LUISS
Department of Business and Management

BI-LUISS Joint Master of Science in Marketing

Chair of Performance Marketing

Man vs. Machine: Do Humans Still Return Higher Satisfaction Ratings Than Chatbots in Customer Service?

Prof. Francisco J. Villarroel Ordenes

THESIS SUPERVISOR

Vilius Baltrukonis

CANDIDATE

Prof. Stefania Farace

THESIS CO-SUPERVISOR

757851

MATRICULATION NUMBER

Master Thesis

Academic Year 2022/2023

Content

Table of Contents

Acknowledgements

First and foremost, I want to thank my thesis supervisor, Professor Francisco Villaroel Ordenes, for guiding and assisting me through the writing process by providing knowledgeable insights. Secondarily, I want to thank all the other professors that I had the pleasure to learn from in these two years. Tertiarily, I want to thank all my friends, amici, venner, and amis in this joint BI-LUISS Master program that made the experience stupendous. Finally, I want to thank my parents, Darius Baltrukonis and Lina Baltrukonienė, for always calming me down when I struggled and cheering me on when I succeeded.

Abstract

The technological boom has impacted many industries and AI is taking more and more jobs. While the news that chatbots will save companies up to $80 billion by 2026 is exciting for many managers that struggle to make ends meet, switching to chatbots has its downsides. This research compares the satisfaction ratings of customers that solely talked to a human customer service employee (HumanOnly) and customers that first talked to a chatbot and then were transferred to a human (Handover). The results show that the average HumanOnly interaction is significantly higher rated than the average Handover interaction. To support this, "bot", as a moderator variable, had a negative impact on the customer becoming a "Promoter" in a logistic regression model. Additionally, the research looked at how "competence" as a linguistic feature impacted customer satisfaction ratings. The results differ between the HumanOnly and Handover interactions and extend the literature on a script's performance depending on whether it was an interaction with a chatbot or a human. Managers are advised to economically investigate the implementation of a chatbot and urge their customer service departments to run tests on different scripts to maximize satisfaction.

Keywords: Customer Service, Chatbot, Text mining.

Introduction

Have you ever entered a customer service chat and been met by a chatbot asking you what it may assist you with today but then not really gotten assisted with the problem at hand? The chances are that you have, and the use of chatbots is only set to increase in the coming years. The customer service part of a company started with call centers and with increasing digitalization and product complexity, almost all companies have introduced one. This is a solution that helps companies quickly assist customers in need. With time, customer service chats also arrived. Not all companies employ this yet, but having a chat option can often save time and allow the employee to assist several customers at the same time. McKinsey and Company (2021) point out that a well-constructed contact center can save 5-10% in channel costs for a company. Additionally, call centers and employees are a cost to a company and the development of chatbots has been increasing as an attempt to reduce this cost. In fact, it is projected that using chatbots will save companies up to $80 billion by 2026 (Quach, 2022).

As with any AI, however, chatbots need to be trained and it is important to understand what leads to customer satisfaction in such interactions. In studies about humans in customer service, a recent article found that concreteness can increase both customer satisfaction and expenditure (Packard & Berger, 2021) and another found that it increases engagement because of processing ease (Berger et al., 2023). Another study found that competent, as opposed to warm, language increased customer satisfaction (Marinova et al., 2018). Finally, Golden and his peers (2022) found that empathy and active listening play crucial roles in customer service interactions. When it comes to chatbots in service, Eren (2021) found that the performance of the chatbot, the perceived trust, and the corporate image of the company increased satisfaction with chatbots but did not dive deeper into the linguistic properties of the actual conversation.

As the knowledge about what leads to customer satisfaction is only increasing and the financial benefits of employing chatbots are clear, it is important to assess whether customers rate interactions with chatbots or humans higher. Crolic and colleagues (2022) found that higher anthropomorphism not only leads to lower satisfaction when the customer was in an angry state but also decreased the overall evaluation of the firm and the future purchase intention. This means that companies must be wary when employing chatbots, as saving on expenditure is not worth it if it is costing the company significant amounts of reputation.

The first identified gap in the literature was a direct suggestion by Crolic and colleagues (2022), he mentions how there is a need for exploring the differences in satisfaction levels when a customer interacts with a human and when a bot is present. Hence, this Master Thesis aims at comparing the satisfaction levels between human-only interactions and interactions that also involve a bot.

The thesis will further examine the effect of a linguistic feature in interactions where a bot is involved. Sands and peers (2021) found that human service employees outperform bots with an educational script, but not with an entertaining one. Their findings indicate that linguistic features that work for human service employees might not be as effective for chatbots. Testing a linguistic concept on bot-involving data will, therefore, give an indication of whether such findings can be generalized from human service employees to chatbots. This research will focus on competence as the linguistic feature as there has not been any research directly tackling competence within chat customer service and chatbots, giving a unique opportunity to test a new linguistic feature in both types of customer service and compare its impact. The Stereotype Content Model (Cuddy et al., 2008) describes "competence" as having the ability to perform a certain task and is synonymous with being confident and skillful. The Stereotype Content Model (SCM) frames competence as something positive and there is no good reason why it should not also be positive in the customer service industry as it guides a customer to a solution.

The research contributes to the existing literature by (1) examining the satisfaction level differences when interacting with humans versus with bots, (2) testing the effect of competence on customer satisfaction when talking with chatbots, and (3) seeing whether a linguistic feature has the same effect in when bots are involved and when they are not.

The two research questions to be answered through this Master Thesis are:

1. Do customers prefer human-only or bot-initiated interactions in the customer service field?
2. Does competence have the same impact on the NPS when a bot is involved in the customer service interaction versus when not?

Literature Review

In the literature review, I will go into detail about how chatbots in general work and then explain the specifics of the chatbot that is used by the company that offered their data for this research. Followingly, I will explain why we should expect higher satisfaction in the human-only data by citing recent findings in the current and other fields and by backing it up with psychological insights. Finally, I will cover concreteness; explaining what it is, how it can be measured, and what effect is to be expected in the coming studies.

Chatbots

A chatbot can be defined as *a conversational application that aids in customer service, engagement, and support by replacing or augmenting human support agents with artificial intelligence (AI) and other automation technologies that can communicate with end-users via chat* (BasuMallick, 2022). This research will be focusing on chatbots within customer service. There are many different types of chatbots, and they range in complexity from a simple FAQ bot to an actively learning, AI chatbot that improves based on previous conversations.

Chatbots are coded to mimic normal human dialogues as if those were with an actual customer service employee. And even though companies spend many resources on developing them, people are still reluctant to adopt them. Sheehan and peers (2020) found, not surprisingly, that error-free chatbots are preferred to ones that make errors and elicit higher adoption willingness. An interesting finding there, however, was that chatbots that sought clarification by asking questions like "Sorry, I did not understand, could you say that again" also elicited higher adoption willingness than just regular error-prone chatbots that did not include these terms. Additionally, they found that customers with a higher need for human interaction preferred higher anthropomorphism. Adding to this, Crolic (2022) found that anthropomorphism can have a negative effect on customer satisfaction, however, only if the customer is already angry. Eren (2021) found that it was the performance of the chatbot, the perceived trust, and the corporate image that drove the willingness to adopt.

When it comes to ways of speaking, Kull and peers (2021) investigated what could elicit the highest brand engagement in conversations initiated by a bot. They found that a warmer tone elicited the highest engagement, then a competent tone, and finally a neutral tone. Sands and peers (2021) found that chatbots perform better with an entertaining service script

than with an educational one and the chatbot also was not outperformed by a human customer service employee when using an entertaining script.

Relations Between Humans and Bots

Humans tend to connect with robots and give them human-like attributes and even involve them in our own ceremonies. The BBC describes how the US military held a funeral and awarded it with a purple heart and a bronze star when it "died" (Gorvett, 2018). Another news story talks about a retirement party thrown for the five mail delivery office machines (Whalen, 2017). Certainly, the robots did not receive this treatment by doing nothing, one had saved many lives, while the others had delivered mail for 25 years, but actions like that bring to light that humans might not be that opposed to connecting with robots.

