Norwegian University
of Life Sciences

Master's Thesis 2020 30 ECTS
School of Economics and Business

Uncovering Lost Potential: The Shortcomings of DNBs Chatbot

Cecilie Augensen Nilsen
Master in Business Administration

Acknowledgement

This thesis concludes a two year Master of Science in Business Administration at the Norwegian University of Life Sciences (NMBU). The thesis consists of 30 credits and was written in the Spring semester 2020.

First, I would like to thank my supervisor Mike Riess for always challenging me and for his valuable guidance and feedback during this process. Even when some aspects of this thesis seemed overwhelming, Mike always guided me back on track.

I would like to thank the team at DNB IT Emerging Technologies for the opportunity to write this thesis and for all the support. Thanks to Yahya Maweed for his input and knowledge in the process of delineating this study. A special thanks to Kamal Singh and Mateo Caycedo Alvarez for help with the Python programming codes and for sharing their knowledge.

Thanks to my family and friends who dealt with my stress and frustration throughout the process. A special thanks to my sister, Christine, for every encouraging words and knowledge. I am forever grateful for her positive energy and love.

Finally, I would like to thank Collegium Alfa and Studentsamfunnet i Ås for making my years as a student at NMBU absolutely amazing. These memories will last a lifetime.

Cecilie Augensen Nilsen

Hvaler, May 2020

Abstract

In 2018 DNB Bank ASA (DNB) launched their chatbot, Aino, an advanced virtual banking agent. Aino handles 55% of all the incoming chat traffic for DNBs Customer Center and is continuously being trained by AI trainers to increase the percentage of messages it can respond to. The former CEO of DNB, Rune Bjerke, stated in 2017 that by 2020, 80% of all incoming chat traffic would be handled by chatbots. However, to get closer to this target, DNBs AI trainers will have to make some priorities in the development process.

The purpose of this study is to contribute to the decision-making process of which types of problems, and intents the AI trainers should prioritize to reduce DNBs costs. The data basis is conversational logs from conversations between customers of DNB and Aino, in addition to structural interviews with four DNB employees with significant knowledge of Aino. This thesis is a mixed-methods study that consists of both statistical analyses to determine group effect, structured interviews, quantitative content analysis, statistical analyses of chatlogs, as well as analysis of economical impact.

The results from the statistical analyses reveal that “Insurance” and “Funds” are the two top-level intents with the weakest performance, where “Insurance” had the weakest overall performance. In contrast, “Funds” had the weakest performance in both classification accuracy and customer satisfaction score. From the structured interviews and quantitative content analysis, eight factors were selected to be tested if the factors had a significant effect on classification performance. Of the eight tested factors, only the number of words, language, and typos in the customer’s messages seem to have a significant effect. With the assumption that an increase in correct predictions can result in an accepted higher automation rate, the results indicated that a 5% increase in correct predictions caused significantly lower costs. The analysis concluded that an increase for “Insurance” performance would give higher cost reductions than an equal increase for “Funds.” However, the amount of resources required to gain a 5% increase is assumed to be significantly less for “Funds.”

Aino has been a great success for DNB, but there is still potential areas of improvement, which could have a significant impact for the company. Further research with larger sample size and a financial model that includes several aspects of Aino´s business case is suggested.

Sammendrag

I 2018 lanserte DNB Bank ASA (DNB) sin chatbot, Aino, en avansert virtuell bankagent. Aino håndterer 55% av all innkommende chat-trafikk for DNBs kundesenter og blir kontinuerlig opplært av AI-trenere for å øke prosentandelen av meldinger den kan svare på. Den tidligere konsernsjefen i DNB, Rune Bjerke, uttalte i 2017 at innen 2020 ville 80% av all innkommende chat-trafikk bli håndtert av chatbots. For å komme nærmere dette målet, vil DNBs AI-trenere imidlertid måtte gjøre noen prioriteringer i utviklingsprosessen.

Hensikten med denne studien er å bidra til beslutningsprosessen for hvilke typer problemer, og intensjoner AI-trenerne bør prioritere for å redusere DNBs kostnader. Datagrunnlaget er samtalelogger fra samtaler mellom kunder av DNB og Aino, i tillegg til strukturerte intervjuer med fire DNB-ansatte med betydelig kunnskap om Aino. Denne oppgaven er et kombinasjonsstudie som består av både statistiske analyser for å bestemme gruppeeffekt, strukturerte intervjuer, kvantitativ innholdsanalyse, statistisk analyse av chatlogger, i tillegg til analyse av finansiell påvirkning.

Resultatene fra de statistiske analysene viser at "Forsikring" og "Fond" er de to intensjonene på topp-nivå med den svakeste ytelsen, der "Forsikring" hadde den svakeste samlede ytelsen. Derimot hadde "Fond" den svakeste prestasjonen både i klassifiseringsnøyaktighet og kundetilfredshet. Fra de strukturerte intervjuene og den kvantitativ innholdsanalysen ble åtte faktorer valgt for å testes om faktorene hadde en signifikant effekt på klassifiseringsevnen. Av de åtte testede faktorene ser det bare ut til at antall ord, språk og skrivefeil i kundens meldinger har en signifikant effekt. Med antakelsen om at en økning i riktige prediksjoner kan føre til en akseptert høyere automatiseringsgrad, indikerte resultatene at en økning på 5% i riktige prediksjoner forårsaket betydelig lavere kostnader. Analysen konkluderte med at en økning for "Forsikrings" prestasjoner ville gi høyere kostnadsreduksjoner enn en lik økning for "Fond." Imidlertid antas mengden av ressurser som kreves for å oppnå en økning på 5% å være betydelig mindre for "Fond."

Aino har vært en stor suksess for DNB, men det er fortsatt potensielle områder for forbedring, som kan ha en betydelig økonomisk innvirkning for selskapet. Det foreslås ytterligere forskning med større utvalgsstørrelse og en finansiell modell som inkluderer flere aspekter av Ainos kostnadsbilde.

Table of content

1. Introduction

In recent years, financial technology, known as "fintech," has become a significant area of development and competition within the financial sector. The term comes from startups competing with traditional financial organizations by giving customers speed and flexibility in their financial services (Nicoletti, Nicoletti, & Weis, 2017). For traditional financial organizations to keep up with startups they need to develop new technological solutions so that they can free up staff while still supplying their customers with better services and support and keeping costs low.

DNB Bank ASA (DNB) is Norway's largest financial services group and one of the largest in the Nordic region in terms of market capitalization. DNB offers a full range of financial services, including loans, savings, insurance and advisory services for both retail and corporate customers. The bank has 2 100 000 retail customers in Norway, where 1 500 000 are active Internet bank users and 1 060 000 use the mobile banking services actively (DNB Bank ASA, 2020).

Digitalization permeates every aspect of DNBs services, where one example is DNBs ongoing development within customer service. Even though customers still can contact customer service by phone or visit a local office, a substantial amount of interactions between customers and DNB is now through a chat-window on DNBs website. Some of these interactions include questions that are so complex or unusual that they need responses from a human agent. Yet, most questions are repetitive and straightforward and can be solved with the use of automation (Nordstrøm, 2019).

Aino is a chatbot, an advanced virtual banking agent, which is a software you can talk to in a chat interface (Stene, 2018). Aino is based on machine learning and artificial intelligence and its performance is continuously increased by the AI trainers that work with Aino. AI trainers are employees with backgrounds from customer support roles in DNB that are now working on increasing Aino’s performance (boost.ai, 2019b).

This thesis has been written in cooperation with DNB IT Emerging Technologies, the department within DNB that monitors and improves Aino. The thesis has used logs of conversations between Aino and customers to get a deeper understanding of Aino´s performance and the financial impact of an increase in this performance.

1.1 Background

Chatbots have only recently made their entry into homes and phones around the world, but they are far from a recent invention. In 1966 Joseph Weizenbaum introduced the first chatbot ever made, Eliza. More than a decade later, the American company Apple, led by Steve Jobs, launched a chatbot revolution when they released the personal assistant Siri. Khan & Das explain the revolution in chatbots as a series of factors such as growth of internet users and advancement in technology, which has significantly increased availability of chatbots (Khan & Das, 2018). Artificial intelligence accelerated in the banking and financial industry in 2010, and chatbots transformed the way banks operate and deliver their services.

In 2017 *Jupiter Research* released a report which projected that chatbots will be responsible for over $8 billion annual cost savings by 2022 for Banking and Healthcare Sectors. This report also stated that chatbots would be handling 85% of all customer service interactions by 2020 (Woodford, 2020). Bank of America introduced their chatbot Erica in 2018 to send notifications, provide balance information and help customers with simple transactions. Since then Erica has expanded as an advanced virtual assistant, and can now send personalized recommendations, offers and advice after analyzing the customer data. Wells Fargo´s chatbot uses artificial intelligence and Facebook Messenger to provide customers with information about their account balance, most recent transactions, and the location of the nearest ATM, among others (Marous, 2018).

Out of the large Norwegian banks, SpareBank 1 SR-bank was the first to launch their chatbot. DNB started using chatbots within Facebook Messenger in cooperation with Convertelligence in 2017. The same year CEO Rune Bjerke stated that he thought chatbots would handle 80% of all incoming traffic within 2020 (Bakken, 2017). One year later DNB launched Aino in cooperation with Boost.ai. Within four months, Aino had automated 50% of all incoming chat traffic, and today Aino handles 55% of all the incoming traffic (boost.ai, 2019b). Aino has not yet reached Rune Bjerke's target and is still only capable of answering simple and repetitive customer questions. For Aino to follow the chatbots of the large American banks and provide a further personalized chat experience and increase the number of tasks it can perform, it is essential to investigate the current performance, what causes Aino´s underperformance and the financial impact of its underperformance for DNB.

1.2 Problem statement and research questions

The development and improvement of Aino is a continuous and ongoing task for DNB IT Emerging Technologies. Aino has contributed to significant cost reductions for DNB, but to increase resources for development, DNB IT Emerging Technologies needs to present new cost reductions for DNB management. This thesis seeks to contribute to the decision-making process of which types of problems, and intents the AI-trainers should prioritize to reduce costs most effectively. Intents are in this thesis defined as all the different topics that the customers might have questions about and that DNB wants the chatbot to give an answer. The overall problem statement is, therefore, formulated as:

What is the financial impact for DNB of an under-performing customer service chatbot?

This problem statement will be enlightened through answering the following three research questions:

1. *What is the level of performance across top-level intents?*
2. *What characterizes the cases where the chatbot underperform?*
3. *If performance were to be improved for the under-performing top-level intents, what is the estimated financial impact?*

The purpose of the first research question is to understand which top-level intents that under-performs, and should be prioritized for further development. The results from this research question will also build hypotheses to investigate in the analyses of the second research question.

The second research question's purpose is first to find key factors that affect the performance of the chatbot with structured interviews and quantitative content analysis. Secondly, this research question seeks to give estimates for these factors effect on Aino's performance.

The third research question aims to estimate the financial impact of an increase in the number of correct predictions. Answering this research question will indicate which top-level intents to prioritize if the goal is to reduce costs.

With regards to the first research question, this thesis argues that Aino´s performance is significantly affected by top-level intent. The top-level intent “Funds” seems to under-perform at both model classification performance and customer satisfaction score, while “Insurance” has the weakest overall performance. The findings will show that the number of words, language and typos have a significant effect on the model classification performance. This thesis argues that even though an increase in “Insurance” performance would give a higher cost reduction, “Funds” might require less resources to gain an equal increase in performance.

2. Related theory

In this chapter, the theory and previous research that is relevant for the problem statement and the three research questions will be presented. The chapter is separated into eight parts: First, machine learning is presented, followed by Natural Language Processing. Then, chatbot is explained, followed by a section explaining intents and their hierarchy. Then, previous research concerning customer satisfaction is presented, followed by a section about Natural Language Processing Classification challenges. Last, a section explaining how performance is defined due to the use in the problem statement, and a section explaining cost- and confusion matrix.

2.1 Machine Learning

Machine learning is the capability that enables artificial intelligence systems to acquire their knowledge, by extracting patterns from data (Goodfellow, Bengio, & Courville, 2016). The introduction of machine learning enabled computers to solve problems that required real-world knowledge and make subjective decisions. Machine learning evolved as a subfield of artificial intelligence. It uses self-learning algorithms in order to build predictions, instead of requiring humans to derive rules and build models manually. Machine learning can be divided into three types; supervised learning, unsupervised learning, and reinforcement learning (Raschka & Mirjalili, 2017).

Supervised learning is when the model trains on a labeled dataset. The outcome variable guides the learning process, and therefore the process is supervised. Examples of supervised learning algorithms are linear regression, logistic regression, nearest neighbor, decision tree, and random forest (Friedman, Hastie, & Tibshirani, 2001).

Unsupervised learning is training an algorithm using datasets that have no labels or targets, where the algorithm is acting on the information without guidance. Examples of unsupervised learning algorithms are clustering and association (Friedman et al., 2001).

In reinforcement learning, the goal is to develop a system that improves its performance based on interactions with the environment which maximizes the reward. There are two types of reinforcement, positive and negative, where the reward of the event defines the type (Friedman et al., 2001).

2.2 Natural Language Processing

Dr. Michael J. Garbade described Natural Language Processing (NLP) as “the technology used to aid computers to understand human’s natural language”. NLPs’ ultimate objective is to read and understand human language in a valuable manner. NLP is the branch of artificial intelligence that makes it possible for computers to read the text, hear speech, and extract central information (Garbade, 2018).

Language is a highly unstructured data source. There are hundreds of languages and dialects, unique sets of grammar and syntax rules, terms, and slang. If that was not enough, humans often misspell words or fail with grammar. When humans speak, we have accents, we mumble, stutter, and mix terms from different languages (Garbade, 2018).

2.3 Conversational Agents

Khan and Das define conversational agents (chatbots) as “a computer program that processes natural-language input from a user and generates smart and relative responses that are then sent back to the user”. Chatbots are powered by rules-driven engines or by artificial intelligence engines. Chatbots can be used individually on a business’s website or popular platforms like Facebook Messenger, Slack, or Skype. Most often, the chatbot is a text-based interface where the user sends text messages to the chatbot and gets a text-message reply, even though voice-based assistants use some of the same technology (Khan & Das, 2018).

Chatbot increases customer experience by streamlining interactions between customers and services. Also, by reducing the costs of customer service, chatbots offer new opportunities to improve customer engagement processes and operational efficiency. A successful chatbot can effectively perform both the improvement of customer engagement and operational efficiency (Expert System, 2018).

Development of chatbot since Eliza, the first chatbot ever made, has included several design approaches. The design approaches often determine the chatbot's purpose, and have evolved to align with the current needs of the market. The early chatbots like Eliza and Parry used pattern matching, where on the other hand, recently developed chatbots use Long Short Term Memory Network or Sequence to Sequence Neural Network Model (Ramesh, Ravishankaran, Joshi, & Chandrasekaran, 2017). For competitive reasons Boost.ai does not share Aino’s design approach.

2.4 Intents

Aino´s classification model can sort customer messages based on the question type, called intent. Standard terminology in the literature is classes instead of intents, but this thesis will use intents to simplify the connection to boost.ai data.

Each message from a customer is classified according to its intent. The classification of intents allows for different formulations of questions, as long they share some similarities.

The AI trainers in DNB create new intents. When a new intent is created, the AI trainer also generates training data for the new intent and synonyms for that intent. The training data should include different ways a customer can ask about the intent, without any typos. Each intents training data consist of between 20 and 25 sentences. After the training data is created, another AI trainer creates test data for the new intent. The test data consist of about ten sentences, which can also include typos.

The AI trainers have created a total of 2929 intents, and they are structured in a hierarchical structure that will be described in 2.4.1.

