Norwegian School of Economics

Bergen, Fall 2021

Chatbots for customer service

A quantitative research

Per Olav Bomann Fosseide & Lars Vattøy

Supervisor: Prof. Ivan Belik

Master thesis: Business Analytics (BAN)

NORWEGIAN SCHOOL OF ECONOMICS

Preface

This master thesis was written as a part of our MSc in economics and business administration with a specialization is within business analytics (BAN) at the Norwegian School of economics. Throughout this thesis, we have gained valuable knowledge about an increasingly popular technology. We have learned how plan and execute a five-month long research study where we have applied data analysis and empirical methods to investigate a phenomenon.

There are many contributors to this thesis, that made it possible to carry out the research we have conducted in the last five-months. First of all, we want to thank our supervisor, Ivan Belik, for both good advice and feedback. We also want to thank DNB, with all the people involved to take time out of their busy schedule to help us with access to data, insights, and professional opinions.

Per Olav Boman Fosseide and Lars Vattøy

Norges Handelshøyskole, December 2021

Executive summary

In this thesis we have investigated the advantages and disadvantages of chatbots for customer service. Chatbots gained popularity because of their potential to partly automate customer service centres. Their success depends on the company's capability to implement the solution, and the customers motivation of using the chatbot. Previous research suggests that cost-cutting, setting a reliable scope, transparency of chatbot capabilities and well written responses, are key contributors to chatbot success. The customers motivations for utilizing chatbots are efficiency, productivity, previous experience, human-likeness, and trust in chatbots. We performed five types of experiments in order to challenge previous research on established truths regarding chatbots. Our experiments included comparative analysis, sentiment analysis, regression, and random forest. Our results indicate that customers are less patient with chatbots than human agents. We found that the users' sentiment towards chatbots is more negative compared to conversations with human agents. Furthermore, we found that when customers use less time and has to write fewer messages to get their inquiry resolved it has a positive effect on customer satisfaction. Asking a chatbot uncomplicated questions instead of complex questions also had a positive effect on customer satisfaction. Based on findings from previous research, our research, and interviews, we recommend several measures to managers that are considering acquisition of chatbots. These are: 1) gradual increase in functionality, 2) include human-likeness, and 3) transparency.

Table of contents

List of Tables

List of figures

Abbreviations

AI	Artificial Intelligence
CSI	Customer satisfaction index
ML	Machine Learning
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
OLS	Ordinary Least Squares
RSS	Residual Sum of Squares
TSS	Total Sum of Squares

1. Introduction

Automating manual tasks is an important success factor for companies in 2021. Since the introduction of the internet and the advancement within data storage, computing power, machine learning (ML), and artificial intelligence (AI), companies are finding new ways to use these technologies to do tasks better and faster. One of these technologies is chatbots. Almost every company that has a customer service function is either using or planning to implement chatbot as the first line of incoming inquiries. In 2019 the number of accessed chatbots were a record breaking 4.2 billion, and the total cost savings enabled by chatbots were 164 million dollars (Woodford, 2020). There is no doubt that chatbots have revolutionized retail, and the forecasted growth of this technology suggests that they will be even more dominating in the future.

Chatbots have existed since the early 1960's but have recently had a significant popularity increase. In the early stages of chatbot development, the machines could only process simple questions and often replied with non-meaningful answers. Since then, chatbots have continuously improved and reduced the difference between machine and human interaction. Through time chatbots have developed from responding with simple pre-programmed answers to understand and improve based on experience. ML techniques, such as deep learning, natural language understanding (NLU) and natural language processing (NLP), allows chatbots to understand human conversations and mimic human intelligence to the point where we often can't separate the two. In addition to technological advancement, chatbots have become less expensive. Previously, companies who wanted to utilize this technology had to invest considerable resources and acquire expert developers. Today the same companies can acquire a chatbot in less than a day from a specialized vendor. The acquiring company only has to personalize and finetune their bot over time, which is much less costly than developing a chatbot from scratch.

Despite the major improvements in chatbot technology, pricing, and the popularity within businesses to acquire such technology, customers are reluctant to use chatbots. Most people do not believe that a conversational AI can outperform a human. One reason is grounded in psychology, where humans don't trust the answers given by a bot. This could also be affected by customers previous experience with chatbots and the fact that chatbots for customer service is fairly new. The other aspect of this is whether a chatbot actually can outperform human

expertise in customer service. We can ask ourselves if the technological improvements have come far enough for chatbots to replace human agents.

Previous research on chatbots for customer service has mostly been conducted through interviews or surveys on customer experience. Research regarding technological advancement is often disconnected from the business aspect of chatbots and whether chatbots can fully function as a replacement for human agents. In our thesis we intended to explore the difference in customer satisfaction between chatbots and human agents. We want to explore whether chatbots are a good investment for business owners and give realistic expectations in what a chatbot can assist with and which tasks a human agent is more capable. In particular, this thesis aims to answer the following research question:

What are the drawbacks of utilizing a chatbot for customer service, and for which tasks are chatbots outperforming human agents?

In order to answer the research question, we have supplemented information from the literature with quantitative research on actual chatbot data. Additionally, we conducted interviews with DNB, to get a third perspective on how well chatbots are functioning as a replacement for human agents. Following this introduction, is a theory and literature review section that 1) gives the reader a fundamental introduction to chatbots, and 2) covers previous research that contributes to answer the research question. Subsequently, a methodology section will present our research approach and how we conducted our own experiments. After the methodology section we present our results from quantitative research of chatbot data and our interpretation of these results. To finalize our thesis, we discuss these results in the context of the literature to answer the research question and explain possible implications and limitations of our research.

2. Theory and literature review

This section contains the necessary information required to understand the capabilities of chatbots and include the theoretical foundation for our approach to answering the research question. The first part defines chatbots and introduce current and historical implementations of the technology. It includes the reasons for why chatbots were invented and the tasks they are assigned to solve. The Second part discusses how chatbots have affected the organizations, managers and the customers who use them. Additionally, it identifies and defines key factors that influence a chatbots ability to accomplish the tasks it was assigned. These factors form the theoretical framework for understanding the current applications and limitations of chatbots in customer service.

2.1 An introduction to chatbots

The purpose of this section is to define chatbots, their properties and give the necessary theoretical foundation to understand them. We first give a historical perspective on what the chatbots purpose and capabilities were when they were developed in the 1960s. Then we work our way through time and technological development to a definition on what chatbots are today. We elaborate on the different types of chatbots that exists in the commercial market and how DNB's chatbot Aino fits with the current available chatbots. Even though voice bots and written conversational chatbots are closely related in both technology and usage, our research mainly focuses on written conversational systems.

2.1.1 History of chatbots

Throughout history we have seen major technological advancements of chatbots. The concept of conversation between computer programs and humans has existed since the 1950s when Alan M. Turing introduced the "Turing test". The goal of his purposed exercise was to figure out if machines are capable of thinking like a human being (Turing, 1950). The test is called the imitation game and to pass the test, a machine had to "successfully convince a significant portion of jurors that it is a human player". (Moody & Bickel, 2016)

One of the earliest known chatbots was called Elizabot (Eliza). Eliza was made for the purpose of passing the Turing test. The bot was developed by Joseph Weizenbaum in 1966, and its goal was to demonstrate natural language between humans and computers. To demonstrate human language, Eliza analyzed and decomposed the input from the user according to a set of pre-defined rules. The responses were a result of decomposed input that was restructured into a question (Weizenbaum, 1966). Eliza had major difficulties in terms of passing the Turing test. These included limited responses and lack of logical reasoning capabilities. This often resulted in inappropriate and non-meaningful responses. (Nuruzzaman & Hussain, 2018)

There is a significant gap between Eliza and the next noticeable technological improvement. The introduction of the internet and collaboration across countries made it possible for Richard S. Wallace to develop Alicebot (Alice) in 1995. This bot was created as a research experiment to pass the Turing test by appearing human-like. For many researchers, Alice was looked upon as an extension of Eliza because of the similarities in how they responded with the intention to mentally stimulate the user of the chatbot. However, Alice was a big upgrade from previous chatbots because it used artificial intelligence markup language (AIML) and was trained with vast quantities of natural language samples. That made a big difference in the capabilities of the bot, as Alice could handle a significantly larger portion of knowledge categories. On launch Eliza had about 200 knowledge categories, while Alice had 40 000 in 2009. In addition to having a larger information database, Alice introduced supervised learning that is still used in most chatbot technology today. A person called the botmaster monitored the chatbot conversation and improved the future responses to be more accurate, appropriate, and believable (Wallace, 2009). Despite winning the Loebner prize for most human-like chatbot in 2000 and 2001, Alice had some major limitations. One of these drawbacks was that personality traits like attitude, mood and emotion had to be integrated manually by the botmaster. The second drawback was lack of NLU, sentiment analysis and grammatical analysis which could result in inappropriate responses. (Nuruzzaman & Hussain, 2018)

The next evident technological improvement was the IBM Watson chatbot. In similarity with most other technologies, the Third Industrial Revolution increased the speed in which chatbots were improved. This explains the shorter gap between technological advancements in chatbots. The idea of the IBM Watson was first pitched to management in 2006 and subsequently launched in 2011. The intention of the Watson was to be a question and answering machine that would push science forward in the field of AI (Lohr, 2021). Compared to previous mentioned chatbots, Watson could analyze extra parameters like phrase structure,

grammar, names, dates, and geographical location. With more parameters, and improved machine learning techniques, IBM's chatbot could provide more accurate answers than its predecessors (Nuruzzaman & Hussain, 2018). The improved technology resulted in a chatbot that could understand natural human language well enough to outperform two previous Jeopardy champions in a quiz competition. (Adamopoulou & Moussiades, 2020)

The success of Watson led to an attempt to commercialize the chatbot. This was a paradigm shift in the usage of chatbots, that until around 2010 had been exclusively utilized for research purposes. IBM turned to the healthcare sector and targeted cancer treatment medical centers which worked with big quantities of data. The new assignment for the bot was to improve medical treatment based on information input regarding a patient. Despite having success in research and academia, Watson did not turn out to be a success in the business world. The problem was high maintenance cost, high complexity, and a lack of flexibility regarding missing data (Lohr, 2021; Nuruzzman & Hussain, 2018). Despite the lack of commercial success, Watson was still a sizable leap towards the chatbots we know in the commercial market today.

2.1.2 Updated definition and chatbot categorization

This thesis has so far described the progress in chatbot technology through time and continues with an updated definition on what chatbots are today. After the updated definition, we describe the different types of chatbots, in addition to placing DNB's chatbot Aino and its peers into the context of these types and definitions. The commercial bots and their traits will be the reference for which types of chatbots we have investigated throughout our thesis.

According to Wang and Petrina (2013) a chatbot can be defined as a computer program that simulates a human-like conversation using natural language. Juniper research (2020) has a similar definition but includes the automated process triggered from the conversation (Woodford, 2020). Another definition of chatbots describes them as software systems intended to mimic human communication while interacting with the user. Furthermore, chatbots are powered by artificial intelligence in order to recognize natural language, emotions, identify meaning and construct responses that are meaningful to the end user (Nuruzzaman & Hussain, 2018). The common idea among these definitions is that a chatbots are computer programs that aims to mimic human conversation, and followingly respond or carry out an action requested by the user. This is the general definition of what is referred to as a chatbot for the

rest of this paper. However, there are multiple chatbot types that can be placed within this definition. We can further distinguish them through 1) type of interaction, 2) purpose, 3) rule-based or AI, and 4) which domain they are supposed to cover (Hussain et al., 2019). In the introduction it was mentioned that this thesis only covers written dialog systems. As a result of that choice, we do not cover other types of interaction.

One way to categorize chatbots is their domain. There are open-domain and closed-domain chatbots and they differ in access to knowledge and their underlying data sources that they are trained on. Open-domain chatbots are programmed to retrieve all sorts of information through the internet. This type of chatbot performs best on general topics. Examples of a use case for open domain chatbots would be to ask about the weather for the upcoming days (Budulan, 2018). Typical examples of open-domain chatbots are Apple Siri or Amazon Alexa. Closed-domain or specific-domain chatbots focus on a particular knowledge domain. All information required to answer the user is in the bot's database (Nuruzzaman & Hussain, 2018). The closed-domain chatbots are designed to answer specific questions in simple scenarios. A chatbot found at a specific webpage will generally be a closed-domain chatbot that can answer questions regarding the business connected to that website. There is not a clear-cut distinction between the two types of chatbot domains, and it should therefore be seen as a scale with two extremes. The more knowledge and scenarios that a specific-domain chatbot can handle, the closer it is to an open-domain chatbot. (Budulan, 2018)

Another classification of chatbots describes whether it serves a specific task or not. The purpose of a chatbot can be summarized into two categories: 1) task-oriented and 2) non-task-oriented. Task-oriented chatbots are conversational systems made for helping the user solve a specific task. In other words, they are designed for dealing with specific scenarios like access to certain information or placing an order. Task-oriented chatbots performs best in specific-domains and cannot assist with general knowledge. Non-task-oriented chatbots, also called chit-chat bots are intended for longer conversations. The purpose of these conversations is to mimic human to human interaction, for fun and entertainment.

The last classification of chatbots is whether they are self-learning or simply follow predefined rules in order to reply. Rule based models rely on input that matches the predefined rules. If not, the chatbot is ineffective in answering the question from the user. These bots perform well on uncomplicated questions but composing the rules necessary to cope with every intended situation can be an infeasible job. The self-learning or AI chatbots use machine learning

algorithms to learn from previous conversations. We can further split self-learning chatbots into two categories: 1) Retrieval-based models and 2) Generative models (Thorat & Jadhav, 2020). Retrieval-based models are limited to predefined replies. The responses are retrieved from a database with techniques such as keyword matching, machine learning and deep learning in order to choose the best possible response from a repository. In essence, these models do not generate new output. Generative models on the other hand can generate new dialog based on large quantities of training data. A combination of techniques such as supervised and unsupervised learning is used to generate proper responses during a conversation. Even though this sounds like a better option than the retrieval-based systems, generative models are still fairly new. They can often sound repetitive and are unable to support stable conversation. Generative models are some of the most advanced chatbots today and are mostly seen in research. (Fainchtein, 2020)

To summarize this section, it is important to note that one chatbot can be a hybrid of different categorizations. We mentioned a scale in the domain categorization, and the same applies for the purpose of a bot and the self-learning vs rule-based systems. It should also be emphasized that chatbots consists of a combination between categorizations. For example, a chatbot can be specific-domain, task-oriented and retrieval-based chatbot simultaneously. The combination mentioned is also the most common chatbot for businesses today. As we mentioned in our introduction, our thesis includes quantitative experiments on DNB's chatbot Aino, which is not an exception from that specific combination. Hereafter we have put emphasis on this combination: specific-domain, task-oriented and retrieval based chatbots when answering our research question in terms of literature and experiments in the upcoming parts.

2.2 Chatbots for customer service

The core intention of this thesis is to discover what makes chatbots desirable in addition to the possible shortcomings of this technology. We have elaborated on how chatbots has evolved through time and defined what they are today. The upcoming section aims to answer the research question by utilizing previous research. There are mainly two stakeholders affected by the performance of a chatbot: 1) The company that implement it and 2) the customers that wants their inquiries resolved. Companies care about efficiency, cutting costs while providing good customer service. Users on the other hand want their requests solved with the least

amount of time spent. Because of the different motivation between company and customers, this section will be divided into two parts where each perspective is examined. To conclude the theory and literature review, we close the gap between theses perspectives and explain how our own research contributed to answering the research question.