The more human traits and behaviors a robot showcases, the more likely it is to receive human-like treatment (Goldberg, 2022). This subsequently means that the better it can mimic us, the higher likelihood that it will be accepted into our society and adopted. Interestingly, Eyssel and Hegel (2012) found that humans will even apply human gender stereotypes to robots if they are assigned a gender. Supporting this, Gorvett (2018) mentions that *humans will feel sorry for our non-human colleagues when things go wrong, project personalities onto them, give them names and even debate over their gender*.

It has been found that customers even prefer human interaction especially when they strongly identify with the product (Leung et al., 2018), this way resisting automation. Not surprisingly, humans then also prefer their doctor to be a human, as they are reluctant to use medical advice provided by an AI (Longoni et al., 2019). Mende and peers (2019) find that robots can be perceived as eerie, increase food consumption and expenditure on identity-based, expensive items because of compensatory behavior, and that higher anthropomorphism is not always beneficial. On the contrary Castelo and peers (2019) find that anthropomorphism is viewed as positive and that bots are more likely to be preferred in objective situations, while Logg and peers (2019) found that there is a general preference for advice given by an algorithm. Another study found that there is an overall preference for humans when the product in question is symbolic, theorizing that the reason is higher product uniqueness (Granulo et al., 2021).

Additionally, there is now also a growing literature about humans' views on algorithms. The phenomenon has become more widely known as "algorithm aversion"; a name that

indicates a negative impact. However, Jussupow and peers (2020) provided a comprehensive review of prior research, showing that one cannot, categorically, say that humans either are always or never algorithm averse. Most of the studies indicate aversion, but plenty were inconclusive or indicated appreciation. They highlight that algorithm agency, performance, perceived capabilities, human involvement, human agents' expertise, and social distance greatly influence aversion.

With these findings and the previously mentioned findings of customers in a non-angry state (Crolic et al, 2022) and customers with a higher need for human interaction (Sheehan et al., 2020) preferring more highly anthropomorphized service robots, it is clear that segmentation is important when choosing where to use robots and that more industries should be directly comparing the impact of their human employees and their robots on customer satisfaction.

Competence

Competence is a widely studied concept and one of the two fundamental dimensions in SCM. In this model, competence is used to measure one's capacity to act on a certain intention (Cuddy et al., 2008). Having competence is synonymous with being confident, capable, competent, and skillful (Cuddy et al., 2009). Cuddy and his peers (2009) found that competence is viewed as mostly similar in all cultures, which means that the stereotype it casts onto the person that is perceived to have competence is close to universal. Having high competence is naturally better than having low competence and the casted stereotype is shown to have an especially great effect on transactional outcomes like increasing the share of wallet (Güntürkün et al., 2020).

When it comes to customer service, supporting your statement means that you supply the customer with reasoning for why they should do something to solve their problem. As an example, if you wanted someone to restart their phone to check if that would improve their phone service, you would be likely to accompany that with a little bit of reasoning. If you suggested resetting the network settings, you might explain why to ensure the customer they are not losing valuable personal information and give directions to make it easy. In other words, providing reasoning is a way of showing competence. Coined as "polite argumentation", this approach has been shown to have a positive effect on customer satisfaction (Okumus & Unal, 2012; JananJohnson et al., 2014).

Service Recovery

There are many mistakes that can be made when a product changes hands between the company and the customer. In those instances, it is important to make sure that the customer service is on point and provides an adequate recovery to make sure that the customer is kept satisfied and his attitudinal loyalty does not go to waste. Yüksel and Rimmington (1998) that customer satisfaction is crucial to increase or keep customers returning (attitudinal loyalty) and DeWitt and his peers (2017) proved this further by finding that most service recovery processes apart from a simple apology significantly boost satisfaction level.

Gebrich (2010) found that informational support can replace monetary compensation after a service failure. By explaining how the service failure happened at what exact actions were taken to prevent it, the customer gains empathy for the customer service employee and anger decreases. Further, Ozuem and peers (2021) characterized "empathizers" as one of the three main groups of customers, a customer group that responds well to any effort done by the service employee to recover from the service failure. The other two groups of customers are called "blanders" and "churners". Blanders expect good or even exceptional recovery, while churners focus on how the service recovery is delivered and how good the service recovery of that company is. Providing competent service recovery by being accurate and quick and creating some empathy in the process is how companies should reach to service failure.

Attitudinal Loyalty: NPS

There are two types of loyalty: attitudinal and behavioral (Ipsos Encyclopedia, 2016). Behavioral loyalty is about repeat behavior like purchases, while attitudinal loyalty is about consistent attitudes towards a company. As mentioned, attitudinal loyalty is important because any company stands to benefit from positive attitudes towards itself and the benefit of the doubt when things do not go as planned and service failure or brand crisis occurs.

Another reason why it is important is that customer satisfaction can lead to positive Word of Mouth, which is important as people tend to trust other consumers over the company itself. Word of Mouth is a consequence of attitudinal loyalty and one of the best ways of measuring it is with Net Promoter Score (NPS) (De Haan et al., 2021). The net promoter score is on a scale from 0 to 10 and segments the customer based on the score. 0-6 are the detractors (will talk negatively about your company), 7-8 are the neutrals (will not talk about the brand), and 9-10 are the promoters (will talk positively and promote your brand to others). With these

definitions, you then go on to calculate your overall company NPS by subtracting the % of detractors from the % of promoters (Salesforce, 2023).

In customer service, a customer is often asked to leave a review after the interaction with a customer service employee, this score is a net promoter score. By looking at the net promoter score and, analytically, looking for themes and patterns that lead to either a high or a low score, companies can get a grasp of how to improve their customers' attitudinal loyalty and ultimately improve their bottom line.

The following table contains a summary of relevant research articles within customer service. Specifically, these are articles within customer service where chatbots are involved.

Research Title	**Main Findings**
Customer service chatbots: Anthropomorphism and adoption (Sheehan et al., 2020)	Error-free chatbots are preferred to error-prone ones. Fewer errors lead to a higher willingness to adopt. Bots that seek clarification in their speech elicit a higher willingness to adopt and increase anthropomorphism. Customers with a higher need for human interaction prefer higher anthropomorphism.
Determinants of customer satisfaction in chatbot use: evidence from a banking application in Turkey (Eren, 2021)	The performance of the chatbot, the perceived trust, and the corporate image of the company increase satisfaction with chatbots.
Blame the Bot: Anthropomorphism and Anger in Customer–Chatbot Interactions (Crolic et al., 2022)	Anthropomorphism negatively impacts customer satisfaction when the customer is in an angry state but impacts it positively the customer is not.
How may I help you? Driving brand engagement through the warmth of an initial chatbot message (Kull et al., 2021)	Chatbot messages that included a warmer tone (friendly) elicited higher brand engagement than both neutral and competent (capable) messages. Competent messages elicited higher brand engagement than neutral messages.

Managing the human–chatbot divide: how service scripts influence service experience (Sands et al., 2021)	An educational service script increases satisfaction and purchase intention more with human service employees than with chatbots. The effect is mediated by emotion and rapport, showcasing the importance of human connection. However, these findings are not replicated with an entertaining service script; there the human service employees and chatbots perform equally well.

Table 1: Short summaries of all articles about chatbots within customer service

Hypotheses

Although humans are becoming more and more accepting of algorithms, it seems to be that there is an overall preference for humans over bots. This is especially true when they identify with a product or it is unique to them (Leung et al., 2018; Granulo et al., 2021). And although humans are not always algorithm averse, there are more situations in which we are averse than not (Jussupow et al., 2020). However, in objective situations, bots can be preferred to humans (Logg et al., 2019). In customer service, a customer comes to the company with an individual problem that makes the company's service fail. As the problem is perceived to be individual and unique, the customer is likely to prefer a human as their assistant as it will feel more like a subjective matter to the customer. Hence, the first hypothesis for this research is:

- **H1:** *Customer satisfaction is higher when no bot is involved.*

Competence is viewed as a positive concept in human-to-human interactions. It is synonymous with being skilful, providing adequate help, and being confident. With these positive connotations, it is no wonder that it increases customer spending (Güntürkün et al., 2020). In addition, providing polite argumentation when solving a problem also has a positive effect on customer satisfaction (Okumus & Unal, 2012; JananJohnson et al., 2014). Although competence seems to be a positive concept throughout, there is reason to doubt that it will extend to bots. The first reason for doubt is that the competent tone was not the one that elicited the highest engagement (Kull et al., 2021). A second reason for this doubt is that chatbots were outperformed by humans with an educational script (Sands et al., 2021). Finally, high anthropomorphism has been found to have a negative impact on customer satisfaction when a customer is angry (Crolic, 2022). The second hypothesis is as follows:

- ***H2:*** *Linguistic competence has a different effect for human, vs human+bot interactions such as:*

- ***H2a:*** *Linguistic competence has a positive effect for human only interactions.*
- ***H2b:*** *Linguistic competence has a negative effect for bot+human interactions.*

The conceptual model to answer, confirm, or deny the hypotheses looks like this:

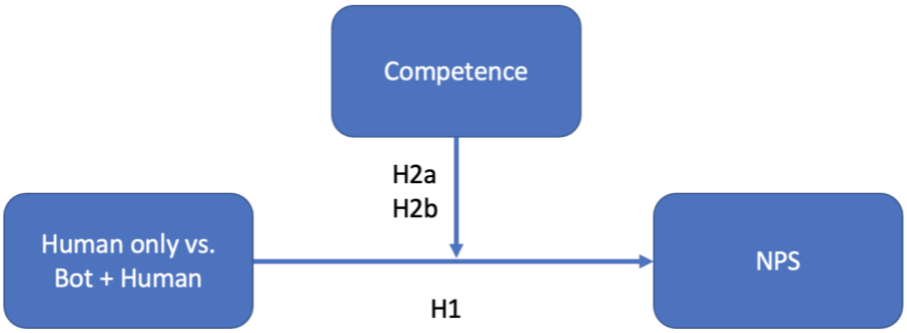


Figure 1: Conceptual model

Methodology

Chatbots can differ vastly based on how they are programmed, hence the company which provided the data also provided a description of how the chatbot they use is programmed. The bot has a database of information with appropriate answers for the customer. The chatbot finds the correct information to answer with by matching what it is asked with the different category-triggering sentences it has been fed. This is called an intent-hierarchy and it is continuously getting fed new information to keep improving. This is not quite automated yet, as the chatbot is fed new data based on what the coders read and learn from previous conversations with customers of the company; but the important part is that it gets new information continuously and keeps improving.

Data

To perform the analyses, I have used data from a Norwegian telecom company with 5,357 customer-company chat interactions that started with a chatbot, of which all are in Norwegian and another 5,005 interactions that are human-only; meaning that the customer only talked to a real human.

The data has two main properties that will be essential for the analysis: (1) the textual conversation from both parties and (2) the NPS for the interaction (range: 0-10). The conversation process when a bot is involved is as follows: a customer opens a chat, talks to the chatbot until either they ask to talk to a human or the chatbot suggests it, the conversation is

sent to a customer service employee, they talk until a solution to the problem is found (ideally, but some chats result in no solution or a cut-off conversation), and after the chat is closed the customer can choose to leave a score of the interaction they just had.

The data is a random sample taken from all conversations between customers and customer service in a big telecom company in Norway. The data for the Human-only interactions were delivered in 13 txt files and is cleaned through a KNIME workflow without much trouble. The sample of Handover-data (conversations that started with a bot and were later handed over to a real human) was delivered in 6 different txt-files: “January-February”, “March-April”, “May-June”, “July-August”, “September-October”, “November-December”. These were also cleaned through KNIME, but the process was not fully automated.

The data was unstructured (Balducci & Marinova, 2018). In particular, there was no automatic way of determining whether the sender of a message was a customer or a customer service employee in the Handover dataset and had to be performed manually. The work was split between two coders and took each party about 60 hours of work. Going through the data in this fashion gave some qualitative insights and understanding in addition to the quantitative insights obtained later by using text mining on the data.

I will be using KNIME workflows to annotate the data and analyze the differences between human-only and bot-involving interactions and insert dictionaries to measure the level of competence.

In the table below, two examples of customer service chats are provided to give the reader a better understanding of the data and give context. Some of the messages are not presented in the table for the Handover example, but those messages consist of the customer describing their problem and the bot trying to provide help (unsuccessfully). Both conversations were rated 10/10 on the NPS score by the customer. All names have been taken out of the messages alongside any personal information.

Sender	Handover	Sender	HumanOnly
Chatbot	Hello! I'm the chatting robot! Ask me what you're wondering, and I'll answer as best I can!	Human Customer Service Employee	Hi, and thank you for contacting us. My name is …, what can I do for you?
Some messages are exchanged where the problem is not solved	…	Customer	Hi, I want to remove my plus insurance, how do I do it?
Customer	I want to talk to a human agent	Human Customer Service Employee	Of course I can help you with that :) Could I just have gotten your date of birth and address first?
Human Customer Service Employee	Hi, and thank you for contacting us. My name is ..., what can I help you with?	Customer	**Provides the needed information for identification*
Customer	Hi, I have a child subscription, but we do not remember the PIN code.	Human Customer Service Employee	Thank you very much! :) See there is a Plus insurance Medium here. Then I cancel this right away :)
Human Customer Service Employee	Of course I will have a look at this! Can you first just confirm your date of birth and address?	Customer	Yup, thank you so much for the help.
Customer	*Provides the needed information for identification*	Human Customer Service Employee	You are welcome:) Have a nice day
Human Customer Service Employee	Thank you. Just enter the wrong PIN code for the SIM card a total of 3 times, and then use PUK code: xxx to create a new PIN code of your choice. :)	Customer	Thank you, same to you!
Customer	Thank you very much for that :) Then we will arrange it that way. Have a nice day		
Human Customer Service Employee	No problem! The least I could do. ;) If there was nothing else, I thank you for the chat and wish you a really nice day ahead!		

Table 2: Examples of a Handover and a HumanOnly conversation. Both rated 10/10.

Measurement and Development of Competence

To measure the construct within the data, a top-down approach (Villarroel Ordenes & Zhang, 2019) is used with self-made dictionaries that come from the LIWC (Linguistic Inquiry and Word Count) and own knowledge obtained by going through thousands of conversations.

Dizier (2020) presents a list of words that could indicate an argument. Polite argumentation and showing competence could be close to the same thing when it comes to customer service as you are guiding the customer from a mistake to a solution or just explaining the specifics of how something works. Coupling it with the knowledge obtained from reading all handover interactions, 56 words were picked out and placed into the self-made dictionary. Additionally, the certitude dictionary from the LIWC (2022) was added. As certitude represents words that indicate being certain about something, it fits perfectly into the concept. 87 of the words from that dictionary were picked out and placed into the self-made dictionary. Additionally, I also added 10 words based on the knowledge acquired by reading all the handover conversations.

Modelling

The human-only data and the handover data were separately concatenated. Each dataset will be grouped by the conversation ID, and then all the text will be concatenated and subsequently turned into a document. To separate the two datasets and put them together a new, binary column was created where human-only received the value 0 and handover received the value 1. The two datasets will then be concatenated into a single file and the customer satisfaction ratings per conversation will be compared with each other. As these are different samples, the independent groups t-test node will be used to test the means against each other and to provide other descriptive statistics.

To examine competence within each conversation, a self-made dictionary will be plugged into the workflow. Followingly, a bag of words will then be created to calculate term frequency (TF) and inverse document frequency (IDF). TF allows us to understand how often a word appears, while IDF calculates the rarity of a word. As NPS does not have a normal distribution, is not an ideal approach. To rectify this issue, a new, categorical variable "Categorical NPS" will be made where ratings "9" and "10" will be coded as "Promoter" and ratings 0 through 8 will be coded as "Detractor". The group "Neutrals" will be a part of this model as detractors because the variable needs to be binary to fit into a logistic regression

(Malhotra, 2019, p.595-597). In the logistic regression model, competence will be the independent variable and the impact will be measured on the dependent variable: Categorical NPS. This procedure will be done to both datasets and the effect of competence will be recorded for both types of interactions and control variables will be added to make sure that the same reference group is used when concluding.

Additionally, the HumanOnly and Handover data will be separated binarily, and the variable "bot" will be created. The coefficient will indicate how the variable "Bot" moderates the likelihood of a customer being a "Promoter".

Comparing Human-Only Data with the Handover Data

In addition to looking at the effect with a logistic regression, an independent groups t-test will be performed. All the cleaned data from the different months will be concatenated into one table and put into the test. If there is a difference between the two groups, this will prove that bot involvement has a moderating effect on customer satisfaction. Using the independent groups t-test will be an additional test of the moderating effect bot involvement has on customer satisfaction.

Control Variables

10 additional variables are included as a control (and to have reference categories): These were "Owning statements", "Competence", "Confirmation", "Anxiety", "Number of terms", "Bot", "Bot*Owning", "Bot*confirm", "Bot*Competence", and "Bot*Terms". "Bot" is a binary variable created to determine whether a conversation is from the HumanOnly or Handover dataset. "Number of terms" is a count of words (terms) in a full conversation. The remaining constructs are measured by calculating TF and IDF with a dictionary as a basis and then multiplied with the variable "Bot" to measure it in both datasets.

Owning statements (or owned messages) are "I-messages", "owning thoughts and feelings" and "speaking for self". In essence, it is the use of first-person pronouns to acknowledge personal responsibility and subjectivity (Proctor II & Wilcox, 1993). Confirmation is a process where there is agreement. Expressing agreement, which is a confirming message, can be done in various ways, such as "You're right", "You're correct", "Understand" and "Agree" (Welchlin, 2017). Anxiety is reverse-coded relaxation and is

defined as the absence of relaxation, something that weakens interpersonal communication (Titlebaum, 1998).

Results

The tests ran through two different models but were performed on the same dataset. Hence, there is only a single study in this research. However, the study had two main objectives: measuring the satisfaction difference between HumanOnly and Handover and measuring the impact of competence on satisfaction for HumanOnly and Handover. The first objective was tested with the logistic regression and the independent groups t-test, while the second objective was measured solely with the logistic regression. As regression uses the existing created variables to determine the effect of an independent variable on a dependent variable, the aforementioned control variables were added to have reference categories. To determine which are eligible for inclusion a variance inflation factor test was performed.

Descriptive Statistics and Variance Inflation Factor Test

Before the initiation of the analysis, descriptive statistics were collected and measured for both "Competence" and control variables. The descriptive statistics for HumanOnly and Handover were also collected to provide a holistic overview.

Variable	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness
Owning statements	0,00	0,18	0,05	0,02	0,00	1,15
Competence	0,00	0,22	0,02	0,02	0,00	0,75
Confirmation	0,00	0,14	0,01	0,01	0,00	1,39
Anxiety	0,00	0,04	0,00	0,00	0,00	10,91
Number of terms	7,00	3883,00	205,59	165,54	27402,62	4,35
NPS	0,00	10,00	8,17	3,40	11,53	-1,72
Bot	0,00	1,00	0,52	0,50	0,25	-0,07
Bot*Owning	0,00	0,15	0,03	0,03	0,00	0,85
Bot*Confirm	0,00	0,14	0,01	0,01	0,00	2,53
Bot*Competence	0,00	0,22	0,01	0,02	0,00	1,47
Bot*Anx	0,00	0,04	0,00	0,00	0,00	13,84
Bot*terms	0,00	2168,00	99,17	144,06	20752,72	2,63
Dataset	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**	**Variance**	**Skewness**
NPS HumanOnly	0,00	10,00	8,33	3,29	10,81	-1,87
NPS Handover	0,00	10,00	8,02	3,50	12,22	-1,59

Table 3: Descriptive statistics

In addition to obtaining the descriptive statistics before running the regression, a test to calculate the variance inflation factor (VIF) was performed to see which concepts should be excluded to avoid multicollinearity. A VIF value over 5 indicates that the researcher should consider removing the variable (James, 2014). There were no variables that had a VIF value of over 5, hence none of the variables were excluded in the logistic regression analysis.

Logistic Regression Analysis

The logistic regression had "Categorical NPS" as the binary dependent variable. Variables excluded from the test were: ID, NPS (numerical 0-10), and "bot*owning". "Bot*owning" was excluded to add another reference category. To calculate the effect and make it more interpretable, all coefficients were exponentiated where 1 means no effect, anything below 1 represents a negative effect and anything above 1 represents a positive effect. The formula [*Exp(coefficient)-1] *100%* gives the change in odds. The confidence level of the test was 95%.

Logit	Variable	Coefficient	Std. Error	z-score	P >|z|	Beta-coefficient
Promoter	Owning statements	-0,433	0,027	-16,286	0,000	0,649
Promoter	Competence	-0,163	0,038	-4,247	0,000	0,850
Promoter	Confirmation	0,639	0,045	14,195	0,000	1,894
Promoter	Anxiety	0,217	0,077	2,828	0,005	1,243
Promoter	Number of terms	-0,113	0,032	-3,489	0,000	0,893
Promoter	Bot	-0,145	0,054	-2,664	0,008	0,865
Promoter	Bot*Confirm	-0,065	0,053	-1,213	0,225	0,937
Promoter	Bot*Competence	0,044	0,051	0,854	0,393	1,045
Promoter	Bot*Anx	0,045	0,082	0,551	0,582	1,046
Promoter	Bot*terms	-0,061	0,041	-1,498	0,134	0,941
Promoter	Constant	1,193	0,026	46,687	0,000	3,296

Table 4: Results of logistic regression

The impact of Variables "Bot*confirm", "bot*competence", "Bot*Anx", and "Bot*terms" was statistically insignificant, while all other remaining variables were statistically different from 0. Followingly, "Owning statements", "Competence", "Number of terms", and "Bot" negatively impacted the probability of a customer becoming a promoter, while "Confirmation" and "Anxiety" positively impacted the probability of a customer being a promoter. The constant's beta-coefficient being 3.296 indicates that all else being equal, a customer is 229.6% more likely to be a promoter than a detractor.

The concept "Competence" had a negative coefficient, this indicates that it would have a negative impact on NPS. With the beta-coefficient, the calculated effect of using competence in the language decreased the probability of a customer being a promoter by 15%. This effect is significant at the 95% confidence level (p-value < 0.001). "Bot*Competence" had a slightly positive coefficient, indicating a positive effect, however, the p-value is above 0.05 (p-value = 0.393), which means that it is not statistically significant from 0 at the 95% confidence level.

The binary variable bot, like "competence", had a negative coefficient. The variable statistically significantly decreases the likelihood for a customer to be a promoter (p-value = 0.008). In fact, the customer is 13.5% less likely to become a promoter if a bot is involved in the interaction.

As a result of the logistic regression, **H1** is accepted because including a bot in the conversation decreases satisfaction. **H2** is accepted because the effect differs between HumanOnly and Handover (negative vs neutral). **H2a** is rejected because the impact of competence is negative towards making a customer a promoter (and not positive). Since the effect of competence for Handover interactions (Bot*competence) was not significant at the 95% confidence level, it is not statistically different from 0 (and hence not negative). **H2b** is rejected.

Independent Groups t-test

KNIME was used to perform the independent groups t-test. The data for HumanOnly and Handover were separately filtered and given either 1 (Handover) or 0 (HumanOnly). The two datasets were then concatenated into a single dataset and put through the test to test the difference between the mean NPS ratings (range: 0-10). The confidence interval for the test was 95%.

Variable	Group	N	Mean	Standard Deviation	Standard Error Mean
NPS	1	5357	8,016	3,495	0,048
NPS	0	5005	8,331	3,288	0,046

Table 5: Descriptive statistics of the HumanOnly and Handover datasets

The hypotheses for the independent groups t-test are:

$$H_0: \mu_{HumanOnly} = \mu_{Handover}$$

$$H_A: \mu_{HumanOnly} \neq \mu_{Handover}$$

The groups were binarily divided, HumanOnly getting 0 and Handover getting 1. The HumanOnly chat interactions had an average rating of 8.331, with a standard deviation of 3.288, and a standard error mean of 0.046. The Handover chat interactions had an average rating of 8.016, with a standard deviation of 3.495, and a standard error mean of 0.048. There were 352 more chats in the Handover dataset than in the HumanOnly dataset. There were no missing rows, columns, or values in the final output of the test.

Test Column	Variance Assumption	t	df	p-value (2-tailed)	Mean Difference	Standard Error Difference	Confidence Interval Probability
NPS	Equal variances assumed	-4,720	10360,000	0,000	-0,315	0,067	95%
NPS	Equal variances not assumed	-4,730	10359,519	0,000	-0,315	0,067	95%

Table 6: Results of the independent groups t-test

As foreshadowed by the descriptive statistics, the difference in means was statistically significantly different from 0. In fact, the difference in means was 0.315. As the p-value is below 0.05 (p-value < 0.001), there is ample statistical evidence to reject H_0. Hence, the alternative hypothesis of the population means being different can be accepted with 95% certainty (and any other reasonable confidence level). As a result, **H1** is again accepted, there is a difference in satisfaction when no bot is involved versus when a bot is involved in a customer service interaction.

Results Conclusion

The logistic regression showed that Competence has a significant, negative effect in HumanOnly interactions. The effect is different when a bot is present as it goes from being negative to being neutral. Additionally, there is a statistical difference between chatbot interactions that are HumanOnly and Human+Bot. Hence, **H1** and **H2** are accepted, while **H2a** and **H2b** are rejected.

General Discussion

Through this research, two out of four created hypotheses were correct. The logistic regression and the independent groups t-test both led to the conclusion that interactions, where bots are involved, have a lower satisfaction level than where there is no bot at all. Furthermore, the logistic regression proved that the participation of a bot negatively moderates the likelihood of a customer becoming a "Promoter". These findings show that humans still have a vital role in customer service and that the technology is not ready to automatize the field yet. Additionally, the sampled data comparison being done between data from the same company and in the same period makes the results more robust as they reflect and showcase a comparable reality between the two methods within the field. Finally, the findings help set numbers for the differences in satisfaction and can lead to easier decision-making within a company.

The results regarding competence were against expectations. Even though H2 was accepted because the impact of competence differed between HumanOnly and Handover, the findings were that competence negatively impacted satisfaction when no bot was involved and

neutrally impacted satisfaction when a bot was involved, while the expectation was that it would impact HumanOnly interactions positively and Handover interactions negatively.

The reason for this expectation was twofold. Firstly, most literature about competence showed that it was an overall positive linguistic concept that should showcase how skilful an employee is and should lead to them solving a problem. Secondly, Sands and his colleagues' (2021) findings about scripts having differing effects between bots and humans inspired an opposing hypothesis for the Handover interactions. However, the findings were almost the opposite of the expectation.

One reason why competence could have had a negative impact in HumanOnly interactions is that an increased amount of competence is likely to be a consequence of an increased number of terms. This means that the conversation is longer, which in turn means that the problem is more complex and, hence, less likely to be solved in a quick and satisfying manner. The reason why the effect might not be the same for Handover interactions and is neutral might be because the problems not solved by the bot can be simple, but just not interpreted right by the algorithm. Hence, the customer service employee shows competence by answering relatively simple questions and leaves the customer more satisfied than another customer that directly spoke to a human agent and had a complex issue.

As a result, the findings about competence being a negatively impacting construct should not necessarily lead to the conclusion that customer service employees should avoid using competence as a linguistic feature. These findings open room for further research where the difficulty of an issue is controlled for. Regardless, the results support the mentioned findings by Sands and his peers (2021) and indicate that a company should consider differentiating between the manual handed out to the human service employees and the manual fed to the chatbot algorithm.

There are two central take-home points from this research. Firstly, one should be considerate when replacing humans with chatbots and expect a slight decrease in satisfaction ratings. Secondly, if one chooses to implement a chatbot, the script one feeds to the algorithm should be carefully reviewed and tested by the company itself to determine what is most effective.

Managerial Implications

Companies, just like the rest of the world, are becoming increasingly digital. Digitalization has shortened the distance between a company and a customer, especially the communication distance. Not only is it easier to contact a company, but customers expect a quick response time, where 90% see instant response time (10 minutes) as either crucial or very important (SuperOffice, 2023). 46% expect an answer within by 4 hours and 12% expect one within 15 minutes (SuperOffice, 2023). Keeping up with these expectations is neither easy nor cheap and handing over a part of this job to a computer is a natural step for many companies, but should managers resort to this?

This research showcases that bot-involving chat interactions result in a 0.3-point lower satisfaction on average. Managers should know the monetary value of losing a satisfaction point and base the implementation on the overall impact on company profits long term. It is important that managers do not get short-sighted and forget the value of customer loyalty that is built through positive company-customer interactions. After all, the wanted average is above 9, as it is those that give nines and tens that become "Promoters" of your brand.

It is also important to remember that there was always a human to resolve issues the chatbot was not able to solve. This means that if customer service is solely left to an algorithm, the satisfaction ratings would be even worse. Managers should, therefore, be especially careful if they plan to hand the customer service chat to AI. Supervision is vital and there should always be a system in place that acts as a security net when the AI fails.

Additionally, linguistic features incorporated in the script that is fed to the algorithm should be reviewed. Managers should advise and urge the department responsible for developing the customer service manuscript to run several smaller tests to find what manuscript produces the highest satisfaction level. A/B tests to smaller samples is one approach.

All in all, this research has shown that AI is not too far from performing at the level of a human being, but it still moderates satisfaction negatively. Managers must, therefore, keep in mind that employing AI will result in reduced satisfaction and can hurt the bottom line and the brand reputation in the long run. Hence, it is important that they thoroughly consider and test the script fed to the algorithm to minimize the loss in satisfaction. All managers should

economically measure the impact of lost satisfaction ratings and urge their customer service department to test and update the chatbot script.

Future Research

Firstly, future research should do an experimental approach. I would suggest developing several chatbots by feeding them specific scripts and then measuring the satisfaction and looking for differences. By doing this, the researcher will be able to control the amount of a linguistic feature. As the interaction will have no human involvement, it will be possible to measure the exact impact of a linguistic construct on satisfaction in the chatbot field.

Secondly, future research should compare chatbot-only data with human-only data. The data used in this research indicated a difference between HumanOnly and bot-involving interactions. However, all conversations included a human and it is not unlikely that the difference in satisfaction would be even greater in if no human was in place to perform service recovery after the bot was unable to assist with a task.

Thirdly, as mentioned in the general discussion, future research could produce research on competence where the complexity of the issue is predetermined and, therefore, controlled for.

Additionally, future research should continue the research on different linguistic features' impact on customer satisfaction. This research indicated that linguistic concepts do not have equal effects when a bot is involved versus when it is not, but there is a need for further research to determine how the researched concepts for human customer service perform when coming from a bot. Concreteness, personal pronouns, and warmth are just some suggestions for what could be researched.

Finally, using the managerial implications, future research should focus on developing econometric models to calculate the impact lowered satisfaction levels in customer service have on the bottom line. Exact numbers on gains and losses will make investment allocation within customer service an easier job for managers.

Bibliography

- Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*, 557-590.Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-chatbot/
- BasuMallick, C. (2022, May 25). *Chatbot Working, Types, Examples - Spiceworks*. Spiceworks; https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-chatbot/
- Berger, J., Moe, W. W., & Schweidel, D. A. (2023). EXPRESS: What Holds Attention? Linguistic Drivers of Engagement. Journal of Marketing, 0(ja). https://doi.org/10.1177/00222429231152880
- Castelo, N., Bos, M. W., & Lehmann, D. (2019). Let the machine decide: When consumers trust or distrust algorithms. *NIM Marketing Intelligence Review*, *11*(2), 24-29.
- Crolic, C., Thomaz, F., Hadi, R., & Stephen, A. T. (2022). Blame the bot: anthropomorphism and anger in customer–chatbot interactions. *Journal of Marketing*, *86*(1), 132-148.
- Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype conte
- Cuddy, A. J., Fiske, S. T., Kwan, V. S., Glick, P., Demoulin, S., Leyens, J. P., ... & Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British journal of social psychology*, *48*(1), 1-33.
- De Haan, E., Verhoef, P. C., & Wiesel, T. (2021). Customer Feedback Metrics for Marketing Accountability. In *Marketing Accountability for Marketing and Non-marketing Outcomes* (Vol. 18, pp. 49-74). Emerald Publishing Limited.
- DeWitt, T., Nguyen, D. T., & Marshall, R. (2008). Exploring customer loyalty following service recovery: The mediating effects of trust and emotions. *Journal of service research*, *10*(3), 269-281.
- Eren, B. A. (2021). Determinants of customer satisfaction in chatbot use: evidence from a banking application in Turkey. International Journal of Bank Marketing.
- Eyssel, F., & Hegel, F. (2012). (s) he's got the look: Gender stereotyping of robots 1. *Journal of Applied Social Psychology*, *42*(9), 2213-2230.
- Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

- Gelbrich, K. (2010). Anger, frustration, and helplessness after service failure: coping strategies and effective informational support. *Journal of the Academy of Marketing Science*, *38*, 567-585.
- Goldberg, P. (2022). *Studies show humans prefer interacting with hyper-realistic human-like robots*.
- Golder, P. N., Dekimpe, M. G., An, J. T., van Heerde, H. J., Kim, D. S., & Alba, J. W. (2022). Learning from Data: An Empirics-First Approach to Relevant Knowledge Generation. *Journal of Marketing*.
- Gorvett, Z. (2018, May 31). *How humans bond with robot colleagues*. Bbc.com; BBC. https://www.bbc.com/worklife/article/20180530-how-humans-bond-with-robot-colleagues
- Güntürkün, P., Haumann, T., & Mikolon, S. (2020). Disentangling the differential roles of warmth and competence judgments in customer-service provider relationships. *Journal of Service Research*, *23*(4), 476-503. HERE. https://www.here.com/learn/blog/how-human-like-will-the-robots-of-the-future-need-to-be
- Ipsos. (2016, May 15). *Ipsos Encyclopedia - Attitudinal Loyalty*. Ipsos. https://www.ipsos.com/en/ipsos-encyclopedia-attitudinal-loyalty
- Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion.
- Kull, A. J., Romero, M., & Monahan, L. (2021). How may I help you? Driving brand engagement through the warmth of an initial chatbot message. *Journal of business research*, *135*, 840-850.
- LIWC (2022). Linguistic Inquiry and Word Count: LIWC-22. [PDF].
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90-103.
- Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). *Resistance to medical artificial intelligence*. Journal of Consumer Research, 46(4), 629-650.
- Malhotra, N. K. (2019). *Marketing Research: An Applied Orientation, Global Edition* (7th ed., p. 888). Pearson.
- McKinsey & Company. (2021). *A new growth story: Maximizing value from remote customer interactions*. McKinsey & Company.

https://www.mckinsey.com/capabilities/operations/our-insights/a-new-growth-story-maximizing-value-from-remote-customer-interactions

- Mende, M., Scott, M. L., van Doorn, J., Grewal, D., & Shanks, I. (2019). Service robots rising: How humanoid robots influence service experiences and elicit compensatory consumer responses. *Journal of Marketing Research*, *56*(4), 535-556.
- Ozuem, W., Ranfagni, S., Willis, M., Rovai, S., & Howell, K. (2021). *Exploring customers' responses to online service failure and recovery strategies during Covid-19 pandemic: An actor–network theory perspective*. Psychology & Marketing, 38(9), 1440-1459.
- Packard, G., & Berger, J. (2021). How concrete language shapes customer satisfaction. Journal of Consumer Research, 47(5), 787-806.
- Proctor II, R.F. & Wilcox, J.R. (1993). An exploratory analysis of responses to owned messages in interpersonal communication. *A Review of General Semantics, 50(2*), 201-220. https://www.jstor.org/stable/42577446
- Quach, K. (2022, September). Goodbye, humans: Call centers "could save $80b" switching to AI. Theregister.com; The Register. https://www.theregister.com/2022/09/01/call-center-ai-gartner/
- Salesforce. (2023). *Why net promoter score (NPS) is important and how to use it*. Salesforce. https://www.salesforce.com/eu/learning-centre/customer-service/calculate-net-promoter-score/#:~:text=facts%20and%20FAQs-,What%20is%20the%20Net%20Promoter%20Score%3F,to%20a%20friend%20or%20acquaintance.
- Singh, H. (2006). The importance of customer satisfaction in relation to customer loyalty and retention. *Academy of Marketing Science*, *60*(193-225), 46.
- SuperOffice. (2023). *7 Ways to Reduce Customer Service Response Times*. Superoffice.com. https://www.superoffice.com/blog/response-times/#:~:text=Nearly%20half%20of%20all%20customers,they%20need%20customer%20service%20assistance
- Titlebaum, H. (1998). Relaxation. *Journal of Evidence-Based Integrative Medicine, 4(2),* 123 146. https://doi.org/10.1177%2F153321019800400206
- Villarroel Ordenes, F., & Zhang, S. (2019). From words to pixels: text and image mining methods for service research. *Journal of Service Management*, *30*(5), 593-620.

- Welchlin, K. (2017, 4. Aug). Interpersonal Communication: What Are Confirming Messages?. *Welchlin*. https://welchlin.com/interpersonal-communication-what-are-confirming-messages/#:~:text=There%20are%20three%20main%20categories,value%20of%20the %20other%20person.
- Weply. (2021). Do you prefer chatbots over humans? Weply. https://weply.chat/blog/do-you-prefer-chatbots-over-humans#:~:text=As%20mentioned%20before%20only%200,had%20bad%20experiences%20with%20chatbots
- Whalen, J. (2017, September 29). *"Goodbye, old friend": CBC sends off mail robots*. CBC. https://www.cbc.ca/news/canada/toronto/goodbye-mailbots-party-1.4313207
- Yüksel, A., & Rimmington, M. (1998). Customer-satisfaction measurement. The Cornell Hotel and Restaurant Administration Quarterly, 39(6), 60-70. https://doi.org/10.1016/S0010-8804(99)80007-X

Summary

This Master Thesis explores the impact of chatbots on customer satisfaction and the effectiveness of linguistic features in chatbot interactions. The use of chatbots in customer service is becoming increasingly common, with companies using them to save on costs and quickly assist customers in need. McKinsey and Company (2021) report that a well-constructed contact center can save 5-10% in channel costs for a company. Additionally, it is projected that using chatbots will save companies up to $80 billion by 2026 (Quach, 2022).

However, chatbots need to be trained, and it is important to understand what leads to customer satisfaction in such interactions. Previous research has found that factors such as concreteness, competence, empathy, and active listening play a crucial role in customer service interactions.

One identified gap in the literature is the need to explore the differences in satisfaction levels when a customer interacts with a human versus when a bot is present. This thesis aims to address this gap by comparing satisfaction levels between human-only interactions and interactions that also involve a bot.

The thesis also aims to examine the effect of a linguistic feature, competence, on customer satisfaction when interacting with a chatbot. This research will focus on competence as there has not been any research directly tackling competence within chat customer service and chatbots, giving a unique opportunity to test a new linguistic feature in both types of customer service and compare its impact.

The research contributes to the existing literature by (1) examining the satisfaction level differences when interacting with humans versus bots and (2) testing the effect of competence on customer satisfaction when talking with chatbots.

The two research questions to be answered through this Master Thesis are:

1. Do customers prefer human-only or bot-initiated interactions in the customer service field?
2. Does competence have the same impact on the Net Promoter Score (NPS) when a bot is involved in the customer service interaction versus when not?

The literature review had five main components: chatbots, relations between humans and bots, competence, service recovery, and attitudinal loyalty.

Chatbots are applications that assist with customer service, engagement, and support by using artificial intelligence (AI) and automation technologies to communicate with users via chat. This research focuses on chatbots within customer service, which can range from a simple FAQ bot to an actively learning, AI chatbot that improves based on previous conversations. Despite significant resources being invested in chatbot development, people are still hesitant to adopt them. Sheehan et al. (2020) found that error-free chatbots are preferred, but chatbots that ask for clarification also elicit higher adoption willingness. Additionally, customers who prefer human interaction prefer higher anthropomorphism. Kull et al. (2021) found that a warmer tone in conversations initiated by a chatbot elicits the highest engagement, followed by a competent tone and a neutral tone. Sands et al. (2021) found that chatbots perform well with an entertaining service script and are not outperformed by human customer service employees.

Humans tend to anthropomorphize robots, assigning human-like attributes and behaviors to them, even going so far as to hold ceremonies for them (Gorvett, 2018). This phenomenon is driven by the degree to which robots can mimic human behavior and is supported by studies showing that humans assign gender stereotypes and personalities to robots (Eyssel & Heggel, 2012). However, the degree to which humans prefer robots over human interaction varies depending on the context. For example, customers may prefer human interaction when they strongly identify with a product (Leung et al., 2018), while they may be more willing to accept medical advice from an AI (Longoni et al., 2019). Additionally, studies have found that higher anthropomorphism is not always beneficial and that algorithms may be viewed positively or negatively depending on factors such as their performance and human involvement (Jussupow et al., 2019). Ultimately, businesses should consider segmenting their use of robots based on customer preferences and directly comparing the impact of human employees and robots on customer satisfaction.

Competence is a fundamental concept in the Stereotype concept model (SCM), measuring one's capacity to act on a certain intention (Cuddy et al., 2008). It is synonymous with confidence, capability, and skillfulness, and has a universal stereotype across cultures. In

customer service, providing reasoning for suggested actions, or "polite argumentation" is a way of demonstrating competence and has been shown to increase customer satisfaction. This approach involves supplying the customer with a rationale for suggested solutions, such as explaining why resetting network settings is a safe and effective way to improve phone service. "Polite argumentation" has been proven to positively impact transactional outcomes, such as increased share of wallet (Okumus & Unal, 2012; JananJohnson et al., 2014).

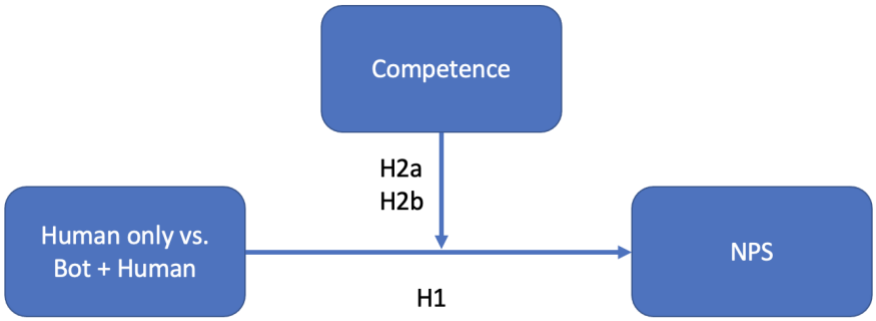
To keep customers coming back, it's essential to provide good service recovery when mistakes happen. Yüksel and Rimmington (1998) showed that customer satisfaction leads to attitudinal loyalty, and DeWitt et al. (2017) found that effective service recovery processes increase satisfaction levels. Gebrich (2010) found that explaining how the service failure occurred and what was done to prevent it can replace monetary compensation, and Ozuem et al. (2021) identified empathizers as a group that responds well to recovery efforts. Competent service recovery means being accurate, quick, and creating empathy to reach customers effectively.

Attitudinal loyalty is essential for companies as it leads to positive attitudes towards the company and benefits of the doubt in case of service failures or brand crises. Customer satisfaction resulting in positive Word of Mouth is another reason why it is crucial. Net Promoter Score (NPS) (De Haan et al., 2021) is one of the best ways to measure Word of Mouth. NPS is segmented into detractors (0-6), neutrals (7-8), and promoters (9-10) based on scores from 0 to 10. Companies can improve attitudinal loyalty by analyzing NPS scores for themes and patterns that lead to a high or low score. Analyzing these scores helps companies improve customer satisfaction and ultimately their bottom line.

With the literature review as a basis, two hypotheses with two sub hypotheses for the second hypothesis were formed:

- **H1:** *Customer satisfaction is higher when no bot is involved.*
- ***H2:*** *Linguistic competence has a different effect for human, vs human+bot interactions such as:*
    - ***H2a:*** *Linguistic competence has a positive effect for human only interactions.*
    - ***H2b:*** *Linguistic competence has a negative effect for bot+human interactions.*

And, followingly, a conceptual model was created:



The programming of chatbots varies, but the key function is to match customer inquiries with the correct information in a database, using an intent hierarchy to continuously improve. This process involves feeding the chatbot new information based on the analysis of previous conversations. The data for analysis is taken from a Norwegian telecom company, consisting of 5,357 customer-company chat interactions that began with a chatbot and 5,005 interactions that were human-only. The dataset includes the textual conversation from both parties and the Net Promoter Score (NPS) for each interaction.

The analysis of the data uses KNIME workflows to annotate the data and to compare the differences between human-only and bot-involving interactions. Competence is measured using a top-down approach, involving self-made dictionaries from the Linguistic Inquiry and Word Count (LIWC) and knowledge obtained from reading thousands of conversations. The dictionaries include words that indicate argumentation, certitude, and knowledge of customer issues. The human-only and handover data are concatenated and grouped by conversation ID, then turned into a document to separate the two datasets. The customer satisfaction ratings per conversation are then compared using the «independent groups t-test» node.

Competence is examined in each conversation by plugging a self-made dictionary into the workflow and creating a bag of words to calculate term frequency (TF) and inverse document frequency (IDF). A new categorical variable called "Categorical NPS" is created to rectify the issue that the NPS does not have a normal distribution. The logistic regression model is used to measure the impact of competence on the dependent variable: Categorical NPS.

The dataset's unstructured nature (Balducci & Marinova, 2018) presented a challenge, as there was no automatic way of determining the sender of a message in the handover dataset. This task was performed manually by two coders, who spent approximately 60 hours each going through the data, giving qualitative insights in addition to the quantitative insights obtained later using text mining.

The process for a conversation involving a chatbot typically involved the customer interacting with the chatbot until either they requested to talk to a human or the chatbot suggested it. The conversation was then handed over to a customer service employee who worked with the customer until a solution was found. After the chat was closed, the customer had the option to leave a score for the interaction.

KNIME workflows were used to annotate the data and insert dictionaries to measure the level of competence. The analysis of the data provided both qualitative and quantitative insights into the performance of the chatbot and the effectiveness of the handover process from the chatbot to the human representative.

A top-down approach (Villarroel Ordenes & Zhang, 2019) using self-made dictionaries from LIWC (2022), and knowledge gained from reading thousands of conversations is used to measure the data. A list of words indicating argumentation and certainty was selected and added to the dictionary, totaling 153 words.

The data was split into human-only and handover data, and a binary column was created to differentiate between them. The datasets were then concatenated, and customer satisfaction ratings per conversation were compared using the independent groups t-test node. To measure competence within each conversation, a self-made dictionary was used to create a bag of words that calculated term frequency (TF) and inverse document frequency (IDF). A new categorical variable was created to analyze the impact of competence on customer satisfaction, using logistic regression. The variable “bot” was also created to determine how it moderates the likelihood of a customer being a “Promoter.” Control variables were added to ensure consistent reference groups. This procedure was done for both datasets, and the results were recorded.

Ten additional variables are used as control and reference categories, including "Owning statements," "Competence," "Confirmation," "Anxiety," "Number of terms," and

"Bot," which is a binary variable created to distinguish between the HumanOnly and Handover datasets. TF and IDF are used to measure these constructs based on a dictionary and are multiplied by the "Bot" variable to assess them in both datasets. Owning statements use first-person pronouns to acknowledge personal responsibility, while confirmation indicates agreement, and anxiety is the absence of relaxation, which weakens communication. "Bot*Owning," "Bot*Confirm," "Bot*Competence," and "Bot*Terms" are the remaining constructs that are measured.

Before running the tests, descriptive statistics were obtained:

Variable	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness
Owning statements	0,00	0,18	0,05	0,02	0,00	1,15
Competence	0,00	0,22	0,02	0,02	0,00	0,75
Confirmation	0,00	0,14	0,01	0,01	0,00	1,39
Anxiety	0,00	0,04	0,00	0,00	0,00	10,91
Number of terms	7,00	3883,00	205,59	165,54	27402,62	4,35
NPS	0,00	10,00	8,17	3,40	11,53	-1,72
Bot	0,00	1,00	0,52	0,50	0,25	-0,07
Bot*Owning	0,00	0,15	0,03	0,03	0,00	0,85
Bot*Confirm	0,00	0,14	0,01	0,01	0,00	2,53
Bot*Competence	0,00	0,22	0,01	0,02	0,00	1,47
Bot*Anx	0,00	0,04	0,00	0,00	0,00	13,84
Bot*terms	0,00	2168,00	99,17	144,06	20752,72	2,63
Dataset	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**	**Variance**	**Skewness**
NPS HumanOnly	0,00	10,00	8,33	3,29	10,81	-1,87
NPS Handover	0,00	10,00	8,02	3,50	12,22	-1,59

The logistic regression had “Categorical NPS” as the binary dependent variable. Variables excluded from the test were: ID, NPS (numerical 0-10), and bot*owning. To calculate the effect and make it more interpretable, all coefficients were exponentiated where 1 means no effect, while anything below represents a negative effect and anything above represents a positive effect. The formula [*Exp(coefficient)-1]*100%* will give the change in odds. The confidence level of the test was 95%.

Logit	Variable	Coefficient	Std. Error	z-score	P >|z|	Beta-coefficient
Promoter	Owning statements	-0,433	0,027	-16,286	0,000	0,649
Promoter	Competence	-0,163	0,038	-4,247	0,000	0,850
Promoter	Confirmation	0,639	0,045	14,195	0,000	1,894
Promoter	Anxiety	0,217	0,077	2,828	0,005	1,243
Promoter	Number of terms	-0,113	0,032	-3,489	0,000	0,893
Promoter	Bot	-0,145	0,054	-2,664	0,008	0,865
Promoter	Bot*Confirm	-0,065	0,053	-1,213	0,225	0,937
Promoter	Bot*Competence	0,044	0,051	0,854	0,393	1,045
Promoter	Bot*Anx	0,045	0,082	0,551	0,582	1,046
Promoter	Bot*terms	-0,061	0,041	-1,498	0,134	0,941
Promoter	Constant	1,193	0,026	46,687	0,000	3,296

With logistic regression, **H1** is accepted because the variable statistically significantly decreases the likelihood of a customer being a promoter. In fact, the customer is 13.5% less likely to become a promoter if a bot is involved in the interaction. Followingly, **H2** is accepted because the effect differs between HumanOnly and Handover (negative vs neutral). **H2a** is rejected because the impact of competence is negative towards making a customer a promoter. Since the effect of competence for Handover interactions was not significant at the 95% confidence level, it is not statistically different from 0. **H2b** is rejected.

Then an independent groups t-test was performed. The data for HumanOnly and Handover were separately filtered and given either 1 (Handover) or 0 (HumanOnly). The two datasets were then concatenated into a single dataset and put through the test to test the difference between the mean NPS ratings (range: 0-10). The confidence interval for the test was 95%.

Variable	Group	N	Mean	Standard Deviation	Standard Error Mean
NPS	1	5357	8,016	3,495	0,048
NPS	0	5005	8,331	3,288	0,046

The difference in means was statistically significantly different from 0. In fact, the difference in means was 0.315 and the p-value was below 0.05 (p-value < 0.001). As a result, **H1** is again accepted, there is a difference in satisfaction when no bot is involved versus when a bot is involved in a customer service interaction.

All in all, **H1** and **H2** are accepted, while **H2a** and **H2b** are rejected.

This research aimed to investigate the impact of chatbots on customer satisfaction and competence in customer service interactions. Two out of four hypotheses were proven correct, indicating that interactions involving bots have a lower satisfaction level than those without bots and that the participation of a bot negatively moderates the likelihood of a customer becoming a “Promoter”. The findings highlight the vital role that humans still play in customer service and suggest that the technology is not yet ready to fully automate the field.

Surprisingly, the results regarding competence were against expectations. The findings showed that competence negatively impacted satisfaction when no bot was involved and

neutrally impacted satisfaction when a bot was involved. The reason for this unexpected finding might be that an increased amount of competence is likely to result in longer conversations, making the problem more complex and hence less likely to be solved quickly and satisfactorily. However, the effect might not be the same for interactions involving bots because the problems not solved by the bot may be simpler, making the customer receive a lot of competence in the texts while solving simple problems.

The take-home message from this research is that companies should be considerate when replacing humans with chatbots and expect a slight decrease in satisfaction ratings. And if one chooses to implement a chatbot, the script fed to the algorithm should be carefully reviewed and tested by the company to determine what is most effective. Managers should also remember that customer loyalty is built through positive company-customer interactions, and the monetary value of losing a satisfaction point should be known and considered before using a chatbot solution. Employing AI may result in reduced satisfaction and can hurt the bottom line and brand reputation in the long run. Therefore, it is important to thoroughly consider and test the script fed to the algorithm to minimize the loss in satisfaction and create econometric models that estimate the monetary losses of NPS points.

There are five suggestions for future research coming from this thesis. Firstly, researchers should consider taking an experimental approach, creating several chatbots with specific scripts and measuring satisfaction levels to identify differences. Secondly, it is important to compare chatbot-only data with human-only data to identify any differences in satisfaction levels where no human is present to save the interaction if the chatbot fails. Thirdly, future studies could measure and control for the complexity of an issue beforehand for a more rigorous approach. Fourthly, future research should examine the impact of more linguistic features on customer satisfaction. Lastly, developing econometric models to identify the economic impact of lowered satisfaction levels on the bottom line can help managers allocate investments effectively. By identifying the exact impact in terms of gains and losses, managers can make informed decisions about where to invest resources in customer service.