2.4.1 Intent hierarchy

In the product package from boost.ai, the company created a hierarchy tree to create a structure for all the different intents. A hierarchy tree is a way of representing data where the elements relate to “branches,” and each item only has one “parent.” The intent hierarchy is enormous because of the number of intents. Figure 1 shows some of the intents connected to the top-level intent “Pension.” As you can see from the figure, this example has five levels, which is the correct number of levels for this top-level intent.

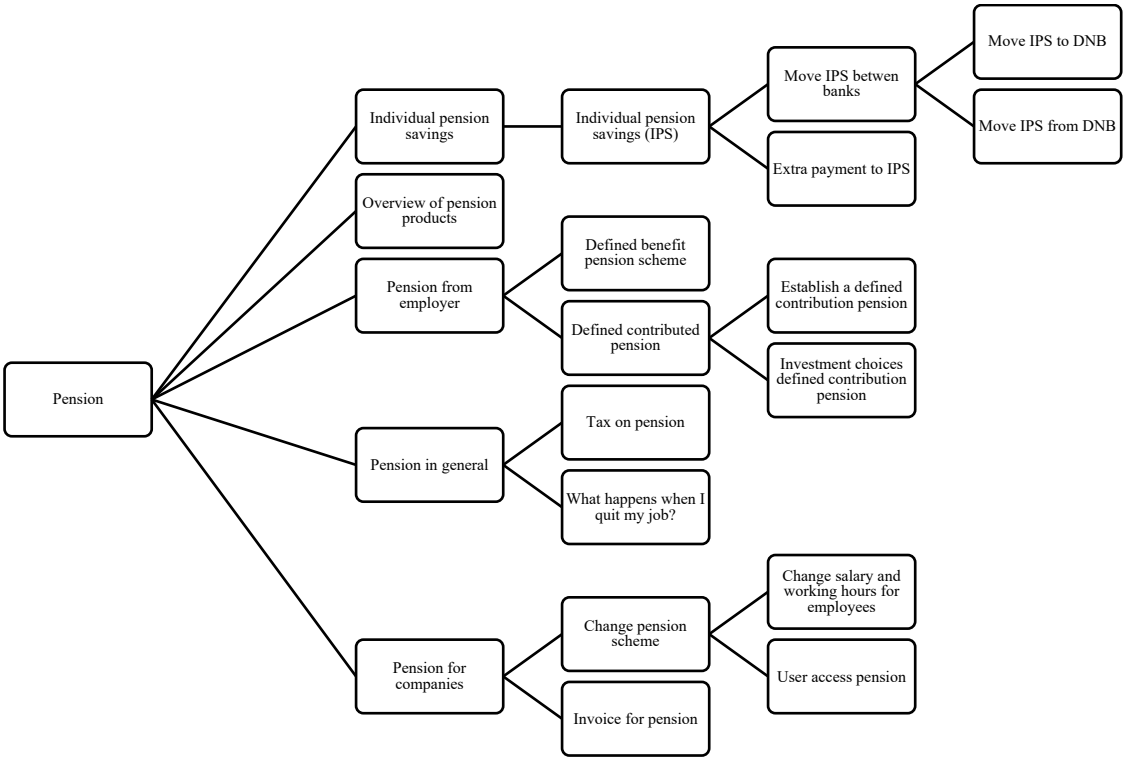


Figure 1 The figure shows some of the intents connected to pension.

The top-level in the hierarchy tree consist of 29 intents, with all the remaining intents being descendants of these. For the full list of top-level intents see appendix 8.5. The top-level intents mostly consist of DNBs product categories, such as insurance or loans, but also general questions, guardianship, tax return, etc. The maximum depth of the hierarchy tree is seven, and the average level of the tree at which an intent is identified is 3,56. The number of intents located at each tree level is shown in Figure 2.

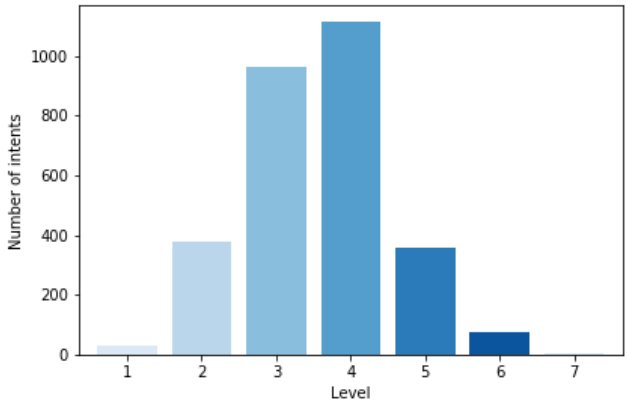


Figure 2 The figure show the number of intents connected to each level in the hierarchy tree.

2.5 Customer Satisfaction with Customer Service Chatbots

A study published in September 2019 by Luo, Tong, Fang, and Qu, stated that AI chatbots can provide several unique business benefits. The study states that undisclosed chatbots are as effective as proficient workers and four times more effective than inexperienced workers in engendering customer purchases. This discloses the financial impact of chatbots, and how they contribute to satisfied customers. In addition to being effective, chatbot can converse in a friendly way, even with humor, and they don´t have bad days, get angry, or become tired like humans, which also contributes to customer satisfaction.

Even though chatbots provide benefits for the supply side, the demand side seems to have more negative associations with talking to chatbots. Because of this, businesses face a dilemma when they launch a customer service chatbot. The analysis by Luo et al. found that disclosure of chatbot identity before the machine-customer conversation reduces purchase rates by more than 79.7%. This negative effect seems to be driven by humans holding a negative perception of machines, and that many customers may feel uncomfortable talking to a computer program. In other words, when customers know that they are talking to a chatbot they purchase less because they perceive the bot as less knowledgeable and less empathetic (Luo, Tong, Fang, & Qu, 2019).

For customers to be satisfied with a customer service chatbot, it is essential that customers feel that they can trust the chatbot. A study published by Følstad, Nordheim, and Bjørkli explores customers' trust in chatbots in an exploratory interview study (2018). This study found that users' trust in chatbots was affected by the quality of its interpretation of requests and advice, its human likeness, its self-presentation, and its professional appearance. The identified factors also suggest that the service context which the chatbot is situated in is important. These contextual factors are for example perceived security and privacy, and the general risk perceptions concerning the topic of the request. Two obvious benefits of a chatbot are that there is no wait time and it is open 24/7. However, the study also found a few surprising benefits of chatbot; some customers found it more relaxing to talk to a chatbot and that the chatbot was non-judgmental when customers asked questions about simple issues. On the other hand, customers reported challenges about the chatbot not being able to interpret the users' requests. The study concludes that to realize chatbots' full potential, they need to be trusted by customers (Følstad, Nordheim, & Bjørkli, 2018).

To measure customer satisfaction is difficult, and businesses use different techniques to get a picture of their customers' opinion of their services. Feine, Morana and Gnewuch published a study where they combined two techniques to measure customer satisfaction regarding chatbots (2019). The study has three different research methods, a comparison of sentiment analysis methods, correlation analysis between sentiment scores and Customer Service Encounter Satisfaction (CSES) values, and exploratory analysis of sentiment scores and CSES values. The first method compared all selected sentiment methods. Sentiment scores for each dialog and utterance were calculated to investigate whether sentiment scores from each tool are similar on a dialog and utterance level. This method revealed that sentiment methods using similar methodologies to identify the expressed polarity in each text provide somewhat similar results. The second method tested the correlation between sentiment scores and CSES values. The results from this method stated that sentiment scores seem to be a primarily better predictor for positive than for negative CSES values. The third method investigated the minimum number of utterances required to show a correlation between sentiment scores and CSES values. The results of the investigation revealed a significant weak positive correlation between sentiment scores and CSES values. The results of this study have design implications for customer service chatbots. As customers express frustration or anger in written language, future chatbots may continuously perform sentiment analyses to identify dissatisfied customers and transfer those customers to a human agent (Feine, Morana, & Gnewuch, 2019).

2.6 Natural Language Processing Classification challenges

This section will present some of the most common challenges for Natural Language Processing Classification.

2.6.1 Multiple intents

For a chatbot's success, it is crucial to be able to detect the customer's intent. In a typical human-to-human dialogue, it is common that a sentence can consist of more than one intent, which will make the conversation smooth and more natural. On the other hand, for a human-to-chatbot conversation, the chatbot assumes that each sentence only consists of one intent. Multi-intent sentences can, therefore, be a challenge for a chatbot to handle (Xu & Sarikaya, 2013). Boost.ai, the developing company of Aino, states that they have solved this issue and that their conversational AI can easily distinguish between the multiple variables (Boost.ai, 2018). Boost.ai´s statement aligns with the article "Multi-Intent Hierarchical Natural

Language Understanding for Chatbots" where Rychalska et al. uses an hierarchical model to precise tagging of multiple intents (2018).

2.6.2 Sarcasm

When classification algorithms are introduced to sarcasm, they tend to get confused and produce false predictions (Kumar & Kaur, 2020). Classification of sarcastic sentences is a difficult task due to representation variations in the textual form sentences. Because sarcasm transforms the polarity of an apparently positive or negative utterance into its opposite, this affects the chatbot's ability to understand the nature of the customer's problem (Dave & Desai, 2016).

2.6.3 Variable length

Generally, the developers of chatbots tend to provide a minimal number of utterances per intent, which makes the classification task difficult. What makes it even more difficult is when the length of the training sentence is short and the customer's sentence is long. Variable length of customers' messages to the chatbot can, therefore, make it difficult for the chatbot to predict correctly (Shashavali et al., 2019).

2.6.4 Multilingual chatbots

Creating chatbots for a multilingual audience presents new challenges. The challenges with multiple languages are not only to understand the language but also written shorthand, abbreviations, and cultural considerations depending on the customer's region. Using the right degree of formality, appropriate conversations, and the correct writing system can be complicated. Language can also affect the chatbot's ability to predict the correct intent because there might be differences in training data and the developer's understanding of different languages (Trippe, 2018).

2.6.5 Homonyms

Natural human language includes words with the same meaning and words with multiple meanings. These words make sense when humans interact with humans, but when humans interact with chatbots, it can become a problem. For a chatbot to succeed in predicting intents, finding keywords in sentences is essential. However, it also needs to determine which of the keywords are most relevant to the customer's question (Agarwal, 2017).

2.7 Performance evaluation

This thesis defines performance through three different metrics that are used to measure the chatbots' performance. This thesis will include customer satisfaction score, automation rate, and classification accuracy as parameters for performance. These three metrics are chosen because they evaluate different aspects of the chatbot, but all three metrics have financial influence for DNB. This thesis assumes that there is a robust connection between customer satisfaction and automation rate, but both are also influenced by the classification accuracy. If the amount of correct predictions increases, customers will most likely be more satisfied. Thus, the automation rate can increase without customer satisfaction score dropping.

2.7.1 Customer Satisfaction Score

Customer satisfaction score (CSAT) is a customer experience metric that measures if customers' needs are fulfilled through a customer satisfaction survey that asks: "How satisfied were you with [company/service]?" Customer satisfaction score is a qualitative key performance indicator, and a way business can measure user feedback. Customer satisfaction is necessary for companies that want to increase their customers' loyalty and enhance business performance (Gronholdt, Martensen, & Kristensen, 2000). A study by Naumann indicates that a satisfied customer is efficient, both in time, money, and resources (1995). As Naumann argues , it costs five times more to attract a new customer, than to keep an existing one (Naumann, 1995).

There are advantages and disadvantages with all methods to measure customer satisfaction. One of the strengths of CSAT is its simplicity, where the customer can click one button to rate the service. Another advantage is that CSAT includes only a few questions, which results in a high response rate by the customers. On the other hand, satisfaction is a subjective word, and "satisfied" can have different meanings for different people. Another disadvantage is that customers in the "neutral" category often skip filling out surveys; therefore, it is likely to get a biased sample. The biggest downside to CSAT is probably that it often creates a knowledge-gap for business. They only receive information on whether the customer is satisfied or not, and not the reason why. The customers rating can be affected if customers are satisfied with the conversation, but unsatisfied with the company or other services the company provides (Birkett, 2018).

CSAT is called “Kundetilfredshetsindeks” (KTI) by DNB IT Emerging Technologies, but this thesis will use the term CSAT instead to conform with the terminology used in the literature. Regardless, KTI can appear in some tables in the appendix.

2.7.2 Automation rate

The International Society of Automation defines automation as “the creation of technology to monitor and control the production and delivery of products and services.” (The International Society of Automation, n.d.). DNB measures automation with an automation rate, and this thesis will define the automation rate as “The proportion of conversations that did not require human agent assistance.” The automation rate will be calculated as the ratio of the number of conversations handled by only the chatbot to the total number of conversations.

$$Automation\ rate = \frac{Number\ of\ conversations\ handled\ by\ chatbot\ only}{Total\ number\ of\ conversations}$$

There are advantages and disadvantages to using the automation rate as a performance metric. The automation rate is a quantitative metric that is very useful in analyzing return on investment and other financial measures of the chatbot project. On the other hand, DNB can increase the automation rate by making it harder for customers to transfer from chatbot to human agents. This change will probably make many customers dissatisfied with the chat-service, and that is why it is crucial to balance satisfied customers and automation rate while developing a better chatbot.

2.7.3 Classification accuracy

Classification accuracy is the ratio of correct predictions to the total number of predictions (Pizer & Marron, 2017). Each message from a customer is classified as an intent, which is referred to as a prediction. If the chatbot is not able to find an intent, then the prediction type is unknown. Each prediction can be classified as True positive (TP), True negative (TN), False positive (FP) or False negative (FN) , where True positive and True negative are classified as correct predictions. Accuracy is a common way of evaluating machine learning algorithms.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy has a clear disadvantage, even though it is a frequently used evaluation metric. The metric only works well if there are an equal number of samples in each class. An unequal number of samples in each class where there is a minor and a major class, accuracy can give the impression of achieving a high degree of accuracy. The algorithm can predict that every sample belongs to the major class, and therefore achieve a high degree of accuracy if the minor class is minimal (Mishra, 2018).

2.8 Confusion and Cost Matrix

Confusion Matrix is a performance measurement tool for machine learning classification problems where output can be two or more classes (Narkhede, 2018). The number correct and incorrect predictions are summarized with count values and give insight into what types of errors that are being made.

	Positive Predicted	Negative Predicted
Positive Actual	TP	FN
Negative Actual	FP	TN

Table 1 Confusion Matrix

The four different prediction types are in this thesis defined as:

True Positive (TP): The chatbot predicts a top-level intent, and it's correct.

False Positive (FP): The chatbot predicts a top-level intent, but it's wrong.

True Negative (TN): The chatbot predicts unknown, and the chatbot´s knowledge bank does not contain a top-level intent for the customers' question.

False Negative (FN): The chatbot predicts unknown, but it should have predicted a top-level intent.

The Cost Matrix is similar to the Confusion Matrix, except the Cost Matrix summarizes the cost of the different kinds of error (Jain, 2018). Cost Matrix is often called Profit/Loss matrix, and can contain both profit and loss at the same time. The product of multiplying the confusion and cost matrix is the net loss or gain for the model.

	Positive Predicted	Negative Predicted
Positive Actual	Cost(TP)	Cost(FN)
Negative Actual	Cost(FP)	Cost(TN)

Table 2 Cost Matrix

The total cost of the model is calculated as the cost of each prediction type multiplied with the number of conversations connected to each type:

$$Cost = Cost(TP) * TP + Cost(FN) * FN + Cost(FP) * FP + Cost(TN) * TN$$

3. Methodology

This chapter describes the methods used for answering the research questions. The first part presents the choice of method. Secondly, all the different methods of analysis will be introduced. Third, the data will be presented, with sources, pre-processing, sample, and quality. The dataset does not consist of any ground truth. For the analysis in this thesis it is necessary to have a ground truth, and therefore, a sample is chosen to be classified.