2.2.1 Business and manager perspective

Even though chatbots as human substitutes have had mixed success, commercial chatbots have experienced an exponential growth from a market size of under 200 million dollars in 2016 to 5 billion dollars in 2020. Furthermore, the market is expected to grow at an annual rate over 20% until 2025. (Markets & Markets, 2020)

The financial industry and especially the banking industry has adopted chatbot technology as a core part of its customer service. Chatbots' ability to solve straightforward user requests have allowed the customer service department to offload the repetitive simple tasks and use more time to handle the complex issues. The Norwegian bank Sbanken reported having chatbots handle over 40% of all incoming customer inquiries. This amounts to the equivalent of 31 full time employees. Chatbot implementation also increased Sbankens capacity for customer service by 175%. The report further claims that the chatbot successfully answered 4 out of 5 questions without the need for human support (Boost.ai, 2020). The potential of replacing existing staff and reduce the need for new employees makes chatbots an attractive option for many businesses. Operational cost savings from using chatbots is expected to reach 7.3 billion dollars worldwide in the banking sector by 2023. (Juniper Research, 2021)

Chatbots does not only create an opportunity to reduce staff. They also present an opportunity to offer better customer service or expand the customer service opening hours to offer support that is available at all times. There are three main factors in customer service that influence customers willingness to pay more. These are 1) availability, 2) efficiency, and 3) speed (PWC, 2018). Hallowell (1996)studied the relationship between customer satisfaction and profitability in banking and found a positive correlation, suggesting that increasing customer satisfaction could improve profitability. Hallowell still cautioned against attempting to satisfy every potential customer. The level of customer service required to satisfy everyone could be more costly than the potential benefit of keeping them as a customer. Similar findings about the positive effects of customer satisfaction were presented in research conducted by Islam et al. (2020), that also found that there was a significant positive correlation between customer

service quality and customer satisfaction. They also found a significant correlation between customer satisfaction and customer loyalty in banking.

It can therefore be argued that chatbots could increase profitability by improving customer service efficiency and availability. There are however some limitations of the technology that may limit the kinds of tasks a company would want a chatbot to handle. The lack of a one-size-fits-all product means that significant resources are needed to tailor off-the-shelf chatbots to meet an organization's needs. Organizations that seek to automate its customer service through chatbots are often required to write the chatbots responses to each specific topic, and for infrequent and complex topics this can require significant investments with small returns in customer service efficiency. For a business, such considerations would need to be made when considering the acquisition of chatbots for customer service. (Zhang et al., 2021)

In an interview study, Zhang et al. (2021) summarized the experiences of 14 managers into 3 main lessons learned. These lessons where: 1) understanding the chatbot technology, 2) acknowledging that chatbots do not eliminate the need for customer service personnel, and 3) the lack of one-size-fits-all solutions. Understanding the chatbot technology was reported to be essential for successful chatbot implementation. It was important for an organization’s ability to 1) set an appropriate scope, 2) estimate the resource requirements and 3) set the evaluation criteria to measure the chatbot performance. Not understanding the technology and having unrealistic expectations could therefore lead to difficulties when deploying a chatbot. The second lesson puts empathies on the fact that companies would still need customer service personnel. The job of the customer service personnel did, however, change and many of the respondents reported the change as a positive one. Their tasks generally became less repetitive and some of the personnel were used as AI-trainers. The last lesson was that there was no way of implementing chatbots that would work for all companies and that the implementation should therefore be adapted to fit a given organization. (Zhang et al., 2021)

The implementation of chatbots requires large investments in technology, infrastructure, and training. Zhang et al. (2021) found that there were significant requirements of technological understanding for organizations that wanted to implement chatbots. The role of the manager as change leader and resource manager was empathized as a success criterion. The manager of the customer support department would need to lead the employees through the transition from doing repetitive work to a more challenging and autonomous role. Implementation of chatbot technology has caused concern about job security amongst the customer service

personnel as some fear that it would outcompete them. Changes in the chatbot market has also led to a reduction of in-house software developers needed to implement and maintain chatbots. Most of the companies who were surveyed had purchased a platform solution from a specialized vendor. Many of these vendors offers “no-code” platforms where the training of the bots can be performed by employees without software developer skills. The quality of customer-bot conversations is therefore increasingly determined by the quality of the data that is used for training the algorithms. Training the bot was often done by the companies themselves by employees who had previously worked in the customer service departments. (Zhang et al., 2021)

The employees who trained the bots to understand the inputs from the users and make sure that it responds with the correct information were called AI trainers. Zhang et al. (2021) identified three core skills that were prerequisites for the AI trainer role. These were: 1) prior experience with customer service, 2) good writing skills and 3) analytical abilities. AI trainers are considered critical for chatbot implementations as they provide adjustments to the chatbots to ensure that it performs as intended. They are also responsible for writing the responses that the chatbot provide. The topics or *intents* that a chatbot can handle is therefore limited by the available training data and the adjustments made by AI trainers. (Kvale et al., 2020)

As chatbots are currently not able to handle all requests, setting a realistic scope is important for a chatbot to be successful. Kvale et al. (2020) studied the chatbot dialogues of the Norwegian telecom company Telenor to better understand the conversational abilities of chatbots for customer service. The study involved manually rating 406 dialogues on its ability to resolve customer request and the quality of the conversations. The study showed that the chatbot only resolved 24% of the conversations without any human input. 25% of the queries were immediately handed over to human customer service personnel, 13 % were classified as irrelevant customer inputs while the rest were either eventually handed over when the chatbot did not understand the user’s questions or abandoned by the customer when the chatbot failed to answer the questions correctly.

The chatbot’s dialogue quality was further rated on a 5-point scale from poor to excellent. Only 36% of the chatbot-interactions were marked as satisfactory, while 64% of the conversations were marked as having room for improvement. The main causes of failure were that the chatbot either predicted the wrong topic or did not understand the customer's intent. Missing topics not yet added to the scope by the AI trainers was a common source of chatbot

failure. However, researchers saw this as less problematic, because the chatbot could transfer the customer to a human agent if the question was outside of the scope. Cases where the chatbots misunderstood the question were seen as more severe, as the chatbots failure would be harder to detect without manual review. Such failures could therefore hinder further improvement and cause frustrations among the customers. Furthermore, the customers' ability to clearly formulate a specific need was a characteristic of successful dialogues. The success of chatbot implementations is dependent on the customers willingness and ability to communicate with the bot in a way that it understands. Additionally, the chatbots will often fail if the customer's questions are either complex, phrased in a way that is difficult for the machine to understand or contain mixed topics. Chatbots also have problems answering customer who are dissatisfied with the answers they were given. (Kvale et al., 2020)

Chatbot technology has the potential provide both cost savings and efficiency gains companies. Whether it can replace the customer service personnel depends on several factors. The important factors a company should consider when deciding whether to implement chatbots can be summarized as: 1) Are the customer inquiries both uncomplex and concerning few topics? 2) Is the organization ready to invest the required resources in personnel and infrastructure required for a successful implementation? 3) Are the customers willing to use chatbots? The topic of user motivation and willingness to use chatbots is further explored in the next section of the theory and literature review.

2.2.2 User experience

This section elaborates on research connected to the user experience with chatbots and how it further explains the capabilities and limitations of chatbots for customer service. An important aspect of what makes chatbots successful from a user perspective is their motivation of utilizing chatbots instead of human customer service. First, we will explain how efficiency motivates customers to use chatbots. Next, we elaborate on how demographics, expectations and previous experience affects customer satisfaction. To finalize this section we explain how transparency, trust, emotion, and human-likeness are traits that make chatbots more desirable.

Two connected motivations for using chatbots as a preferred way of contacting customer service, is efficiency and productivity. A study conducted by CGS (2019) describes these two traits as the customer's main motivation of using a chatbot. Similar results were found in an interview study conducted by Brandtzaeg and Følstad (2019), where 24 participants shared

their experiences with chatbots. In this study it was found that customers preferred chatbots because it allowed them to get quick answers, without having to wait for a human agent. Some of the participants compared talking to a chatbot with internet searches, because the customer does not have to locate the information themselves. In a survey of 500 U.S and UK participants chat, and written messages were ranked as the most popular way to reach customer service. (CGS, 2018)

In addition to efficiency and productivity other motivations were mentioned. The chatbots were preferred by some of the respondents because they were seen as non-judgmental and would not judge them for asking what they considered as stupid questions (Brandtzaeg & Følstad, 2019). Despite being perceived as less judgmental, most people prefer human agents over chatbots because they are perceived as less competent (CGS, 2018). However, this perception differs based age and gender. Surveys conducted by CGS (2019) show that women are on average more likely to prefer chatbots over human representatives and were more likely to believe that the chatbot could resolve their issues. People under 34 were also more likely to declare that “chatbots and virtual assistants make it easier to get their issues resolved”. (CGS, 2019)

Expectations of what a chatbots provide in terms of problem solving is another factor that determines chatbot performance. Most people have realistic expectations regarding the chatbot's ability to provide customer service and understand that the bots may have limited capabilities in understanding and solving complex issues (Følstad & Skjuve, 2019).Despite having realistic expectations, people tend to have a negative attitude towards chatbots. They are often perceived as less knowledgeable and lack empathy. The negative attitude towards chatbot capabilities can be mitigated by exposure and experience with AI and chatbots (Luo et al., 2019). Gümüş & Çark (2021) found that people who considered chatbots easy to use were more likely to have a positive customer experience and use them again. Their findings suggest that there exists a positive feedback loop where previous experiences encourage future use.

Customers also prefer that the chatbot is transparent and clearly states its limitations and assumptions of what topic of the conversation is. Allowing the user to change what the chatbots believe that the customers want, could compensate for poor topic prediction or cases where the chatbot has misunderstood the context of the conversation. (Jain et al., 2018). There are however some potential negative consequences of transparency. Luo et al. (2019) found

that disclosing that the customer is speaking with a bot decreases the purchase rate by 70%. The effect of disclosure is less severe if the chatbot reveals itself to be a bot at the end of the conversation. This suggests that chatbot performance is directly affected by customers initial distrust. (Luo et al., 2019)

For automation and chatbots, trust can be defined as "*the extent to which a user is confident in, and willing to act on the basis of the recommendations, actions, and decisions of an artificially intelligent decision aid*" (Madsen & Gregor, 2000). A study from the Norwegian research center SINTEF (2019), explored the topic of users' trust in chatbots through semi structured in-depth interviews of 14 participants. Trust in technology is a significant factor when determining whether a customer would want to use chatbots. Lack of trust could make chatbots less effective and would therefore be a limiting factor. The chatbots ability to understand the user messages and retrieve the correct information was the most common factor that affected a user's trust. Another feature that increased trust was how the bot wrote and presented itself, where most of the participants wanted the bot to appear human-like. Examples of human traits mentioned were informal language like humor or a human avatar. Other research of how the human-like qualities affect the user's willingness to use chatbots support this view. (CGS,2018; Nordheim et al., 2019)

The ability to recognizes the users emotional state and to express empathy helps the chatbots to generate responses that resembles human conversation (Agarwal et al., 2021). Giving customers the option to select among several possible chatbot interpretations of the topic could compensate for lack of training or scope in the chatbots capabilities (or "knowledge"). This can help the conversation and avoid some of the situations where the chatbot might misunderstand the conversations topic (Jain et al., 2018). Understanding the emotional cues of messages and responding appropriately could increase the users trust in the bot. The customer could also be transferred to a human agent if the topic is perceived as to sensitive for a chatbot to handle. Human traits and emotional intelligence could therefore increase the capabilities and performance of chatbots. (Følstad et al., 2018). However, human-traits are not desirable in every situation. Ng, et al. ((2020) studied the relationship between disclosure of financial information and human-traits for the two chatbots named XRO23 and Emma. Emma was described as empathetic with human qualities, while the description of XRO23 empathies its robotic efficiency. People who were presented with both chatbots were more likely to share financial information with XRO23 than with the chatbot Emma. There are also ethical concerns with introducing human traits to commercial bots. Feine et al. (2019) studied the

gender-features like name, avatar, and description of 1375 chatbots. Most commercial chatbots were gendered and 83% of customer support chatbots had female specific gender traits. They raised concerns about how gendered chatbot design could promote gender specific stereotypes. Whether chatbots should have human-traits like names and other features, should therefore be dependent on the task it is assigned to solve and the environment it operates in. Bridging the gap between humans and chatbots in terms of the customer experience might introduce some of societies prejudice to the chatbots which could cause unintended consequences for the company.

Throughout section *2.2 Chatbots for customer service*, we have elaborated on chatbot success factors and limitations. We have explored these advantages and disadvantages in two sections: 1) business and manager perspective and 2) user perspective, because the two groups interest are not necessarily aligned when it comes to utilizing chatbots. In the business and managers perspective, we found that a key contributor to the investment in chatbots is automation and cost savings. Expanding the availability and at the same time reducing the customer service personnel is the desired result. However, the implementation of chatbots is both costly and time consuming. Good management and resource allocation are two important factors when deciding whether a company can gain value from utilizing chatbots for customer service. Companies must understand the technology and set an achievable scope. By failing to do so, chatbots could prove to be a disadvantage rather than an advantage.

Expectations are important for customers as well. If customers have unrealistic expectations in what a chatbot can assist with, it could negatively impact their willingness to talk to chatbots in the future. Realistic expectations can be achieved through experience, but also the transparency from the company in what a chatbot can do. Another key component of chatbot success is how good the chatbot predictions are in combination with responses written by AI-trainers. As long as the chatbot understand the user intention, it can either respond or hand over the conversation if the question is outside of the chatbot scope. How well the chatbot understand an inquiry is also highly dependent on the customer. Chatbots are better at prediction intention for well formulated questions that is not to complex or consists of multiple topics. From the user side of chatbots, efficiency and productivity drives customers willingness to use chatbots. Chatbots are available 24-7 and could prove very useful to navigate complicated websites. A clear drawback of chatbots is their limited capability of interpreting emotion. That is an area where human agents are superior thus far.

After investigating the capabilities of chatbots from research, we were left with an impression that most of it was done through qualitative metrics. Surveys and interview studies dominate the current research in disclosing chatbot capabilities. The advantage of qualitative research on chatbots is that it provides an understanding of each interviewee's perception of chatbots. They can elaborate on why they prefer or dislike utilizing chatbots. However, qualitative research can be subjective and give vague conclusions. In our research we therefore wanted to focus on numbers, and specifically how different quantifiable metrics influence the chatbot performance. Our goal is to either confirm or reject established "truths" in addition to finding other influential factors. The upcoming section intends to elaborate on our research approach, the data used, the reasoning behind every experiment and how the different experiments could assist in answering the research question.

3. Methodology

The intention of this section is to further explore the drawbacks of chatbots, and for which tasks they are outperforming human agents. A qualitative research approach was chosen in order to complement and add to previous research on this topic. The experiments presented in our research has been inspired by the lack of quantitative analysis on chatbot data from the literature. First, we elaborate on our research design and why that specific design was chosen. Following the research design will include data collection and how we prepared the data to carry out the experiments. Following the data collection, we continue with data evaluation. The data evaluation includes strength and weaknesses with the dataset, insights in quantitative metrics of chatbot and a breakdown of why the chatbot received low scores in august 2021. The last section elaborates on the methods used in each of the five experiments.

3.1 Research design

The field of chatbots is widely researched and continues to be a topic of interest because of its increased relevance and business application. The chatbot capabilities are developing rapidly, and new research is therefore necessary. Previous research on chatbots for business has mostly been conducted through survey and in-depth interview studies. That gave us inspiration to contribute with quantitative research that is less common when evaluating chatbot performance. We are focusing on the phenomenon of chatbot performance and what makes them inferior or superior to human customer service.

One of the key attributes of quantitative research is that it establishes statistically significant conclusions. In order to do so, it requires a representative sample of the population that is studied. The population can be broad or narrow but requires that every individual data-point fits the description of the group being studied. Because of impractical reasons of including everyone in a population, it is common practice to choose a representative population. Our thesis consists of both descriptive and experimental research. Experimental research determines if independent variables have a causal effect on a dependent variable. Causality refers to a how the independent variables influence the dependent variables. Descriptive research is used to describe the characteristics of the researched population. Contrary to qualitative approaches, quantitative research is establishing causality because it happens in a controlled environment and provides more precise measurements. (Lowhorn, 2007)

3.2 Data Collection

3.2.1 Primary data

The datasets used for the empirical part of this research are from DNB's internal databases that is not available for the general public. One of the datasets consists of chatlogs where we could choose a time interval between 2018 and 2021. The other dataset is within the same timespan but includes customer evaluations. The two datasets are connected through a conversation identifier. The datasets do not include any personal information about the customers or the employees which makes it possible to use the dataset without any concerns about privacy. The chatbot datasets used for our experiments was from September 2020 and consists of 191 735 unique conversations with a total of 2 212 675 messages sent. We found that how a customer could evaluate a conversation changed over time. We therefore chose to use chatlog data from September 2020, because of the consistency in customer evaluation metrics. The variables in the dataset are described in the following tables:

Chatlog dataset

Table 1: Chatlog data set: variables, descriptions, and values

Variable name	**Description**	**Value**
Timestamp	Time of when the message is sent.	Year-Mont-Day-Hour-Minute-Second **Datatype:** timestamp
From	Which participant that is sending the message	User, Agent, Bot **Datatype**: String
Conversation ID	An id that is common for every message in a conversation	Any unique number 1,2...Maxint **Datatype:** Integer
Message ID	Unique number for each message sent.	Any unique number 1,2...Maxint **Datatype:** Integer
Message	Conversation sent from bot, agent, or customer.	Text input **Datatype:** String
Prediction description	Chatbots prediction of how well it predicts the user intent.	Ex: Very certain, uncertain... **Datatype:** String

Context intent	The topic predicted by the chatbot	Reason for the inquiry. Ex: problem with bankID **Datatype:** String
language	Which language the bot predicts the conversation to be in	English, Norwegian, Danish, Swedish **Datatype:** String
Language ID	Numerical description of the language variable	1.0 = English, 2.0 = Norwegian, 3.0 = Danish, 4.0 = Swedish **Datatype:** String
Message description	Type of message	Message, Action buttons, Link clicked, customer question etc. **Datatype:** String

Customer evaluation dataset

Table 2: Evaluation data set: variables, description, and values

Variable name	**Description**	**Value**
Conversation ID	An id that is common for every message in a conversation	Any unique number 1,2... Maxint **Datatype:** Integer
Customer satisfaction index (CSI)	Rating given by customer of overall performance	CSI = (1,2,3,4,5,6) **Datatype:** Integer
Task completion	Rating given by customer of overall performance	Yes / No **Datatype:** String
Chat duration	Measuring time of the conversation	seconds 1,2.3 … end of conversation **Datatype:** String
Chat mode	Whether the evaluation is for a chatbot or human agent	Agent, Bot **Datatype:** String

3.2.2 Software

The programming languages used to carry out our analysis has been R and SQL. We used RStudio on a virtual machine in order to get access to the desired datasets described in *3.2.1 primary data*. All data querying, preparation, cleaning, and analysis has been done with various packages in RStudio and can be viewed in section *R packages utilizised* in the appendix.

3.2.3 Repairing and merging datasets

In order to work on the dataset that contains messages from users, we had to repair some of the conversations. Furthermore, the dataset contained encoding errors from language specific characters for the Nordic countries. Because of continuity, we were able to replace the encoding errors with the intended Nordic characters. For analysis purposes, we removed unnecessary whitespaces, one-letter words, punctuation, and digits. Next, we ordered the chatlogs so that each conversation is separate and in order. This made it easier to read full conversations, and was useful as an extra measure to see if our functions and methods did what was intended. Rows that did not include any messages, and conversations that did not include any message from a user was removed. Finally, the Norwegian words “ja” and “nei” was substituted with “yes and no” to get a universal standard for the task completion variable.

After repairing necessary messages in the chatlog data set, we connect the two datasets. This was done by the left-join on the variable *Conversation ID* that identifies each unique conversation. The variable-columns depicted in *Table 3* and *Table 4* was merged into a common dataset that was used for the rest of the experiments.

3.2.4 Tokenization and Document-term matrix

Before the sentiment analysis could be performed, the data had to be first tokenized then arranged into a document-term matrix. The size of the dataset combined with processing and memory limits of the AWS instance the analysis was performed on constrained the size of the dataset that could be analysed. Measures to decrease processing time and ram requirements was therefore implemented.

The first step was tokenization, which involves dividing the messages into individual words and removing all punctuation and any numbers in the chats. The algorithms used in the

analysis does not consider the order that the words are written or any of the punctuation that is used. The next step was to remove words that provided no benefit for the analysis. Words containing 2 or less characters and stopwords were therefore removed. Stopwords are words that don't provide any information for classification purposes or have little or ambiguous meaning. These words provide no information for further analysis and increases processing times. Lemmatization of the words was considered as that would have reduced the number of words needed for the analysis. Lemmatization is a form of text-normalization where contextual and grammatical information is used to find the lemma (root of the word) For example: "go" is the lemma of "goes", "going", "went", and "gone". (Hofmann & Chisholm, 2020). We ultimately decided to not lemmatize the words as the dictionary that was used in the sentiment analysis contained the inflected forms of the words.

The R packaged *SentimentAnalysis* (Proellochs & Feuerriegel, 2021) was used to perform the sentiment analysis. The utilized functions required that the data was either a vector of words, document-term matrix, or a corpus. We chose to use a document term matrix because of their storage efficiency and search and retrieval speed. The document term matrix represents the data in the form of a matrix where the rows correspond to the documents (Messages), and the columns corresponds to the terms (Words). Each cell in the matric then represents the frequency of a certain term in a document. (Hofmann & Chisholm, 2020)

3.2.5 Secondary data

In total we conducted two semi structured interviews, based on an interview guide. The semi structured interviews had the goal of revealing insights from a company's perspective on chatbots and disclosing the knowledge that the informants had on the topic. The interviews were conducted in a relatively free manner with open questions and the possibility to ask follow-up questions. The interviews have been an important asset to get an additional perspective on our research question. The questions asked in the interviews are summarized in appendix under *Interview 1* and *Interview 2*.

The interviewees are three people that work closely related with chatbots. In order to gain valuable insights, we interviewed employees with both a technical and managing experience. All of the interviewees have many years of experience in customer service and with technical solutions connected to customer service. Their names and concrete roles have not been

collected and are excluded for privacy reasons. When referencing to the interviews in the section *5. Discussion*, we refer to informant number 1 and 2 from *Interview 1*, and informant number 3 from *Interview 2*.

3.3 Data evaluation

3.3.1 Data characteristics: weaknesses and strengths

The data collected for this thesis has both strengths and weaknesses. One of the good features is that our dataset contains a large quantity of conversations. A large dataset mitigates some of the problems of diverse selection and having a representable dataset for the researched population. Another good feature of the dataset is that each conversation and message have unique identifiers. In combination with a timestamp that is accurate to the second level. This made it easy to sort conversations. Furthermore, topics and language are identified with numbering and text. For research connected to topics, it enabled us to include every language in the source material, because the topic numbering did not change across languages. The dataset also contained all the messages that each user's ether sent or retrieved. This both allowed for text analysis of user messages and the opportunity to manually review outliers in the data.

The most prominent data weakness of the dataset is the messy chatlogs that must be processed to be used for text and statistical analysis. This could however be fixed with observing the errors and insert the intended characters. Another weakness is that the datasets does not include when a chat was disconnected due to time-outs, loss of connections or closing of the chat-window. That makes it hard to evaluate the chatbot performance based on abandonment rate. If a customer abandons a chat, it could either suggest that they got the information they wanted, or that they don't want to continue their conversation. Another metric that could be included in the upcoming section is the number of repair utterances. That is how many times the chatbot has to ask the user to rephrase their question, because it did not understand the intention. The utterances are varying over time, and not standardized which makes it infeasible to cover all of these. A final weakness that should be mentioned is the amount of customer ratings. Only about 10% of customers leave a review of their conversations. This could imply that the ones who leave a review are either very dissatisfied or very satisfied with their inquiry. This could cause some bias as the data might not be representative of the entire population.

3.3.2 Quantitative metrics of chatbot

In order to answer our research question, we need some framework to evaluate the performance of the chatbot. In this chapter we want to give some clarifications on why good evaluation metrics are important, and which metrics we want to use when evaluating chatbots through literature and through our own research.

Chatbot evaluation is important for many reasons. The book: *Evaluating Dialog Systems* lists the following reasons why chatbot evaluation is important: 1) Developers want to know whether the system performs as anticipated, 2) For users it is important to see if the chatbot meets the user needs in terms of understanding and achieves the user goals efficiently for task-oriented systems. For the non-task-oriented systems, whether the chatbot gives an enjoyable experience. 3) Lastly for the researchers it is important to establish whether the aim of the research is met or whether the chatbot shows improvement. (McTear, 2021)

Since our research revolves around task-oriented systems and quantitative data. We will present some quantitative metrics that will contribute to our way of answering the research question. The book "Conversational AI" presents the following quantitative metrics for measuring chatbot performance.

Table 3: Quantitative evaluation metrics

Quantitative metrics	**Explanation /Description**
Task success	If the chatbot solved the task prompted by the user. This is possible to measure through user ratings at the end of chats where customers answer the question: "did this resolve your problem".
Task duration	How long it took (time) to resolve the matter.
Number of system turns	How many messages the chatbot had to send in before the problem was resolved.
Correct transfer rate	Did the chatbot correctly transfer the customer to a human customer advisor?
Containment rate	Percentage of chats that does not transfer to a human agent.

Abandonment rate	Number of hang-ups or chat-disconnects before the task is completed.

Source: (McTear, 2021)

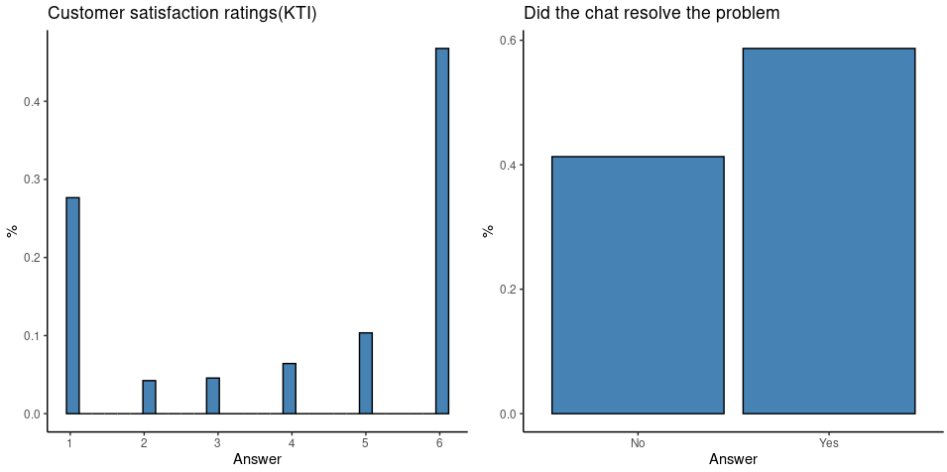
Abandonment rate is hard to measure with the dataset we are using in our research. Although this is an important metric to evaluate a chatbot system, the dataset does not contain a clear identification of when a user disconnects, times-out, miss-clicks or abandons the chat.

Correct transfer rate is also an important way to evaluate how well a chatbot system is working. When a chatbot cannot help, customers want a system that can seamlessly connect them with the correct competence. However, this metric is not possible to measure with the available dataset and will therefore be left out of the analysis part.

Customer satisfaction index (CSI) is a metric that is not mentioned in the table above. This is a way for the customer to rate the overall performance of the chat service. The index is created from post chat reviews where the customer can leave a rating between 1 to 5. There are several implications of using this as a metric of how well the chatbot is working. First of all, the summary statistics shows that customers rarely leave anything other than a 1 (Worst) or 6 (Best). Furthermore, the customers only rates approximately 10 % of all chats. We assume that the ones who rate the chat is either very satisfied or very dissatisfied with their customer service experience. This could affect how representable the CSI score is for the entire population.

A better metric of the overall chatbot experience is the “Did this resolve your problem” customer feedback. It fits better with the fact that customers are either satisfied or not satisfied. Similarly, to the CSI, this binary valuation of rating the chatbot is only answered by approximately 10% of the customer base. We refer to this metric as task completion for the rest of this thesis. In the *Figure 1*, we give an overview of the two metrics: CSI (left picture) and task completions (right picture).

Figure 1: Rating of the chat-service in general for the chosen time period. Answers on the x-axis and % portion of that answer is marked on the y-axis (includes both ratings of customer service agents and chatbot).



There are two ways of measuring user satisfaction, these are user and expert ratings. According to Ultes et. al (2013) , they both have their advantages and disadvantages. Unlike user ratings described in the two previous paragraphs, expert review could give more insight in identifying specific issues. As mentioned in the literature review, humans are better at understanding emotions and the context of conversation and could therefore give developers a better insight into what went wrong. Despite their differences, the correlation between user and expert ratings is quite high (Ultes et al., 2013). This means that the best metric to use is the one that is most easily accessible. We used user-ratings because they were the most accessible in our situation.

3.3.3 Review of low score conversations

DNB wrote a report on all the user feedback that was given in august 2021. In the report they divide the feedback of the users who gave the lowest rating to the chatbot into 10 categories. There were 1819 comments where the user also rated the bot on a scale from 1-6 (1 is worst, 6 is the best). Only 2% of all the customers who interacted with the chatbot chose to leave a comment and the majority of comments came from customers who gave the worst rating to the chatbot. As mentioned, this can create bias, as the distribution of negative and positive ratings are polarised where most users rate the chat either as either 1 or 6. This means that we can’t extrapolate the breakdowns here to the entire dataset, but it can give an indication of what the main causes of chatbot failure are.

Table 4: Aino chatbot breakdown August 2021 (where CSI = 1)

Customer did not get an answer to their question	388
Customer did not understand how to contact a human representative	115
The customer was kicked out of an active chat	71
the customer had to wait a long time before speaking to a human representative	59
The customer had a negative attitude towards of chatbots	46
There were technical issues	28
The customer was dissatisfied with DNB or DNB's products	28
The customer pressed the wrong rating	27
The customer complained about dnbs availability and opening hours	22
Total	835

From the table we can observe that the most common reason for poor customer ratings is caused by the chatbot not answering the customers questions correctly, while 4 of the top 6 reasons are caused by the customer not reaching a human agent or not wanting to speak with the chatbot.

The significant number of customers who did not want to interact with chatbot suggests that many of DNBs users don't trust the chatbot or have experienced that the chatbot has failed to resolve their issues. Many of the customers reported that they were annoyed that they were forced to interact with a technology that they did not trust or wanted to use. From customer surveys in the U.K and US we know that there is significant part of population that prefers human over chatbots and a many are concerned that companies are moving too fast towards AI driven customer service. 67% of UK customers over 65 feared that this development would make it harder to reach a human representative. It is therefore most likely a large percentage of DNB's customers who also have a negative view of chatbots and do not want to interact with them. There is however an important caveat when considering customers with a negative view on chatbots. 75% of the same people who were surveyed on chatbots said that they were not willing to pay more to access human representatives (CGS, 2018).

Most of the negative customer feedback was caused by the chatbot failing to answer or help the customer with their inquiries. DNB states that such failures could cause the customer to choose another medium to contact customer service like phone or electronic messages. When discussing measures for this issue DNB suggests both improving the capabilities and to be more transparent with the capabilities of the bot.

3.3.4 Reliability

Reliability of research is "*the ability of separate researchers to come to similar conclusions using the same experimental design or participants in a study to consistently produce the same measurement"* (Lowhorn, 2007). Our research is mainly focusing on quantitative measures that inherently makes it strong in terms of reliability. The data collected for this thesis is not available to the general public. However, there are examples of where similar datasets have been used for different research purposes. It is therefore likely that future research could utilize similar datasets. Our datasets contain large amounts of data that makes the reliability of our research less vulnerable to biases. In order to ensure good data selection, we have applied our experiments to subsections of data as well as different timeframes and achieved nearly identical results. Repeating our experiments with a chatlog and evaluation datasets should yield the same results within statistical uncertainties and the results can therefore be generalized.

3.3.5 Validity

The validity of research is the used instrument's ability to quantify the intended measure. In our thesis we intend to measure chatbot performance on many different metrics to understand the overall performance of a chatbot. This helps us to understand what the drawbacks of chatbots for customer service are, and on which metrics the chatbots performs better than human agents. If one of the metrics used in our experiments does not measure chatbot performance, the research would not be valid. To ensure validity of our experiments we have carefully stated what we would like to achieve with each experiment, and how it helped us in answering the research question. (Lowhorn, 2007)

Internal Validity

Internal Validity refers to the truthfulness of the executed study, or in other words how the research establishes the cause-effect relationship for the experiment and outcome. It also includes the researcher's degree of confidence of how the variation in the dependent variable is influenced by the independent variables (Lowhorn, 2007). There are risks of encountering correlation between the independent variables and the dependent variable without any causal relationships between them. We therefore carefully assessed our variables and ensured that the ones that were included in the experiments had its effect documented in other research. There is also a risk that other variables effect our dependent and independent variables, without being present in the dataset. Interviews and other secondary data were therefore collected to understand how the variables could be affected. This data will be used when discussing the experiments to account for bias.

3.4 Methods

The main objective for this thesis was to understand the capabilities of chatbots for customer service. To understand which tasks chatbots are capable of handling, several forms of analysis were used. The methods were chosen because they alone provided insight into different aspects of chatbots and combined provide a holistic view of the capabilities of the technology.

First, we compare evaluation of chatbot and human customer in an experiment called comparative analysis. The chatbots evaluation metrics will be assessed to understand how the chatbots performs compared to human agents. Exploring how a large banks chatbot implementation handle user requests, grants insight into what capabilities contemporary chatbots have. These metrics reveal how chatbots influence the time and effort required by users when interacting with the customer service department. The user experience will be further explored trough sentiment analysis. Measuring if the user's sentiment and use of emotionally charged words differs between agents and chatbots

Further analysis is done through multiple and logistic regression where the goal is to understand how chatbot metrics influence customer satisfaction and which topics the chatbot can handle. Regression allows us to establish if there are any linear relationships between the chatbots metrics and the user ratings. Regression analysis will also reveal the magnitude of the relationship and whether its statistically significant. Statistically significant results can be applied to the entire population when the assumptions of the regression are satisfied.

Finally, a random forest model is applied to the data. The random forest model complements the regression analysis well, as it accounts for non-linear relationships and interactions between the independent variables. This method requires less assumptions about the variables in the dataset, and the possible variables that are not present. Comparing the predictive power and influential variables of the random forest to the regression analysis will grant greater certainty of which variables influence task success. Therefore, it could compensate for some of the potential weaknesses in regression analysis.

To conduct the regression and random forest analysis, new variables were created from the variables depicted in *Table 1*: Chatlog data set: variables, descriptions, and values and *Table 2*: Evaluation data set: variables, description, and values. Followingly, these variables are listed and the theoretical reasoning behind each variable is explained:

Time: The *time* variable gives us insight in how long each conversation is. From studying previous literature, we experienced that one of the core motivations for utilizing chatbots are efficiency. We therefore use time spent in conversation with chatbots to see whether it has an impact on the achieved customer index satisfaction and task completion.

Number of Messages per conversation: The number of messages per conversations was used as it measures the effort needed to resolve the issue. The amount of time between each user message might vary per person and it is possible that someone that write many messages quickly will encounter the same frustration as users who send the same number of messages over a longer time period.

Agent Available: The chatbot is always available, while agents can only be reached within the opening hours. This variable is used to control for whether a chatbot is able to forward users to a human agent. The reasoning behind this variable is that when the chatbot can’t receive help from humans it has to resolve more inquires. Some inquires requires transferring the user and can’t be resolved outside of opening hours.

Topic: Including what topic or intent the bot assumes that the user have is important for two reasons. From previous research, we know that chatbots often struggle with complex topics. Understanding which topics on average achieve a higher user rating, can reveal the capabilities of the chatbot. The second reason is that the variable controls for the *duration* and *number of messages.* Some topics requires more input and effort from the users, and this needs to be accounted for in the analysis.

Prediction description: The variable controls for the customers' ability to interact with the bot. Complex and badly written user messages can influence whether a chatbot can understand the message. Badly written messages are independent of conversation topic and is therefore included in our analysis.

User Sentiment: Sentiment measures the emotional state of a user when interacting with the chatbot. It can show how satisfied or dissatisfied they are with the customer service. As the interpretation of this variable is rather complex, it will be further discussed later in section *3.4.2 Sentiment analysis.*

3.4.1 Comparative analysis

As a first part of our own experiments, we conducted analysis of the chatbot performance and measured how it compares to a human agent. The first part of the comparison described how long conversations are in general with a chatbot compared to a human agent. In measuring how long the conversations where we had two available metrics: 1) time spent in chat, and 2) the number of messages between chatbot-customer and agent-customer. This research can help us to understand how quickly a chatbot resolves an inquiry compared to human agent. As mentioned in section *3.2.3 Repairing and merging datasets,* we removed empty conversations and messages in order to make this statistic more representable.

In the second part of the comparative analysis, we looked at the customer satisfaction index and the task completion evaluation criterion. In this type of analysis, we can figure out how the chatbot are rated compared to a human agent. From our literature review on chatbots, we found that in general customers prefer talking to humans and view them as superior. We expected to see the same results in customer ratings.

For our comparative analysis we use task completion instead of customer satisfaction index because it better represents if the customers are satisfied or dissatisfied. Furthermore, we assume that it is not possible even for human agents to achieve 100% task completion. This assumption is grounded in possibility of getting impossible questions that cannot be resolved. Another assumption for the comparative analysis is that the opening hours can affect the results for the chatbot when the possibility of seamless handover is absent. Customers appreciate the possibility of talking to a human agent in case their inquiry is not resolved through the chatbot.

3.4.2 Sentiment analysis

Evaluations of the chatbot usually happen only at the end of the chat, and most people don't answer the questionnaires. One method for measuring customer satisfaction in real time is sentiment analysis. The logic is that the general sentiment of a customer is reflected in the text that they produce.

Sentiment analysis is a branch of natural language processing that aims to extract the polarity of a document (Positive, Negative, Neutral) (Farzindar & Inkpen, 2020). The method of sentiment analysis that was selected, involves comparing the words to a list of terms that are marked as having either a positive or negative sentiment. The terms can be weighed on a scale on the intensity of the sentiment or simply be binary: positive/negative. The document is then evaluated by creating a combined score that is the document sentiment. This score can be calculated by simply summarising the terms or with a formula that can account for other things like document length. The accuracy of lexicon-based sentiment analysis is dependent on the quality of the lexicon. Some terms' sentiment differs based on the topic or domain it discusses, and accuracy can often be increased by adjusting the lexicon based on the domain of the documents. (Farzindar & Inkpen, 2020)

Feine et al. (2019) found that several sentiment analysis methods had a high correlation with the ratings given by human evaluation of the text and therefore was a good approximation. They concluded that sentiment analysis could help companies understand their customers emotions. They further suggested that the chatbot messages and handover to human agents could in part be guided by the sentiment of the customers to avoid poor customer experiences.

To perform the sentiment analysis, we had to create a dictionary of words with either positive or negative sentiment. We used a Norwegian lexicon created by Øvrelid et al. (2020) based on an English lexicon generated from customer reviews. We chose a binary scale for each term where it is either -1 for negative or +1 for positive, neutral terms have no value. The package *SentimentAnalysis* was used to calculate each message score. The custom lexicon was supplied to ensure that it only considered Norwegian words. The sentiment of each conversation was the sum of all sentiment score because that accounted for conversations were the user sent multiple negative or positive messages. (Proellochs & Feuerriegel, 2021)

3.4.3 Analysis of emotion

When discussing human vs robot communication, one topic that often is discussed is empathy or the ability to understand another human's emotions. Our interviews revealed that while the bot was allowed to handle many tasks, there were some tasks that was restricted to only human agents. Questions regarding inheritance and estate management were thought of as being of such a sensitive nature that it would be better handled by a human. To better understand how the users interacted with the bots and agents we measured the general sentiment and the amount of emotionally charged words. According to Luger & Sellen (2016), people use simpler language when chatting with bots. However, we did not know if users were on average more positive or negative when communicating with a chatbot.

3.4.4 Regression

One of the main goals of the thesis was to find which tasks chatbots can perform to a satisfactory level. We further wanted to understand which factors impacts performance and customer satisfaction. Regression analysis was therefor used to attempt to discover the true relationship between customer satisfaction and the chatbot metrics found in our chosen variables.

Regression analysis is used to find the relationship between the dependent variable (Y) and the independent variable (X). Mathematically this relationship between Y and X can be described by the function:

$$Y = \beta_0 + \beta_1 X$$

Where β_0 and β_1 are two unknown constants that represents the intercept and the slope of the model. These are the *coefficients* of the model, and they can be used to predict the value of Y based on values of X after the model is fitted to the data. The goal is to write a function where the predicted value of the dependent variable $\hat{y}$ equals the dependent variable Y for all values of X (x). The difference between $\hat{y}$ and Y is called the *residual* ϵ. This thesis uses ordinary least squares regression (OLS) to find this function. OLS or simple linear regression estimates the coefficients by finding the function that minimizes the sum of the squared residuals (RSS). (James et al., 2013)

When estimated, the function for $\hat{y}$ then becomes:

$$\hat{y} = \beta_0 + \beta_1 x$$

Where we expect that the true function of Y is:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Here β_0 or the *intercept* is the value of Y when X equals 0. β_1 is the *slope* which describes the average effect of how one unit increase in X, increases the value of Y. The error term ϵ describes the effect of all the other variables that affects Y, that is not accounted for in the model. It also accounts for measurement error and variations in X that is not linear. The error term is assumed to be independent of X. To measure how much of the variation in Y the model can explain, the R^2 statistic is calculated. To calculate the R^2, one need to calculate the total sum of squares (TTS). TTS is the sum of squared differences between each observation of the response variable and the mean of the response variable. The function for R^2 is then:

$$R^2 = \frac{TTS - RSS}{TSS}$$

R^2 is always a number between 0 and 1, where an R^2 of 0 mean that the model explains zero of the variation in Y. An R^2 of 1 implies that the model explains all of the variation in Y.

Multiple Regression

This thesis will use multiple regression to find causal relationships between chatbot metrics and customer satisfaction, while controlling for other factors. The function form and fitting of multiple regression is similar to simple regression. The multiple linear regression function for *n* independent variables is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_n X_n + \epsilon$$

Interpreting the coefficients for multiple regression is similar to regression with only one independent variable. The coefficient for the independent variables represents the average increase in Y based on a one unit increase in X with all other variables being equal. (James et al., 2013)

Logistic Regression

While a linear regression model works well with a continuous dependent variable, it is not well suited for classification purposes. This meant that a different model would be needed when modeling how the variables selected affect task completion, because it is a binary variable. Logistic regression was therefore selected. The benefit of logistic regression over other forms of classification is that it is both easy to fit to data, and the output is easy to interpret. Logistic regression shares many similarities with linear regression. "*Rather than modeling this response Y directly, logistic regression models the probability that Y belongs to a particular category*". (James et al., 2013)

Another difference between logistic and linear regression is the interpretation of the coefficients. In logistic regression the coefficients represent the average increase in the log odds of the observation belonging to a certain category or class. (James et al., 2013)

Model interpretation and evaluation

When evaluating the coefficients of regression models, there are two main statistics that are evaluated. That is the F statistic for the entire model, and the T statistic for each coefficient.

The F statistics measures whether there is a relationship between the response variable and any of the independent variables where the null hypothesis is that

$$\mathrm{H0} : \beta_1 = \beta_2 = \cdots = \beta_n = 0$$

Versus the alternative hypothesis:

$$Ha: At\ least\ one\ coefficents\ is\ non - zero$$

If the F statistic is too low to disprove the null hypothesis, then one can't say with certainty that any of the independent variables affect Y.

The T-statistic measures whether an independent variable affects the dependent variable when adjusting for the other variables. The null hypothesis for each coefficient is:

$$\mathrm{H0} : \beta_i = 0$$

And then alternative hypothesis is:

$$\mathrm{Ha} : \beta_i \neq 0$$

If the T statistic does not meet the threshold that is set for the model, then we can't say that an independent variable affects the dependent variable. It is common to set the threshold for both the F- and T-statistic to a level where there is a 5% or less chance of falsely rejecting the null hypothesis. (Wooldridge, 2009)

There are 5 assumptions that must hold for the regression model to provide an unbiased estimate of the coefficients. These are: 1) Linear in parameters, 2) random sampling, 3) no perfect collinearity, 4) zero conditional mean and 5) Homoskedasticity (Wooldridge, 2009). Unbiased in this case mean that the estimated coefficients are equal to the true model. More specifically an unbiased estimator equals what the actual effect of one variable has on the dependent variable in the real world. Likewise, a biased estimator is not equal to the real effect the variable has on the dependent variable.

Linear in parameters: The effect of one unit increase of X in Y is linear and the true model can be described as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_n X_n + \epsilon$$

Random sampling: The data used to fit the model is randomly sampled in a way that makes it representative of the entire population and therefore fit to estimate the coefficients. In section *3.3.4 Reliability*, we elaborated on our data sampling.

No perfect collinearity: The sample data cannot contain any independent variables that are either constant or have an exact linear relationship with each other.

Zero conditional mean: The error term has an expected value of 0 for any value of the independent variables.

Homoskedasticity: The variance of the error term is the same for all values in any of the independent variables.

Linearity, collinearity, zero conditional mean and homoskedasticity in our regression models, will be presented in *4.3 OLS Regression* and *4.4 Initial attempts* of fitting a multiple regression function to the dataset showed results that indicated that the model contained bias. This introduced uncertainty of whether the results could be interpreted causally. There were especially issues concerning the zero conditional mean assumption as the residuals of the model where not normally distributed and the expected mean was not zero. As explain previously in the thesis, the CSI scores are influenced by factors that is not in the dataset. Most of the ratings were also either the 1 or 6 which made the dataset unbalanced. We expect that these factors were some of the reasons as to why the residuals were not normally distributed. The results from the OLS regression fits well with what prior research states. We expected that the number of messages would negatively affect satisfaction and that the sentiment score would have a positive correlation with user satisfaction. The magnitude of the coefficients is however likely biased. This means that the exact effect that is described in the regression summary in *Table 5* is most likely not the same as the effect in the true model. The task completion variable was therefore a better representation of the population, and we therefore considered it as the better variable for further analysis. As the task completion is a binary variable it already accounts for the customer rating being either 1 or 6. .

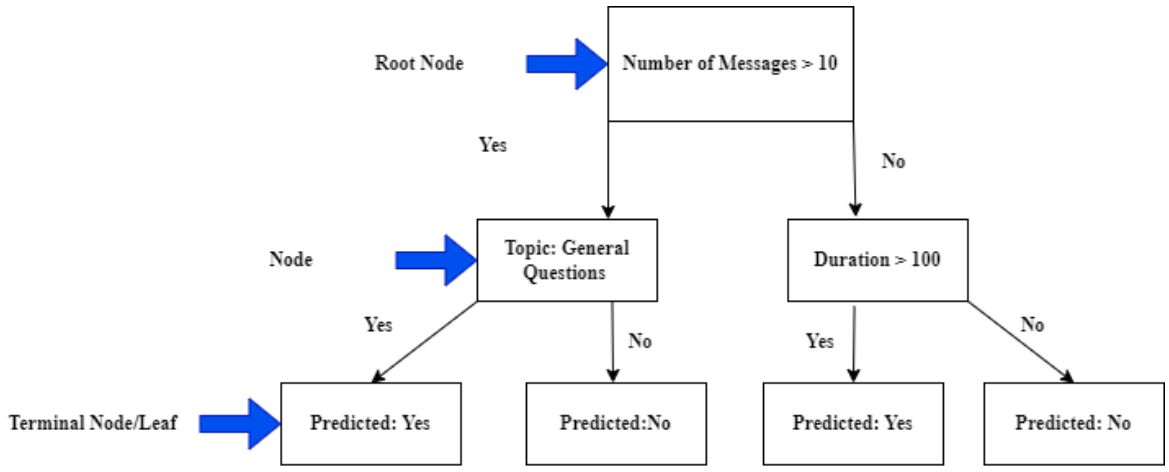
3.4.5 Random forest

Random forest is an ensemble learning method that was created by Tin Kam Ho in 1995. The method involves creating many decision trees based on a randomly selected subset of variables for each tree. Decision trees for classification separates the data based a categorical variable where the goal is to maximise the proportion of a single category in each leaf. A decision tree follows a flowchart-like structure where the data is split into subsets based on predefined rules. The variables that best divides the data into separate categories are chosen first as the root of the ‘tree’. Subsequent variables are chosen based on their discriminative power until a limit of nodes are reached or the data is sorted. The end nodes represent the decision of which class the observation belongs to. An example of a decision tree can be found in *Figure 2*. Decision

trees are however prone to overfitting, meaning that they often perform poorly on data that it has not been trained on. The random forest algorithm solves this issue by using a random subsection of the variables and creating many trees. If the model is created for classification purposes, then the class of an object is decided by a majority vote from all the created decision trees. (James et al., 2013)

The benefit of random forest is that it is less susceptible to overfitting while often performing well with classification. Randomly selecting variables decreases the variance as it avoids using the strongest predictor as the root of every decision tree. Random forest is also able to model interactions between independent variables and non-linear relationships with the dependent variable. The random forest model does not provide coefficients similar to what regression analysis does. It can however provide measures of which variable is most influential when dividing the data into specific classes (James et al., 2013). For the analysis in R, the packages *caret* (Kuhn, 2008) and *ranger* (Wright & Ziegler, 2017) were used. *Ranger* is a fast implantation of the random forest algorithm in R and was chosen for its speed and its compatibility with the *caret* package. *Caret* was used for training and tuning of the model.

Figure 2: Decision tree



4. Experiments

This section presents and describes the results of our experiments. The order of the experiments will follow the same sequence as illustrated in the *Methods* section. The impact of the results is discussed in section *Discussion*.

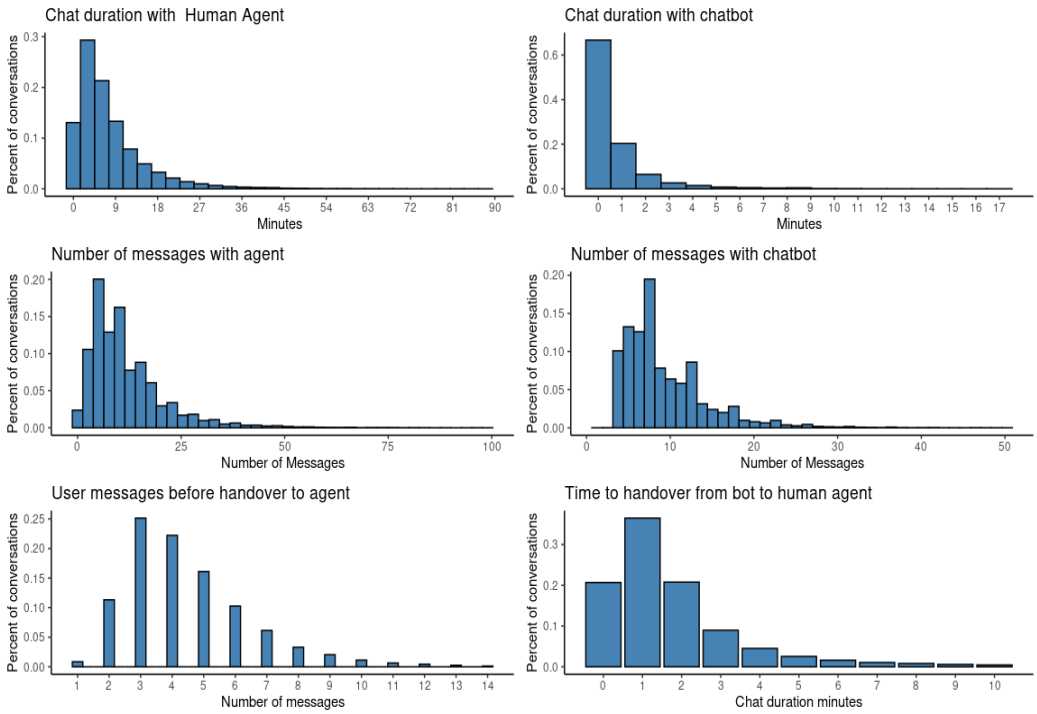
4.1 Comparative analysis

The comparative analysis consists of 2 sub-parts: 1) Length of conversation and 2) Task completion. Length of conversation investigates the time spent in chat with a chatbot vs a human agent. The task completion experiment investigates customer satisfaction for chatbots vs human agents.

4.1.1 Length of conversation

In order to generate the plots in *Figure 3*: Comparing length of conversation between chatbot and human agent.*Figure 3*: Comparing length of conversation between chatbot and human agent., the data was split into 3 parts. The first part contained conversations were the chatbot handled the entire conversation. The second part included user and bot messages, and the last part only included user-human messages.

Figure 3*: Comparing length of conversation between chatbot and human agent.*



Common for all of the graphical presentations in *Figure 3*: Comparing length of conversation between chatbot and human agent. are the y-axis that represents the percentage distribution of conversation that is included in each range. The x-axis show shows the duration of the conversation in minutes or number of messages.

Figure 3: Comparing length of conversation between chatbot and human agent. shows the comparison between human agents and chatbots in terms of time and number of messages. The time that customers spent with chatbots and human agents are presented in the top left and top right corners of *Figure 3*: Comparing length of conversation between chatbot and human agent.. From these plots we can see that conversations are significantly shorter with chatbots than a human agent. For chatbots, the conversations vary between 0-10 minutes where most conversations are under 3 minutes. Comparatively for human agents, the conversations last between 0-36 minutes. The majority of the human agent conversations are under 18 minutes.

When comparing the number of messages written to chatbots vs human agent (middle left and middle right graphs in *Figure 3:* Comparing length of conversation between chatbot and human agent.), we find a similar trend. Customers-chatbot conversations are between 3-30 messages. Most chats contain between 3-10 messages. Comparatively, the customer- agent conversations vary between 0-40 messages. The majority of customer-agent conversations are between 2-20 messages. In general, the trend shows that customers-agent conversations are longer in both time and number of messages compared when talking to a chatbot.

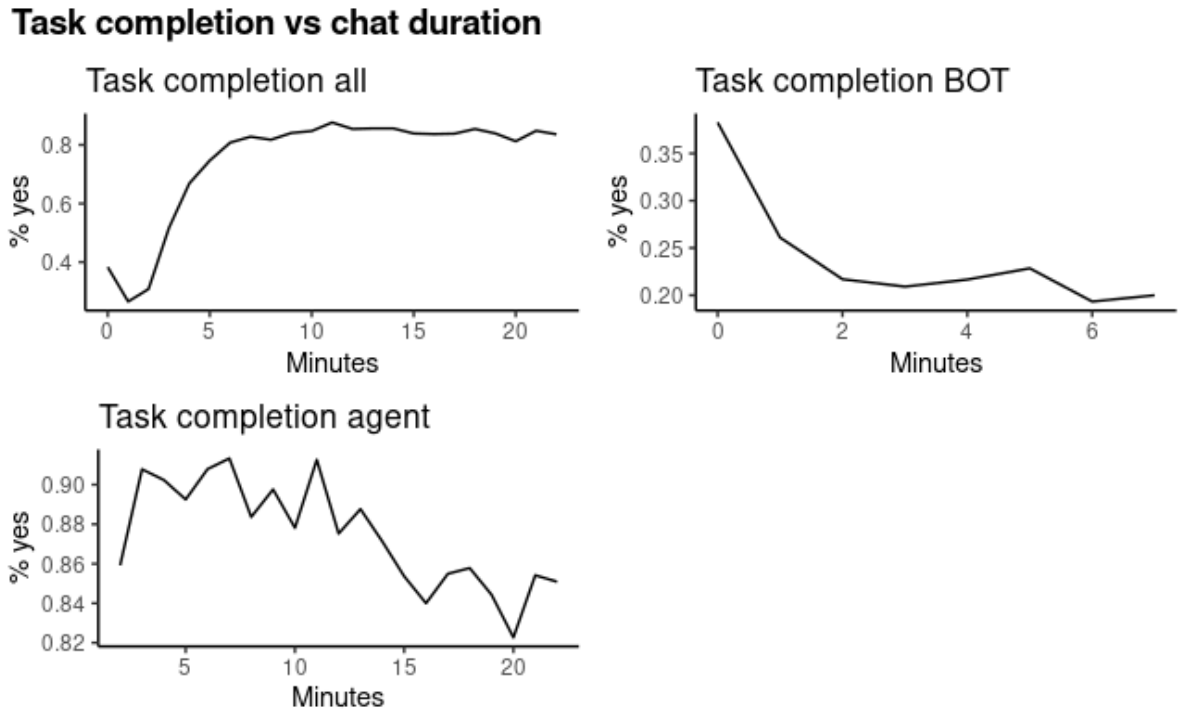
The bottom two graphs depicted in *Figure 3*: Comparing length of conversation between chatbot and human agent., show how long the conversations are before the chatbots transfer the user to a human agent. This indicates how long users are willing to try before being transferred. Most customers spend less than 3 minutes and write less than 5 messages before being transferred to human agents.

4.1.2 Task completion

The datasets used to produce the plots in *Figure 4*: Comparing task completion variation over time for chatbots and human agents. and *Figure 5*: Comparing task completion during time of day for chatbots and human agents., are the same parts as described in the previous plot.

Task completion vs chat duration

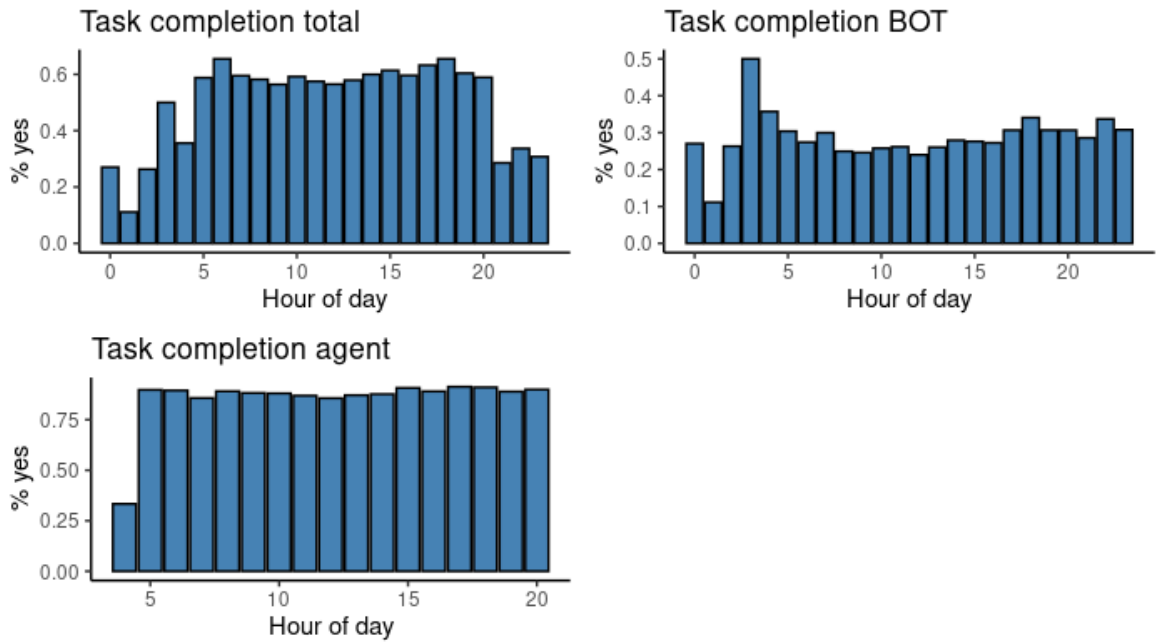
Figure 4: Comparing task completion variation over time for chatbots and human agents.



For the second graphical presentation we compare the variation in task completion for human agents and chatbot. The duration of the chat is represented by the x-axis, and the % of yes to the whether their inquiry was resolved is represented on the y-axis. In the top-left figure we include both chatbot and agent conversations. We can see a fairly quick increase in task completion for the first 5-6 minutes before it levels out. If we isolate the conversations between chatbot-human, we can see that the task completion decreases with 50% if the chatbot does not solve the inquiry within 2 minutes. Furthermore, the chatbot task completion rate over time seems to vary between 20 and 40 %. Comparatively, human agents have a task completion rate between 86 and 95 %. Task completion between human agent and customers also decline over time, but not as drastically as the chatbot-customer conversations.

Task completion vs hour of day:

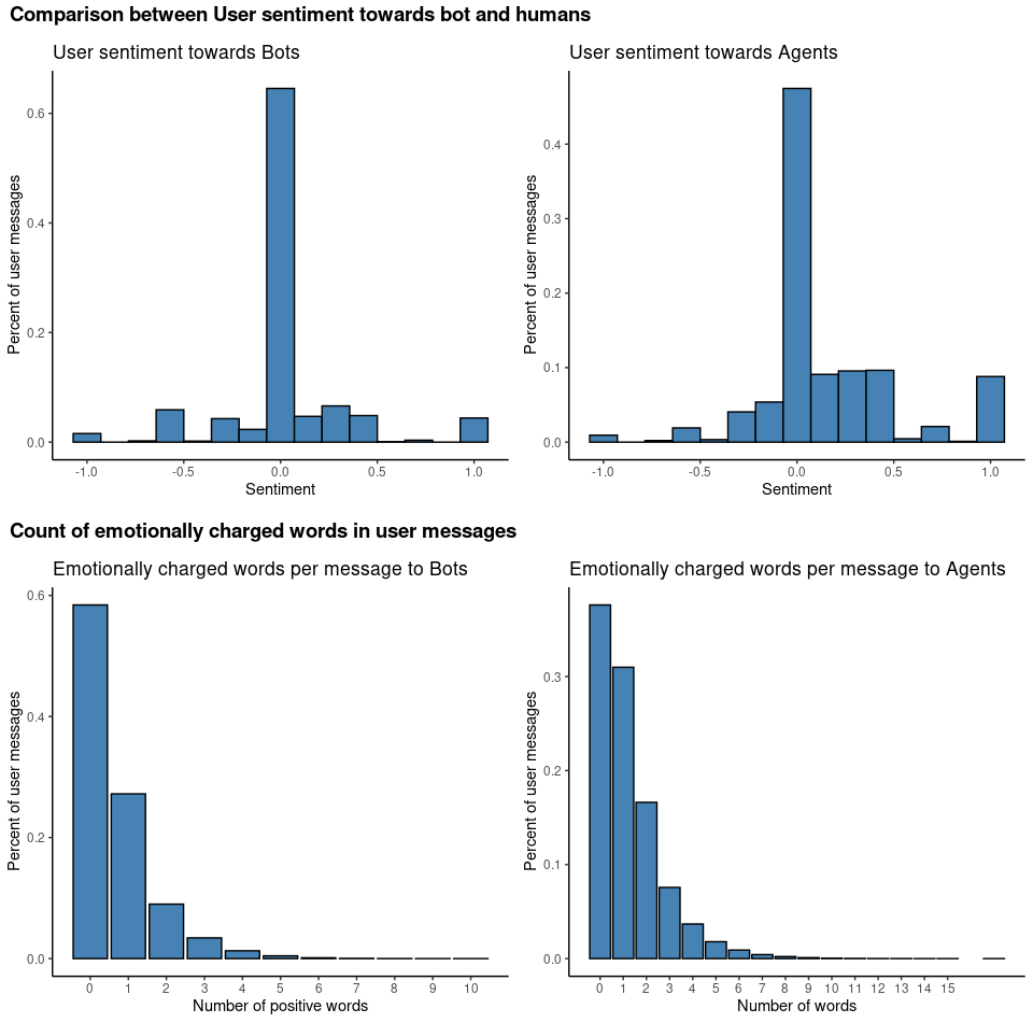
Figure 5: Comparing task completion during time of day for chatbots and human agents.



For the next part of the comparative analysis between chatbot and agent performance, we investigate the task completion by hour of day. The x-axis in all of the figures is showing time of day, and the y-axis shows the percentage of customers who got their inquiry resolved. We can see that human agents perform on a very stable level of approximately 85% task completion. The task completion of chatbots is varying more but have an average task completion rate of approximately 30%. From the task completion total statistic, we see that customer ratings are worse outside of the opening hours of the human agents.

4.2 Sentiment Analysis

Figure 6: Comparing sentiment of chatbots vs human agents.



The average sentiment-value for messages to human agents were 0.153 with a standard deviation of 0.361. For bots the average sentiment score was 0.0333 with a standard deviation of 0.316. This indicates that customers use a neutral language towards chatbot and human agents. The users sent on average more positively charged messages to human agents than to bots. Users also on average send more emotionally charged words per message to agents. Messages to agents are on average more likely to include words with a positive or negative sentiment and are more likely to contain multiple emotionally charged words.

4.3 OLS Regression

Before the models could be fitted to the data, some adjustments had to be made to the topic variable. Some categories of the topic variable had too few observations because the topic variable consisted of every variable the bot predicted for the conversation. The performance of initial models with all topic variants were therefore poor. All combinations of topics that had less observations than the top 20 topics were therefor set to "Other".

The intension of conducting the OLS regression analysis was to find factors that has a causal relationship between the customer satisfaction rating and the chatbot metrics. Conversations with human agents was therefore excluded from the analysis. The data was summarised to one row per conversation as the intension of the analysis was to explain the rating of the entire conversation. Chat containing other languages were excluded as the sentiment lexicon that were used included only Norwegian words. The task completion variable was excluded as the rating of chats happen after the chat is completed. We also expected that the task completion variable would have a high correlation with the customer satisfaction score. The following OLS Model was therefore fitted with all the bot only conversations, where the user left a rating of the conversation.

$$KTI = \beta_0 + \beta_1 Messages + \beta_2 Chat\ Duration + \beta_3 Agent\ Available + \beta_4 Prediction + \beta_5 Topics + e$$

1. Messages is the number of that were sent.
2. Time is the Chat duration from the first messages sent to the last.
3. Agent Available: did the chat happen within customer service opening hours.
4. Prediction is how certain the bot is of the topic.
5. Topics is what topic(s) the bot predicted for the conversation, there can be multiple topics per conversation. For this analysis combinations of two or more topics are considered a separate topic independent of its respective parts.

There are three main points of interest that we can observe from the regression results. 1) The coefficient for the number of messages is significant and negative, even when we control for the topic, time of day, bot certainty, sentiment, and chat duration. This indicates that the number of messages sent has a negative correlation with customer rating .2) the user's sentiment coefficient is positive and significant and 3) Some of the topics are also significant which indicates that some topics on average earn a higher or lower rating than the base topic(stocks). All topics that were a combination of general questions and other topics where significant and positive, suggesting that they on average get higher ratings than the base topic. The coefficient for the topic: asking for human representative was significant and negative indicating that these conversations on average earn a lower rating. The chat duration coefficient was not significant, and we can therefore not claim that it affects the customer satisfaction score. The coefficient for whether an agent is available is negative and significant. This indicates that the mean score of the chatbot is lower within the opening hours of the customer service department. The models explanatory power (R^2) was 9.6 %, but when we adjust it for the number of independent variables in the function, it decreased to 9.0%.

Initial attempts of fitting a multiple regression function to the dataset showed results that indicated that the model contained bias. This introduced uncertainty of whether the results could be interpreted causally. There were especially issues concerning the zero conditional mean assumption as the residuals of the model where not normally distributed and the expected mean was not zero. As explain previously in the thesis, the CSI scores are influenced by factors that is not in the dataset. Most of the ratings were also either the 1 or 6 which made the dataset unbalanced. We expect that these factors were some of the reasons as to why the residuals were not normally distributed. The results from the OLS regression fits well with what prior research states. We expected that the number of messages would negatively affect satisfaction and that the sentiment score would have a positive correlation with user satisfaction. The magnitude of the coefficients is however likely biased. This means that the exact effect that is described in the regression summary in *Table 5* is most likely not the same as the effect in the true model. The task completion variable was therefore a better representation of the population, and we therefore considered it as the better variable for further analysis. As the task completion is a binary variable it already accounts for the customer rating being either 1 or 6.

4.4 Logistic Regression

The intention of performing logistic regression on this dataset was to understand how the variables that we mapped affected the likelihood of a user reporting that they got an answer to their question. We considered it possible that a customer could have a good interaction with a chatbot while not receiving an answer to their question and that the inverse of that could also be possible.

The logistic regression was fitted on the same dataset as the OLS regression with the only difference being that the response variable was task completion, and that the customer satisfaction score was excluded.

$$Task\ Completion = \beta_0 + \beta_1 Messages + \beta_2 Chat\ Duration + \beta_3 Agent\ Available + \beta_4 Prediction + \beta_5 Topics + e$$

From the regression output we can observe that the coefficients are similar to the OLS Regression. The coefficient for the number of messages per conversation is negative and significant. This indicates that there is a negative correlation between the number of messages and the reported task completion. User sentiment has a positive coefficient that is significant which indicates that the customers who use more positively charged language has on average higher task completion. The coefficient for whether the chat is within customer service opening hours is not significant at the 5% but it is close with a p-value of 8.1%.

A short form of both the OLS and logistic regression table is listed in Table 5. The full tables are listed in the *Complete regression table* in the appendix.

Table 5: OLS and logistic regression models

OLS And Logistic Regression Models Complete model in appendix						
	OLS Model			Logistic Regression		
	Coef.	**Std.Error**	**P.Value**	**Coef**	**Std.Error**	**P.Value**
(Intercept)	2.229**	0.752	0.003	−0.695	0.924	0.452
Messages	−0.050***	0.005	0.000	−0.077***	0.008	0.000
Chat duration (Seconds)	0.000	0.000	0.473	0.000	0.000	0.748
User's sentiment[1]	0.227***	0.042	0.000	0.287***	0.054	0.000
Agent available: Yes	−0.460**	0.152	0.003	−0.292+	0.168	0.081
Num.Obs.	0.096			Num.Obs	4498	
R2	0.090					
R2 Adj.	18891.7					
AIC	19090.5			AIC	5368.7	
BIC	−9414.873			BIC	5561.0	
Log.Lik.	16.408			Log.Lik.	−2654.348	
F	0.096					
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001						
1:Chat duration is measured in seconds from the first message to the last						

4.5 Random Forest

The random forest model was fitted with the same data as the logistic regression model with the same model formula. Before the model could be fitted, the data was split into training and testing subsections. The model was first tuned to find the optimal amount of randomly drawn variables. Cross-validation was used increase the certainty of the final model being the optimal model. The model was run 10 times with a randomly drawn subsection of the training data for

each amount of randomly drawn variable. The optimal number of variables was 3, which indicates that there is a significant amount of interaction between the variables. The variable importance plot show which variables has the greatest ability to divide observations into task completed and task not completed. The variables in the importance plot are similar to the significant variables from the regression models. The largest difference is that the *chat duration* variable which was not significant in the regression analysis. A reason for these results could be that *chat duration* is a good approximation for the *number of messages* and serves the same purpose in decision trees where the *number of messages* are not included.

Figure 7: Random forest variable importance plot

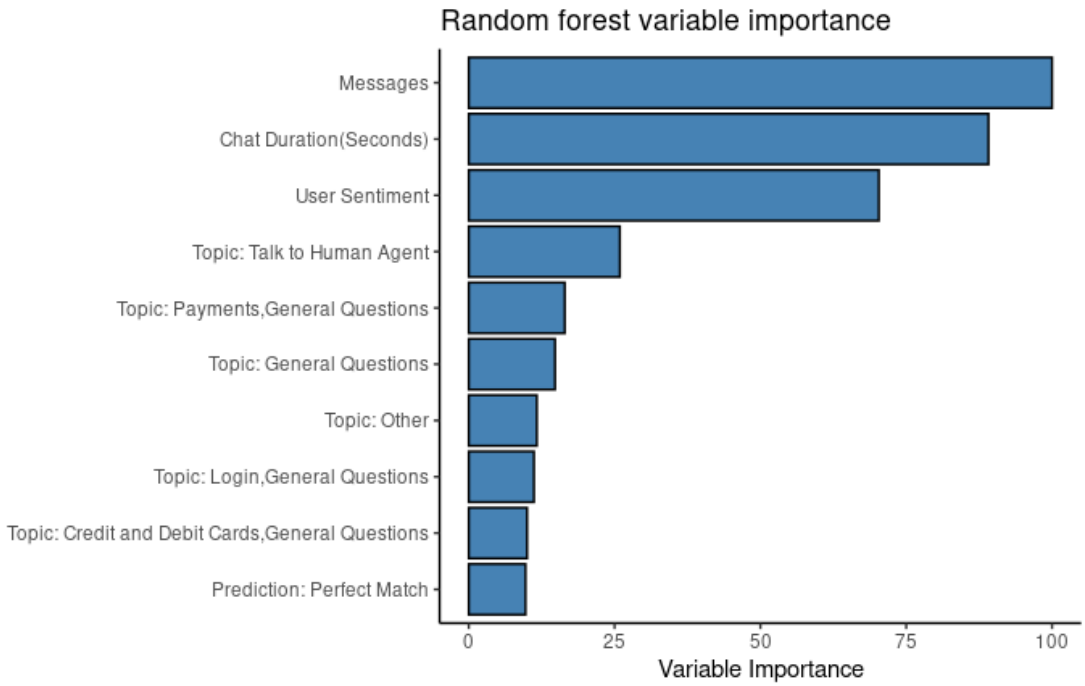


Table 6: Random forest predictions vs actual user rating

Random forrest prediction of task resultion		
	No	**Yes**
Predicted: No	1408	45
Predicted: Yes	654	142

The tuned model achieved a balanced accuracy of 70% when predicting task completion on the testing dataset. The model was on average equally good at predicting when a user responded “yes” to the question concerning task completion, as when the user responded with “no”.

5. Discussion

In this section we discuss the different drawbacks and advantages of utilizing chatbots for customer service. The discussion includes arguments from the literature review, our experiments, interviews, and the opinions from the authors of this thesis. The discussion structure mimics the structure of *2.2 Chatbots for customer service*. The intention is to discuss each weakness and strength separately. First, we discuss how automation of customer service is a key contributor of why chatbots are used for customer service. Followingly, we will discuss how efficiency and productivity are making chatbots superior to human agents. Next, we include the managers role and how strategy can distinguish systems that are working efficiently from the ones where chatbot becomes a liability. From the manager role we continue with chatbots ability to provide better customer service, expectations, and human traits that all contribute to answering our research question. Before concluding, we answer what chatbots today are missing in order to replace human agents in customer service. This section ends with a conclusion that summarized our findings that answers the research question.

Automation, efficiency, and productivity are repeatedly mentioned as the main motivation of utilizing chatbots for customer service. This applies to both the users and the businesses. From our literature review, we found that that company's motivation of using chatbots for customer service are driven by automation purposes. By automating their customer service, businesses can reduce the need of customer service personnel and in turn reduce cost. The chatbot might require restructuring of the customer service department and an initial investment. However, over time as the chatbot is customized for the desired purpose, scalability could reduce long term salary costs. Reported numbers from Sbanken in 2020 indicated that 40% of their customer inquiries are handled by their chatbot. One of our interview objects can confirm that their chatbot handles 60% of all incoming inquires (Informant #3, 2021). There is no doubt that utilizing chatbots are less costly than their human counterpart. The chatbot can also respond instantaneously, which makes them superior as long as they give the correct answers. With the reported numbers collected from multiple sources, we argue that in terms automation and cost chatbots superior are to human customer service.

Our experiments show that chatbots resolve their conversations much faster than humans, and with less messages. Over 80% of the chats where the bot handled the entire conversation were resolved within 2 minutes. In most of the cases were the chatbot could not answer the user's question, the conversations were forwarded after less than 4 minutes. Our research show that DNB reaches their goal of automating 60% of user conversations. This shows the potential of the technology as their chatbot handles the majority of their conversations, that in turn leads to significant efficiency gains. The waiting time and conversations time is also reduced which increase customer satisfaction. The fact that chatbot resolve inquiries faster than human agents, make them superior in terms of speed.

While chatbots are faster, the comparison is not entirely fair. Human agents handle the tasks that are more complex, and these tasks require more time and messages to resolve. Furthermore, chatbot task completion drops by 50% after the first 2 minutes. This indicates that if the inquiries are not quickly resolved, the customers will be less satisfied and more prone to leaving the conversation. This view was further strengthened by both the regression and random forest models. Our experiments give an indication that increasing the time and messages reduces customer satisfaction. The duration of the conversation, and the number of messages were strong predictors of task success. The chatbots inability to handle more complex conversations that contains more messages, is a drawback of current chatbots for customer service.

Our findings suggest that the Chatbot is less capable in terms of task completion. Of all conversations that were not forwarded to a human agent only 27% were marked as completed by the users. Human Agents task completion were close to 90%. These numbers were based on user rating which only covered 10% of the dataset. Based on previous research and our conducted interviews, we suspect that negative ratings are more likely to appear in these ratings and could introduce a negative bias. Breakdown of low score conversations found in *Table 4* showed that only half of the low scores were caused by the chatbot not being able to help the user. There are therefore reasons to assume the bot task completion is higher, but there is still a significant gap between the chatbot and human agents. There are also tasks which the chatbot is restricted from handling, as company policy or legal requirement restricts the chatbot. Our interviews revealed that some tasks in the financial industry requires the customer service agents to be certified before offering some forms of financial advice. Regulations concerning user data also restricts the chatbots ability to perform some tasks. Finally, there are some topics that DNB wants humans to handle as they are sensitive of nature, or they want the

customer to talk to a human to ensure that they get the appropriate customer service experience (Informant #3, 2021). We argue that chatbots are more effective than what is shown through the data collected on customer ratings. Companies also choose specific topics that should not be handles by a chatbot due to legal regulations or internal strategy. Therefore, we argue that chatbots are more effective than depicted by customer ratings. However, these limitations cannot explain the entire difference between human agent and chatbot capabilities.

In order for chatbots to perform better than human customer service, we argue that good management and step-wise transition determines success. Realistic expectations and knowledge of chatbot capabilities could determine the outcome of chatbot investments. From our conducted interviews, we learned that a good strategy for successful implementation is to start small and work on small improvements at a time. Making sure that new features work before implementing it, reduces customer frustration and is more feasible in terms of funding. (Informant #3, 2021). The argument of not being too ambitious at the start is backed up by research. From section 2.2.1 Business and manager perspective, we learn that setting an appropriate scope for what the chatbot should do is important. We also learn that chatbots are not a one-size fits all type of product, and therefore has to be tailored to each company before use. To summarize, we see the manager role and strategy of implementation as essential for mitigating drawbacks of chatbots. By implementing a full-size all-in-one solution without testing would certainly lead to a lot of problems for the chatbot and resulted in frustrated customers who would prefer human agents instead.

As mentioned, chatbots do reduce the need for customer service personnel. However, there are new jobs created upon implementation of chatbots. We argue that the AI-trainers job is a key contributor to better customer service through chatbots. The type of chatbots used for customer service today are not able to generate human language. As a result of that, chatbots selects the best response out of a database of possible responses. The responses are written by AI-trainers, and the appropriateness of the response is therefore reflected in their competence. From section *2.2.1,* we learn that AI-trainers competence depends on prerequisites like writing skills, customer service experience and analytical abilities. A good response should give the customer the necessary information to either 1) solve their problem instantaneously, 2) guide the customer to where the necessary information is or 3) seamlessly transfer the inquiry to a human agent. The response should also be well written and not leave any room for confusion. If these conditions are met, the chatbot would certainly outperform their human counterpart for topics within the scope of the bot and avoid most of the frustrations caused by inappropriate

responses. The chatbots also have the advantage of 24/7 service with very quick response times.

Good management and competent AI-trainers is the foundation for chatbot success. If the scope of the bot is well thought out, and responses are well written it could result in a better customer service experience. Unlike human agents, chatbots can answer multiple customers simultaneously. Chatbots requires no waiting, and therefore performs better given that the customer inquiry is resolved. The counterargument is that chatbots percentagewise solves less inquires than human agents. But for the questions that a chatbot can solve, they are much faster. Our regression and random forest analysis indicate that customers are significantly happier when asking general questions. This fairs well with the impression we got from the interviews, where we are told that the chatbot is intended to answer uncomplex and repetitive questions (Informant #1 and #2, 2021). We conclude that chatbots are outperforming human agent because of their scalability and speed. However, this only applies to uncomplicated questions that is within the scope of the chatbot.

Chatbots perform better than human customer service given that the questions asked are within the scope of the bot. Despite of this, customers still see chatbots as inferior to human customer service regardless of the situation. This can be explained by previous bad experiences, or the fact that chatbot technology is fairly new. From our interviews, we found that customers give chatbots lower rating despite having their problems resolved. The perception of chatbots has changed for the better but is not yet on an equal level of human customer service (Informant #3, 2021). This is also confirmed by previous literature. We elaborated on chatbot disclosure in *section 2.2.2*. We found that chatbots for sales are 70% less effective when the customer know that they are talking to a chatbot instead of a human sales representative. The perception of chatbots and that they perceived are less capable is a clear drawback of chatbots for customer service. This reduces their success in aiding customers at the same level as their human counterpart.

There are measures to deal with the negative perception of chatbots. Even though disclosure of chatbots is negative, disclosure of their abilities has a positive effect on customer satisfactions. If chatbots are transparent in what they can assist with, we believe that customers would utilize chatbots to a greater extent than today. From research referred to in *section 2.2.2*, we also find that changing the assumptions made by a chatbot could be useful for customers in situations where the chatbot prediction are not correct. We further argue that sensible

customer expectations could make chatbots more effective. Too high expectations in the abilities of chatbots leads to frustration and low customer ratings. Alternatively, if customers have experience and knows the capabilities of the chatbot they are talking to, they can easily decide whether the chatbot is sufficient to answer their question. Furthermore, research suggests that previous good experiences make customers more likely to resolve their questions through chatbots in the future. We conclude that transparency of chatbot abilities and giving customers the ability to change the assumptions made by the chatbot, can be very useful in mitigating the negative perception drawbacks. Furthermore, realistic expectations and experience with chatbots are also a significant contributor to making chatbots more successful.

Research of people's perception and experiences of interacting with chatbots revealed that chatbots are often perceived as cold and emotionless. This negatively impacted the customer experience and decreased the user motivation for using chatbots. Previous interview studies often mentioned that people want the chatbots to present a human persona and exhibit human traits like empathy and humor. Through our conducted interviews, it was mentioned that DNB have over time changed the language of the bot to become less formal. They further claimed that they at first had a perception of chatbots needing to be "impersonal and generic", but this view has changed over time (Informant #1 & #2, 2021). Another mentioned benefit of using less formal language is that it is more accessible for people who are not familiar with industry specific terms and vocabulary. The chatbots robotic presentation is however not without benefits as they are also seen as less judgmental than humans. Chatbots could therefore encourage more users to ask questions that they perceive to be stupid or embarrassing. Lack of emotional intelligence is a drawback of chatbots, but it could also be an advantage because it allows for customers to solve inquiries that they perceive as embarrassing.

Similar arguments can be made about the persona of the chatbots. Repeated studies have found a preference for human persona where the chatbot have a 'human' name and avatar. Such a persona is also relatively easy to implement compared to imitating human conversations and empathy. These factors could explain why most customer service chatbots present human like avatars or have human names. Increasing the user's motivation for using chatbot is beneficial for companies, but there are also potential consequences that could impact user behavior when interacting with 'human' chatbots, as it could decrease the user's willingness to disclose personal or sensitive details. Studies of commercial chatbots found that customer service chatbot avatars are mainly female, but the gender proportions differ based on the industry and the role of the chatbot. Such gender disparities raise ethical concerns for managers who are

implementing chatbots. The DNB chatbot is not gendered in part because of such concerns (Informant #3, 2021). Lack of human traits in chatbots could have negative consequences for user motivation to contact customer service. Measures that intend to humanize chatbots risks introducing some of societies prejudice to the bot which could cause negative consequences for companies.

While a 'human-like' avatar can motivate people to use chatbots, it would still need empathy and emotional intelligence if the quality of the conversation is to be equal of user-human conversations. Our experiments reveal that conversations with bots are on average less positive than conversations with humans. The amount of emotionally charged words are also lower, indicating that the language of the user differs based on who they are talking to. Current chatbots with pre-generated responses where a response is selected based on pattern recognition might lack the ability to respond to complex human emotions. The consequences of lacking empathy could include lower user satisfaction and task completion. The regression analysis indicated that the sentiment could correlate with user ratings. Establishing causality will however require further study as there is uncertainty whether is it the chatbot malfunctioning that causes negative user sentiment, or if it is the users negative emotional state that causes chatbot malfunction. From the literature review we know that chatbot performance is dependent on the users writing in a manner that the chatbot understand. The chatbots will therefore often struggle with understanding dissatisfied user messages. In that case, transferring customers with a negative sentiment to a human agent would improve task resolution. It is also possible that the user's sentiment is negative because of the quality if the conversation is poor. In that case the causality is reversed and improving the chatbot in general might be the better choice than to transfer users based on sentiment scores.

As stated previously in the thesis, chatbots in customer service are not yet capable of replacing humans. The technology requires further improvements before bots can handle all request without any human backup. While implementing and improving current technology would increase the amount of traffic the chatbots could handle. It would most likely not be enough to handle every conversation successfully. If the bot is trained on data to recognize intents, then a key limitation would be topics where there is little data, or each conversation differs greatly between each other in terms of language and sentence structure. Chatbots not being able to produce its own language and relying on prewritten responses also limits its ability to handle infrequent topics as the time investment of writing a response to every possible question would be higher than the efficiency gains. People have also shown a substantial

preference for humans in customer service, and they might react negatively if the possibility of contacting a human is removed.

6. Conclusion manegerial implications

In this thesis we have investigated chatbots and their abilities within customer service. Our intention has been to clarify misconceptions about chatbots and give a realistic insight in the limitations of chatbots. Specifically, we aimed to answer our research question:

What are the drawbacks of utilizing a chatbot for customer service, and for which tasks are chatbots outperforming human agents?

In order to answer that question, we have elaborated on how chatbot technology has evolved through time and specified what it can be defined as today. Followingly, previous research was used to partly answer our research question. In the absence of quantitative research, we conducted our own experiments on chatbot data and conducted interviews in order to confirm or disclaim established truths. Our research can be summarized into the following advantages and drawbacks from utilizing chatbots for customer service.

Chatbots are outperforming human agents in terms of costs. The scalability and speed of chatbots make them less costly and more efficient than their human counterpart. Chatbots are also solving problems at a faster pace than human agents. This was true for both the number of messages and time used to solve customer inquiries. With correct management, expertise, and scope of chatbot implementation, we see chatbots as superior to human agents when answering general and uncomplicated questions. Motivation and trust are also an important part in making chatbots more desirable. We found that if companies are transparent with their customers in the capabilities of chatbots, they are more willing to utilize chatbots when their questions are within the scope of the bot.

The drawbacks of utilizing chatbots for customer service revolves around the negative perception and previous bad experiences with chatbots. Unrealistic expectations leave customers frustrated and make chatbots less effective. Bad management, unrealistic scope and lack of knowledge would lead to overoptimistic chatbot implementation. This is a costly option and would most likely leave customers unsatisfied. Chatbots are in general bad at interpreting emotional state of the user. This is a features that customers prefer, and therefore a clear drawback of customer service chatbots. As a last concluding remark, we see natural language generation as a potential gamechanger. If customer service chatbots had the ability to generate their own sentences, it would make personalized customer service achievable

through chatbots. However, this technology is not yet mature enough for business implementation. As of yet chatbots for customer service use pre-defined responses that makes it hard to give an appropriate response for every situation.

Followingly, we summarize what our findings mean in terms of actions for managers. Our findings suggest that companies and their managers should be well educated in terms of chatbot capabilities before deciding if chatbots are beneficial to their company. The strategy of implementation should be to start small and expand the functionality gradually. This strategy has been proven to cost less and give better customer service. Managers should also understand that the current chatbots do not replace human customer service, but rather compliments it. Chatbots should be used for low-complexity inquiries that in time will reduce the amount of traffic for customer service centres. The emotional and human-like aspect of customer service should not be underestimated, as it is proven to be essential for the motivation of customers to use chatbots. Human-like features should be included in order to encourage chatbot use. However, the absence these features in situations where the inquiry requires financial information could be beneficial. We suggest a hybrid solution that account for different situations. Lastly, we suggest that companies are honest with their customers, and reveal the capabilities of chatbots within the conversation. This could improve customer expectations and make chatbots more efficient.

7. Limitations and future research

While the main objective for this thesis was to evaluate chatbots on a holistic level some limitations of the scope were imposed by lack of additional data and computational power. The thesis also focuses on chatbots for customer service, and we expect that most of the key findings can be transferred to other forms of bots and personal assistants. There will however be differences that will extend past the scope of this thesis. The analysis was based on data from a single source. Expanding the analysis to other industry sectors and firms would grant more insight in how chatbot performance is influenced by the industry it operates in. Worth mentioning is the absence of the variables age, gender, location, or any personal or demographic features of the customers. While research on motivation for using chatbots show differences based on demographic traits, reached into how it affects chatbots success is not sufficiently explored. Further exploration of text analysis and the contents of the customers messages would also provide more depth to chatbot evaluation frameworks, that otherwise focuses mainly on macro metrics. The field of chatbots is rapidly evolving and this thesis makes use of several papers that has implemented new features that show promising results. However not all these features have been implemented on a large scale. In the case of disclosure of financial details to ‘human’ or robotic chatbots, we would encourage future research to include further research on an actual chatbot implementation. This would further strengthen the claim made in the paper.

Bibliography

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications, 2*, 3. Retrieved from https://doi.org/10.1016/j.mlwa.2020.100006

Agarwal, A., Maiya, S., & Aggarwal, S. (2021). *Evaluating Empathetic Chatbots in Customer Service Settings.* Cornell University. Retrieved from https://arxiv.org/abs/2101.01334

Brandtzaeg, P. B., & Følstad, A. (2019). Why people use chatbots. *International Conference on Internet Science*, 377-392.

Budulan, S. (2018, 03 26). *Chatbot Categories and Their Limitations*. Retrieved from Dzone: https://dzone.com/articles/chatbots-categories-and-their-limitations-1

CGS. (2018). *Chatbots deliver speed, but consumers want humans. Are we moving too quickly to automation?* CGS. Retrieved from https://www.cgsinc.com/en/resources/CGS-2018-ERP-Report

CGS. (2019). *2019 CGS Customer Service Chatbots & Channels Survey,*. Retrieved from https://www.cgsinc.com/en/resources/2019-CGS-Customer-Service-Chatbots-Channels-Survey

Fainchtein, L. (2020, 06 28). *Cloudboost*. Retrieved from Generative vs Retrieval Based Chatbots: A Quick Guide: https://blog.cloudboost.io/generative-vs-retrieval-based-chatbots-a-quick-guide-8d19edb1d645

Farzindar, A. A., & Inkpen, D. (2020). *Natural Language Processing for Social Media: Third Edition* (3 ed.). Morgan & Claypool Publishers.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). *Gender Bias in Chatbot Design.* Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, Institute of Information Systems and Marketing (IISM). doi:https://doi.org/10.1007/978-3-030-39540-7_6

Følstad, A., & Skjuve, M. (2019). Chatbots for Customer Service: User Experience and Motivation. *Proceedings of the International Conference on Conversational User Interfaces (CUI 2019)*.

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. *Internet Science*. Retrieved from https://doi.org/10.1007/978-3-030-01437-7_16

Gümüş, N., & Çark, Ö. (2021). The effect of customers' attitudes towards chatbots on their experience and behavioral intention in turkey. *Interdisciplinary Description of Complex Systems*.

Hallowell, R. (1996). The relationships of customer satisfaction, customer loyalty, and profitability: An empirical study. *nternational Journal of Service Industry Management*(7), 27–42. Retrieved from https://doi.org/10.1108/09564239610129931

Hofmann, & Chisholm. (2020). *Text Mining and Visualization: Case Studies Using Open-Source Tools* (1st ed.). Chapman and Hall/CRC.

Hussain, S., Sianaki, O. A., & Ababneh, N. (2019). A Survey on Conversational Agents/Chatbots Classification and Design Techniques. In L. Barolli, M. Takizawa, F. Xhafa, & T. Enokido (Eds.), *Web, Artificial Intelligence and Network Applications* (pp. 946-956). Springer Nature Switzerland.

Islam, R., Ahmed, S., Rahman, M., & Asheq, A. A. (2020). Determinants of service quality and its effect on customer satisfaction and loyalty: an empirical study of private banking sector. *The TQM Journal*.

Jain, M., Kota, R., Kumar, P., & Patel, S. (2018). Convey: Exploring the Use of a Context View for Chatbots. *The 2018 CHI Conference*, (pp. 1-6). doi:10.1145/3173574.3174042

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R* (1st ed.). New York: Springer. doi: 10.1007/978-1-4614-7138-7

Juniper Research. (2021). *Juniper Research*. Retrieved from Bank Cost Savings via Chatbots to Reach $7.3 Billion by 2023, as Automated Customer Experience Evolv: https://www.juniperresearch.com/press/bank-cost-savings-via-chatbots-reach-7-3bn-2023

Kuhn, M. (2008). Caret Package. *Journal of Statistical Software, 28*. doi:10.18637/jss.v028.i05

Kvale, K., Sell, O. A., Hodnebrog, S., & Følstad, A. (2020). Improving Conversations: Lessons Learnt from Manual Analysis of Chatbot Dialogues. *CONVERSATIONS 2019. Lecture Notes in Computer Science, vol 11970* (pp. 187-200). Springer, Cham.

Lohr, S. (2021, 07 16). What Ever Happened to IBM's Watson? *New York Times*. Retrieved from https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html

Lowhorn, G. L. (2007). Qualitative and Quantitative Research: How to Choose the Best Design. *Academic Business World International Conference* (p. 5). Nashville, Tennessee: SSRN. Retrieved from https://ssrn.com/abstract=2235986

Luger, E., & Sellen, A. (2016). "like having a really bad pa". *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.*

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science*.

Markets & Markets. (2020). *Conversational AI Market.* Markets & Markets. Retrieved from https://www.marketsandmarkets.com/Market-Reports/conversational-ai-market-49043506.html

McTear, M. (2021). Evaluating Dialog Systems. In M. McTear, & G. Hirst (Ed.), *Conversational AI: Dialog Systems, Conversational Agenst, and Chatbots* (pp. 91-120). Toronto: Morgan & Claypool.

Moody, L., & Bickel, W. K. (2016). Substance Use and Addictions. In J. K. Luiselli, & A. J. Fischer, *Computer-Assisted and Web-Based Innovations in Psychoology, Special Education, and Health* (p. 159). London: Elsevier Inc.

Ng, M., Coopamootoo, K. P., Toreini, E., Aitken, M., Elliot, K., & Moorsel, A. v. (2020). Simulating the Effects of Social Presence on Trust, Privacy Concerns & Usage Intentions in Automated Bots for Finance. *IEEE European Symposium on Security and Privacy Workshops.* Genoa, Italy: IEEE. doi:10.1109/EuroSPW51379.2020.00034

Nordheim, C. B., Følstad, A., & Bjørkli, C. A. (2019). An Initial Model of Trust in Chatbots for Customer– Findings from a Questionnaire Study. *Interacting with Computers*. Retrieved from 10.1093/iwc/iwz022.

Nuruzzaman, M., & Hussain, O. K. (2018). A Survey on Chatbot Implementation in Customer. *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)* (pp. 54-61). Xi'An (China): IEEE. doi:https://doi.org/10.1109/ICEBE.2018.00019

Proellochs, N., & Feuerriegel, S. (2021, 2 18). *SentimentAnalysis*. Retrieved from https://github.com/sfeuerriegel/SentimentAnalysis

PWC. (2018). *Experience is everything: Here's how to get it right.* PWC. Retrieved from https://www.pwc.de/de/consulting/pwc-consumer-intelligence-series-customer-experience.pdf

Thorat, S. A., & Jadhav, V. (2020). A Review on Implementation Issues of Rule-based Chatbot Systems. *Thorat, Sandeep A. and Jadhav, Vishakha, A Review on ImpleProceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020.* doi:http://dx.doi.org/10.2139/ssrn.3567047

Turing, A. M. (1950, October). Computing machinery and intellegence. *Mind*, pp. 433-434.

Ultes, S., Schmitt, A., & Minker, W. (2013). On Quality Ratings for Spoken Dialogue Systems – Experts vs. Users. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies. Association for Computational Linguistics, (pp. 569-578). Ulm, Germany. Retrieved from https://aclanthology.org/N13-1064.pdf

Wallace, R. S. (2009). The Anatomy of A.L.I.C.E. In R. Epstein, G. Roberts, & G. Beber, *Parsing the Turing Test* (pp. 181-210). Springer, Dordrecht. Retrieved from https://doi.org/10.1007/978-1-4020-6710-5

Weizenbaum, J. (1966). ELIZA: a computer program for the study of natural language communication between man and machine. *Communications of the ACM,* 36-45.

Woodford, S. (2020). *Why chatbots are essential to retail whitepaper.* Juniper Research. Retrieved from https://www.juniperresearch.com/whitepapers/why-chatbots-are-essential-to-retail

Wooldridge, J. M. (2009). *Introductory Econometric: A Modern Approach* (5th ed.). Thomson South-Western.

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*. Retrieved from https://doi.org/10.18637/jss.v077.i01

Zhang, J. J., Følstad, A., & Bjørkli, C. A. (2021, 08 31). Organizational Factors Affecting Successful Implementation of Chatbots for Customer Service. *Journal of internet commerce*. Retrieved from https://doi.org/10.1080/15332861.2021.1966723

Øvrelid, L., Mæhlum, P., Velldal, J. B., & Velldal, E. (2020). A Fine-grained Sentiment Dataset for {N}orwegian. *Proceedings of the 12th Edition of the Language Resources and Evaluation Conference.* Marseille, France.

8. Appendix

Complete regression table

	OLS Model			Logstic Regression		
	Coef.	**Std.Error**	**P.value**	**Coef**	**Std.Error**	**P.value**
(Intercept)	2.229**	0.752	0.003**	−0.695	0.924	0.452
Messages	−0.050***	0.005	0.000***	−0.077***	0.008	0.000***
Chat duration(Seconds)	0.000	0.000	0.473	0.000	0.000	0.748
predictionMulti-Intent	0.025	0.958	0.979	−0.429	1.201	0.721
predictionPerfect Match	1.205+	0.700+	0.085+	0.719	0.867	0.407
predictionPossible Missing Intent	0.409	0.706	0.562	−0.422	0.876	0.630
predictionRegular/Valid Answer	0.990	0.699	0.157	0.469	0.866	0.588
predictionUnknown	0.506	0.721	0.482	0.009	0.895	0.992
User's sentiment	0.227***	0.042	0.000***	0.287***	0.054	0.000***
topicsAccount	0.248	0.254	0.329	0.375	0.294	0.203
topicsBank Services	0.118	0.296	0.692	0.287	0.342	0.401
topicsBSU	0.443	0.331	0.181	0.434	0.376	0.248
topicsCredit and Debit Cards	0.719**	0.240	0.003**	0.780**	0.278	0.005**

topicsCredit and Debit Cards,General Questions	1.949***	0.349	0.000***	2.023***	0.397	0.000***
topicsGeneral Questions	1.051***	0.258	0.000***	1.436***	0.297	0.000***
topicsInsurance	0.439+	0.254	0.084+	0.643*	0.291	0.027*
topicsInvestment Funds	0.087	0.322	0.788	0.410	0.367	0.264
topicsLoans	0.216	0.246	0.381	0.507+	0.286+	0.076+
topicsLoans,Talk to Human Agent	−0.419	0.371	0.259	−0.835	0.592	0.158
topicsLogin	0.531*	0.237	0.025*	0.749**	0.276	0.007**
topicsLogin,General Questions	1.681***	0.332	0.000***	1.851***	0.373	0.000***
topicsOnline Banking	0.304	0.323	0.347	0.366	0.371	0.324
topicsPayments	0.451+	0.249	0.070+	0.642*	0.288	0.026*
topicsPayments,General Questions	2.702***	0.412	0.000***	3.123***	0.560	0.000***
topicsPension	−0.667+	0.384	0.082+	−1.219+	0.658	0.064+
topicsSavings	−0.321	0.401	0.423	−0.636	0.521	0.222
topicsTalk to Human Agent	−0.840**	0.258	0.001**	−1.101***	0.330	0.001***
topicsVipps	1.044*	0.418	0.013*	1.360**	0.452	0.003**
topicsOther	0.218	0.235	0.354	0.388	0.276	0.161
Agent_availableYes	−0.460**	0.152	0.003**	−0.292+	0.168	0.081+
Num.Obs.	4498			4498		

R2	0.096					
R2 Adj.	0.090					
AIC	18891.7			5374.4		
BIC	19090.5			5566.7		
Log.Lik.	−9414.873			−2657.191		
F	16.408					
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001						
*Chat duration is measured in seconds from the first message to the last						

Interview 1

Interview with informant #1 and #2

In general, we asked open questions in order to get unexpected answers. The questions were also meant as a conversation starter where we asked some follow up questions when we deemed it necessary.

Intro:

- Introduction of our research and what the goal of the interview was.
- Explain how the result of the interview is going to contribute to the thesis.
- Give them the ability to pull statement post interview.

Initial question to interviewee

1. Is it ok that we utilize the answers given from this interview in our thesis?

About chatbots and why the company utilize chatbots

1. What is the goal of the chatbot and why does the company utilize chatbots?
2. What are the characteristics of a good chat and when has the chatbot done what it was intended for?
3. What are the characteristics of when a chat with the chatbot is not successful?
4. Do you have any internal goals in terms of task completion rate for the chatbot?
5. What can be reasons why a chatbot is not qualified to help a customer, and what is typical questions that is outside the scope of the chatbot?
6. Can chatbots replace human agent and perform as good?
7. What are the future goals for the chatbot that you are utilizing?
8. Is there something that internal review, or something that employees pick up that is not covered by the customer satisfaction index?
9. Do you see any difference in task completion for different demographics: Age, gender, preferred language etc.?
10. How should the language of chatbots be?
    - Formal or informal?

Interview 2

Interview with informant #3

In general, we asked open questions in order to get unexpected answers. The questions were also meant as a conversation starter where we asked some follow up questions when we deemed it necessary.

Introduction:

- Introduction of our research and what the goal of the interview was.
- Explain how the result of the interview is going to contribute to the thesis.
- Give him/her the ability to pull statement post interview.

Initial question to interviewee

1. Is it ok that we utilize the answers given from this interview in our thesis?

Motivation of utilizing chatbots for customer service

1. Could you give us an explanation of how DNB's chatbot Aino works?
2. What is DNB's motivation of chatbot use, and has it changed over time?

Implementation

1. What does it take to implement a well-functioning chatbot today?
2. Has the implementation of chatbots led to any problems for DNB?`
3. What do you think is missing from chatbots today in order to replace people in customer service?
4. Do you know of any technology included in other chatbots that could be desirable to implement in DNB?
5. How do customers perceive the chatbots, and has this changed over time?
6. Which laws and regulations can be of hinderance for the chatbot in solving customer inquiries?
7. What are the employees view on chatbots and has it changed over time?
8. What are the factors of making an optimal chatbot ?
9. What has to be included in order to make customers happy when utilizing a chatbot?
10. Is there anything that you think is important, that we did not cover in this interview?

R packages utilizised

Richard Iannone, Joe Cheng and Barret Schloerke (2021). gt: Easily Create Presentation-Ready Display Tables. R package version 0.3.1. https://CRAN.R-project.org/package=gt

Marvin N. Wright, Andreas Ziegler (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1-17. doi:10.18637/jss.v077.i01

Julia Silge, Fanny Chow, Max Kuhn and Hadley Wickham (2021). rsample: General Resampling Infrastructure. R package version 0.1.1. https://CRAN.R-project.org/package=rsample

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-90. https://CRAN.R-project.org/package=caret

Nicolas Proellochs and Stefan Feuerriegel (2021). SentimentAnalysis: Dictionary-Based Sentiment Analysis. R package version 1.3-4. https://CRAN.R-project.org/package=SentimentAnalysis

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ingo Feinerer and Kurt Hornik (2020). tm: Text Mining Package. R package version 0.7-8. https://CRAN.R-project.org/package=tm

Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J. Reproducible summary tables with the gtsummary package. The R Journal 2021;13:570–80. https://doi.org/10.32614/RJ-2021-053.

Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot

Sam Firke (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. https://CRAN.R-project.org/package=janitor

Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2021). skimr: Compact and Flexible Summaries of Data. R package version 2.1.3. https://CRAN.R-project.org/package=skimr

Hao Zhu (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra