the research questions.
4. Answer research questions and subresearch questions with resulting themes.

In the first step, raw data is gathered from interviews in the form of interview notes and transcripts from the interview. As this data is unstructured and it does not clearly show the intention of the statements of the interviewees, this raw data is converted into more concrete data with the use of codes. These codes are short sentences describing the meaning of statements of the interviewees. After that, these codes are analyzed and similar codes will be grouped into so-called themes which represent all codes in this group. Finally, these themes represent different answers to the interview questions in a clear way. Finally, with these themes the research and subresearch questions could be answered.

In order to figure out whether answers from different interviewees are similar or whether more interviews are required, the resulting themes and codes can be used. As codes result into themes and themes are used to answer the research questions, it can be analysed whether each theme does have enough corresponding codes. If there are many themes with only one or two corresponding codes, then it would be useful to find more participants since in this case the variation between answers of interviewees is too much.

3.3. Designing the evaluation framework

At this point of the research, we have done a formative study where we gained knowledge about what the important dimensions and metrics are when evaluating a chatbot in customer service. We gained this knowledge from both the literature study and interviews during the formative study. Extensive results of the formative study can be found in chapter 4

In order to design the evaluation framework with this knowledge, we compare the knowledge found from the literature study and interviews and find the intersection between these two to be the final list of dimensions and metrics. After that, we analyze the resulting dimensions and metrics and determine what relations there are between the different dimensions and how each of these dimensions is important for the value of a chatbot in customer service. Finally, based on the dimensions and its relations, we will design an approach to compute the practical value of a chatbot in customer service based on these dimensions and metrics.

3.4. Validating the evaluation framework

The final step in this research is to validate the evaluation framework. With the previous three steps, an evaluation framework has been designed. However, it should still be validated that this evaluation framework conforms the informativeness and operationalizability property, meaning that the evaluation framework captures the stakeholders values and that it is possible to apply and execute the designed evaluation framework.

In order to validate the evaluation framework, a case study will be conducted on the customer service of water utility companies in The Netherlands. This section describes the design of this case study and how we plan to collect data for the study.

3.4.1. Case study design

The final part of this research is a case study. During this case study we will apply the designed evaluation framework to the real world case of customer service of water utility companies of The Netherlands.

For the case study, we collect data from this case in order to apply the evaluation framework. In order to gather data from the chatbot, two different approaches will be used as not all required data can be acquired through only one of the two approaches.

1. Generate mock data from complaints data set from WGB and generate results with this test data.

2. Design pilot with the chatbot and let stakeholders and users try out the chatbot. Then design a survey to collect useful data for the evaluation framework.

3.4.2. Collecting information

In order to compute some dimensions of the evaluation framework, some business information must be collected. It would not be possible to apply the evaluation framework if there is no reliable estimate of data of certain business processes of water utility companies in The Netherlands. Examples of such processes are the cost of a phone call to the customer service and the cost of asking a question to the chatbot. Also other information should be collected, such as an approximation of wages of employees in the required positions of customer service.

3.4.3. Mocked data

The first approach to gather results for the evaluation framework is to use mocked data. In this approach, first mocked data (complaints) is generated using the complaints data set from WGB. Then, manually, we will ask these complaints to the chatbot, and it can be analyzed whether the chatbot is able to solve the complaints or not. With this data, it can be estimated how it would affect the efficiency of the chatbot.

This approach does however have the disadvantage that the mocked data does not contain questions or complaints customers would actually ask to a chatbot as the data is retrieved from complaints submitted through email or contact forms. Therefore, the result might not be a good representation of how the chatbot would perform in production.

Therefore, this mocked data is only used to gather some data which could then be used to estimate the sample size of the actual pilot. The mocked data would not be used for the actual evaluation of the evaluation framework.

3.4.4. Pilot data

The second approach to gather data for the evaluation framework is to conduct a pilot study. For the pilot, water utility people and potential customers will be asked to use the chatbot for some of the implemented use cases. After their interactions, they will fill in a form asking which question they asked and whether the chatbot resolved their complaint. The survey also contains questions about customer satisfaction. This pilot results in data about whether the chatbot can resolve certain types of complaints or not and customer satisfaction. The results can be used to compute both costs and the customer satisfaction scores.

Compared with the mocked data, the pilot data contains more accurate and representative data about the quality of the chatbot and whether the chatbot is able to solve different types of complaints as the received feedback is the actual opinion of users. The data is more representative as the questions are actual questions that customers would ask to a chatbot and not questions generated from a complaints data set. Also, customer satisfaction data can be generated which is not be possible with the mocked data. Finally, with this pilot, potentially, feedback is collected about the chatbot itself, which can be used to improve the chatbot.

Therefore, this data is very useful to perform the case study on. This data will be used to apply the evaluation framework on.

3.4.4.1. Pilot sample size

One important aspect to consider when designing the pilot is the sample size. To determine the sample size of this pilot, it first needs to be considered what a good sample size is.

To find such a sample size where both the success rate and customer satisfaction are reliable, we use the formula [1] calculating what the sample size should be if we expect a certain confidence. As one aspect we want to compute is the customer satisfaction, we base the sample size on the customer satisfaction. When estimating the customers satisfaction, we want an accurate customer satisfaction score. We determined an error of maximum: $\epsilon = 0.5$. Next, we should determine the variance. Based on our own experience when filling in customer satisfaction, we can see that there is a high variance in the entered customer satisfaction and the actual satisfaction. Therefore we will use a variance of $s^2 = 3^2$. Finally, we want a confidence of 90 percent. This results in a z-value of 1.645. Finally, this results in the following sample size:

$$\frac{1.645^2 \cdot 3^2}{0.5^2} = 97$$

This results in a sample size of 97 complaints. As there are seven scenarios, we need between the 10 and 20 data points for each scenario.

3.4.4.2. Pilot design

During the design of the pilot, there are a few important steps to take.

1. Describe objectives and aim of the pilot study [28]
2. Decide on the participants
3. Design the pilot
4. Gather results

The **objective** of this pilot is to gather data for the case study. It is not sufficient to only use mocked data as this does not represent actual user behaviour on the chatbot. Also, with the use of this pilot, some data about the customer satisfaction can be gathered to be used in the evaluation framework.

Second, for the participants, there are two different type of participants for the pilot. The first type of participants are people from customers service from water utilities. These are people who will work with the chatbot, and therefore their customer satisfaction level is a good indicator of how they think the chatbot will perform in their water utility. The other type of participants are potential customers of water utility. As they might actually use the chatbot in the future, their questions will be representative to the questions actual customers will ask to the chatbot. Also their customer satisfaction metrics will be useful for the analysis with the evaluation framework.

Third, the actual flow of the pilot should be designed. The flow of the pilot is as follows:

1. Invite participant to join the pilot
2. Inform participant about the pilot and its goals

3. Inform participant about the types of complaints the chatbot can resolve
4. Ask participant to choose x **scenarios** and simulate them with the chatbot
5. For each complaint asked to the chatbot, ask the pilot to fill in a survey. The survey contains the following questions
    (a) What complaint did you ask the chatbot?
    (b) How did the chatbot try to resolve the complaint?
    (c) Has your complaint been resolved?
    (d) How happy are you with the service of the chatbot?

Scenarios

1. There is low water pressure in your shower and you want to ask customer service what the problem is and potentially get it solved.
2. Your invoice is not correct. The amount is too high and you want to ask the customer service to fix it.
3. You have problems submitting your meter reading and want to ask a short question on how to submit the meter reading (i.e. your submission is not accepted).
4. You want to report a disturbance in your area (i.e. there is no water or low water pressure in your area).
5. You want to complain about bad service from one of the employees (i.e. the employee was not interested in helping you).
6. You have issues logging in.
7. You just moved and want to change your address.

Finally, the results from the pilot are gathered. First, it will be gathered whether the chatbot has resolved the complaint. This is gathered from the answers given on question one and three. With question one, the complaint can be categorized and with question three, it can be determined whether the complaint has been resolved correctly. Then to measure the customer satisfaction, we will use the answer from questions 4.

4

Formative study and design

This chapter describes the results from step one until three. First, the results from the formative study are described. This consists of two parts. The first part is a literature study on customer service. The result of this literature study is a list of dimensions and why they should be in the evaluation framework according to the literature study. Then, the results from the interview are shown. Finally, once we have the results from the literature study and interviews, we will design the evaluation framework.

4.1. Formative study: Literature study on customer service

This subsection describes the resulting dimensions and the corresponding metrics based on the literature study:

Dimensions	Metric	Goal	Description	Stakeholders	Metrics
Quality	QoS	How does a chatbot affect the quality of the customer service. Does the quality improve or worsen?	Within the quality dimension, it is important to measure the extent to which the chatbot is able to meet the requirements from the customer and business owner and whether it works as expected [30]	• Customer • Business owner	• Success rate
Efficiency	QoS QoBiz	Figure out if efficiency improves and therefore whether the business will actually be able to be more efficient.	Efficiency is about what the input-output ratio is [30]. Therefore, with this dimension, it should be measured what the input-output ratio is and how this is affected by the chatbot. For this case of the chatbot, the input is employee (hours) and output is the number of processed complaints.	• Business owner	• Time spent per complaint • # employees required to process x complaints per day
Implementation effort	QoBiz	How much does it cost to implement and maintain the chatbot. This is important to compare against the advantages of the chatbot and whether the advantages outweigh the costs.	With the implementation effort dimension, it should be measured how much time and effort it takes to implement and maintain the chatbot.	• Business owner • Employee	• Implementation costs • Maintenance costs

Table 4.1: Results literature study

Dimensions	Metric	Goal	Description	Stakeholders	Metrics
Customer satisfaction	QoE	How does the chatbot affect customer satisfaction. Does it not get to the point where we are not able to actually retain customers.	Another important dimension is customer satisfaction. When measuring customer satisfaction, we want to know how a chatbot affects the customer satisfaction as customer retention should not be lost.	• Business owner • Customer	• Customer satisfaction score • NET Promotor score • Perceived value
Employee satisfaction	QoE	How does the chatbot affect employee satisfaction. Are they still satisfied with their work? Does it not get to the point where employees do not like their job anymore.	When measuring employee satisfaction, it should be measured if the employees are still satisfied with their work at the customer service even though the workflow is now changed when incorporating the chatbot.	• Business owner • Employee	• Perceived workload • Employee satisfaction score

Table 4.2: Results literature study

4.2. Interviews

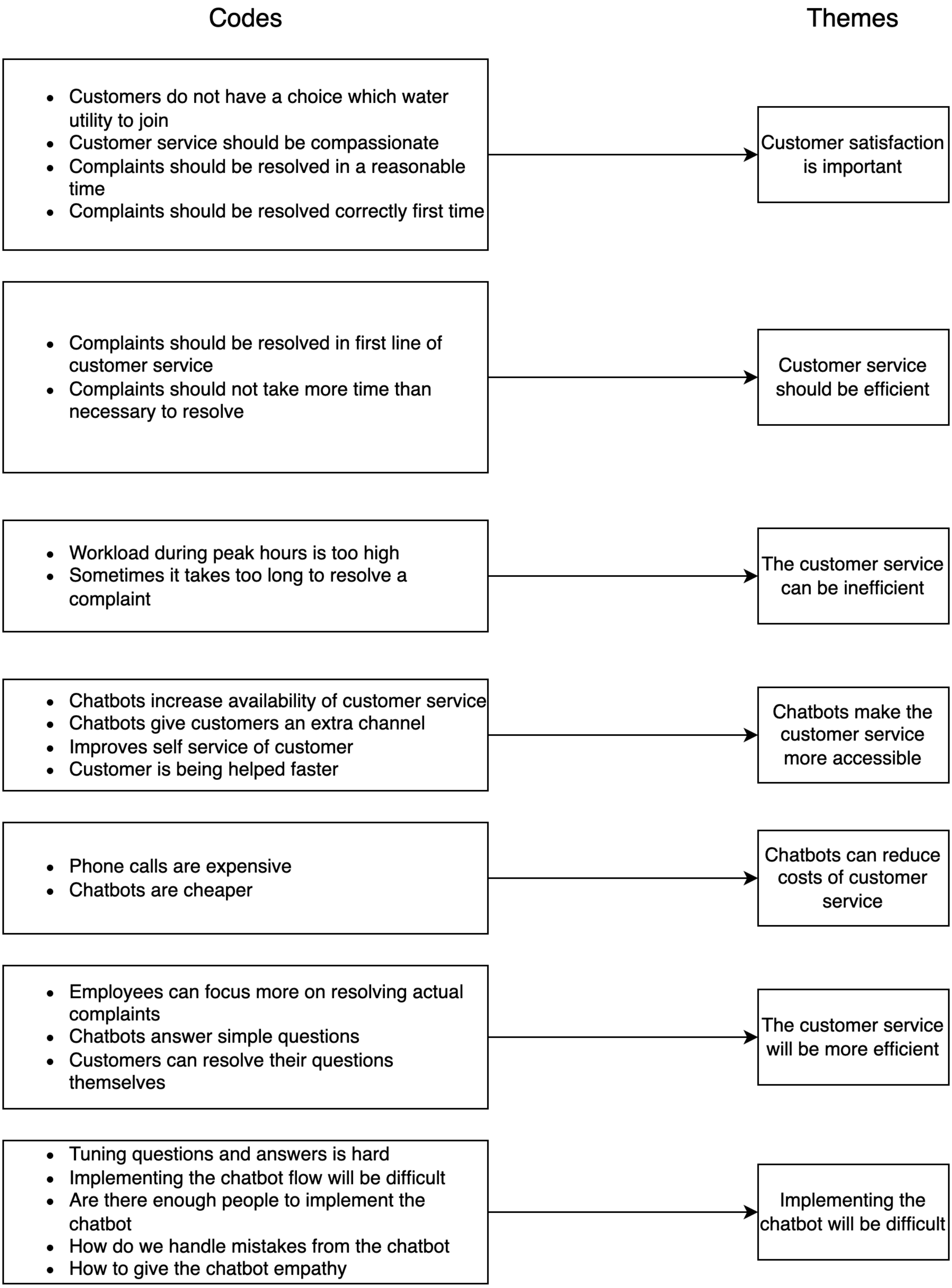
4.2.1. Interview results

Finally, in order to validate and determine the final set of dimensions and metrics, interviews have been conducted. In this section, the results from the interview are discussed. First, the process of analyzing the data is discussed. Second, the results from the several steps in the data analysis are shown. After that, it is discussed how given the result of the data analysis, the research and sub-research questions can be answered.

To gather data from the interviews, they were recorded and notes were taken during and after the interviews. At this point, the data is still in a raw, unstructured format.

After the data has been gathered, codes are extracted from the raw data to create more productive data. An overview of the extracted codes can be found in the figure 4.1 and 4.2.

After this stage, the data is in a useful form to be interpreted. However, it is still not clear what the actual answers from different interviewees are and whether the answers differ a lot or whether there are many commonalities. Therefore, in the next step of data analysis, codes are grouped in themes which represent the meaning of different codes which give a similar answer for different research questions. The resulting themes have been visualised in the figure below.



Codes

Themes

- Customers do not have a choice which water utility to join
- Customer service should be compassionate
- Complaints should be resolved in a reasonable time
- Complaints should be resolved correctly first time

Customer satisfaction is important

- Complaints should be resolved in first line of customer service
- Complaints should not take more time than necessary to resolve

Customer service should be efficient

- Workload during peak hours is too high
- Sometimes it takes too long to resolve a complaint

The customer service can be inefficient

- Chatbots increase availability of customer service
- Chatbots give customers an extra channel
- Improves self service of customer
- Customer is being helped faster

Chatbots make the customer service more accessible

- Phone calls are expensive
- Chatbots are cheaper

Chatbots can reduce costs of customer service

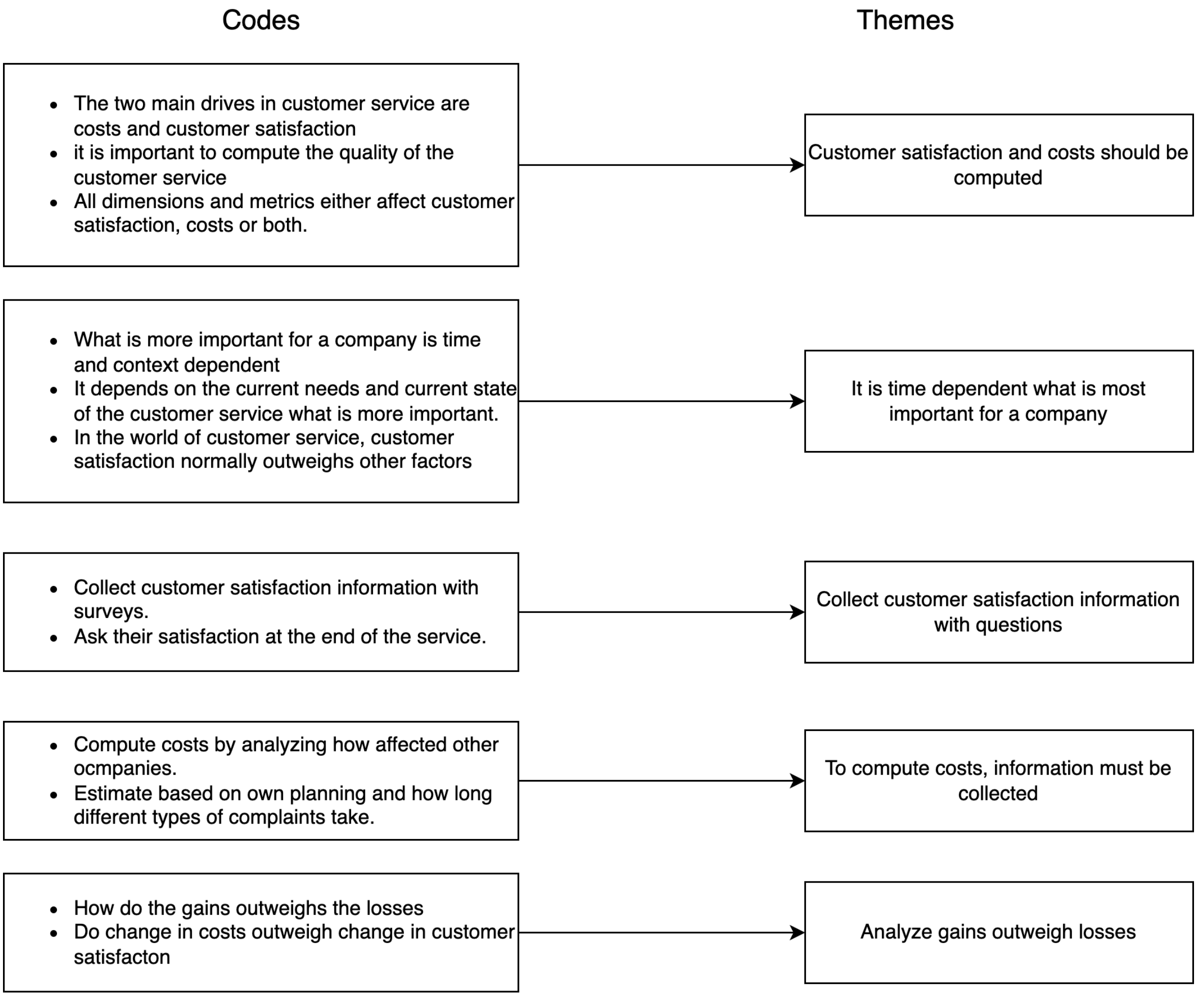
- Employees can focus more on resolving actual complaints
- Chatbots answer simple questions
- Customers can resolve their questions themselves

The customer service will be more efficient

- Tuning questions and answers is hard
- Implementing the chatbot flow will be difficult
- Are there enough people to implement the chatbot
- How do we handle mistakes from the chatbot
- How to give the chatbot empathy

Implementing the chatbot will be difficult

Figure 4.1: Codes and relating themes first part of the interview



- The two main drives in customer service are costs and customer satisfaction
- it is important to compute the quality of the customer service
- All dimensions and metrics either affect customer satisfaction, costs or both.

Customer satisfaction and costs should be computed

- What is more important for a company is time and context dependent
- It depends on the current needs and current state of the customer service what is more important.
- In the world of customer service, customer satisfaction normally outweighs other factors

It is time dependent what is most important for a company

- Collect customer satisfaction information with surveys.
- Ask their satisfaction at the end of the service.

Collect customer satisfaction information with questions

- Compute costs by analyzing how affected other ocmpanies.
- Estimate based on own planning and how long different types of complaints take.

To compute costs, information must be collected

- How do the gains outweighs the losses
- Do change in costs outweigh change in customer satisfacton

Analyze gains outweigh losses

Figure 4.2: Codes and relating themes for dimensions and metrics

Code	Description
Customers do not have a choice which water utility to join	According to one interviewee, water utilities are different than normal companies as customers are obligated to join a water company and they do also not have the choice which water utility they can join as there is only one in the region. Therefore, customer satisfaction is a very important aspect for water utilities
Complaints should be resolved correctly first time	Multiple different interviewees showed that they had one aspect that is very important in the customer service. This aspect is the first time right aspect. It is very important to resolve complaints correctly the first time. First, if not correctly resolved, this will result more time spent on the complaint and therefore increase costs. Second, customers will not be satisfied with the service as they have to contact the customerservice again
Workload during peak hours is too high	One problem found in customer service, is that there are several peak hours/days when the amount of calls to customer service is the highest. This occurs since there are moments when all customers must submit their meter readings. As the workload is high during these hours, regular work will get behind.
Employees can focus more on resolving actual complaints	Multiple interviewees expressed their interest in having a chatbot to resolve simple questions. Except that it will decrease the workload of customer service, it also allows employees to focus more on actual interesting complaints instead of answering simple questions. This can result in both employees having more fun with their work and more complex complaints being resolved better as employees have more time to resolve it.
Tuning questions and answers is hard	Water utilities have datasets containing complaints and questions asked by customers through their customer service. These could be used to develop a chatbot. However, customers ask different type of questions when communicating with a chatbot than using for example an online form. Therefore, the data is not representative to the behaviour the chatbot will actually receive and the chatbot should be tuned based on data actually received.
What is more important for a company is time and context dependent	In water utilities, customer satisfaction is the most important aspect to consider. However, as mentioned by Han, what is actually more important could also depend on the time and state the water utility is in. For example, if customer satisfaction is very high but the costs are also too high, for that moment it might be more important to decrease costs and a small loss of customer satisfaction might be acceptable.

Table 4.3: Table with important codes and a better in depth explanation based on the interviews

Finally, given the resulting themes, all sub research questions can be answered. The answer of each question can be found below:

What are the most important aspects in customer service
When analyzing the results from the interview, it showed that all interviewees had similar aspects which are important in customer service. First, customer satisfaction is a very important factor. It showed that espe-

cially in the water utilities, it is very important that the customer satisfaction is high. Also, during all interviews it stood out that one particular aspect is very important in customer service. That is that complaints should be resolved in one try. This means that the complaint should be resolved correctly and in time such that it does not result in extra complaints. This means that the quality of the customer service should be high and that the process should be efficient such that complaints can be resolved in time.

What problems do they want to solve with chatbots
In general, there are a few problems to be solved with chatbots. First, chatbots can allow the customer service to filter out simple questions. This reduces the amount of interactions customers have to make with actual employees, therefore reducing time employees have to spend on simple questions/complaints. It also allows companies to to reduce time customers have to wait for interaction with the customer service as the interaction will be instant.

What (aspects in customer service) do they want to improve with chatbots
With chatbots, multiple aspects can be improved. First, the availability of customer service will be improved. As a chatbot is available 24/7, customers will be able to ask for help anytime. Even though a chatbot is limited in what it can do, it still improves the customer service as it is able to help customers at any time. As the chatbot is also able to resolve simple complaints, employees will be able to focus better on more difficult complaints. The quality of the service itself will therefore be improved as well. On the employee side, the efficiency will also be improved as a subset of questions/complaints can be resolved by the chatbot. However, the chatbot could also result in additional questions as mentioned by one of the interviewees. Therefore it should still be computed whether the customer service could still do more with the same amount of resources if we incorporate a chatbot.

How do they want to improve their customer service with chatbots?
To improve the customer service with chatbots, chatbots are going to be used in a few different aspects. First, chatbots can be used to answer simple repeatable questions. Often, answers to such questions can also be found on the website itself, but customers still call customer service. However, if a chatbot first checks if it knows how to answer the question, a lot of effort can be reduced from the customer service. Chatbots can also be used to improve self service. A chatbot is very accessible and easy to communicate with. Therefore, a chatbot can be used to improve self service and assist customers with tasks such as submitting watermeter readings.

What difficulties are expected when incorporating a chatbot
There are multiple difficulties that the stakeholders expect to encounter when incorporating a chatbot in their business. First, it takes a lot of time to implement the chatbot. There are a lot of important decisions to make, such as which framework to use and how to implement the chatbot. After that, the chatbot behaviour should be implemented. Implementing the behaviour takes a lot of time as it takes effort continuously to built the chatbot and keep improving it based on current behavior. As customers ask different questions to a chatbot than to a customer service, the chatbot behaviour should be updated based on previous conversations. Finally, interviewees mention that it should be taken into account how the chatbot deals with questions it does not know the answer of. How are these low confidence questions being recognized and how does the chatbot and customer service handle these questions afterwards is also something that takes a lot of effort and that should be taken account of.

4.3. Design of the evaluation framework

Given the results of the literature search, it showed that in general there are ten important analytical dimensions when evaluating RPA. These are: efficiency, availability, scalability and flexibility, costs, quality, compliance, interoperability, implementation effort, employee satisfaction and customer satisfaction. However, when specifically discussing Machine Learning models, some dimensions are not applicable as they are not affected by the Machine Learning model, but by how it is implemented within the organization. Also, the interviews showed that not all dimensions are applicable in the use case of the current context. This finally resulted in the following dimensions to analyse: efficiency, costs, quality, implementation effort and customer satisfaction. Figure 4.3 summarises how the interviews have influenced the resulting dimensions and metrics from the literature study. In the next few subsections, the reasoning behind these dimensions will be discussed.

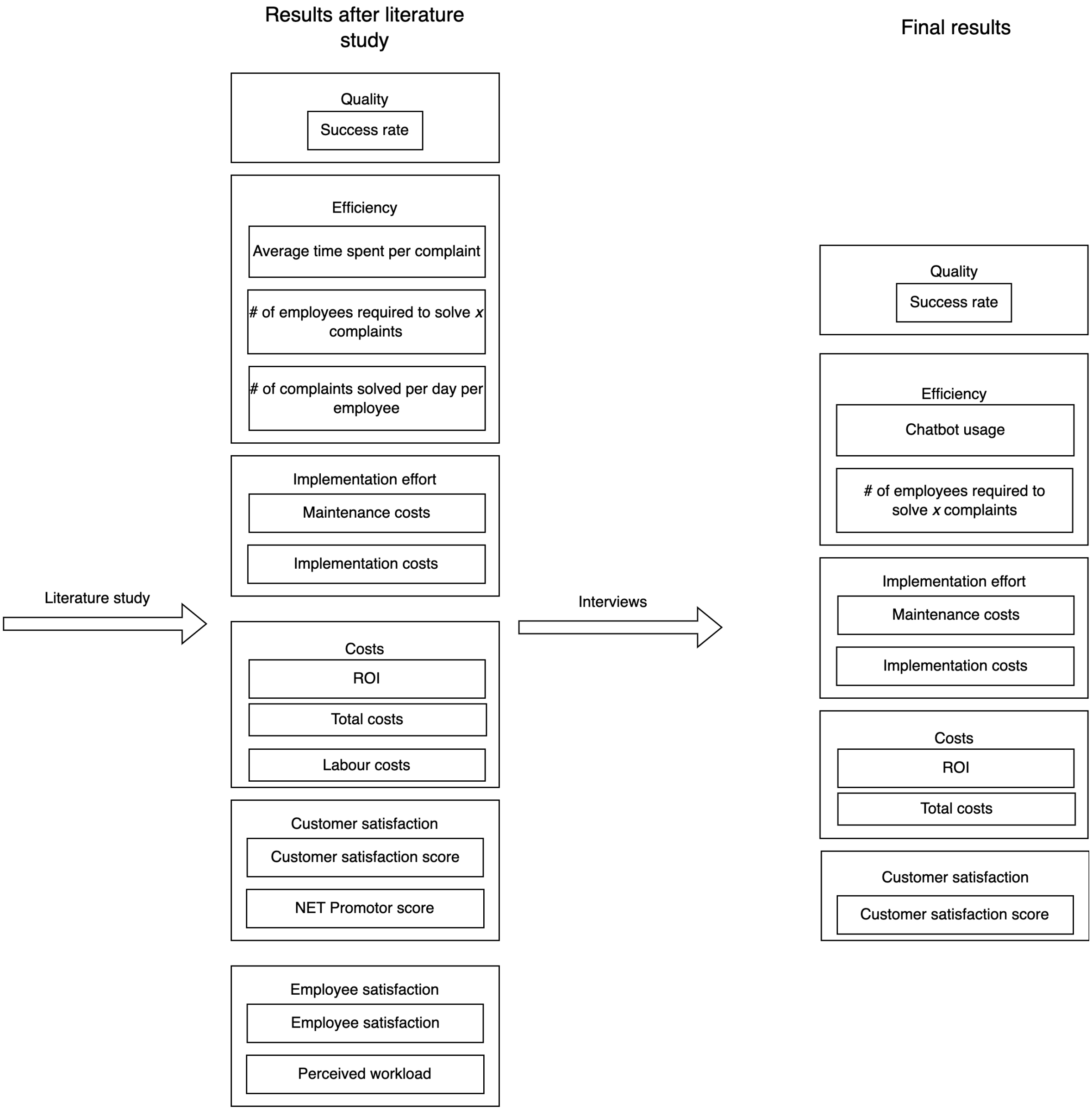


Figure 4.3: Set of dimensions and metrics after literature review and after interviews

4.3.1. Metrics

Based on the literature search and interviews, we determined that the following are the dimensions for the evaluation framework:

1. Quality
2. Efficiency
3. Implementation effort
4. Costs
5. Customer satisfaction

Dimension	Description
Quality	When computing the value of ML models, it is important to know how it affects the quality. Here, it is important to know how the quality of the tasks are affected by the ML model. The ML model can both positively and negatively affect the quality of the chatbot. Knowing the quality of the chatbot can be used to determine the value of the chatbot as change in quality is a very important factor.
Efficiency	When discussing efficiency, one would like to evaluate how the ML model affects the efficiency of the business processes. Here we want to calculate whether the organization can do more with less employees or in less time. This dimension is very important as it reflects the improvement within business processes.
Implementation effort	When incorporating a ML model within an organization, you have to take into account how much it costs in terms of both money and effort to implement and maintain the model. If it takes more effort to implement and maintain the model than what you would actually gain with the chatbot in the future, it should be considered whether it would be valuable to incorporate the chatbot into the organization.
Costs	Analysing the costs for the process we want to use the chatbot for is very crucial. One of the main goals of using a chatbot is to reduce costs as it should automate tasks previously done manually by employees. However, we cannot automatically assume that the chatbot will reduce costs. Therefore, it is important to compute how the costs will be affected by incorporating a chatbot.
Customer satisfaction	It is important to take into account customer satisfaction. When incorporating a chatbot, the flow for the customer changes. This affects the customer satisfaction. How this influences the customer satisfaction is important to take into account. For example, with the new system with the ML model, we could greatly improve the efficiency, but if it results in not being able to retain customers, then it might actually not be valuable to incorporate the chatbot.

Table 4.4: Final dimensions of evaluation framework

The figure below shows every dimension and how they are related to the final value of a chatbot.

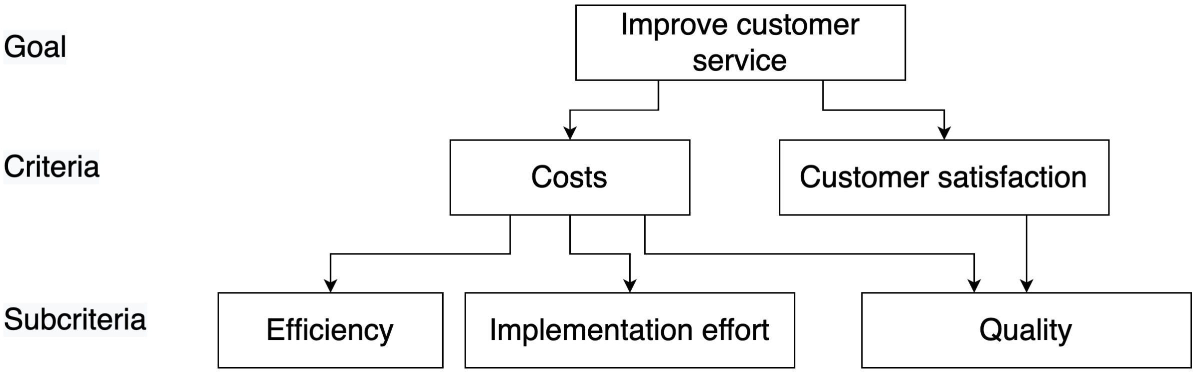


Figure 4.4: Relation of different dimensions to the practical value

For each dimension it has to be decided how they will be computed. For that, first the actual metrics are being determined and then the method of computing the metrics is determined.

For the **quality** dimension, according to the interviewees, the most important aspect is that complaints are being resolved right the first time. Therefore, the metric to be computed is the **success rate** of the chatbot meaning how many complaints the chatbot is able to process correctly.

For the **efficiency** dimension, there are a few different metrics to compute. First, it should be computed what the **chatbot usage** is. It should be computed if people are actually using the chatbot channel instead of other channels like phone or email. As this metric does not immediately reflect the efficiency improvement, a different metric should also be computed which actually shows whether the company is more efficient with the chatbot. The metric to compute this is **the amount of employees required to resolve x complaints in one day**. This metric shows whether we need less employees to resolve the same amount of complaints, thus being more efficient. To compute this metric, we should first estimate how long it takes to process all complaints in one day given we have a chatbot. Then we can simply divide this duration by the amount of hours one employee works each day (which is eight). Based on the result of this calculation, we have an estimation of how many employees we would need to resolve all the complaints of one day.

The next dimension is **implementation effort**. There are many different aspects that should be taken into account when computing the implementation effort. However, based on the interviews it showed that there are two aspects which are most important. These are the **cost to implement** the chatbot and **cost to maintain** the chatbot. Both should be computed in terms of money. To compute these two costs, it needs to be estimated how much time it costs to implement the chatbot. For the maintenance costs, we also need to estimate how many hours we spend per time-span to maintain the chatbot.

To compute the **costs**, there are two metrics we need to compute. These are the **total costs** and **Return of Investment (ROI)** metric. The total costs should be computed such that the company knows how much in total a chatbot costs to have in their business. With this number, companies know whether their budget would allow such an investment. To compute the total costs, both variable and constant costs have to be taken into account. Variable costs are costs which change based on certain variables in the customer service, such as number of complaints per day. Constant costs are costs which only have to be paid once or once every period. However, these costs should stay the same given different amount of complaints. Example of variable costs are the implementation cost of the initial version of the chatbot and server costs.

The total costs shows whether an investment fits within the budget of the company. However, it is also important to know whether and in which time span the the company would have actually earned back the money it invested. To compute this, we need to compute the ROI metric, where it is being compared how the gains of the investment weight against the investment costs. To compute this, we must compute how much money we save each year with the investment and how much money we have to spend each year to maintain the chatbot.

Finally, for the **customer satisfaction**, we compute how happy the customers are with the service. There are many other approach to take, such as having large questionnaires with many different questions. These are very useful to figure out how to improve the chatbot, however, they are not very useful when computing the value of a chatbot. Therefore, only measuring the **customer satisfaction score** of customers using the chatbot is sufficient to measure.

To conclude, this results in the following set of metrics for each of the different dimensions.

1. Quality

   (a) Success rate

2. Efficiency

3. (a) Chatbot usage

   (b) Amount of employees required to resolve *x* complaints in one day

4. Implementation effort

   (a) Cost to implement

   (b) Cost to maintain

5. Costs

   (a) Total costs

   (b) ROI

6. Customer satisfaction

   (a) Customer satisfaction score

4.3.2. Computing the metrics

1. To compute the **success rate** of the chatbot, it must be computed how many complaints the chatbot is able to resolve. In order to compute this metric, a set of complaints is generated which are used to evaluate the chatbot. For each complaint, it must be labeled whether the chatbot has solved the complaint. Based on the labels from all these complaints, we can compute the success rate.

2. **Chatbot usage** is a metric which is difficult to compute as you cannot beforehand compute this without implementing the chatbot in the business. However, it can be estimated and predicted how many people would use the chatbot based on pilots and an analysis of the customer base. Also, customers could be forced by the company to use the chatbot channel (i.e. enforce to always first contact chatbot before calling customer service).

3. To compute the **amount of employees required to resolve x complaints in one day**, it must first be estimated for each complaint category how long it would take to resolve the complaint. For each category, we have to estimate two durations. These are how long it would take to solve that type of complaint with and without chatbot. When we have this information, the success rate and the amount of complaints per day, we can compute how long we would have to spend to solve all complaints. Given the hours one employee works per day, we can compute how many employees would be required.

4. Computing the **cost to implement** requires estimating two factors of the implementation process. First, it must be estimated what resources are required to implement the chatbot and how much that costs. Second, it must be estimated how much time it cost to implement the chatbot. It can then be computed how much the chatbot would cost in terms of hiring developers.

5. To compute the **cost to maintain** the chatbot, it must be specified what tasks are still to be executed when the chatbot is in production. This includes tasks such as checking the server and updating the chatbot based on current data. For each of these tasks it must be specified what type of employee is required, the frequency of the task and how long it would take. Finally, it can be calculated how much it costs to hire the required employees for the tasks.

6. Computing the **total costs** is important as it will show whether a chatbot would actually fit in the budget of the company. To compute the cost, the results from the implementation effort can be aggregated. However, other costs such as license costs or server costs should also be included in the total costs. It should also be computed what the costs of the customer service would as a whole given the costs of processing complaints. In order to do this, costs for processing a complaint for each type of complaint should be specified for both a customer service with and without a chatbot.

7. Computing the **ROI** shows whether an investment is favourable for a company. The formula for calculating the ROI is as follows:

$$\frac{ValueOfInvestment - CostOfInvestment}{CostOfInvestment}$$

   The *ValueOfInvestment* is the amount of money being saved by the company when they invest the chatbot. To compute this, the results from the efficiency dimension can be used as that dimension shows if employees can resolve more complaints with a chatbot. This result should then be used to compute how much a company will save in terms of money given a certain time span. The *CostOfInvestment* is the total costs made to incorporate the chatbot, which are the implementation costs, maintenance costs and all other investments made for the chatbot.

8. The final metric to compute is **customer satisfaction score**. It is difficult to estimate the customer satisfaction score without actually exposing the chatbot to real customers. Therefore, to compute the customer satisfaction score, a pilot is required, where potential customers are asked to use the chatbot with fake complaints and ask whether they would be satisfied with the result of the chatbot. Then this is used to estimate the customer satisfaction score given the current state of chatbot.

4.3.3. Computing the final practical value

Finally, when each dimension has been analyzed, the practical value has to be determined based on the results. However, based on literature and interviews, it showed that there is no straightforward method to determine the practical value directly based on the results from the evaluation framework.

When determining the final practical value of the chatbot, there are two main performance indicators that have to be compared with each other. These are the cost and customer satisfaction indicator. How every dimension relates to these two indicators is visualized in figure 4.4. Trying to compare these two indicators is not easy as costs and customer satisfaction are being measured in different metrics and customer satisfaction is more of an intangible indicator as a change in customer satisfaction does not necessarily immediately show its impact in the businesses.

In order to define the practical value of the chatbot, an approach similar to the Analytical Hierarchy Process (AHP) us used. This process is used to compare different criterias in decision making [22]. In this process, a goal and criteria and subcriteria related to this goal are specified. Then for each criteria and subcriteria, weights are assigned and based on the values of the criteria it can be decided which option is most desirable to achieve the goal.

In this use case of a chatbot in customer service, there are two alternatives to compare. These are a customer service with and without chatbot. The different criteria and sub-criteria are all defined dimensions, where costs and customer satisfaction are the main criteria and efficiency, implementation effort and quality are sub-criteria. We used the metrics for each of these dimensions to create an AHP model. The given AHP hierarchy model is shown below in figure 4.5:

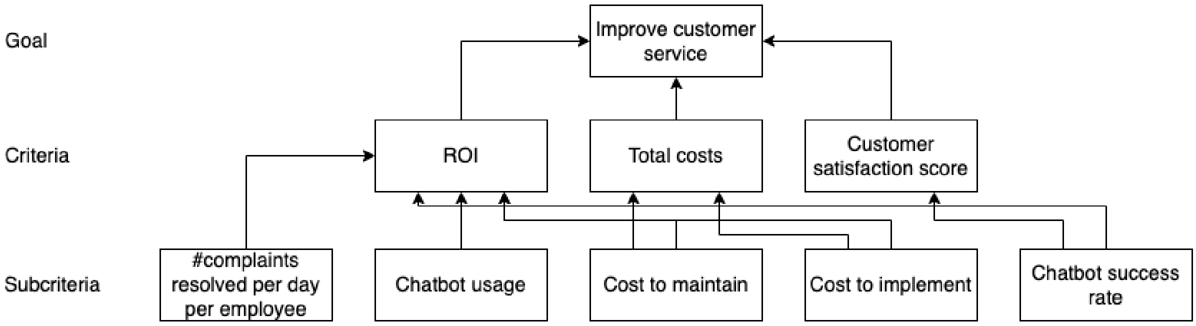


Figure 4.5: AHP hierarchy model of the evaluation framework

Normally, in AHP, for each of the criteria, weights should be assigned which represent the importance of each criteria in the current context. Then based on these weights and the results of each of the criteria, a total score can be assigned to each of the alternatives and these scores can be compared [46]. However, based on the shape of this AHP model, all these computations with weights add unnecessary complexity to the evaluation framework. Instead, each of the criteria should be computed and then based on the results, discussions with the stakeholders can decide the value of incorporating the value of the chatbot in the customer service. For this discussion, the three criteria can be used as insights. The first criteria, ROI, indicates how long it takes until the money of the investment has been recouped. The second criteria, total costs, shows how much money is spent for the customer service. Having this metric for both a customer service with and without chatbot, it can be compared whether costs are actually reduced when using a chatbot in the customer service. The final criteria, customer satisfaction score, shows how the customer satisfaction is compared with a customer service with and without chatbot. This is important to analyze as customer satisfaction is an important aspect in customer service and a decrease in customer satisfaction should be explainable.

5

Case Study

The final step of this research is to perform a case study. The case study has two main goals. First, the case study is used to show whether it is feasible to execute the evaluation framework in a real world case to show its operationalizability. Second, the case study shows that the actual result of the evaluation framework does indeed represent the value the chatbot in the customer service. This shows that we conform to the informativeness property. Overall, the goal of this case study is to validate that the evaluation framework works conforms to both the informativeness and operationalizability property.

This case study focuses on the customer service of water utilities in The Netherlands. A chatbot is developed to handle customer complaints received by water utilities in The Netherlands. After that, information is being collected about the customer service of water utilities in The Netherlands. Finally, using the chatbot and information about customer service of water utilities in The Netherlands, we conduct a case study to validate that the evaluation framework conforms to the informativeness and operationalizability property.

5.1. Implementation of the chatbot

5.1.1. chatbot flow in the business

As described in section 2.4.1, before incorporating a chatbot in a business, we have to decide what level of automation we want and how we want to implement the hybrid human-AI workflow. The current state-of-the-art of chatbot shows that it is possible for a chatbot to resolve multiple tasks. As it is also possible for chatbots to reject messages if it is not sure what to answer, we have decided to use the following workflow show in figure 5.1.

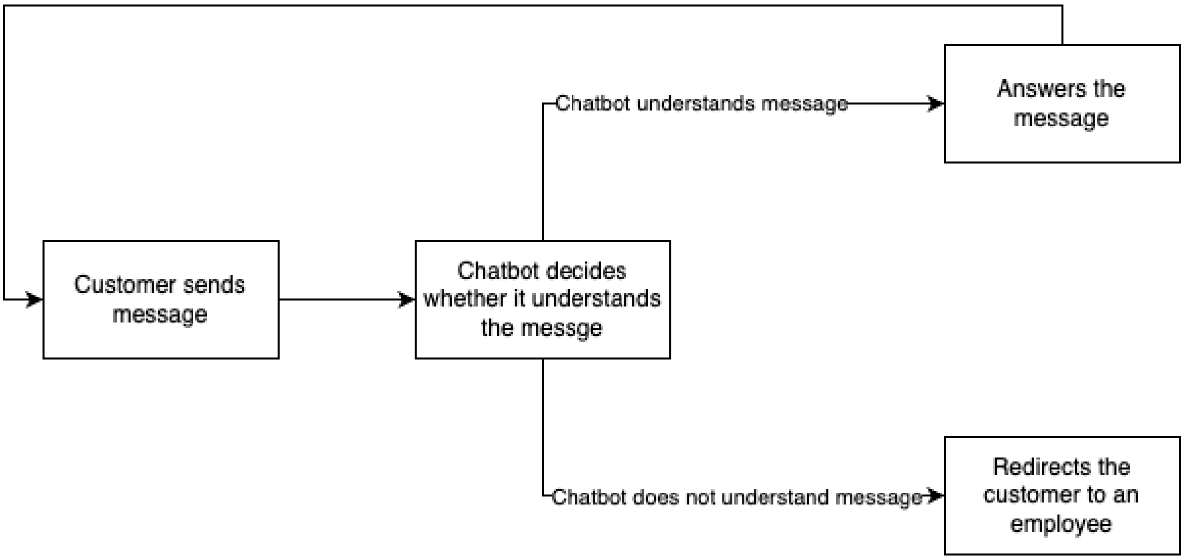

Answers the message
Chatbot understands message
Customer sends message
Chatbot decides whether it understands the messge
Redirects the customer to an employee
Chatbot does not understand message

Figure 5.1: Chatbot flow in this case study

In the designed flow, it begins with a customer starting the conversation with the chatbot with a single message. With the use of the threshold method as described in section 2.3, the chatbot determines whether it understands the question. If the chatbot understands the message, it will answer accordingly and continue the conversation. Otherwise, we will abort the conversation with the chatbot and redirect the customer to an actual employee.

5.1.2. Data

Before decisions can be made about the design of the chatbot, first, the data has to be analysed and altered where necessary. The available data consists of complaints collected by water utility companies in The Netherlands either through their website or telephone. For most complaints it is described what the complaint is and how the complaints have been resolved. Some complaints have also been classified in categories by the water utility.

When analyzing the data, there were initially three fields for each complaint which could be very useful for the development of the chatbot. These fields are the complaint itself, the category and how it was resolved. However, when further analyzing the data, it showed that the descriptions of how the complaint was resolved, was most of the time a description mentioning that they called the customer and that they have resolved it. No further explanation was given about what has been done to resolve it. Therefore, this field turned out to be not that useful after all.

5.1.3. Implementation

To implement the chatbot for the case study, the RASA chatbot framework is used. This framework fits the use case of this case study as it is relatively easy to implement the chatbot with RASA and it does have the capability to implement a fully functioning chatbot. RASA has already implemented all required components described in section 2.2. To implement the chatbot with RASA, there are in general two tasks to complete. First, intents for chat messages must be specified. In this case, the intents will be different categories of complaints. Second, stories have to be defined. Stories define how the chatbot will interact for different intents or sequences of intents. It should be defined how the chatbot will resolve all different types of complaints.

To specify intents, there are two aspects which are important. First, intents should not be too specific as that makes it difficult for the Machine Learning algorithm to classify the messages. Second, the intents should be classified in such a way that each of the complaints can be solved in a similar way for every complaint classified in an intent.

Specifying the stories is not as difficult as specifying the intents as it is straightforward how complaints should be resolved. For some types of complaints however, the stories might be a bit more complicated as there is more information required to actually resolve the complaint. RASA has functionality which easily allow to extract information from text messages.

5.2. Data collection and chatbot analysis

The first step in conducting this case study is to collect data. Data is collected using two different methods. First, data is collected using mocked data from a water utility in The Netherlands. Second, a pilot is being conducted to collect real-world data. The results of these two data collection methods are described below.

First, mocked data from the water utility is collected. Using a script, all data is being evaluated against the implemented chatbot. We evaluate whether the chatbot correctly finds the intent of the message. The results are labeled as either correct, incorrect or redirected, where redirected means that the chatbot does not know the answer and redirects the answer to a human employee. The results from the mocked data can be found below in figure 5.1.

Category	Correct	Incorrect	Redirected	Total
Administration assign employee	0	5	4	9
Administration change data	0	1	2	3
Administration help logging in	3	0	2	5
Administration change address	3	0	0	3
General complaint no action required	10	1	3	14
Complaint payment action required	22	0	3	25
Submit meter reading action required	17	0	1	18
Problem at home	28	3	8	39
Report general disturbance	5	2	9	16

Table 5.1: Results mock test

As what we label as a correct response from the chatbot might not be actually perceived as a correct response by an actual customer, it is also important to conduct a pilot study. The goal of this pilot study is to receive some actual data from potential customers. This will better reflect how a chatbot will work when actually implemented in a business and give a better indication of its capabilities of solving a customer complaint. The results of this pilot study can be found below. The pilot study does not have all categories from the mock test as not all categories were useful to test during the pilot. The results from the pilot can be found below in figure 5.2.

Category	Correct	Incorrect	Redirected	Total
Problem at home	3	1	12	16
Administration help logging in	3	0	10	13
Complaint payment action required	0	0	15	15
Submit meter reading action required	9	0	1	10
Report general disturbance	4	0	9	13
General complaint no action required	3	0	10	13
Administration change address	4	1	7	12

Table 5.2: Results pilot

5.3. Applying the evaluation framework

5.3.1. Collecting required information

In order to apply the evaluation framework, some information is required to compute all of the metrics in the evaluation framework. The results of this information can be found in the tables below:

Costs table

Complaint category	without chatbot	with chatbot
Administration assign employee	10	10
Administration change data	10	3
Administration help logging in	10	1
Administration change address	10	2
General complaint no action required	10	0
Complaint payment action required	10	10
Submit meter reading action required	8	1
Problem at home	15	5
Report general disturbance	10	3

Table 5.3: Costs in terms of minutes of different types of complaints with and without chatbot

Amount of complaints per day
Based on analysis of the complaints data set and information provided by the water utility companies, the amount of complaints is assumed to be **800** complaints per day. The data-set showed about 290.000 complaints in the whole year. Therefore, the customer service receives about 800 complaints per day.

Employee wages
Hourly wage customer service employee: **17.30 euro**
Hourly wage AI engineer: **32.18 euro**

Implementation time
To estimate the implementation time, we will use our own experience when implementing the chatbot for this case study. We analysed how long it took us to implement a simple chatbot covering a few simple use cases. We estimated that implementing a basic chatbot with only a few use cases would take about **200** hours.

Maintenance time
Maintenance of the chatbot consists of two tasks. First, the chatbot should be monitored. It should be monitored that the server is up and running and that there is no suspicious behaviour. Second, during the maintenance, chatbot input should be analyzed to update and improve the chatbot and implement new use cases based on frequent questions asked to the chatbot. Monitoring should be a task of the current ICT team. Therefore, no additional costs occur. Updating the chatbot would take a small team to work on the chatbot continuously. Given the work we had to do to implement the chatbot (analyse data, update chatbot behaviour) we would estimate that this team would consist of two AI engineers working **16 hours per a week**.

Chatbot usage
To simplify the process of computing the costs of implementing a chatbot in a customer service of water utilities, it will currently be assumed that the chatbot usage is **100%** meaning that each customer will first try to contact the chatbot before using a different communication channel. In reality however, it would not be possible to have a chatbot usage of 100%. Therefore, in the formula, it will be shown how different the total cost can be calculated for a chatbot usage of less than 100%. Without actually providing the chatbot service to the customers, it is not possible to make an accurate prediction of chatbot usage as it depends on a lot of different factors (age, consumer conservatism [14]).

Customer satisfaction
8,5/10 (retrieved from a water utility from The Netherlands)

Complaints	
Number of complaints	800 per day
Wages	
Customer service employees	17.30 euro per hour
AI engineer	32.18 euro per hour
Investments	
Server	144 euro per year
Laptops	400 euro per year
Implementation duration	
Implementation time	200 hours
Maintenance time	32 hours per week
Other	
Chatbot usage	100%
Customer satisfaction	8.5

Table 5.4: Retrieved information for case study

5.3.2. Computing all metrics

Total variable costs of customer service

To compute the total costs of the customer service with a chatbot in terms of minutes, we use the following approach. For each complaint category, we will take the amount of complaints correctly resolved, redirected and incorrectly resolved and multiply that by the corresponding cost. Taking the sum of all complaint categories, we will get the total cost for the customer service with chatbot given the amount of complaints.

Computing total with chatbot costs in terms of minutes:

$$\sum_l C_{l,chatbot,corr} * SUCCESS_{l,succ} + C_{l,chatbot,redir} * SUCCESS_{l,redir} + C_{l,chatbot,incorr} * SUCCESS_{l,incorr}$$

Computing total without chatbot costs:

$$C_{l,no\_chatbot} * SUCCESS_{l,total}$$

1. l stands for each complaint category
2. C is a matrix containing the costs in minutes processing the complaint with and without chatbot for each complaint category l.
3. $SUCCESS$ contains the results for each complaint category l. It shows for each category how many complaints have been successfully process, redirected and unsuccessfully processed. $SUCCESS_{l,total}$ shows for complaint category l, how many complaints there are in total.
4. corr = correct
5. incorr = incorrect
6. redir = redirect
7. succ = success

Using this formula, we have the following cost in terms of minutes per day:

- Without chatbot: **133.33 hours per day**
- With chatbot: **107.999 hours per day**

Total employee costs of the customer service

Variable costs in this context are costs that change depending on the amount of complaints received by the customer service. In each of these computations, we expect a chatbot usage of 100%. Based on estimation, in this case, the amount of complaints received per day will be estimated to 800 complaints per day. Based on this estimation, the following results can be found.

- 133.33 hours · 17.30 euro = **2,306.61 euro per day without a chatbot**
- 100%·(107.99 hours·17.30 euro)+0%·(133.33 hours·17.30 euro) = **1,868.23 euro per day with a chatbot**

Total other costs of the customer service
Fixed costs are costs that do not change over time. Such costs are costs for computers to develop the chatbot, servers and eventually license costs for the chatbot. For these three costs, the following numbers have been estimated per year.

- 2 laptops (1000 euro, 20% amortization): **400 euro/year**
- Server: **144 euro per year** (digitalocean.com)
- Implementation costs: 200 hours · 32.18 euro = **6,436 euro**
- Maintenance costs: 52 · 32 · 32.18 = **53,547.52 euro**

Total costs of the customer service
The total costs of the customer service consists of two aspects. These are the variable costs and the fixed costs which have both been computed in the previous paragraphs. At the end, a total cost is computed for the whole year for a customer service with and without chatbot.

- Costs without chatbot: 365 days · (133.33 hours · 17.30 euro) = **841,912.29 euro per year**
- Costs with chatbot: 365 days·(107,99 hours·17.30 euro)+400+144+6436+53,547.52 = **742,430.38 euro**

ROI
Given the formula of ROI:

$$\frac{ValueOfInvestment - CostOfInvestment}{CostOfInvestment}$$

The following is the ROI for an investment of the chatbot

1. The value of the investment is the amount of money that is being saved based on the differences in cost between a customer service with and without chatbot. $ValueOfInvestment = 841,912.285 - 742,430.375 = 99,481.91$ euro
2. The cost of investment is the total cost of the investment per year. These includes cost for implementation and maintenance as well as buying servers. $CostOfInvestment = 400 + 144 + 6,436 + 53,547.52 = 60,527.52$ euro

$$\frac{99,481.91 - 60,527.52}{60,527.52} \cdot 100 = 64.4$$

Therefore, with the current calculations, we have an ROI of **64.4 percent** for the first year.

As the first year has an additional cost for the additional implementation, we also calculate the ROI for the upcoming years excluding the first year.

$$\frac{99,481.91 - 54,091.52}{54,091.52} \cdot 100 = 83.9$$

Therefore each year after the first year, we have an ROI of **83.9 percent**.

Customer satisfaction
Taking the average of all complaints from the pilot, the customer satisfaction is a 5.7. Here, we also take into account the complaints that are being redirected to employees. However, this result is a little bit biased as the redirected participants did not get their complaint actually resolved, so the resulting customer satisfaction score might be a bit lower than expected, as the participants never actually got their complaints resolved.

If we only take the complaints which have been resolved, the resulting customer satisfaction is **7.9**.

If we only look at complaints that have been redirected or not been solved, a customer satisfaction of 4.9 can be found.

When we compare that to the customer satisfaction when there is no chatbot (from figure 5.4), we can see a slight decrease of the customer satisfaction score. However, depending on the company, the customer satisfaction score can still be high enough.

Metric	Result without chatbot	Result with chatbot
Success rate	-	28%
Chatbot usage	0%	100%
ROI	-	64.4%
Total costs	841,912.285	742,430.375
Customer satisfaction score	8.5	7.9
#employees required to resolve 800 complaints	17	14
Cost to maintain	0	54,091.52
Cost to implement	0	6,436

Table 5.5: Results metrics from case study

5.4. Concluding on the practical value

Finally to conclude on the practical value, the results from all criteria should be taken into account to conclude on the practical value. For total costs, the costs should be compared between a customer service with chatbot and without chatbot. For ROI, it should be analyzed whether the resulting ROI is acceptable for the company or whether it is expected that the ROI should be higher. ROI is the most important criteria when analysing the money aspect of the chatbot. Finally, for customer satisfaction, depending on the company and its goals, there are two approaches to take to analyze its practical value based on the customer satisfaction. First, it could simply be compared if the customer satisfaction improves and add value to the chatbot based on this improvement. Second, it could be checked if the customer satisfaction is still above a certain threshold. In that case, it could be concluded that the chatbot would still be valuable for the water utility company reasoning from a customer satisfaction perspective.

For this chatbot, we see that the chatbot does have some positive impacts on the customer service of water utility companies in The Netherlands. First, the costs are reduced and we have a positive ROI. The customer satisfaction score decreases. However, the score is still sufficient. Therefore, we can conclude that this chatbot does have a positive value on the water utility companies in The Netherlands.

5.5. Validating the practical value

In order to validate that this evaluation framework does actually result in a correct practical value of the chatbot, the following approach is conducted. Since the practical value would differ from business to business, it is difficult to actually validate the practical value. In this approach, to validate the evaluation framework, we derive insights from the results of the evaluation framework. Based on these insights, stakeholders themselves can conclude on the practical value of the chatbot from their own perspective.

Based on the results of this case study, we derived the following insights:

Insight 1: developing the chatbot takes away most of the benefits gained with chatbots

If we look at the costs of the customer service with and without a chatbot and we exclude maintenance costs, we can see that maintenance costs uses the most money. Therefore, in order to improve the advantages of using a chatbot, we should look at how we can optimize the maintenance costs such that these costs can be lowered. However, we can assume that maintenance will become less intense at further stages of the chatbot and therefore it is expected that the benefits in terms of money will improve over the course of months/years.

Insight 2: Using a chatbot will decrease the average customer satisfaction score. The score is however still at a desirable level

Incorporating a chatbot shows that its customer satisfaction score gets decreased. While some other channels receive a customer satisfaction score of 8.7 and 8.5, the chatbot channel does receive a customer satisfaction score of 8 when the complaint gets resolved. Although it is lower than some of the customer satisfaction score of some other communication channels, the customer satisfaction score is still desirable for a customer service

Insight 3: Incorporating this chatbot in a water utility does have financial benefits
If we purely look at the ROI of this chatbot, it shows that this chatbot does have a positive ROI given our estimations and calculations. This means that this chatbot would be profitable and that we would earn 64.4 percent of the investment that we made each year. For a chatbot that is still in the early stages of its developments, the ROI should increase when the success rate of the chatbot increases.

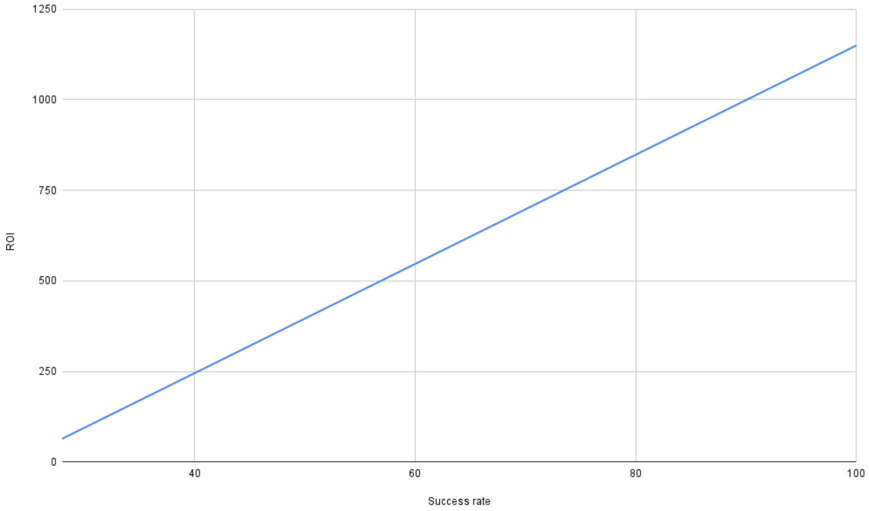


Figure 5.2: How ROI changes over different success rates of the chatbot

Insight 4: The total investment required per year is 60,527.52
Based on our calculations, it shows that this chatbot requires an investment of 136,536 euro per year. Although the ROI shows that we will earn the money back, it is still important to know how much the investment cost as we should still check if we have the budget.

Insight 5: It is important to keep the chatbot usage as high as possible
In this estimation we used a chatbot usage of 100 percent, meaning that all customers will contact the chatbot first before being redirected potentially to other better communication channels. However, a chatbot usage of 100 percent is very optimistic as that would never be possible. A chatbot is not accessible for all customers as for example older customers do not know how to talk to a chatbot. Therefore, the actual estimated costs would be a bit higher if actually incorporated in a company. If we want to keep the costs of the customer service as low as possible, we should try to keep the chatbot usage as high as possible.

When assuming that the cost linearly changes between a chatbot usage of 0 and 100, figure 5.3 shows that we would need at least a chatbot usage of about 60 percent in order to break-even meaning that the cost of the customer service would be the same as if we did not use a chatbot.

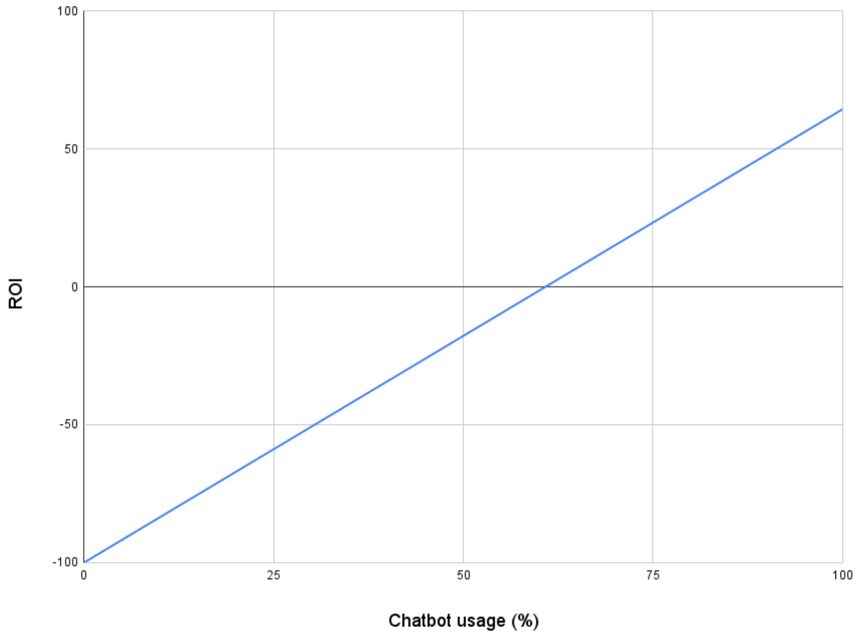


Figure 5.3: How ROI changes over different chatbot usages

Insight 6: There is a strong correlation between the customer satisfaction and whether a chatbot resolved the complaint
Based on the results, it showed that the customer satisfaction score heavily depends on the fact whether the chatbot solves the complaint or not. Even redirecting the complaint results in a significantly lower customer satisfaction score. While in a real customer service, the customer satisfaction score will most likely be higher as during the pilot, the complaint would not be resolved. This difference in customer satisfaction score would still exist in a less extreme version and it shows redirecting a complaint does have a negative effect on the customer satisfaction.

6 Discussion

This chapter discusses the approach of this research and its results. It will be discussed what the limitations of this research are. We will discuss why these limitations existed in this research and how these shortcomings affected the results of this research.

6.1. Summary of findings

In this research, we used the research through design methodology to design an evaluation framework to compute the value of a chatbot in customer service. Based on the results of a formative study consisting of a literature study and interviews, an evaluation framework is designed. This framework conforms to both the informativeness and operationalizability property. It conforms to the informativeness property as the design is based on the interview with stakeholders. The results from the case study show that the evaluation framework provides meaningful information to stakeholders. To show that we conform to the operationalizability property, we performed a case study. Applying the designed evaluation framework on this case study, we showed that it is possible to apply the evaluation framework and that it is not too complex to compute the value of a chatbot in customer service with this framework. We also showed that it is possible to predict what the values of some metrics would be if results of other metrics would be different allowing businesses better make choices of which metrics to improve.

6.2. Limitations

First of all, the amount of interviews conducted in this research is not very high. In many cases, conducting five interviews is not enough for a qualitative study. However, we reasoned that if answers from all five interviewees are in the same direction, then the whole solution space has been explored and more interviews is not necessary. As for all interview questions, the answers were heading in the same direction and there were no outliers of interviewees giving completely different answers compared to other employees, we concluded that we discovered the whole solution space and therefore more interviews were not necessary. Conducting more interviews however could have allowed us to get a better understanding of the context this research is conducted in and could have resulted in better fine-tuned metrics for this specific use case. Therefore, to what extend we conform to the informativeness property is questionable. However, based on the interviews and literature study, we can conform to the informativeness property.

Second, the information of this use case is not very precise as it was not easy to retrieve the information from the water utilities in The Netherlands. Therefore, some information had to be estimated based on our own intuition and the results might not be accurate to represent the real use case. Therefore, the case study might not actually be a real representation of the described use case. However, as we use realistic data, the results of the case study shows that the evaluation framework applies to the operationalizability property.

Third, the results of the case study might not actually completely represent how it would look like when the chatbot would actually be implemented in a water utility company in The Netherlands. It is difficult to actually simulate a customer using a chatbot with an actual complaint. While the results might not be accurate, it is still a good indication of how this evaluation framework could be used in practice.

6.3. Implications for research

This research is one of the first to discuss how one could compute the (practical) value of a chatbot in customer service. Computing the practical value of Machine Learning models is still new so this research would be a good first step in showing what approach we could take to compute the value of a chatbot or Machine Learning algorithm in general. This research can be used and built on to design approaches to compute the value of other ML algorithms.

Instead of only designing the evaluation framework, this research also shows how this framework could be applied to the real-world case of water utilities of The Netherlands. It showed that the framework conformed to both informativeness and operationalizability property. It also showed the practical possibilities of this framework and opens up more opportunities for research to figure out what other data (visualisations) would be useful in such a framework.

Analyzing this research, there are still a few gaps that should be resolved in future research. First, it should be researched how we could actually perform this case study when the chatbot is not completely implemented. The goal of computing the practical value of a Machine Learning algorithm is to check whether it would be valuable or not for a business to implement. However, if a business has to implement the algorithm completely to apply the evaluation framework, the advantages of computing this value would be reduced as we have already implemented the algorithm. This would be useful as we could analyze how we could improve the algorithm to increase the practical value. Finding a way to get a good estimate of the practical value while keeping the efforts as low as possible would be a good direction for future research.

6.4. Implications for practice or applications

This research would be able to help businesses figuring out whether they should incorporate chatbots in their business. Currently, chatbots are already used a lot by many businesses, but some are still not useful as there has not been a lot of thought behind the advantages of the chatbot in a specific business. Therefore, chatbots are sometimes not very efficient for some businesses.

This research has two implications for businesses that want to incorporate a chatbot. First, businesses could use this research to compute the practical value of their chatbot and therefore knowing what benefits a chatbot has to their business. Second, as sometimes, businesses implement chatbots in their business without a lot of thought, the use of this evaluation framework would trigger such businesses to think about the value of the chatbot. The quality of chatbots on the web should therefore improve if this evaluation framework is used.

7 Conclusion

In this research, we have designed an evaluation framework for chatbots in customer service. The main goal of this evaluation framework is that it should conform to the informativeness and operationalizability property, meaning that it should capture all needs, desires and values of the stakeholders and that we should be able to apply the evaluation framework in real life.

First, if we analyze the informativeness property, we performed interviews and designed the evaluation framework based on these interviews. If we then analyse the list of insights derived from a case study we performed with the evaluated framework in the context of water utilities in The Netherlands, we see that all important aspects derived from the interviews can be found in the derived insights from the interviews. These insights show that the evaluation framework provides meaningful information to the stakeholders. Therefore, we can conclude that we have conformed to the informativeness property.

Second, we have to analyze the operationalizability property. To show that this evaluation framework conforms to the operationalizability property, we performed a case study to show that it is possible to apply the evaluation framework on a real life case. It shows that it is certainly possible to apply the evaluation framework. A limitation in this case is that some results do not completely reflect what the results would be when the chatbot is implemented in an actual business as some estimations had to be made (for example for the chatbot usage). However, overall, we could say that we have conformed to the operationalizability property.

We conclude that we have designed an evaluation framework for chatbots in customer service conforming to both the informativeness and operationalizability property.

7.1. Future work

This project shows that it is possible to design an evaluation framework for chatbots in customer service conforming to both the informativeness and operationalizability property. There are still some aspects that can be improved.

First, if we analyze how we performed the case study, it showed that we had to do some estimations in order to perform the evaluation. Also, we could not actually test the chatbot on actual customers who have real problems, therefore some results might not be as accurate as it would be in real life. Future research could take a look at this problem. It could be analyzed how we should approach this problem allowing us to get as accurate results as possible for the evaluation framework while making sure that we can get these results as early as possible. We do not want to implement a whole chatbot, concluding at the end that the practical value would actually be negative. We want to be able to perform such evaluation as early as possible such that we can improve the implementation process and/or abort the process as early as possible. Therefore, research could be done in how we could apply such an evaluation framework in early stages of the implementation of the chatbot.

Second, this evaluation framework is designed for the very specific use case of chatbots in customer service of water utilities in The Netherlands. Future work should research how we could design a more general evaluation framework which could be used in other contexts while still capturing the informativeness and operationalizability property.

Third, we could take a better look how the results of this evaluation framework can be used to assist businesses in improving their chatbot. We already showed how an improvement of the chatbot's success rate would affect the ROI. It could be researched what relations between metrics should be studied and how these relations could point businesses in how they should improve their chatbot.

Bibliography

[1] 6.1 - estimating a mean: Stat 415. *PennState: Statistics Online Courses.* URL `https://online.stat.psu.edu/stat415/lesson/6/6.1`.

[2] Rosemarie Anderson. Thematic content analysis (tca). *Descriptive presentation of qualitative data*, pages 1–4, 2007.

[3] Amina Asif et al. Generalized learning with rejection for classification and regression problems. *arXiv preprint arXiv:1911.00896*, 2019.

[4] Soufyane Ayanouz, Boudhir Anouar Abdelhakim, and Mohammed Benhmed. A smart chatbot architecture based nlp and machine learning for health care assistance. In *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, pages 1–6, 2020.

[5] Theophilus Azungah. Qualitative research: deductive and inductive approaches to data analysis. *Qualitative research journal*, 2018.

[6] Emin Babakus and Gregory W Boller. An empirical assessment of the servqual scale. *Journal of Business research*, 24(3):253–268, 1992.

[7] Sarah Elsie Baker and Rosalind Edwards. How many qualitative interviews is enough. 2012.

[8] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.

[9] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[10] Vikram Kamath Cannanure, Timothy X Brown, and Amy Ogan. Dia: A human ai hybrid conversational assistant for developing contexts. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*, pages 1–5, 2020.

[11] John Carruthers. A rationale for the use of semi-structured interviews. *Journal of Educational Administration*, 1990.

[12] Fabio Casati, Pierre-André Noël, and Jie Yang. On the value of ml models. *arXiv preprint arXiv:2112.06775*, 2021.

[13] Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification, 2021. Cost sensitive approach taking into account risks and costs of rejection. Need to calibrate the rejection algorithm.

[14] Peter J Danaher and John R Rossiter. Comparing perceptions of marketing communication channels. *European Journal of Marketing*, 2011.

[15] Kweku Ewusi-Mensaxh. Evaluating information systems projects: A perspective on cost-benefit analysis. *Information Systems*, 14(3):205–217, 1989.

[16] Sarah E Finch, James D Finch, Daniil Huryn, William Hutsell, Xiaoyuan Huang, Han He, and Jinho D Choi. An approach to inference-driven dialogue management within a social chatbot. *arXiv preprint arXiv:2111.00570*, 2021.

[17] Batya Friedman. Value-sensitive design. *interactions*, 3(6):16–23, 1996.

[18] Cheng Gong, Jiuyang Tang, Shengwei Zhou, Zepeng Hao, and Jun Wang. Chinese named entity recognition with bert. *DEStech Transactions on Computer Science and Engineering*, 2019.

[19] Eko Handoyo, M Arfan, Yosua Alvin Adi Soetrisno, Maman Somantri, Aghus Sofwan, and Enda Wista Sinuraya. Ticketing chatbot service using serverless nlp technology. In *2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pages 325–330. IEEE, 2018.

[20] Maximilian Hofer, Andrey Kormilitzin, Paul Goldberg, and Alejo Nevado-Holgado. Few-shot learning for named entity recognition in medical text. *arXiv preprint arXiv:1811.05468*, 2018.

[21] Kenneth F Hyde. Recognising deductive processes in qualitative research. *Qualitative market research: An international journal*, 2000.

[22] Alessio Ishizaka and Ashraf Labib. Review of the main developments in the analytic hierarchy process. *Expert systems with applications*, 38(11):14336–14345, 2011.

[23] Lucija Ivančić, Dalia Suša Vugec, and Vesna Bosilj Vukšić. Robotic process automation: systematic literature review. In *International Conference on Business Process Management*, pages 280–295. Springer, 2019.

[24] Anran Jiao. An intelligent chatbot system based on entity extraction using rasa nlu and neural network. In *Journal of Physics: Conference Series*, volume 1487, page 012014. IOP Publishing, 2020.

[25] John Leslie King and Edward L Schrems. Cost-benefit analysis in information systems development and operation. *ACM Computing Surveys (CSUR)*, 10(1):19–34, 1978.

[26] Chaitrali S Kulkarni, Amruta U Bhavsar, Savita R Pingale, and Satish S Kumbhar. Bank chat bot–an intelligent assistant system using nlp and machine learning. *International Research Journal of Engineering and Technology*, 4(5):2374–2377, 2017.

[27] Tarun Lalwani, Shashank Bhalotia, Ashish Pal, Vasundhara Rathod, and Shreya Bisen. Implementation of a chatbot system using ai and nlp. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST) Volume-6, Issue-3*, 2018.

[28] Gillian A Lancaster, Susanna Dodd, and Paula R Williamson. Design and analysis of pilot studies: recommendations for good practice. *Journal of evaluation in clinical practice*, 10(2):307–312, 2004.

[29] Rohit Binu Mathew, Sandra Varghese, Sera Elsa Joy, and Swanthana Susan Alex. Chatbot for disease prediction and treatment recommendation using machine learning. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 851–856, 2019. doi: 10.1109/ICOEI.2019.8862707.

[30] Anja Meironke and Stephan Kuehnel. How to measure rpa's benefits? a review on metrics, indicators, and evaluation methods of rpa benefit assessment. 2022.

[31] Maali Mnasri. Recent advances in conversational nlp: Towards the standardization of chatbot building. *arXiv preprint arXiv:1903.09025*, 2019.

[32] Grazia Murtarelli, Anne Gregory, and Stefania Romenti. A conversation-based perspective for shaping ethical human–machine interactions: The particular challenge of chatbots. *Journal of Business Research*, 129:927–935, 2021.

[33] Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. 1 2019. URL `http://arxiv.org/abs/1901.10655`.

[34] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3):286–297, 2000.

[35] Dijana Peras. Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, pages 89–97, 2018.

[36] José Quiroga Pérez, Thanasis Daradoumis, and Joan Manuel Marquès Puig. Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, 28(6):1549–1565, 2020.

[37] Bhavika R. Ranoliya, Nidhi Raghuwanshi, and Sanjay Singh. Chatbot for university related faqs. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1525–1530, 2017. doi: 10.1109/ICACCI.2017.8126057.

[38] Adama Samaké and Lahsen Boulmane. Acceptance and rejection zones for a classifier's predictions in deep learning. In *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)*, pages 1–5. IEEE, 2021.

[39] Peter G Sassone. Cost benefit analysis of information systems: A survey of methodologies. In *Proceedings of the acm sigois and ieeecs tc-oa 1988 conference on office information systems*, pages 126–133, 1988.

[40] Burcu Sayin, Jie Yang, Andrea Passerini, and Fabio Casati. The science of rejection: A research area for human computation. In *The 9th AAAI Conference on Human Computation and Crowdsourcing*, HCOMP 2021. AAAI Press, 2021.

[41] Burcu Sayin, Jie Yang, Andrea Passerini, and Fabio Casati. The science of rejection: A research area for human computation. *arXiv preprint arXiv:2111.06736*, 2021.

[42] Matthias Schumann. Methods of quantifying the value of office automation. *Journal of Information Systems Management*, 6:20–29, 1989. ISSN 07399014. doi: 10.1080/07399018908960168. Risk benefit and effects on different levels<br/>.

[43] Daniel W Turner III. Qualitative interview design: A practical guide for novice investigators. *The qualitative report*, 15(3):754, 2010.

[44] Steven Umbrello and Ibo Van de Poel. Mapping value sensitive design onto ai for social good principles. *AI and Ethics*, 1(3):283–296, 2021.

[45] Aad Van Moorsel. Metrics for the internet age: Quality of experience and quality of business. In *Fifth International Workshop on Performability Modeling of Computer and Communication Systems*, volume 34, pages 26–31. Citeseer, 2001.

[46] Jiang-Jiang Wang, You-Yin Jing, Chun-Fa Zhang, and Jun-Hong Zhao. Review on multi-criteria decision analysis aid in sustainable energy decision-making. *Renewable and sustainable energy reviews*, 13(9): 2263–2278, 2009.

[47] Heiko Wersing, David Nebel, Barbara Hammer, Thomas Villmann, and Lydia Fischer. Rejection strategies for learning vector quantization-a comparison of probabilistic and deterministic approaches.

[48] Yonghui Wu, Min Jiang, Jianbo Lei, and Hua Xu. Named entity recognition in chinese clinical text using deep neural network. *Studies in health technology and informatics*, 216:624, 2015.

[49] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3506–3510, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025496. URL `https://doi-org.tudelft.idm.oclc.org/10.1145/3025453.3025496`.

[50] John Zimmerman and Jodi Forlizzi. Research through design in hci. In *Ways of Knowing in HCI*, pages 167–189. Springer, 2014.