3.1 Choice of Method

This mixed-methods study has elements of both qualitative and quantitative research combined for a more in-depth understanding of the problem.

For answering the first research question, *what is the level of performance across intents*, descriptive statistics will be used for a descriptive analysis of the top-level intents' performance. Chi-square test of independence and analysis of variance are used to determine if there is a statistically significant difference between the top-level intents. Both analyses include post-hoc analysis for investigation of what causes the effects and analysis of effect sizes.

For answering the second research question*, what characterizes the cases where the chatbot underperforms*, logistic regression is used, with True positive (TP) predictions as the binary measure for performance. For the process of finding features that will be the independent variables in the logistic regression, the method will be a combination of structured interviews and quantitative content analysis. First, structured interviews with DNB employees will be held to gain a deeper understanding of differences between top-level intents and to gather information about which factors that affect the chatbot's accuracy. Secondly, quantitative content analysis will be used to systematically categorize the occurrence of the features in the conversations in the data sample.

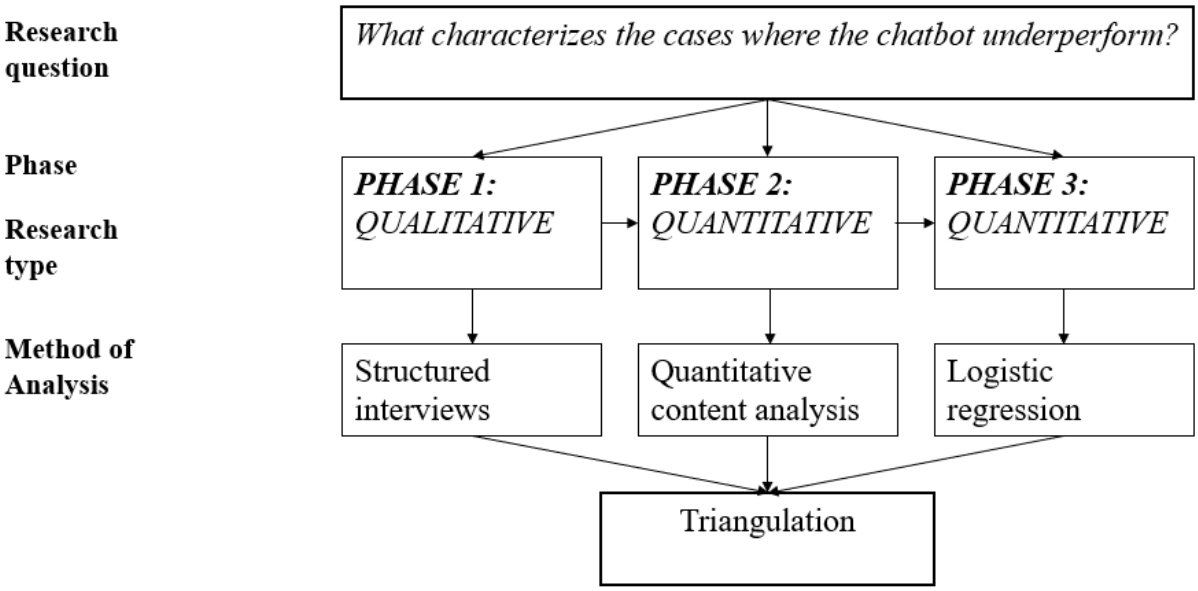


Figure 3 The figure shows the research design for analyzing the second research question.

For answering the third research question, *what is the estimated financial impact*, *if performance were to be improved across weak areas*, the financial impact will be simplified to measure the financial impact of an increase in correct predictions. The method used for this analysis is confusion and cost matrix, which are two frequently used machine learning evaluation tools. The estimated financial impact will be estimated based on the assumption that an increase in True positive (TP) predictions will reduce the need for human labor in the Customer Center.

3.2 Method of Analysis

This section presents each method of analysis used in answering the research questions, ordered by the research questions.

3.2.1 Performance across top-level intents

For analyzing performance across top-level intents three main methods will be used: Descriptive statistics, Chi-square test of independence, and Analysis of Variance.

Descriptive statistics will be performed to analyze the performance across top-level intents, where descriptive tables and plots are used. This will make further analysis easier and will create hypotheses to investigate. Chi-Square analysis is used for determining the relationship between top-level intents and model classification performance, and between top-level intents and automation efficiency. Analysis of Variance (ANOVA) will be used to check for systematic differences in CSAT-score between top-level intents.

The statistical analyses purpose are to investigate the following null hypothesis:

H0: There is no relationship between top-level intent and the chatbots performance

3.2.2 Characteristics of the under-performing cases

For analyzing the characteristics of the under-performing cases three methods will be used; structured interviews, quantitative content analysis and logistic regression.

3.2.1.1 Structured interview

For the purpose of gaining a deeper understanding of what characterizes the cases where the chatbot under-performs, interviews with relevant employees in DNB IT Emerging Technologies are performed. In this thesis, short and structured interviews are used as a method of extracting relevant information from employees that work with the chatbot every day. The technique used for selecting interviewees is a strategic section, where two AI trainers and two Software Engineers were chosen. This selecting technique is used to get as much relevant information as possible and from two different views. The interviews are not the main data collection method and are only used as a supplement for the conversational logs.

Because of the corona-situation, all the interviews were held using Microsoft Teams. There were not any sensitive questions in the interview guide, which indicate that there is no disadvantage to completing the interviews using this communication platform.

The interview guide was developed with the purpose of gaining information about the employee's experience of the chatbot's handling of typical Natural Language Processing issues, and to get a deeper understanding of the findings from the analysis of research question 1. The complete interview guide is in appendix 8.1.

The structured interviews purpose among other, are to investigate the following working hypotheses:

H1: The length of the customer's messages does not affect Funds weak accuracy

H2: The language of the customer's message does not affect Insurance performance

H3: Sarcasm does not affect the chatbot's performance.

3.2.1.2 Quantitative content analysis

In this thesis, quantitative content analysis will be used to look for features that can lead to false predictions of top-level intents. Every conversation in the data sample will be reviewed and for every time one of the features appears it will be noted. The content analysis seeks to look for:

1. The number of words in the customer's message
2. The number of intents in the customers' messages
3. The number of descendant intents of the true top-level intent
4. The number of messages in Norwegian, English, Swedish and Danish
5. The number of messages where the language was misidentified
6. The number of messages containing sarcasm
7. The number of messages containing abbreviations
8. The number of messages containing typos

3.2.1.3 Logistic regression

For the analysis of the characteristics that affect the performance, logistic regression is used to model the probability for a conversation to be classified as True Positive, which means that the chatbot predicts a top-level intent, and it is the correct top-level intent. The chatbot's predictions can be classified as True Positive (TP), True Negative (TN), False Positive (FP) or False Negative (N). However, in this analysis TP will be classified as 1, while TN, FP and FN will be classified as 0. Both TP and TN are correct predictions, but TN implicates that the chatbot does not contain an intent for the customer question. Hence, the chatbot needs human assistance to answer the customer question. For both FP and FN the chatbot needs human assistance since it fails to answer the question correctly. Consequently, TP is the only prediction type of interest in the logistic regression analysis.

The independent variables will be the results from the qualitative content analysis and the structured interviews with DNB employees that work continuously to improve the chatbot's performance. For logistic regression it is important that the independent variables are independent of each other, hence the model has little or no multicollinearity. To check for multicollinearity, a correlation matrix with all input variables is used.

Input/independent variables:

1. Number of words (Numeric): Total number of words in customers' message
2. Number of intents (Numeric): Total number of intents in customers' message
3. Number of descendant intents for top-level intent (Numeric): Total number of descendant intents for top-level intent
4. Language (categorical: Norwegian, English, Swedish): The language of the customers' message
5. Misidentification of language (binominal): The language of the message is misidentified (yes=1, no=0)
6. Sarcasm (binominal): The customers' message contain sarcasm (yes=1, no=0)
7. Abbreviations (binominal): The customers' message contains abbreviations
8. Typos (binominal): The customers' message contains typos

Language will be transformed to dummy variables, which is a method to turn categorical variables into binary variables. Since Norwegian, English and Swedish are dummy variables, Swedish can be removed from the model without losing any information. The final input variables will be number_of_words, number_of_intents, number_of_descendant_intents, language_norwegian, language_english, misidentification_of_language, sarcasm, abbreviations and typos.

The null hypothesis of the logistic regression will be that there is no relationship between the independent variables and the prediction of True Positives. The logistic regression also includes null hypothesis for each independent variable, that including that variable does not increase the fit of the model.

3.2.3 Estimated financial impact of improved performance

For analyzing the estimated financial impact of improved performance, confusion and cost matrix will be used.

3.2.3.1 Confusion and Cost Matrix

Confusion matrix will be used to find the percentage of True positives for the top-level intents, while cost matrix will be used to estimate the financial impact of an 5% increase in the prediction of True Positives for two of the top-level intents that underperforms compared to the other top-level intents.

The costs used in the Cost Matrix comes directly from DNB IT Emerging Technologies Business Case model. These costs are presented in Table 3

Average turnaround chat conversations	12 minutes
Average number of conversations held at the same time	2,1 conversations
Number of working hours per year for a Customer Center employee	1475 hours
Yearly cost per Customer Center employee	850 000kr

Table 3 The table presents information about human agent conversations and the labor costs for the Customer Center.

For the financial estimations in this thesis all other costs than human labor have been excluded. The estimation is also simplified by assuming that when the chatbot predicts the correct top-level intent, that the customer will be satisfied with the answer and not request a human agent to answer their question. This initiates that to reduce the costs associated with the Customer Service Chat, True positives needs to be increased.

3.3 Data

The data used in this thesis is mainly conversations from DNBs chatbot solution. All conversations included are conversations between the chatbot and real customers. It should be noted that the conversations do not contain any personal information about the customers. For example, all names and account numbers mentioned in the messages are shaded. The removal of sensitive information means that there are no applicable restrictions from GDPR or similar directives.

The dataset is conversations taking place from February 16, 2020, until March 17, 2020. This period is chosen because of a new version of the method DNB uses to fetch conversational data from boost.ai. The analyses in this thesis are dependent on the extra features in the latest version.
The first section presents the different sources of data and how the data was received. The second section focuses on the pre-processing of data, while the third explains the selection of sample. The final section will focus on data quality.

3.3.1 Data Source

This thesis has two different data sources, the first one is the conversations from DNBs chatbot solution, while the second is structured interviews conducted with four DNB employees. This data source will be transcripts of the interviews.

The conversational data from DNBs chatbot solution used for analyses is collected from three different data sources stored in one database.

The product package from boost.ai includes everything needed to use the chatbot, such as training and test data, preprogrammed responses to a variety of questions and a classification system for different question types. The classification model for this package can sort customer messages based on the question type, called intent, and then generate a preprogrammed response based on the identified intent. DNB's AI trainers can add information to the preprogrammed responses as well as add new intent categories. The data from conversations between the chatbot and customers is saved within boost.ai's servers. The *Export API* (boost.ai, 2019a) enables DNB to fetch conversational data from boost.ai.

When a customer starts a chat with DNB, the customer will first be introduced for the chatbot. The chatbot will then predict the intent in the customer's question. However, if the chatbot

fails to answer the customer's question or the customer requests a person, the conversation is transferred to a human agent.

The company, Genesys, provides DNB with a contact center (Genesys, n.d.-a). The contact center is responsible for contact between customers and DNBs resources (Genesys, n.d.-b). Since the connection between human agents and customers go through Genesys servers, the conversational data is split between Genesys servers and DNB. The data stored at Genesys servers is later fetched and stored in DNBs database too.

The last data needed for this thesis, the CSAT-data, is gathered by DNB and is therefore stored right in DNBs database.

Figure 4 shows the connection between the three different sources and how they all are collected in DNBs database.

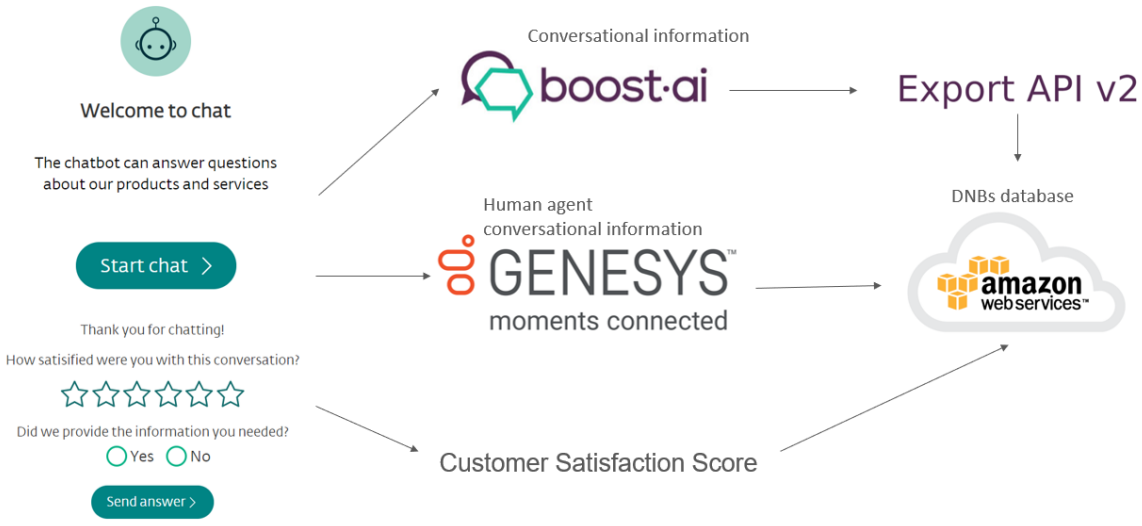


Figure 4 The figure presents the connection between the data sources

The data used in this thesis is received in the form of 47 CSV-files with three different structures; chatbot conversation logs, CSAT-data, and chat conversation parties.

The chatbot conversational logs contain the conversational information and is received in the form of 14 CSV-files. These files contain five columns; a primary key for conversational logs, information about when the entry in chatbot logs table was created, when the entry in chatbot logs table was updated, the chat history, and a unique id for all chatbot conversations. The history column contains 38 columns for each message in the conversation, in the form of JSON objects. These 38 columns are the columns from boost.ai Export API, which contain the messages from the chatbot and customer – and also information about date and time for each message, the language, predicted intent, sentiments, feedback, and more. For the full list, see appendix 8.2. The chatbot logs consist of 171 052 conversations, which take place from February 15, 2020, until March 15, 2020. The conversational logs from February 17, February 26, March 16, and March 17, are not included in the file because of server issues. This indicates that the average number of conversations per day is 6109 conversations.

The CSAT-data files contain information about the CSAT-survey and is received in the form of 18 CSV-files. These files consist of 31 columns, which are presented in appendix 8.3. The survey column contains four columns in the form of JSON-objects; questions, survey id, survey type, and survey type value. The JSON-object, questions, also contain four columns in the form of JSON-objects; question id, question rating value, question ratings, and question text. The CSAT-data files contain both CSAT from human agent chats and chatbot, and therefore contains all the information we need from Genesys. The CSAT-data files consist of 19 821 conversations, which take place from February 16, 2020, until March 17, 2020.

The chat conversation parties' files contain information about which parties are included in the conversation, whether it is just chatbot, chatbot and user, or chatbot, user, and agent. These files are received in the form of 14 CSV-files. The files consist of six columns, which are all presented in appendix 8.4. The conversation parties' files have registered 165 360 conversations, which take place from February 16, 2020, until March 17, 2020.

Because of a significant server issue in DNBs production database, 11 days and 7 666 conversations are lost. Because of this server issue, the dataset consists of 12 155 conversations which take place from February 28, 2020, until March 15, 2020.

3.3.2 Data Pre-Processing

In this section, the steps in the data pre-processing and filtering of data will be described.

The first step was to reduce the number of files from 47 to three files; chatbot logs, CSAT-data and parties´ data. Secondly, the files needed to be joined to connect each conversation with its CSAT-score and the parties. For the full overview of the data pre-processing see appendix 8.6. The further data pre-processing with treatment of missing values and outliers will be described in 3.3.2.1.

3.3.2.1 Treatment of missing values and outliers

This section will describe the treatment of missing values and the different cases that create missing values.

3.3.2.1.1 Treatment of unnecessary columns

After merging chatbot conversation logs, CSAT-data, and chat conversation parties, the data frame contains 66 columns. However, a large part of these columns is not necessary for the analysis in this thesis. Thus, a new data frame is created and using only the columns needed. This data frame contains 14 columns, which means that 52 columns were dropped.

3.3.2.1.2 Treatment of rows without predicted intent

Each row in the data frame will contain a message, either from the chatbot or a customer. The chatbot will predict an intent for all the customers' messages, but for all the chatbot's messages, the columns for predicted intent will be empty. All the rows with empty predicted intent will, therefore, be dropped using Pandas' function for dropping missing values. This reduces the data frame with 423 064 rows and 3 704 conversations.

3.3.2.1.3 Treatment of timestamp outliers

Customer Service have opening hours from 7 am until 11 pm. The period will be limited to the customer service opening hours because automation can not be measured when the only option is an automated chatbot. This makes all the conversations taking place outside the opening hours timestamp outliers, and they will be filtered out. The data frame includes 500 conversations taking place outside of opening hours, which will be removed. This also reduce the data frame with 4 482 rows.

3.3.2.1.4 Treatment of Question id 2

When customers close the chat window, they get two questions with question id 1 and question id 2. The first question is; "How satisfied were you with this conversation?" which

gives the CSAT-score. The second question is; “Did we provide the information you needed?” which is a yes or no answer. This answer is not included in this thesis as a metric, and therefore the rows containing this answer will be removed. Since this is one of two questions, 50% of the rows in the data frame are removed.

3.3.2.1.5 Treatment of duplicate rows

The dataset includes 3 094 duplicate rows. Duplicate rows are removed as not to give that data object a bias when doing analysis.

3.3.3 Data sample

This section will describe how a data sample is created based on the dataset.

After data pre-processing, the dataset consists of 12988 conversations, which take place from February 28, 2020, until March 15, 2020. The dataset consists of 14 columns, containing information about conversation id, date, time, message text, predicted intent, top-level intent, CSAT-score, and more. The number of rows in the dataset is 52 227.

Because of the desire for the sample to represent the dataset, a probability sampling technique is used. Probability sampling starts with an entire population of all the observations in the dataset, before choosing observations for the sample. Simple random sampling is the chosen technique in this thesis, which is a method where all the different observations in the dataset have an equal chance of being chosen. An advantage of this method is that it is a straightforward method of probability sampling, and it reduces sampling bias. However, a significant disadvantage is that it is possible not to select enough observations with specific characteristics. This problem is especially challenging when the number of observations connected to each group is diverse (Kirk, 2011).

3.3.3.1 Sample size

Determining sample size is a statistical concept that involves deciding the number of conversations that should be included in the statistical sample. Sample sizes are used to represent parts of a population, or in this thesis, the dataset. The sample size is found using this formula:

$$n = \frac{N}{1 + \frac{z^2 * \hat{p}(1-\hat{p})}{ME^2 * N}}$$

z = z score

ME = The margin of error

N = Population size

$\hat{p}$ = Population proportion

In this thesis, the sample size is determined with a z-score for a 95% confidence level and a margin of error of 5%. The population is the number of rows in the dataset, which is 52 227 rows. Since the top-level intents in the dataset are predicted, we have no associated percentage of top-level intent, which makes this value 50%.

Using the formula above, the sample size is determined to be 382 rows. To collect the sample from the dataset, Pandas' sample-function is used. This function returns a random sample of the dataset.

3.3.3.2 Final data sample

The data sample consists of 382 rows, 14 columns, and 375 conversations. The conversations take place from February 28, 2020, until March 15, 2020. Figure 5 shows how the number of conversations varies between the dates in the dataset. March 2 is the date with the largest number of conversations with 36 conversations, while March 1, only has six registered conversations.

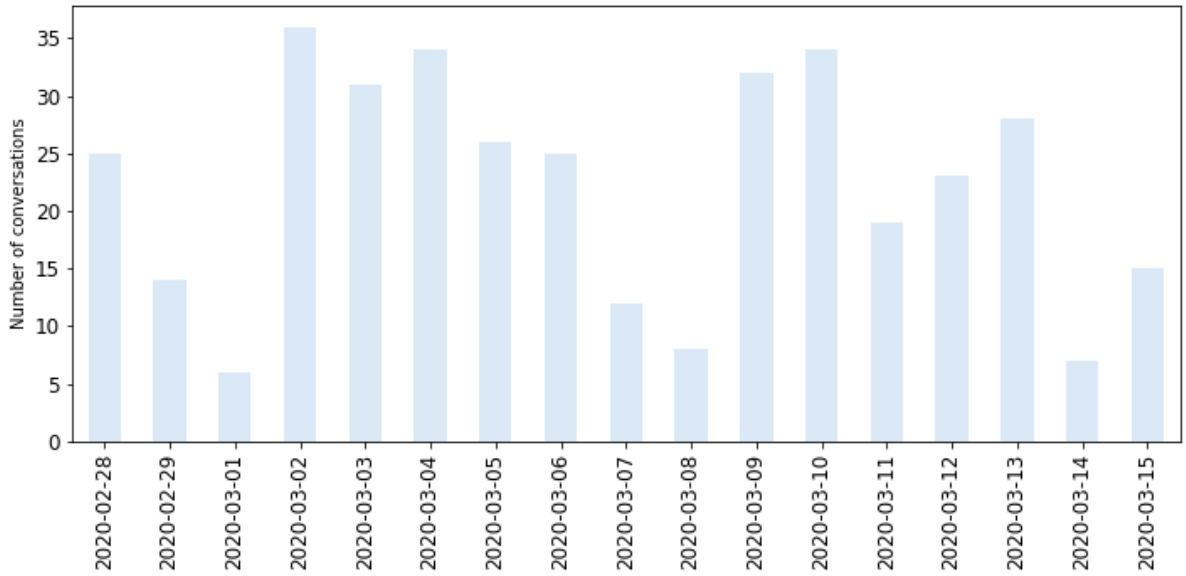


Figure 5 The number of conversations each date in the data sample. .

3.3.4 Data reliability and validity

This section will present two indicators of quality in research; reliability and validity.

3.3.4.1 Reliability

The structured interviews contribute to information that is easy to replicate and therefore has high reliability. The selection of interviewees was focused on employees with considerable knowledge of chatbots' performance and issues. Interviewing employees with this knowledge increases the results' reliability. On the other hand, it was only conducted four interviews, which is a considerable small sample.

The data sample was only reviewed by one person, which decreases the reliability of the data sample because human errors might cause some false classifications of the predictions. If the number of individuals reviewing the sample were increased, the reliability would have been improved.

3.3.4.2 Validity

One of the forms of validation that are particularly appropriate to the logic of qualitative research is triangulation. This is a method where different kinds of data and different methods are compared to see whether they corroborate one another (Silverman, 2014). For analyzing research question 2, both qualitative and quantitative methods are used, with interviews, quantitative content analysis, and logistic regression. The triangulation strengthens the

generalization of this analysis. On the other hand, the period where the conversational data was conducted may weaken the external validation of this thesis.

On March 12, 2020, the Norwegian government announced the most substantial and most comprehensive efforts Norway has had in peacetime. These efforts significantly affected the Norwegian population. The value of the Norwegian currency dropped; the stock market took an enormous hit, and it caused economic and financial uncertainty for Norwegian citizens and their families.

The corona-situation affected DNB and DNBs customers. The customers wanted answers from DNB about payment extension, interest-only period on loan, change in interest rate, travel cancellation, and more. The virus did not just change the customers' questions, but it also increased the number of questions drastically.

As previously mentioned, the dataset used for analyzing all research questions is conversations taking place from February 28 until March 15. As a consequence, the dataset is affected by the coronavirus, which also affects if the data sample is representative of the population of conversations between the chatbot and customers.

4. Results

In this chapter the results found in the analyses are presented, organized by the research questions. The results will be discussed against literature and across the different analyzes in Chapter 5. To be able to analyze the research questions, the data sample needed to be manually reviewed. The results from the manual review of the data sample will be presented in section 4.1

4.1 Manual review of data sample

The data sample contains 375 conversations in the form of 382 rows. The number of rows is larger than the number of conversations because each row is a message in the conversation. The data sample contained 19 different top-level intents. Figure 6 shows the number of conversations for each top-level intent. The top-level intent with the largest number of conversations is “Talk to advisor” with 67 conversations, which is 17,78% of the total number of conversations.

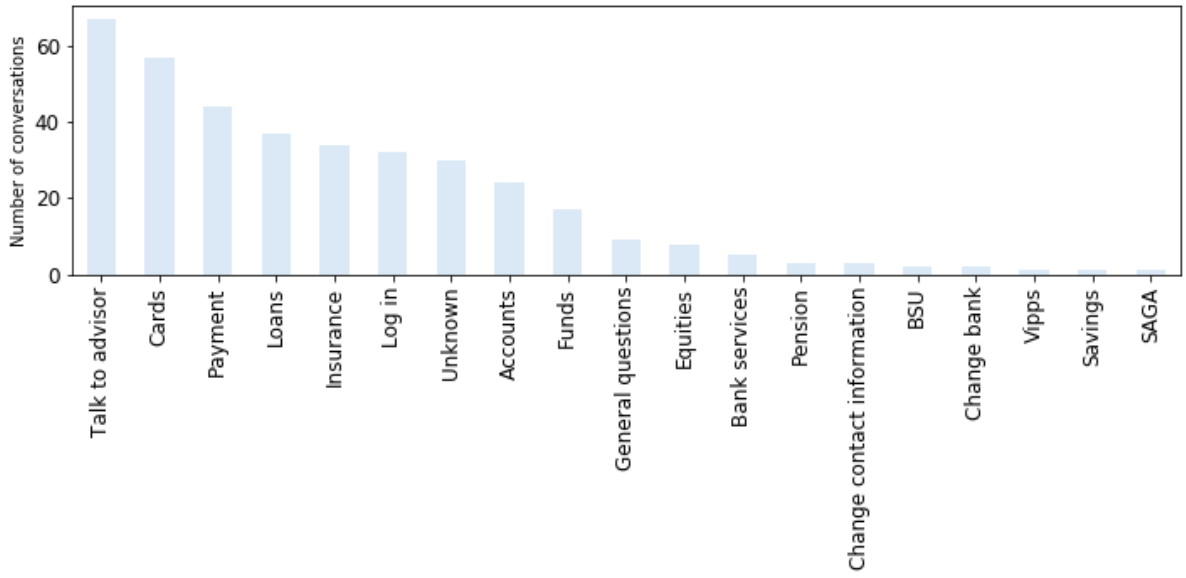


Figure 6 The number of conversations for each top-level intent, sorted according to the total number of conversations.

Of the 19 different top-level intents presented in the figure, only nine of these are connected to more than 10 conversations. The remaining 10 top-level intents are not connected to enough conversations to be analyzed and will therefore be filtered out for the analysis. The two top-level intents “Talk to advisor” and “Unknown” will also be filtered out before the analysis, because of their nature. Many customers want to talk to a human agent when they

contact the Customer Center, and they therefore request this. When a customer requests a human agent, the chatbot predicts the top-level "talk to advisor" and will transfer the conversation to a human agent. In this thesis, there is no advantage of analyzing the conversations connected to this top-level intent. The top-level intent, "Unknown", is connected to all the conversations with context that is not included in the chatbots knowledge bank. Since this top-level intent is a collection of messages it will not add any additional value to this study.

4.2 Performance across top-level intents

This section will present the findings from the analysis of research question 1, *What is the level of performance across top-level intents.* The methods of analyses used to answer the first research question is descriptive analysis of performance, Chi-square test of independence and Analysis of variance. A correlation matrix of the correlation between the three performance metrics will be presented last.

4.2.1 Descriptive analysis of performance

This section presents the top-level intents with their performance in the form of mean accuracy, mean CSAT-score and mean automation rate.

Top-level intent	**Accuracy**	**CSAT-score**	**Automation rate**	**Number of conversations**
Cards	0,679	4,750	0,161	57
Payment	0,705	3,932	0,295	44
Loans	0,757	4,486	0,162	37
Insurance	0,618	3,706	0,235	34
Log in	0,844	3,688	0,469	32
Accounts	0,792	4,417	0,208	23
Funds	0,412	2,941	0,471	17

Table 4 The table shows the top-level intents with their performance and number of conversations.

4.2.1.1 Accuracy

The top-level intents have a mean accuracy of 68,6% and a standard deviation of 14,2%. The top-level intent with the highest accuracy is “Log in” with accuracy of 84,4%, while “Accounts” has an accuracy of 79,2%. The top-level intent with the lowest accuracy is “Funds” with an accuracy of 41,2%.

4.2.1.2 Customer Satisfaction Score

The top-level intents have a mean CSAT-score of 3,89 and a standard deviation of 0,617. The top-level intent with the highest CSAT-score is “Cards” with 4,75, while “Insurance” has an CSAT-score of 4,49. The top-level intent with the lowest CSAT-score is “Funds” with a score of 2,94.

4.2.1.3 Automation rate

The top-level intents have a mean automation rate of 28,6% and a standard deviation of 13,4%. The top-level intent with the highest automation rate is “Funds” with 47,1%, while “Log in” has an automation rate of 46,9%. The top-level intent with the lowest automation rate is “Cards” with 16,1%.

4.2.1.4 Overall performance

The three performance metrics used in this thesis have different scales, while accuracy and automation rate have values between 0 and 1, the CSAT-score have values between 1 and 6. To be able to compare the overall performance, the CSAT-score needs to be normalized. Normalizations means that the values are scaled to a fixed range, and in this case 0 to 1 (Lakshmanan, n.d.). The normalization is done with the following equation:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

In Table 5 the normalized CSAT-score is put in together with the accuracy and automation rate. To the right in the table there is also an overall performance, which is the sum of accuracy, CSAT-score and automation rate. To visualize the overall performance, Figure 7 shows the performance as a stacked bar chart. The figure shows that “Log in” has the highest overall performance, while “Insurance” has the lowest overall performance.

Top-level intent	Accuracy	CSAT-score	Automation rate	Overall performance
Log in	0,844	0,538	0,469	1,850
Accounts	0,792	0,697	0,208	1,697
Payment	0,705	0,683	0,295	1,683
Loans	0,757	0,750	0,162	1,669
Funds	0,312	0,586	0,471	1,469
Cards	0,679	0,541	0,161	1,380
Insurance	0,618	0,388	0,235	1,241

Table 5 The table shows the top-level intents with their accuracy, CSAT-score, automation rate and overall performance.

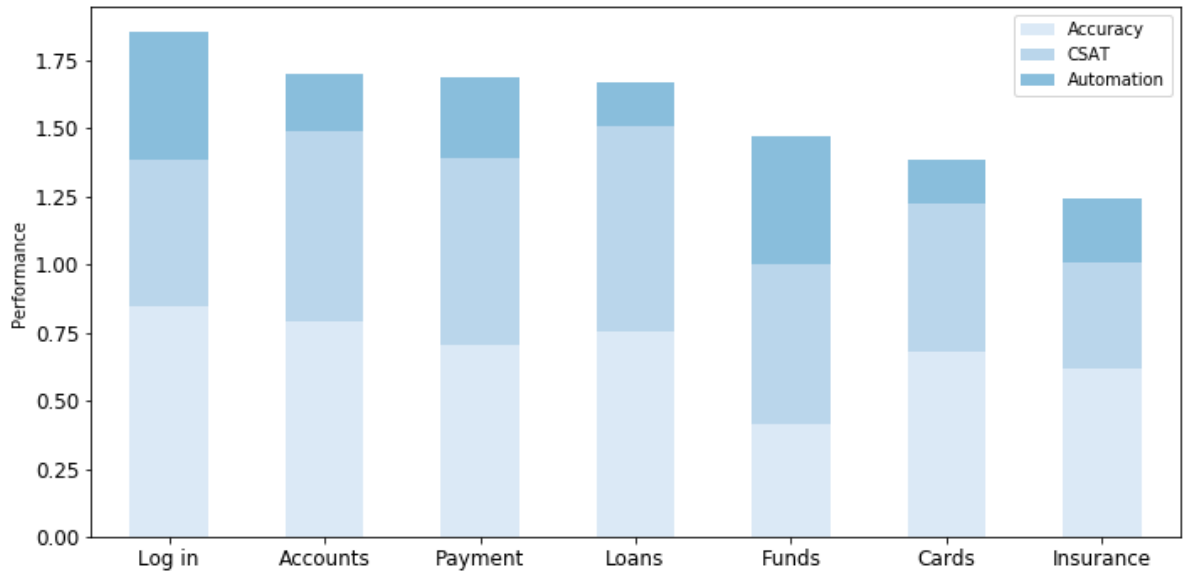


Figure 7 The figure shows the overall performance for the top-level intents.

4.2.2 Chi-square test of independence

The analysis is used to test both if there is a relationship between top-level intent and model classification and if there is a relationship between top-level intent and automation efficiency.

4.2.2.1 Model classification performance

In this section the results from the chi-square test of the independence between top-level intent and model classification performance is presented. The following hypotheses are tested:

H_0: There is no relationship between top-level intent and model classification performance

H_1: There is a relationship between top-level intent and model classification performance

Table 6 display the observed values of the model classification performance. The numbers in the parentheses are the expected values if top-level intent and model classification performance are independent variables.

	Correct top-level intent	**Wrong top-level intent**	**Sum**
Cards	38 (39)	18 (17)	**56**
Payment	31 (31)	13 (13)	**44**
Loans	28 (26)	9 (11)	**37**
Insurance	21 (24)	13 (10)	**34**
Log in	27 (22)	5 (10)	**32**
Accounts	19 (17)	5 (7)	**24**
Funds	7 (12)	10 (5)	**17**
Sum	**171**	**73**	**244**

Table 6 The table shows the observed values, with the expected values in parentheses behind for the relationship between model classification performance and top-level intent.

Chi-Square value = 12,646

Degrees of freedom = 6

Critical chi-square value for 5%: 12,59

The null hypothesis is rejected as the chi-square value is larger than the critical chi-square value at 5% significance level. The results indicate that top-level intent had a significant effect on model classification performance.

A post-hoc analysis to determine what the association between model classification performance and top-level intent might be, is completed with adjusted residuals. Table 7 shows the calculated adjusted residuals for each category.

	Correct top-level intent	**Wrong top-level intent**
Cards	-0,414	0,414
Payment	0,060	-0,060
Loans	0,807	0,807
Insurance	-1,142	1,142
Log in	1,894	-1,894
Accounts	1,024	-1,024
Funds	-2,698	2,698

Table 7 The table show the adjusted residual for the relationship between model classification performance and top-level intent

$$Significance\ level\colon \frac{0{,}05}{14} = 0{,}0036$$

$$Critical\ Z\ value = 2{,}69$$

In this analysis, "Funds" adjusted residuals are the only top-level intent that is above this critical Z value at 2,69. "Funds" has adjusted residuals at 2,698, while the other top-level intents have adjusted residuals in the interval from 0,06 to 1,894. The test revealed that "Funds" is the only top-level intent with a significant difference between correct predictions and wrong prediction, at a 5% significance level.

To indicate the strength of the association between model classification performance and top-level intent Cramér's V is used.

$$Cramér's\ V = 0{,}227$$

The results that Cramér's V is 0,22 indicates that the association between top-level intent and model classification performance is significantly moderate.

The chi-square test of independence concludes that top-level intent and model classification performance showed to have a significant moderate association at a 5% significance level. A post-hoc z-test on the adjusted residuals with Bonferroni correction revealed that only for "Funds" there is a significant difference between the correct and not-correct predictions at a 5% level of significance. This indicates that "Funds" has the highest association with predicting the wrong top-level intent. This analysis will conclude that the level of model classification performance is dependent on the top-level intents.

4.2.2.2 Automation efficiency

In this section the results from the chi-square test of the independence between top-level intent and automation efficiency is presented. This chi-square test has these hypotheses:

H_0: There is no relationship between top-level intent and automation efficiency

H_1: There is a relationship between top-level intent and automation efficiency

Table 8 display the observed values of the automation efficiency. The numbers in the parentheses are the expected values if top-level intent and automation efficiency are independent variables.

	Successful automation	**Required human assistance**	**Sum**
Cards	9 (15)	47 (41)	**56**
Payment	13 (12)	31 (32)	**44**
Loans	6 (10)	31 (27)	**37**
Insurance	8 (9)	26 (25)	**34**
Log in	15 (8)	17 (24)	**32**
Accounts	5 (6)	19 (18)	**24**
Funds	8 (4)	9 (13)	**17**
Sum	**64**	**180**	**244**

Table 8 The table show the observed values, with the expected values in parentheses behind for the relationship between automation efficiency and top-level intent.

Chi-Square value = 16,504

Degrees of freedom = 6

Critical chi-square value for 5%: 12,59

The null hypothesis is rejected as the chi-square value is larger than the critical chi-square value at 5% significance level. The results indicate that top-level intent had a significant effect on the automation efficiency.

A post-hoc analysis to determine what the association between automation efficiency and top-level intent might be, is completed with adjusted residuals. Table 9 shows the calculated adjusted residuals for each category.

Adjusted residuals	**Successful automation**	**Required human assistance**
Cards	-1,969	1,969
Payment	0,552	-0,552
Loans	-1,503	1,503
Insurance	-0,386	0,386
Log in	2,848	-2,848
Accounts	-0,633	0,633
Funds	2,024	-2,024

Table 9 The table show the adjusted residual for the relationship between automation efficiency and top-level intent.

$$Significance\ level\colon \frac{0{,}05}{14} = 0{,}0036$$

$$Critical\ Z\ value = 2{,}69$$

In this analysis, “Log in” adjusted residuals are the only top-level intent that is above this critical Z value at 2,69. “Log in” has adjusted residuals at 2,85, while the other top-level intents have adjusted residuals in the interval from 0,39 to 2,02. The test revealed that “Log in” is the only top-level intent with a significant difference between successful automation and required human assistance at a 5% significance level.

To indicate the strength of the association between model classification performance and top-level intent Cramér’s V is used.

$$Cramér's\ V = 0{,}26$$

The results that Cramér’s V is 0,26 indicates that the association between top-level intent and automation efficiency is significantly high moderate.

The chi-square test of independence conclude that top-level intent and automation efficiency showed to have a significant moderately high association at a 5% significance level. A post-hoc z-test on the adjusted residuals with Bonferroni correction revealed that only for “Log in” there is a significant difference between the successful automation and the required human assistance, at a 5% level of significance. This indicates that “Log in” has the highest association with successful automation. This analysis will conclude that the level of automation efficiency is dependent on the top-level intents.

4.2.3 Analysis of Variance

The analysis is used to test if there is a difference in CSAT-score between the top-level intents.

4.2.3.1 Customer satisfaction

For the analysis of variance, the hypothesis used for testing are:

H_0: All the top-level intents have equal mean CSAT-score

H_1: All the top-level intents mean CSAT-scores are not equal.

Top-level intent	**CSAT-score**
Cards	4,750
Payment	3,932
Loans	4,486
Insurance	3,706
Log in	3,688
Accounts	4,417
Funds	2,941

Table 10 The table show mean CSAT-score for each top-level intent

The calculated overall mean for all the top-level intents is:

$$Overall\ mean = 4{,}119$$

Table 11 is the ANOVA table, containing the statistics for testing the hypotheses.

	Degrees of freedom	**Sum of Squares**	**Mean of Square**	**F-value**	**P-value**
Top-level intent	6	66,306	11,051	2,486	0,024
Residual	237	1053,247	4,444		

Table 11 The table shows the ANOVA statistics used to test the hypotheses

To conclude with the ANOVA statistics, the p-value is used. Table 11 shows that the p-value is 0,024. This p-value says that there is a 2,4% chance to observe these differences or more extreme differences between mean CSAT-score of the top-level intents if the null hypothesis is correct. Since the p-value is less than the significance level at 0,05, the null hypothesis is rejected, which indicates that all the top-level intents do not have equal CSAT-scores.

A post-hoc analysis to determine which mean CSAT-scores that statistically differ from each other is completed with Tukey-Kramer test. Table 12 shows the calculations from the Tukey-Kramer tests. The significance level chosen for the test is 5%.

Group 1	Group 2	Mean difference	Lower	Upper	Reject
Log in	Loans	0,799	-0,7144	2,3124	False
Log in	Cards	1,0625	-0,3267	2,4517	False
Log in	Accounts	0,7292	-0,9637	2,422	False
Log in	Insurance	0,0184	-1,5257	1,5624	False
Log in	Funds	-0,7463	-2,6278	1,1352	False
Log in	Payment	0,2443	-1,2122	1,7008	False
Loans	Cards	0,2635	-1,0646	1,5917	False
Loans	Accounts	-0,0698	-1,7129	1,5733	False
Loans	Insurance	-0,7806	-2,2699	0,7087	False
Loans	Funds	-1,5453	-3,3822	0,2915	False
Loans	Payment	-0,5547	-1,953	0,8437	False
Cards	Accounts	-0,3333	-1,8628	1,1962	False
Cards	Insurance	-1,0441	-2,4071	0,3189	False
Cards	**Funds**	**-1,8088**	**-3,5448**	**-0,0728**	**True**
Cards	Payment	-0,8182	-2,0811	0,4448	False
Accounts	Insurance	0,7108	-2,3822	0,9606	False
Accounts	Funds	-1,4755	-3,4628	0,5118	False
Accounts	Payment	-0,4848	-2,0757	1,106	False
Insurance	Funds	-0,7647	-2,6269	1,0975	False
Insurance	Payment	0,2259	-1,2055	1,6574	False
Funds	Payment	0,9906	-0,7996	2,7809	False

Table 12 The table shows the Tukey-Kramer statistics with mean difference, confidence intervals and whether the null hypothesis should be rejected.

The results from the Tuckey-Kramer method reveals that “Cards” significantly differ from “Funds”. These two are the only top-level intents that significantly differ from each other, and where the null hypothesis can be rejected at a 5% significance level.

To indicate the strength of the effect of top-level intent, eta-squared is used.

$$\eta^2 = \frac{66{,}31}{1119{,}55} = 0{,}059$$

The results that eta-squared is 0,059 indicates that the variance explained by the model is 5,9%. The top-level intents had a significantly small effect on the CSAT-scores.

The analysis of variance concludes that top-level intent had a significantly small effect on CSAT-score at a 5% significance level. A post-hoc Tuckey-Kramer test revealed that only two mean CSAT-scores significantly differ from each other. These two top-level intents are “Cards” and “Funds”. This indicates that “Cards” has the highest average CSAT-score, while “Funds” has the lowest average CSAT-scores. This analysis will conclude that the level of mean customer satisfaction scores is not equal between top-level intents.

4.2.4 Correlation performance metrics

From Table 13 it can be observed a high negative correlation between CSAT-score and automation rate, with a correlation coefficient at -0,859. The correlation between CSAT-score and accuracy is medium positive, with a correlation coefficient at 0,633, while the correlation between automation rate and accuracy is -0,3 which is low negative correlation. This implies that with an increase in automation rate, the CSAT-score and accuracy seem to decrease. When accuracy increases, CSAT-score also seems to increase.

	Accuracy	**CSAT**	**Automation rate**
Accuracy	1,000	0,633	-0,300
CSAT	0,633	1,000	-0,859
Automation rate	-0,300	-0,859	1,000

Table 13 The table shows the correlation between the performance metrics

4.3 Characteristics of the under-performing cases

This section will present the findings from the analysis of research question 2, *What characterizes the cases where the chatbot underperform* . The methods of analyses used to answer the second research question is structured interviews, quantitative content analysis and logistic regression.

4.3.1 Structured interviews

This section will present the results from the interviews with DNB employees, organized by the working hypotheses introduced in section 3.2.1.1. Appendix 8.1 contains the interview guide.

H1: The length of the customer's messages does not affect "Funds" weak accuracy

The results from the interviews suggest that "Funds" weak accuracy is due to the top-level intents' lack of intelligence and focus by the AI trainers, and few descendant intents. The employee's answers indicated that many customers ask more complicated questions about funds to the chatbot than it can handle, and what it is suitable to answer. The chatbot is not suited to give customers advice about funds or to influence their investing decisions.

The four employees agreed that both concise and very long messages from the customers make it difficult for the chatbot to predict the correct intent. For the concise messages, the

customer often gets a top-level intent answer from the customer. Hence, the customers have to answer questions from the chatbot to eventually get to what they need help with. Based on one word, it is not easy for the chatbot to figure out what the customer needs. If a sentence is very long and contains much information, it is often problematic to reduce the text into only one intent, and therefore the message is predicted unknown. The long messages tend to have a connection to multiple intents. The employees informed that when a message contained two intents, Aino answers by asking which intent the customer wants an answer to. None of the employees' experience that this function works optimal, and they experience that the function creates an unnatural conversation. If the number of intents in the message exceeds two, Aino has trouble predicting the correct intent, which leads to unknown intent. One employee also informed that it happens that Aino is 99% certain that one intent is correct, but it answers with an unknown-intent answer or a multi-intent answer. Even though the AI trainers have challenged boost.ai on this issue, boost.ai has not found a solution to the problem.

The results indicate that the length of the customers' messages may affect “Funds” weak performance. The customers tend to ask longer, and more complicated questions related to “Funds”, which affect “Funds” accuracy. “Funds” weak accuracy might also be related to a lack of focus from the AI trainers in developing this top-level intent. The interviews indicate a strong connection between the length of sentences and Aino´s ability to predict the correct intent. Both concise and very long messages tend to lead to false predictions.

H2: The language of the customer’s message does not affect Insurance performance

The results from the interviews suggest that “Insurance” is an area with room for interpretation. Because of this, there is uncertainty among customers to what is covered by each insurance. The employees stated that “Insurance” is demanding and complicated, besides, to be a large area. Since “Insurance” is a very large area it might also affect the prediction. Even though the chatbot predicts the correct top-level intent, it has not caught what the customer asks specific enough. This often leads to the requirement of a human agent for answering the customers questions. One of the employees also stated that one of the reasons for “Insurance” to under-perform might be that Insurance was one of the already created top-level intents in the chatbot-package from boost.ai. Because of limited time and resources before the chatbot was launched, the AI trainers only had time to moderate the already created intents and adding DNB-names for the products. Boost.ai was also a small

startup at this point and, since then, they have renewed and improved their algorithms and best practices. The admin panel where the AI trainers create new intents has also been improved considerably since the beginning. The employee informed that responsibility for each top-level intent is distributed between the AI trainers. The AI trainer with responsibility for “Insurance” has been on maternity leave, which has left “Insurance” without one dedicated AI trainer.

The four employees had different opinions on how language affected the chatbots' ability to predict the correct intent. While one of the employees stated that the chatbot struggles to find the correct language, and therefore sometimes cannot find the correct intent because it lacks the correct language. This employee also stated that when customers write with typos in Norwegian, the chatbot identifies the language as Swedish or Danish. Another of the employees stated that after a clean-up of the English intents, Aino predicts as good in English as in Norwegian. On the other hand, messages in Swedish and Danish tend to lead to wrong predictions because of the lack of synonyms in those languages. The lack of synonyms in Swedish and Danish is explained by Norwegian and English being the chatbots' primary languages. One employee had never noticed specific differences in prediction quality between Norwegian and English. However, the chatbots' answers were significantly better in Norwegian than English, because the AI trainers have used more time on the Norwegian content. The chatbot’s answers also often refer to DNBs website, which has significantly more content in Norwegian than English.

The results from the interviews conducted show no relationship between language and Insurance performance. The employees rather stated that insurance is a large and complicated area as reasons for the under-performance. The interviews also indicated whether the customers write in English or Norwegian does not affect the chatbots' ability to predict the correct intent. At the same time, messages in Swedish or Danish has a more substantial possibility to affect the prediction.

H3: Sarcasm does not affect the chatbot’s performance.

The interviews indicate that sarcasm affects the predictions in the way that Aino will predict the wrong intent because it does not understand that the customer is sarcastic. Sarcastic messages tend to happen at the end of a conversation when a customer, for example, writes,

“Thank you for nothing”, and Aino answers with “You are welcome”. This answer often leads to frustration from the customer and an extra-low CSAT-score.

Other interesting findings from the interviews:

The employee’s answers initiated that the reason for “Cards” low accuracy might be related to the customers' messages containing much additional information about card transactions that might lead to wrong predictions. “Cards” is the only top-level intent where an API is connected to one of the descendant intents, which enables Aino to solve the customers' problem. Aino can also help the customers with a more significant number of their problems connected to “Cards” because the guidance in the online bank is easy to understand. The employees mention these factors as the reason for “Cards” high CSAT-score.

The employees have not experienced that words with different meanings, or the word "How" have caused any difficulties for Aino to predict the correct intent. The word "How" should not be picked up as an important word, and therefore it should not affect the prediction.

From the question "In a customer's message, which factors do you think are crucial for Aino's ability to predict the correct intent?" the employees mentioned a diverse sample of factors. One employee indicated that the customer’s knowledge of Aino is very important, because it makes the customer write questions where Aino easily finds the intent. Another employee stated that the customer’s sentences should be very clear and precise to the point, where the length of the sentence and number of sentences were the main factors for correct predictions. The quality of the training data was also mentioned as a factor for Aino's ability to predict the correct intent, where the number of training sentences and the quality of sentences were essential factors. It was also mentioned that answers based on sentiment analysis of the customer's questions might be something to consider for future development of the chatbot.

4.3.2 Quantitative content analysis

1. *The number of words in the customer’s message*

The quantitative content analysis revealed the average number of words in the customer messages are 7,21 words, with a standard deviation of 6,11. The minimum number of words is one word, while the maximum is 23 words. There is a limit of 110 characters for every customer message, which limits the number of words. Figure 8 displays the distribution of the

number of words for the different top-level intents in the form of boxplots. “Account” has the highest average number of words with 10,04 words, while “Loans” has the lowest average with 5,51 words. For the full overview of the average number of words, see Table 14.

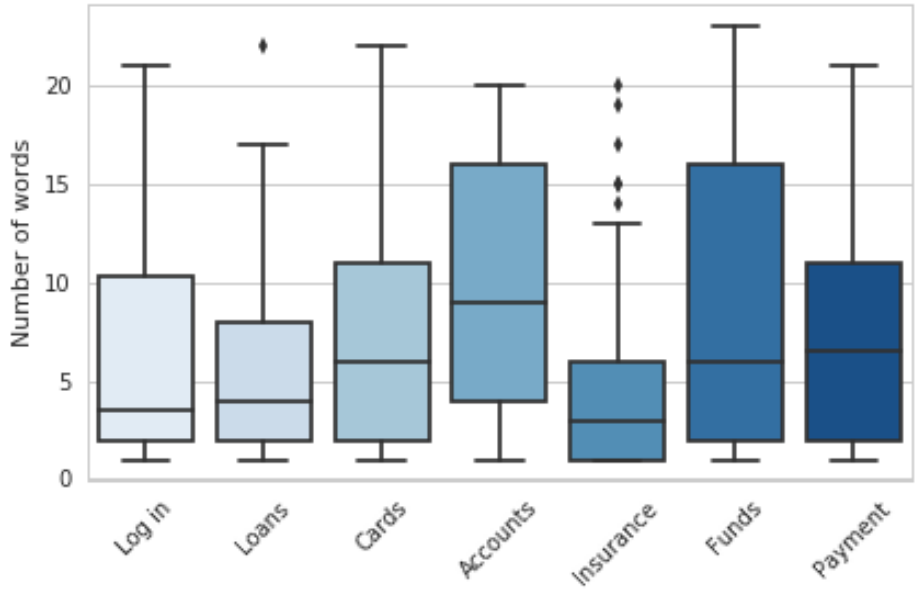


Figure 8 Boxplot of number of words for the top-level intents

Top-level intent	Number of words	Number of intents	Number of descendants
Cards	7,807	1,298	265
Log in	6,438	1,313	230
Payment	7,591	1,295	216
Insurance	6,647	1,176	498
Loans	5,514	1,135	293
Accounts	10,043	1,348	144
Funds	8,706	1,471	30

Table 14 The table presents the top-level intents number of words and number of intents in messages and the number of descendants in the training data.

2. *The number of intents in the customers’ messages*

The data sample has an average number of intents in the customer messages on 1,27 intents, with a standard deviation of 0,54. The minimum number of intents in the sample is one and the maximum is three. “Funds” has the highest average number of intents with 1,47, whereas “Insurance” has the lowest average number of intents with 1,18 intents. Out of 244 conversations, 10 conversations contain three intents and 56 conversations contain two or

three intents. “Cards” has 14 conversations with two or three intents, and “Payment” has 12 conversations. “Loans” has only four conversations with two or three intents.

3. *The number of descendant intents of the true top-level intent*

The data used to find descendant intents are JSON-files containing the entire intent-tree obtained from the Admin panel at boost.ai. The top-level intents in the data sample has an average number of descendant intents on 260,51 intents, with a standard deviation on 116,36. “Insurance” has the highest number of descendant intents with 498 intents, whereas the “Funds” has the lowest number of descendant intents. Since the AI trainers create 20-25 training sentences for every new descendant intent, there is a large variation in size of training data for the top-level intents.

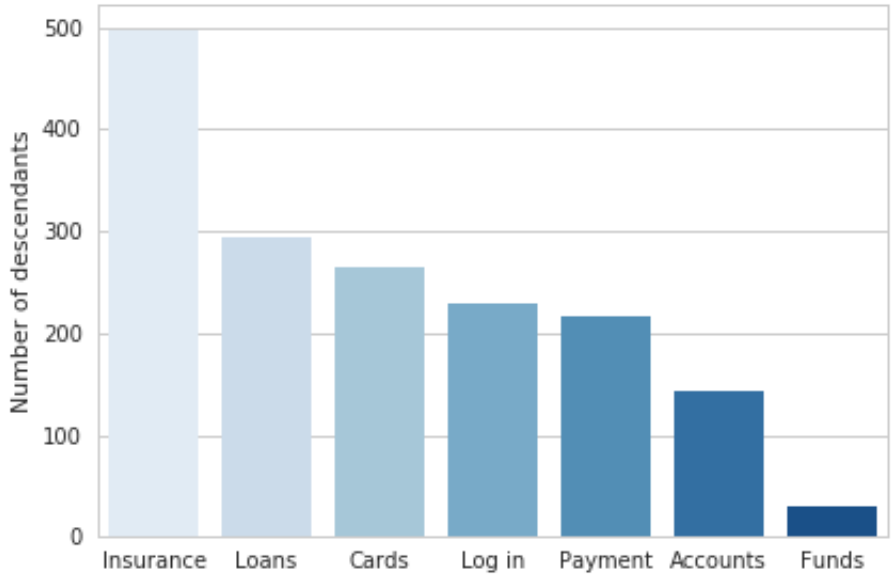


Figure 9 The figure show number of descendants for the top-level intents

4. *The number of messages in Norwegian, English, Swedish and Danish*

The quantitative contents analysis revealed that 226 conversations were Norwegian, 17 conversations were English, and one conversation was Swedish. There were no Danish conversations in the dataset. This makes 92,62% of all the conversations Norwegian.

“Insurance” has the highest percentage of Norwegian conversations with 97,06%, while “Accounts” has only 82,61% Norwegian conversations. Table 15 present the percentage of conversations in Norwegian, English and Swedish for each top-level intent.

Top-level intent	Messages in Norwegian	Messages in English	Messages in Swedish	Misidentified language
Cards	96,49%	3,51%	0,00%	0
Log in	96,88%	3,13%	0,00%	0
Payment	93,18%	6,82%	0,00%	0
Insurance	97,06%	2,94%	0,00%	1
Loans	86,49%	13,51%	0,00%	0
Accounts	82,61%	17,39%	0,00%	0
Funds	88,24%	5,89%	5,88%	2

Table 15 The table presents the percentage of Norwegian, English and Swedish messages, and the number of messages where language is misidentified.

5. *The number of messages where the language was misidentified*

The data sample contained three messages where the language was misidentified. One of these messages the customer writes in Swedish, but the language is identified as Norwegian. While the two other messages the language is identified as Danish, but the customer writes in Norwegian. Two of the messages are connected to “Funds”, while the third is connected to “Insurance”.

6. *The number of messages containing sarcasm*

The data sample contains two messages where the customer writes sarcastic. Both messages are connected to “Cards”, and both are written in Norwegian. This is 3,50% of all the conversations connected to “Cards”. The full overview of conversations with sarcasm, abbreviations and typos is shown in Table 16.

Top-level intent	Sarcasm	Abbreviations	Typos
Cards	2	3	6
Log in	0	0	4
Payment	0	0	4
Insurance	0	1	3
Loans	0	4	7
Accounts	0	1	4
Funds	0	1	3

Table 16 The table presents the percentage of messages containing sarcasm, abbreviations and typos for each top-level intent.

7. *The number of messages containing abbreviations*

The data sample contains 10 messages where the customer writes with abbreviations. Four of these messages are connected to “Loans”, three are connected to “Cards”, while “Insurance”, “Accounts” and “Funds” each have one message containing abbreviations.

8. *The number of messages containing typos*

The data sample contains 31 messages where the customers write with typos. “Loans” have the highest number of messages containing typos with seven messages, this is 18,92% of the messages connected to “Loans”. “Cards” have six messages containing typos, while “Log in”, “Accounts” and “Payment” have four messages. “Insurance” and “Funds” both have three conversations containing typos each.

4.3.3 Logistic regression

This section will provide the results from the logistic regression analysis. The dataset includes 244 conversations and 10 fields, which is presented in 3.3.3.2. The section will be split into two parts, where the first part consists of data exploration, while the other is implementation of the model.

4.3.3.1 Data exploration

The dataset consists of 167 TP and 77 FP, FN and TN, which is shown in Figure 10. The percentage of TP is therefore 68,44%, while the percentage of FP, FN and TN are 31,56%.

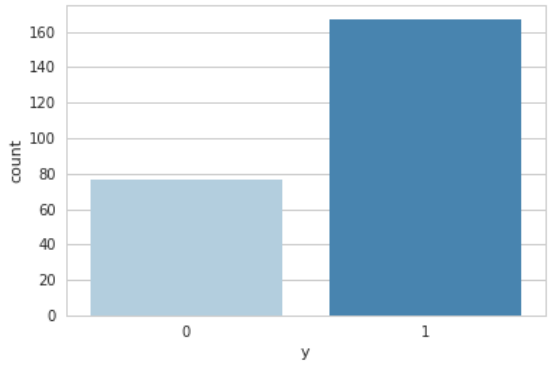


Figure 10 The figure shows the number of TP and the number of FP, FN and TN.

Table 15-17 contains the mean values of the columns in the dataset grouped by Y.

Y	Number of words	Number of intents	Number of descendant intents
0	9,429	1,351	256,844
1	6,192	1,240	262,198

Table 17 Number of words, intent and descendant intents of the true root grouped by Y.

Y	Language Norwegian	Language English	Misidentification of language
0	0,857	0,130	0,013
1	0,958	0,042	0,012

Table 18 Percentage of messages in Norwegian, English and Swedish, and percentage of misidentification of language grouped by Y.

Y	Sarcasm	Abbreviations	Typos
0	0,026	0,052	0,273
1	0	0,054	0,060

Table 19 Percentage of conversations contain sarcasm, abbreviation and typos grouped by Y.

Figure 11 visualizes the mean number of words for TP and for FP, FN and TN with the estimate's uncertainty. The figure indicates that the number of words might be a good predictor of the outcome variable, TP.

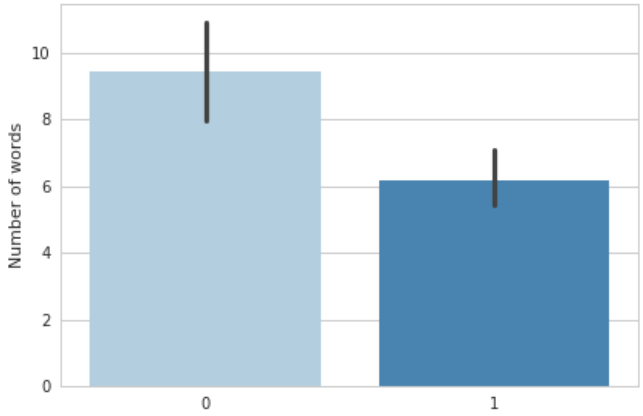


Figure 11 The figure visualizes the mean number of words for TP and FP, FN and TN.

Figure 12 visualizes the mean number of intents for TP and for FP, FN and TN with the estimate's uncertainty. Number of intents in the customer's messages does not seem to be a strong predictor for the outcome variable, TP.

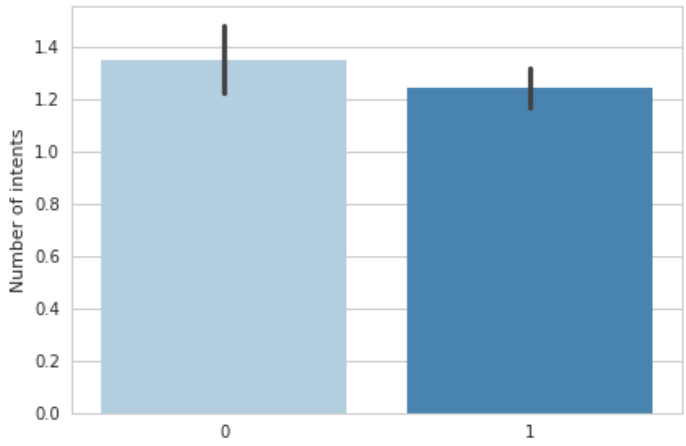


Figure 12 The figure visualizes the mean number of intents for TP and FP, FN and TN.

Figure 13 visualizes the mean number of descendant intents for TP and for FP, FN and TN with the estimate's uncertainty. Number of descendant intents does not seem to be a strong predictor for the outcome variable, TP.

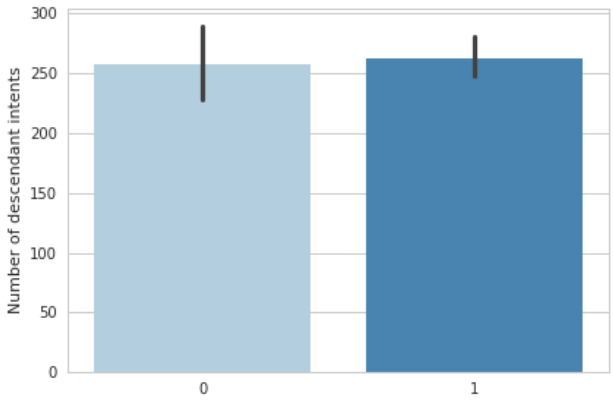


Figure 13 The figure visualizes the mean number of descendant intents for TP and FP, FN and TN.

Figure 14 visualizes the frequency of TP and FP, FN and TN for English, Norwegian and Swedish. The number of conversations in Swedish is too small to see in contrast to the number of conversations in Norwegian and English. Language can be a good predictor for the outcome variable, TP.

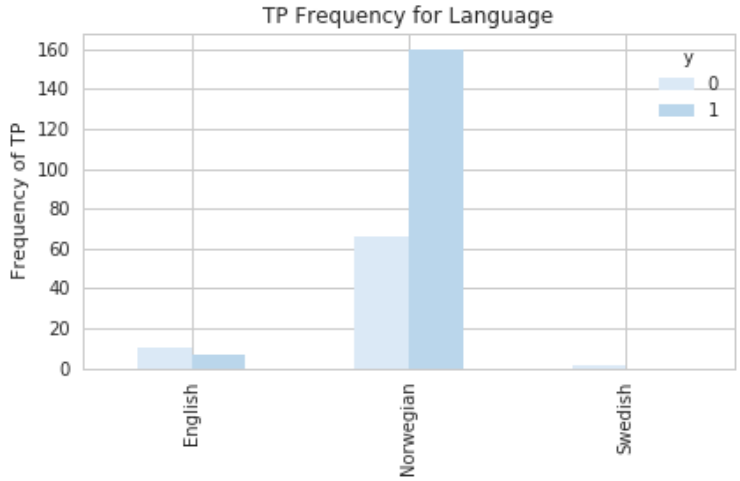


Figure 14 The figure visualizes the TP frequency for language

Figure 15 visualizes the frequency of TP and FP, FN and TN for sarcasm, abbreviations and typos. Both abbreviations and typos might be good predictors for the outcome variable. Because of few conversations containing sarcasm it does not seem to be a strong predictor for the outcome variable, TP.

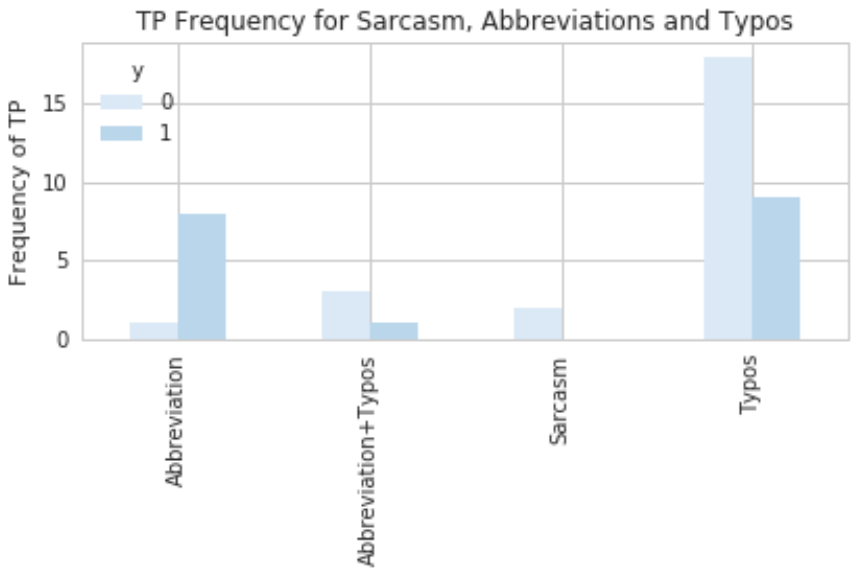


Figure 15 The figure visualizes the TP frequency for Sarcasm, Abbreviation and Typos

4.3.3.2 Implementing the model

Logit from Statsmodels is used for the implementation of the logistic regression model. Table 20 shows the results from the implementation.

	Coef	**Std err**	**z**	**P < |z|**	**[0,025**	**0,975]**
Number of words	-0,0856	0,031	-2,741	0,006	-0,147	-0,024
Number of intents	0,1931	0,346	0,558	0,577	-0,485	0,871
Number of descendant intents	-0,0008	0,001	-0,625	0,532	-0,003	0,002
Language Norwegian	1,6471	0,553	2,980	0,003	0,564	2,731
Language English	1,0745	0,736	1,459	0,145	-0,369	2,518
Misidentification of language	1,2515	1,462	0,856	0,392	-1,615	4,118
Sarcasm	-3,4617	2,349	-1,474	0,140	-8,065	1,141
Abbreviations	0,5844	0,707	0,827	0,408	-0,801	1,970
Typos	-1,5328	0,449	-3,416	0,001	-2,412	-0,653

Table 20 The table shows the implemented logistic regression model

From Table 20, it can be observed that Number of words, Language Norwegain and Typos are the only variables where their p-values are smaller than the significant level at 0,05. This implies that for the variables Number of words, Language Norwegian and Typos, the null hypothesis can be rejected, and a connection between these variables and correct predictions can be claimed.

The model has a R-squared value at 0,1262, which indicates that 12,62% of the variation in Y can be explained by the variables. This means that 87,38% of the variation in correctness of the predictions is explained by other variables than those who are included in this model.

	Coef	**Std err**	**z**	**P > |z|**	**[0,025**	**0,975]**
Number of words	-0,0560	0,023	-2,442	0,0015	-0,101	-0,011
Language Norwegian	1,4705	0,232	6,329	0,000	1,015	1,926
Typos	-1,3581	0,436	0,002	0,002	-2,213	-0,503

Table 21 The table shows the implemented logistic regression model without the non-significant variables.

The fitted model in Table 21 says that for every increase of one word the odds of predicting TP decrease by a factor of 0,95 when holding language at Norwegian and typos at a fixed value. In terms of percentage change, there is a decrease of 5% in the odds of predicting TP for a one-word increase in the number of words in the customer messages. The coefficient for language indicates that when holding the number of words and typos at a fixed value, the odds of predicting TP for Norwegian over the odds of predicting TP for English or Swedish is 4,35, which indicate 335% higher odds for Norwegian than English and Swedish. When holding the of words at a fixed level and language at Norwegian, the odds of predicting TP with typos over the odds of predicting TP without typos is 0,25, which indicates an 74,29% decrease in the odds when typos is included in the customer's message.

The logistic regression analysis indicates that of the 10 independent variables, number of words, language and typos are the three independent variables that significantly affect the chatbot's predictions. Language seems to have the largest effect on performance, and messages where the language is Norwegian seems to have significantly higher odds of being predicted as TP than messages in English and Swedish.

4.4 Estimated financial impact of improved performance

The analyses of research question 1 concluded that the two top-level intents with the weakest performance were “Funds” and “Insurance”. “Funds” performed weakly on both model classification performance and CSAT-score, while “Insurance” had the lowest overall performance.

Out of the 375 conversations in the data sample, 4,53% of the conversations are connected to “Funds”. With an average of 6109 conversations each day, the estimate for the number of conversations connected to “Funds” each year is 101 084 conversations. Figure 16 gives the confusion matrix for Funds of the conversation in the data sample. This gives that 41% of the conversations is predicted TP.

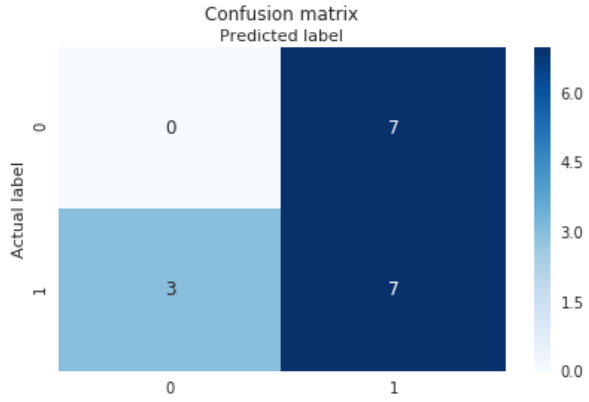


Figure 16 Confusion matrix for Funds

In this estimation of financial impact, we use cost per conversation for a conversation handled by a human agent:

Time used per conversations for human agent: 0,095 hours
Hourly labor costs: 576kr/hour
Cost per conversation: 54,72kr

With an increase of 5% of TP predictions the change in cost will be:

$$Cost\ reduction = number\ of\ conversations\ per\ year * \Delta percentage\ of\ TP * cost\ per\ conversation$$

$$101\ 084\ conversations * 5\% * 54{,}72kr = \ 276\ 565{,}82kr$$

This indicates that with an 5% increase in "Funds" prediction performance, the yearly cost reduction will be 276 565,82kr with this simplified financial model.

"Insurance" is connected to 9% of the conversations in the data sample, which gives an estimate for number of conversations connected to "Insurance" each year on 202 167 conversations. Figure 17 gives the confusion matrix for "Insurance" of the conversation in the data sample. This gives that 62% of the conversations are predicted TP.

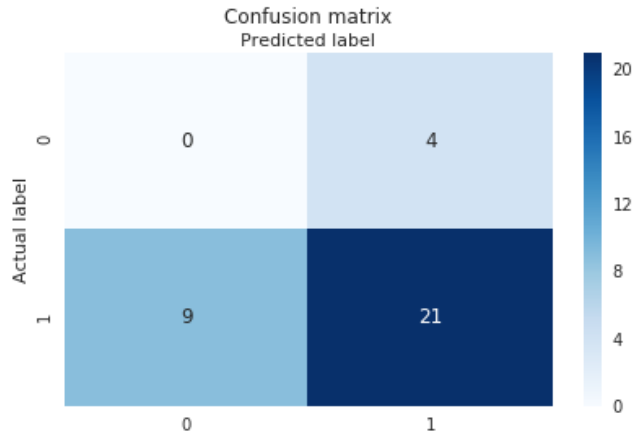


Figure 17 Confusion matrix for Insurance

With an increase of 5% of TP predictions the change in cost will be:

$$202\ 167\ conversations * 5\% * 54{,}72kr = \ 553\ 128{,}91kr$$

This indicates that with an 5% increase in "Insurance" prediction performance, the yearly cost reduction will be 553 128,91kr with this simplified financial model.

5. Discussion

The results indicate that the chatbot's performance is affected by the top-level intent, where top-level intent had a significant effect on both customer satisfaction score, automation rate and classification accuracy. Section 2.4 highlights the connection between the three different metrics, Customer Satisfaction Score, Automation rate and Classification accuracy. The findings from analyzing research question 1 indicate that there is correlation between the three performance metrics, where the correlation is especially strong between Customer Satisfaction Score and Automation rate. This indicates that the customers are more satisfied talking to a human agent than with the chatbot. Naumann suggested that to draw a new customer cost five times more than to keep an existing one. Thus, satisfied customers are efficient, which highlights the importance of focus on customer satisfaction when increasing automation rate (Naumann, 1995).

Shashavali states that long sentences from the customers make it difficult for chatbots to predict correct intents (Shashavali et al., 2019). This matches the findings that "Funds" has a significantly weaker model classification performance than the other top-level intents. In the interviews the employees indicated that customers typically ask long and complex questions connected to "Funds", and therefore the chatbot struggles to predict correctly. The logistic regression analysis reveals that the length of the sentences in customer messages has a significant impact on prediction performance. Multiple intents on the other hand did not have a significant effect on prediction performance, despite the statements from the employees that the multiple intent function did not work optimally and that it created an unnatural user experience. Xu and Sarikaya indicate that multiple intents is a challenge for a chatbot (Xu & Sarikaya, 2013), but boost.ai stated that their algorithm could distinguish between multiple intents. The findings in this study can not reject boost.ai's statement. The analysis of variance also revealed that "Funds" has the weakest CSAT-score of the top-level intents in the data sample, this might be caused by a weak model classification performance. Luo, Tong, Fang and Qu discovered that customers seem to have a negative perception against machines, and that they might feel uncomfortable talking to a computer program. They also discovered that this might lead to less purchases (Luo et al., 2019). The negative perception might be another reason why "Funds" receive such low CSAT-scores, because customer's seek advice about purchases and the stock market, and they feel uncomfortable talking to a machine about these questions. One employee stated in the interviews that the chatbot is not capable or suited to help customers with those questions, and they should be transferred to a human agent.

Kumar and Kaur stated that when classification algorithms are introduced to sarcasm they tend to get confused and produce false predictions, which matches the findings from the interviews (Kumar & Kaur, 2020). The employees said that when a customer writes sarcastic, the chatbot predicts the wrong intent because it does not understand sarcasm. The logistic regression showed no significant effect from sarcasm on the prediction performance. The data sample contained just two conversations with sarcasm, and both conversations were not predicted correctly. Even so, the number of conversations were too small to make a significant result.

Language can affect the chatbot's ability to predict the correct intent when the training sample is unbalanced between languages. Trippe stated that the challenges with multiple languages are the written shorthand, abbreviations and cultural considerations (Trippe, 2018). The employees had different opinions on how language affects prediction performance. Even though one employee stated that the chatbot's prediction performance is not affected by whether the language of the customer question is Norwegian or English, the logistic regression reveals that language has a significant effect on prediction performance. Differences in prediction performance between Norwegian and English might be caused by unbalanced training data, or the unbalanced resources used to increase performance. On the other hand, the underperformance of messages in Swedish and Danish might be caused by lack of synonyms in Swedish and Danish. The quantitative analysis revealed that in three conversations the language was misidentified, but despite the employee's thoughts that misidentification of language leads to wrong predictions, the logistic regression revealed no significant effect of misidentification of language on prediction performance.

According to boost.ai, predictions should not be affected by typos, because they use text-processing to clean up messy and complicated queries into information that the chatbot could understand (Boost.ai, n.d.). However, the logistic regression revealed that typos affect the chatbots predictions.

"Insurance" seems to be the top-level intent with the weakest overall performance. The interviews revealed several reasons why; insurance is a demanding and complicated area, the top-level intent was already created by boost.ai, and the AI trainer responsible for the top-level intent has been absent. The quantitative content analysis revealed that "Insurance" has 498 descendant intents, which makes this top-level intent significantly larger than the other top-level intents. Even though a large number of descendants also leads to a large sample of training data, it also leads to an enormous variation in questions and context. In the data

sample, 9% of all the conversations are connected to “Insurance”. Consequently, an increase of 5% in correct predictions for “Insurance”, might lead to a reduction of 553 128,91kr in labor costs for the Customer Center. On the other hand, “Funds” is only connected to 4,53% of the conversations in the data sample and therefore an increase in “Funds” performance will not lead to an equal cost reduction as “Insurance”. However, “Funds” is a top-level intent that has little content and few descendant intents. Thus, an increased focus on developing “Funds” might lead to a larger increase in performance than an equal increase in focus on developing “Insurance”.

6. Conclusion

To uncover Aino´s lost potential various quantitative and qualitative methods of analysis were executed to increase understanding of which factors contribute to the chatbot's performance. The manual review of the data sample revealed that only seven of the top-level intents in the sample were qualified for further analysis. The three metrics used to measure performance (classification accuracy, customer satisfaction score and automation rate) vary significantly across top-level intents. The results indicate a significant moderate association between top-level intent and model classification performance, and a significant moderately high association between top-level intent and automation efficiency. In addition, top-level intent had a significantly small effect on Customer Satisfaction Score. From the overall performance, "Insurance" seems to have the weakest performance. Whereas, the statistical analyses revealed "Funds" had the weakest performance on both classification accuracy and customer satisfaction score.

The structured interviews contributed a deeper understanding of factors that affected the performance, and factors that contributes to some top-level intents' underperformance. Findings indicate that the factors that have a significant effect on performance is the messages number of words, language and typos. Number of words in the customer's messages are already limited to 110 characters, but the limit could be lowered because of the significant effect on performance. That language had such a significant effect on performance was unexpected, especially due to the employee's statements. Further research on the causes of this effect is suggested. Typos' significant effect on performance was contrary to boost.ai's statements on how the text-processing clean up noise. Further research is suggested on why typos in the customer messages seem to create a problem.

A cost matrix was used to create a simple financial model that estimated the cost reduction of an increase in performance. Calculating the cost per conversation handled by a human agent, revealed that a 5% increase in performance for "Funds" could reduce the human labor cost by 276 566,82kr and an equal increase in performance for "Insurance" could reduce cost by 553 129,91kr. Even though the cost reduction for "Insurance" is twice as great, "Insurance" might also need more resources to increase performance by 5% than "Funds", since "Insurance" is already a large quantity of training sentences.

This study has uncovered that even though Aino has been a great success for DNB, there is still lost potential that could have great financial impact for the company. Due to limitations

of this study, further research with a larger sample size and a financial model that includes several aspects of Aino´s business case is suggested. In addition, a more technical error analysis from a machine learning perspective could be beneficial to get a deeper understanding of errors and their causes.

7. Bibliography

Agarwal, S. (2017). Word to Vectors — Natural Language Processing. Retrieved May 7, 2020, from towards data science website: https://towardsdatascience.com/word-to-vectors-natural-language-processing-b253dd0b0817

Bakken, J. B. (2017). *Her er 56 spørsmål DNBs robot kan svare på*. Retrieved from https://www.dn.no/handel/robotrevolusjonen/dnb-dnb/rune-bjerke/her-er-56-sporsmal-dnbs-robot-kan-svare-pa/2-1-65257

Birkett, A. (2018). What is Customer Satisfaction Score (CSAT)? Retrieved January 31, 2020, from Hubspot website: https://blog.hubspot.com/service/customer-satisfaction-score

boost.ai. (2019a). *Export API v2*.

boost.ai. (2019b). How Norway's biggest bank automated 51% of its online chat traffic with ai. https://doi.org/10.1017/CBO9781107415324.004

Boost.ai. (n.d.). Start simplifying the customer´s experience. Retrieved May 21, 2020, from https://www.boost.ai/conversational-ai-technology

Dave, A. D., & Desai, N. P. (2016). A comprehensive study of classification techniques for sarcasm detection on textual data. *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 1985–1991. https://doi.org/10.1109/ICEEOT.2016.7755036

DNB Bank ASA. (2020). About the Group. Retrieved May 21, 2020, from https://www.dnb.no/en/about-us/about-the-group.html

Expert System. (2018). Chatbot: What is a Chatbot? Why are Chatbots Important? Retrieved April 13, 2020, from https://expertsystem.com/chatbot/

Feine, J., Morana, S., & Gnewuch, U. (2019). *Measuring Service Encounter Satisfaction with Customer Service Chatbots using Sentiment Analysis*.

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In S. S. Bodrunova (Ed.), *Internet Science* (pp. 194–208). Cham: Springer International Publishing.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1).

Springer series in statistics New York.

Garbade, D. M. J. (2018, October). *A Simple Introduction to Natural Language Processing*. Retrieved from https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

Genesys. (n.d.-a). About Genesys. Retrieved April 13, 2020, from https://www.genesys.com/en-gb/company

Genesys. (n.d.-b). DNB. Retrieved April 13, 2020, from https://www.genesys.com/en-gb/customer-stories/dnb

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Retrieved from www.deeplearningbook.org

Gronholdt, L., Martensen, A., & Kristensen, K. (2000). The relationship between customer satisfaction and loyalty: cross-industry differences. *Total Quality Management*, *11*(4–6), 509–514.

Jain, V. (2018, May). Confusion & Cost Matrix helps in calculating the accuracy, cost and various other measurable factors in classification problem. *Medium*.

Khan, R., & Das, A. (2018). Introduction to chatbots. In *Build Better Chatbots* (pp. 1–11). Springer.

Kirk, R. E. (2011). Simple Random Sample. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1328–1330). https://doi.org/10.1007/978-3-642-04898-2_518

Kumar, R., & Kaur, J. (2020). Random Forest-Based Sarcastic Tweet Classification Using Multiple Feature Collection. In S. Tanwar, S. Tyagi, & N. Kumar (Eds.), *Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms and Solutions* (pp. 131–160). https://doi.org/10.1007/978-981-13-8759-3_5

Lakshmanan, S. (n.d.). How, When and Why Should You Normalize / Standardize / Rescale Your Data? *Medium*, 2019. Retrieved from https://medium.com/@swethalakshmanan14/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

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*,

38(6), 937–947.

Marous, J. (2018). Meet 11 of the Most Interesting Chatbots in Banking. Retrieved May 21, 2020, from https://thefinancialbrand.com/71251/chatbots-banking-trends-ai-cx/

Mishra, A. (2018). Metrics to Evaluate your Machine Learning Algorithm. Retrieved April 23, 2020, from towards data science website: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Narkhede, S. (2018, May). Understanding Confusion Matrix. *Medium*. Retrieved from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Naumann, E. (1995). *Customer satisfaction measurement and management: Using the voice of the customer*. Cincinnati, OHio: Thomson Executive Press.

Nicoletti, B., Nicoletti, & Weis. (2017). *Future of FinTech*. Springer.

Nordstrøm, J. (2019, October 14). *Nå får kunder i DNB råd fra en robot*. Retrieved from https://e24.no/privatoekonomi/i/0nr4k2/naa-faar-kunder-i-dnb-raad-fra-en-robot

Pizer, S. M., & Marron, J. S. (2017). Chapter 6 - Object Statistics on Curved Manifolds. In G. Zheng, S. Li, & G. Székely (Eds.), *Statistical Shape and Deformation Analysis* (pp. 137–164). https://doi.org/https://doi.org/10.1016/B978-0-12-810493-4.00007-9

Ramesh, K., Ravishankaran, S., Joshi, A., & Chandrasekaran, K. (2017). A Survey of Design Techniques for Conversational Agents. In S. Kaushik, D. Gupta, L. Kharb, & D. Chahal (Eds.), *Information, Communication and Computing Technology* (pp. 336–350). Singapore: Springer Singapore.

Raschka, S., & Mirjalili, V. (2017). *Python machine learning*. Packt Publishing Ltd.

Rychalska, B., Glabska, H., & Wroblewska, A. (2018). Multi-Intent Hierarchical Natural Language Understanding for Chatbots. *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 256–259. https://doi.org/10.1109/SNAMS.2018.8554770

Shashavali, D., Vishwjeet, V., Kumar, R., Mathur, G., Nihal, N., Mukherjee, S., & Patil, S. V. (2019). Sentence Similarity Techniques for Short vs Variable Length Text using Word Embeddings. *Computacion y Sistemas*, *23*(3), 999–1004. https://doi.org/10.13053/CyS-23-3-3273

Silverman, D. (2014). *Interpreting qualitative data* (5th ed.). 5th ed. Los Angeles, Calif: SAGE.

Stene, C. F. (2018). Hva er en chatbot? Retrieved January 7, 2020, from https://www.techweb.no/blogg/hva-er-en-chatbot

The International Society of Automation. (n.d.). What Is Automation? Retrieved January 31, 2020, from https://www.isa.org/about-isa/what-is-automation/

Trippe, B. (2018). The Challenges of Multilingual Chatbots Are Worth the Reward. *Econtent*. Retrieved from http://www.econtentmag.com/Articles/News/News-Feature/The-Challenges-of-Multilingual-Chatbots-Are-Worth-the-Reward-126105.htm

Woodford, S. (2020). Chatbot Conversations to deliver $8 billion in Cost savings by 2022. Retrieved May 21, 2020, from Juniper Research website: https://www.juniperresearch.com/analystxpress/july-2017/chatbot-conversations-to-deliver-8bn-cost-saving

Xu, P., & Sarikaya, R. (2013). Exploiting shared information for multi-intent natural language sentence classification. *Interspeech*, 3785–3789.

8. Appendix

8.1 Interview guide

1. Some chatbots experience problems when messages contain multiple intents, while boost.ai claim that their solution can handle multiple intents. From your experience, how does multiple intents seem to pose a problem for Aino?
2. Have you experienced that sarcasm affect Aino's ability to predict the correct top-level intent?
3. Variable-length sentences have shown to make it difficult for chatbots to predict the correct intent. How do you experience this issue with Aino's ability to predict intents?
4. Aino is a multilingual chatbot. How do you experience that its ability to predict the correct intent is affected by with language the customer writes?
5. Words can have different meanings, and especially the word "How" can have different meanings in different settings. Do you think when a customer writes a message containing "How", that it affects Aino's ability to predict the correct intent?
6. In a customer's message, which factors do you think are crucial for Aino's ability to predict the correct intent?
7. Of the seven analyzed top-level intents, Insurance is the top-level intent with the weakest overall performance. Which factors do you think is the reason for this weak performance?
8. Cards has a high CSAT-score, but an accuracy of only 67,9%. What do you think is the reasons for that?
9. Funds is the top-level intent with the weakest accuracy. Why do you think that is?

8.2 Columns Export API

Field	Data type	Description
Message_id	Bigint	Message log id
Date	Text	Message created date
Time	Text	Message created time
User_message	Text	Message from end user
Message_text	Text	Message from virtual agent
Message_link	Text	Message from virtual agent including any external links
Message_image_url	Text	URL of image from virtual agent
Message_video_url	Text	URL of video from virtual agent
Conversation_id	Bigint	Conversation id of the current message
Message_type_id	Bigint	Message type id varying from 1-9
Message_type_description	Text	Description of message types
Is_support_human	Boolean	True when message is from human chat
Is_customer	Boolean	True when message if from the end user
Is_human_chat	Boolean	True when message is from human
Language_id	Bigint	Id of detected language ranging from 1-25
Language	Text	Language name
Displayed_action_id	Bigint	Action id for displayed action (if any)
Came_from_action_id	Bigint	Action id for previously displayed action
Prediction_type_id	Bigint	Prediction type id, from 1-10
Prediction_type_description	Text	Description of prediction types
Predicted_intent_id	Bigint	Predicted intent id
Predicted_intent_title	Text	Predicted intent title
Root_intent_id	Bigint	Root intent id from which the predicted intent comes from
Root_intent_title	Text	Root intent title
Context_intent_id	Bigint	Context intent id
Context_intent_title	Text	Context intent title
Sentiment_neutral	Bigint	Number predicted for neutral sentiment
Sentiment_negative	Bigint	Number predicted for negative sentiment
Sentiment_positive	Bigint	Number predicted for positive sentiment
Customer_support_representative_id	Bigint	Human chat agent id
Customer_support_email	Text	Human chat agent email
Source_url	Text	Source URL from the chat API
Id_tolken	Text	Id token from openid connect
Consent_version	Text	Consent version
Last_anonymized	Text	Date and time the message was anonymized
Filter_values	Text	List of filter values

Device	Text	“Computer”, “Tablet”, “Smartphone” or “Undefined”
Feedback	Bigint	1 = thumbs up -1 = thumbs down 0 = no user feedback or user removed feedback

8.3 Columns Customer Satisfaction Score Information

Column name	Datatype	Description
Chat_survey_info_pk	Bigint	Generated number for each row in table
Unique_kti_id	Character varying	Unique id for conversations with KTI
Gms_chat_id	Character varying	Unique id from Genesys-chats
Unique_chat_id	Character varying	Unique id for all chats
Chat_mode	Character varying	Bot or Agent
Chat_duration	Bigint	Number of second the chat lasts
Chat_date	Timestamp without time zone	The date of the chat
Customer_info	JSONb	Private customer or business customer Boolean information about authentication
Bot_conversation_details	JSONb	Private or business chatbot
Surveys	JSONb	Information about KTI answers
Conversation_info_pk	Character varying	The primary key for the conversation_info table
Conversation_createdon	Character varying	Date and time when conversation was created
Conversation_state	Character varying	State of the conversation
Conversation_unique_id	Character varying	Unique id for each conversation
Conversation_updatedon	Character varying	When was the conversation last updated (in chat_conversation_info_table)
Is_active	Boolean	It is being fixed, just to show if the conversation is active or not
Is_auth_conversation	Boolean	If the customer is logged in or not
Tenant_channel_fk	Integer	Which bot the user is talking to, AINO in 2
Bot_conversation_ref_id	Float	Boost admin panel id
Converation_history_pk	Integer	Primary key for chat_conversation_history table
Conversations_history	JSON	Messages with positive or negative feedback
Ipa_posted_date	Character varying	When was the log posted to IPA
Is_posted_ipa	Boolean	If the log was posted to IPA
Logged_time	Character varying	When the chat is logged
Chat_conversation_info_fk	Integer	Foreign key for chat conversation info table
Is_migrated_data	Boolean	If data is migrated
Chat_bot_logs_pk	Float	Primary key on chat bot logs table
Created_date	Character varying	When the entry in chatbot logs table was created
Updated_date	Character varying	When the entry in chatbot logs table was updated
Chat_bot_logs_history	JSONb	History of the conversation with the messages
Bot_conversation_id	Float	Boost admin panel id

8.4 Columns Chat Conversation Parties

Column name	Datatype	Description
Conversation_party_pk	Integer	Primary key for table
Joined_on	Timestamp without time zone	Time for joining tables
Left_on	Timestamp without time zone	Time for left joing tables
Party_ref	Character varying	Id for parties, chatbot, user and human agent
Party_type	Character varying	Title for parties, chatbot, user and human agent
Conversation_info_fk	Integer	Primary key in chatbot logs

8.5 Full list of top-level intents

	Name of intent		Name of intent
1	Accounts	16	Insurance
2	Bank services	17	Loans
3	Become a customer	18	Log in
4	BSU – Home saving for young people	19	Payment
5	Cards	20	Pension
6	Change bank	21	SAGA
7	Chat buttons	22	Savings
8	Change contact information	23	Separation
9	Complaints	24	SoMe-spesifikt
10	Currency	25	Talk to advisor
11	Equities	26	Tax return
12	Funds	27	Vipps
13	General questions	28	Work assessment allowance (AAP)
14	Guardianship	29	Z_Arkiv
15	Inheritance and estate		

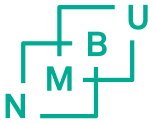
8.6 Data pre-processing

Reduction of the number of files from 47 to three files; chatbot logs, CSAT-data, and parties' data.

1. The 14 chatbot logs were read one by one into Pandas' data frames.
2. The JSON-objects in the chat history column was normalized using JSON_normalize.
3. Each file was made into a data frame and concatenated using Pandas. The concatenated data frame was then written to a CSV file.
4. The 18 CSAT-data files were read one by one into Pandas' data frames.
5. Since both the columns survey and questions contain JSON-objects, JSON_normalize needed to be run twice to get all the columns.
6. Each file was made into a data frame and concatenated using Pandas. The concatenated data frame with the CSAT data was written to a CSV file.
7. The 15 parties' data files were read one by one into Pandas' data frames.
8. The data frames were concatenated into one data frame using Pandas.
9. The concatenated data frame containing parties' data was written to a CSV file.

After reducing the number of files from 47 to 3, the files needed to be joined.

10. The chatbot logs and CSAT-data were joined using Pandas' inner merge, which uses the intersection of keys from both data frames. These two data frames were merged using bot conversation id, which is found in both data frames and therefore, can be used for joining. Because of the inner join with chatbot logs and CSAT-data, only the conversations with customer ratings will be kept. The chat logs contained 171 052 conversations, while the CSAT-data contained 19 821 conversations, which gives us that 11,59% of the conversations in these logs were rated by customers, which means that only 11,59% of these conversations will be kept in the dataset for the analyses.
11. The merged data frame was joined with the parties' data. The two frames were joined using chat conversation info fk and conversation info fk. These two columns are foreign keys from conversation info in both data frames and contain the same number, and can, therefore, be used for joining the data frames.



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway