UTRECHT UNIVERSITY

MASTER THESIS

Quantifying Chatbot Performance by using Data Analytics

Author:
Cas JONGERIUS

Thesis supervisor:
Dr. Matthieu BRINKHUIS

External supervisor:
Joop SNIJDER

Second thesis supervisor:
Dr. Marco SPRUIT

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Master in Business Informatics
Graduate School of Natural Sciences

July 16, 2018

Acknowledgments

I wish to thank various people for their contribution to this master thesis:

- Joop Snijder for personally guiding me during the internship and for arranging all facilities to make life easier. His knowledge of chatbots and his general view on Artificial Intelligence provided me with fruitful insights.
- Dr. Matthieu Brinkhuis for his pleasant supervision of this research projects. His insights and enthusiasm inspired me to challenge myself in exploring data analysis approaches that were unknown to me.
- Dr. Marco Spruit for his oversight on this research project as my second supervisor.
- Willem Meints for sharing his domain knowledge during all stand-up meetings.
- Fellow graduate interns & colleagues at Info Support for enabling a very pleasant work environment. Their words of wisdom, moral support and humor made the graduation period very enjoyable.

Cas Jongerius - 16th of July 2018

Contents

Acknowledgments **iii**

1 Introduction **1**

2 Research Set-up **3**
2.1 Problem Statement and Research Objective 3
2.2 Research Questions 3
2.3 Relevance 4
2.3.1 Academic Relevance 4
2.3.2 Practical Relevance 4
2.4 Research Approach 4
2.4.1 Systematic Literature Review 6
2.4.2 Case Study 7
2.4.3 Survey 8
2.4.4 Data Analysis 8
2.5 Validity 11

3 Chatbot Evaluation Methods **13**
3.1 Questionnaire-based 14
3.1.1 Expert Review 14
3.1.2 User Opinion 15
3.1.3 Combined Approach 16
3.2 Automatic Metrics 16
3.2.1 Metric Attributes 17
3.2.2 Chatbot Attributes 18
3.2.3 Metric Analysis Approach 19
3.2.4 The Automatic Metrics 19
3.2.5 LIWC 22

4 Data Analysis **25**
4.1 Data Collection 25
4.1.1 Open Dataset 25
4.1.2 Case Study Dataset 26
4.2 Data Preparation 26
4.3 Descriptive Statistics 27
4.4 Hypothesis Testing 27
4.4.1 Patterns in Automatic Metrics 28
4.4.2 Relations to Perceived Performance 30
4.4.3 Hypotheses Insights 31
4.5 Predictive Analysis 32
4.5.1 Decision Tree 33
4.5.2 Principle Component Analysis 33
4.5.3 Random Forest 34

5 Conclusion **37**
5.1 Conclusion of Sub-Questions 37
5.2 Conclusion of Main Research Question 38

6 Discussion **39**
6.1 Limitations 39
6.2 Future Work 40

A Appendix: Case Study **47**
A.1 Information Sheet 47
A.2 Informed Consent 48
A.3 Experiment Instructions 49
A.4 Survey 50

B Appendix: Automatic Metrics **51**

C Appendix: Results **55**

List of Figures

2.1 Design Cycle within the Engineering Cycle 5
2.2 CRISP-DM breakdown 9
2.3 CRISP-DM process model lifecycle 10

3.1 Chatbot evaluation methods 14
3.2 Chatbot maturity levels 19

4.1 Dialogue length distribution 28
4.2 User rating distribution 28
4.3 Sentiment of messages 29
4.4 Hypothesis 1 29
4.5 Hypothesis 2 & 3 30
4.6 Hypothesis 4 & 5 31
4.7 Variance explained by principal components 33
4.8 Random forest 34
4.9 Prediction results for DS_3 - User 35
4.10 Random forest prediction results for DS_2 (1^{st} half) - Chatbot 36
4.11 Random forest prediction results for DS_3 (1^{st} half) - User 36

List of Tables

3.1 Chatbot evaluation concepts and underlying categories 15
3.2 Chatbot evaluation characteristics 16
3.3 Automatic metric attributes 18

4.1 Metadata of the datasets 26
4.2 Pearson correlation results per hypothesis 32
4.3 User rating statistics per dataset split 32
4.4 PCA: MSPE user rating per dataset split 33
4.5 Random forest: MSPE user rating per dataset split 34
4.6 Random forest: MSPE user rating per dataset split (1st half) 35

A.1 Case study survey 50

B.1 Automatic Metrics in datasets 51

C.1 Top-9 most predictive variables in random forest 55

List of Abbreviations

NLP	**N**atural **L**anguage **P**rocessing
CRISP-DM	**CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
AI	**A**rtificial **I**ntelligence
CUI	**C**onversational **U**ser **I**nterface
KDD	**K**nowledge **D**iscovery in **D**atabases
SEMMA	**S**ample, **E**xplore, **M**odify, **M**odel, and **A**ssess
QoS	**Q**uality **of** **S**ervice
QoE	**Q**uality **of** **E**xperience
GS	**G**olden **S**tandard
LIWC	**L**inguistic **I**nquiry and **W**ord **C**ount
POS	**P**art **O**f **S**peech
GO	**G**oal **O**riented
G&O	**Q**uestion & **A**nswer
WoZ	**W**izard **of** **O**z
PM	**P**rediction **M**odel
DS	**D**ataset **S**plit
SD	**S**tandard **D**eviation
PCA	**P**rincipal **C**omponent **A**nalysis
Bagging	**B**ootstrap **Agg**regat**ing**
OOB	**O**ut **O**f **B**ag
TN	**T**rue **N**egative
FN	**F**alse **N**egative
TP	**T**rue **P**ostive
FP	**F**alse **P**ostive

1 Introduction

Ever since the rise over the internet, it has influenced the way humans communicate. Instead of locally conversing with a small group of people, the internet provided us with the possibility to reach out to a much greater number of people on a global scale. Starting two decades ago, social media has played an important role in changing the way people communicate (Mihailidis, 2014). However, since the last two years, social networks are losing terrain while messenger applications flourish. The number of monthly active users of the four largest messenger applications surpassed the four largest social networks recently (Statista, 2017). In the last few years, people shifted from social broadcasting to a more personal variant: social messaging.

This trend is not limited to social interactions but can be extended to customer behavior as well. Consumers want access to personalized information on demand, preferably 24/7, and in any language. This trend has strengthened due to the recent advancements in machines, especially in artificial intelligence and mobile internet. As businesses acknowledge the urge to adapt to this trend, a rising number of early adopters consider deploying text-based conversational agents (McTear, Callejas, & Griol, 2016, pp. 51–72), also called chatbots, as a method of communicating with its customers (Van Eeuwen, 2017). These conversational interfaces "enable people to interact with smart devices using spoken language—just like engaging in a conversation with a person" (McTear et al., 2016). Although the utilization of chatbots is not a new development, the recent increase in popularity makes it a hot topic.

In an ideal situation, the intelligent behavior exhibited by a chatbot is indistinguishable from that of a human. In 1950 Alan Turing proposed a test situation in which a computer intelligence's indistinguishability with that of a human can be measured (Turing, 1950). Up to today, the Turing test has not been fully passed. According to Ray Kurzweil, director of engineering at Google, the first Artificial Intelligence (AI) capable of passing the Turing test, will probably not be built before 2029 (SXSW conference in Austin, Texas, 9 March 2017).

For this reason, a chatbot's performance is likely to be divergent from that of a human, meaning businesses deploying chatbots to interact with customers should recognize the likelihood that their customers sense the artificialness of their communication partner. Since the performance of a chatbot directly influences the user experience (Lemon & Verhoef, 2016), businesses should aim at enhancing their chatbot's performance to keep their customers satisfied. According to the continuous improvement and learning curve theory (Zangwill & Kantor, 1998), businesses should improve their services, such as chatbots, by taking improvement actions, evaluating the service, and thereby learning how to improve it further over time.

The narration above indicates that the continuous evaluation of a chatbot's performance is crucial for businesses to keep their customers satisfied. The focus of the research that follows from this proposal is to investigate what evaluation methods exist to assess a chatbot's performance and whether this evaluation can be automated. The intention of this proposal is to provide an overview of how the research will be performed. In chapter 2, the research set-up is described by elaborating on the problem statement, the accompanied research objectives, the project scope and the research approach. In chapter 3, the literature findings are discussed that form the foundation for the rest of this research. Next, chapter 4 describes the data analysis procedure and results. In chapter 5, a conclusion of the research is provided. Finally, the research project is reflected in chapter 6.

2 Research Set-up

2.1 Problem Statement and Research Objective

Since establishing a comprehensive overview of a chatbot's performance is essential, a wide range of evaluation frameworks that measure the quality of chatbots have been proposed in the past. In a recent study, Radziwill and Benton (2017) provided an overview of chatbot quality attributes and assessment frameworks that have been proposed by researchers over the last decades. However, their study concludes that an absolute list of quality attributes for evaluating chatbots is non-existing, due to the wide variety of chatbot types. Moreover, properly assessing a vast majority of the quality attributes is difficult, time intensive or expensive, because doing so requires the opinion of users or experts.

The objective of this research project is to fill this gap by researching what metrics can be automatically measured by analyzing chatbot conversations and how they relate to the chatbot's performance. Automatically quantifying chatbot performance allows for faster prototyping and testing of new chatbot models, requiring less expensive human evaluations.

2.2 Research Questions

Based on this research objectives, the following main research question is defined:

[MRQ] *"How can the performance of chatbots be automatically quantified by analyzing its conversations?"*

To address the main research question, it is subdivided into the following sub-questions:

[SQ1] *"Which approaches exist for evaluating chatbot performance?"*

In the initial phase of this research project, a literature study is performed to create a brief overview of approaches that can be used to evaluate chatbot performance.

[SQ2] *"Which metrics can be automatically measured by analyzing chatbot conversations and with what techniques?"*

Subsequently, the literature study is extended by researching which metrics can be automatically measured by analyzing text conversations. Moreover, for each metric, the measuring techniques are identified and elaborated on.

[SQ3] *"What patterns can be discovered in the automatic metrics and how are they related?"*

Next, two datasets containing chatbot conversations are gathered to perform a data analysis on. The first dataset is collected from an online available source. Moreover, a case study is performed to gather a second dataset that is enriched with perceived performance scores per conversation. Subsequently, the earlier identified automatic metrics are applied to the gathered datasets. The scores of the automatic metrics are analyzed to discover reciprocal patterns and correlations.

[SQ4] *"How can the automatic metrics be related to the perceived performance of chatbots?"*

Finally, the data analysis on the case study dataset is extended to determine how the automatic metrics relate to the perceived performance of chatbots. The goal of this final phase is to create a model, which to capable of analyzing chatbot conversations, to make predictions about the perceived performance.

2.3 Relevance

The relevance of this research project is described from both an academic and a practical point of view.

2.3.1 Academic Relevance

Previous research on conversational user interfaces shows that creating a good performing chatbot is a very sophisticated and challenging task. As mentioned earlier, the current state of chatbots is still far from perfect. Therefore, evaluating chatbots is essential to determine its strengths and weaknesses. In the past, multiple chatbot evaluation frameworks have been created by scientists. However, the primary critique of those frameworks is the necessity of manual input by a human. Instead, these researchers allude to the creation of accurate automatic evaluation procedures (Lowe et al., 2017). We believe this research project is scientifically relevant because it contributes to existing frameworks by further researching the field of automated chatbot evaluations.

2.3.2 Practical Relevance

As the capabilities of chatbot technologies are increasing, so are the possibilities for businesses to apply them in practice. According to Lester, Branting, and Mott (2004), the five major fields of business applications, in which chatbots can play an important role, are customer service, help desk, website navigation, e-commerce, and technical support. The trend of businesses applying chatbots in these fields is in line with the expectations of customers. In a recent customer survey, a majority of the respondents indicated that they are interested in using messaging apps to interact with organizations (HubSpot, 2017). Furthermore, in a similar study, the majority of respondents indicated that they expect organizations to be open 24/7 (Venturebeat, 2016).

However, as mentioned in the introduction, businesses implementing chatbots should recognize the likelihood that the chatbot's performance is divergent from that of a human. For this reason, businesses aim for a continuous improvement of their chatbot's performance while keeping the costs low. Therefore, the demand for an efficient and effective evaluation framework increases. This research contributes to this need by increasing the chatbot evaluation efficiency, saving businesses time and resources.

2.4 Research Approach

This section describes the research methods that are used in this project. The project itself is framed around the Design Science Framework (Wieringa, 2014). Additionally, the systematic literature review, comparison analysis, data analysis, and case study approaches are used to perform numerous research tasks, such as problem investigation, the creation of an artifact, data analysis and the validation of the proposed solution.

In this research, the iterative problem-solving method proposed by Wieringa (2014) is adopted. The method is aimed at solving a problem by creating a new artifact through iterative investigation and design processes. To achieve the research goal, the research activities are broken down into a set of tasks. These tasks are according to the structure of the design cycle, a subset of the engineering cycle, see Figure 2.1. The four tasks of the engineering cycle are:

1. Problem investigation: the design of a treatment is prepared by learning more about the problem to be treated. Stakeholders, desired goals, problems, phenomena, and effects are investigated.

2. Treatment design: requirements are specified, available treatments are identified, and new artifacts are designed for the treatment.

3. Treatment validation: the tradeoffs, effects, and requirements that are satisfied by the artifacts are determined.

4. Treatment implementation: the artifacts are applied in a real-life situation.

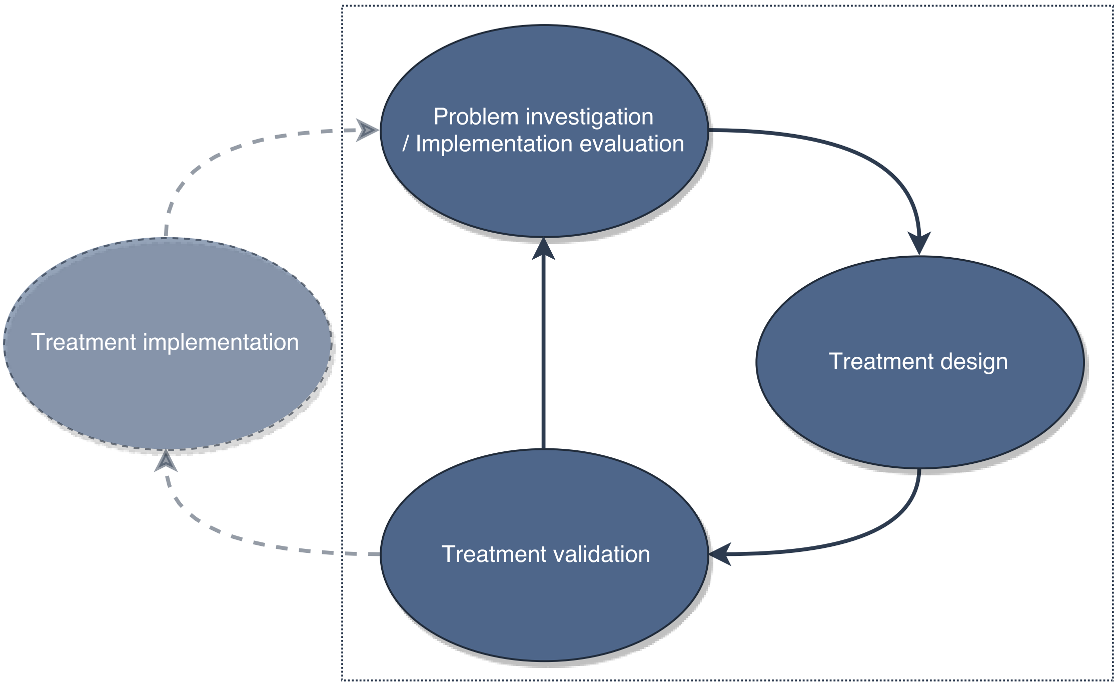


FIGURE 2.1: Design Cycle within the Engineering Cycle

In design science, the last task of the engineering cycle is not performed (Wieringa, 2014). The engineering cycle is often carried out in long-term research projects, for which the research solution can be implemented in a real-life situation. According to Wieringa (2014), design science research projects are restricted to the design cycle because transferring new technology to the market is not part of the research project. Therefore, the design cycle is adopted in this research project. The steps that are taken in each task are listed below. Subsequently, the research approaches that are named are described in more detail in the sections below.

Problem Investigation

The first task is to conduct exploratory research to get an understanding of the problem at hand. To get this understanding, related research from the past is examined (Webster

& Watson, 2002), by conducting a systematic literature search (Kitchenham, 2004). This activity is performed to find the generic approaches for evaluating chatbot performance [SQ1]. Moreover, the systematic literature search is extended to identify metrics that can be automatically measured by analyzing chatbot conversations, including their corresponding measuring technique [SQ2].

Treatment Design

The chatbot evaluation approaches and automatic metrics that are found during the previous task are compared by performing a qualitative research synthesis in the form of a narrative literature review (Baumeister & Leary, 1997). The first aim of the narrative literature review is to construct a brief overview of the most common chatbot evaluation approaches [SQ1]. The second aim is to create a non-exhaustive list of the most relevant automatic metrics that can be measured by analyzing chatbot conversations [SQ2].

Subsequently, a large online available dataset is gathered. Moreover, a dataset with chatbot conversations and corresponding performance scores is collected by performing a case study (Runeson & Höst, 2009), with a survey element included (Robson & McCartan, 2016). Thereafter, data analysis is performed on the gathered datasets by following the CRISP-DM methodology (Wirth & Hipp, 2000). First, the datasets are split into two subsets: a training set and a test set. Next, data mining techniques are used on the training set to discover patterns in the automatic metrics [SQ3] and to determine whether these metrics can be linked to the chatbot's (perceived) performance [SQ4]. These findings result in a prediction model that estimates a chatbot's performance by automatically analyzing its conversations.

Treatment Validation

In the final task, the correlations of the automatic metrics and the prediction model are validated. Correlations between the automatic metrics in the case study dataset are validated on the online dataset. Moreover, the prediction model is validated by testing it on the test set. This is done to estimate the accuracy of the model.

2.4.1 Systematic Literature Review

Because chatbots already exist for many years, an extensive number of studies have been previously performed on the topic of chatbot metrics and chatbot evaluation approaches. Contemporary literature is investigated in this study by performing a systematic literature review, to build a solid understanding of existing artifacts. The aim of the systematic literature review is to identify, critically evaluate and integrate the findings of all relevant, high-quality studies that address one or more research questions (Baumeister & Leary, 1997). The systematic literature review is split up into a systematic literature search followed by a narrative literature review.

Systematic Literature Search

To identify relevant literature, a keyword-search is carried out on the scholarly databases: Google Scholar, ResearchGate, WorldCat, and the Computer Science Bibliography (DBLP). The search is performed with the following input:

- Keywords: `chatbot, conversational user interface, chatbot AND evaluation, conversational user interface AND evaluation, quantitative evaluation AND chatbot, chatbot quality metrics, conversation metrics, conversation notation standard, chatbot AND sentiment analysis, chatbot metrics, chatbot AND perceived performance;`
- Year of publication: 2011 or newer;
- Document types: books, journal articles, conference proceedings, and theses.

The relevancy of the found sources is determined by their scope, objectives, methods, and conclusion subjectively (Budgen & Brereton, 2006). Sources published by IEEE, Springer, ACM and Elsevier are preferred, due to their typical state-of-the-art research articles in the field of Computer and Information Sciences. Moreover, each source should fulfill the following requirements: the source should be available in digital format; the source language should be English; and the source should be a book, journal article, conference proceeding or thesis. Additionally, the literature review is extended by performing both backward and forward searching (Levy & Ellis, 2006), to ensure a complete census of relevant literature. Forward searching identifies more recent work, by reviewing the sources that have cited the found articles. Backward searching is performed to identify high-quality research projects, on which the found articles are based. The two approaches are repeated until no new relevant chatbot quality metrics and corresponding methods of measurements are found.

Narrative literature Review

Subsequent to the systematic literature search, the relevant sources are reviewed by using the narrative literature review approach. The advantage of the approach is that it can integrate results from very different methods and procedures (Baumeister, 2013). According to Baumeister, recognition of methodological diversity is a major advantage of a narrative reviewer. The goal of the review is to critically compare the found chatbot quality metrics and their corresponding methods of measurement.

2.4.2 Case Study

Next, in order to answer the third and fourth research question, a data analysis of chatbot conversations is required. To collect a dataset, the case study research methodology is applied. Although numerous case study definitions exist, multiple pieces of research agree on the following definition: a case study is an empirical method aimed at investigating contemporary phenomena in their context (Benbasat, Goldstein, & Mead, 1987; Robson & McCartan, 2016; Runeson & Höst, 2009; Yin, 2013). Robson and McCartan extent this definition by stating that a case study can be considered a research strategy that uses multiple evidence sources. Moreover, Benbasat et al. state that case studies gather information from few entities and lack experimental control.

In this research, a group of respondents is given the task to chat with a chatbot. The conversations are stored to form a dataset with chatbot conversations from multiple respondents. The dataset collected during the case study is subsequently analyzed to seek new insights and generate ideas for an automatic evaluation model, and can, therefore, be considered as exploratory (Robson & McCartan, 2016). Since the respondents are given instructions but also free to interact with the chatbot as they like, we consider this case study as partially controlled.

Furthermore, besides storing the conversations, the respondent's opinion is captured as well. For this reason, the case study can be considered interpretive because it attempts to understand phenomena through the participants' interpretation of their context (Klein & Myers, 1999). Moreover, the goal of the case study is to collect both a set of chatbot conversations, which can be considered quantitative, and the opinion of the respondents, which can be considered qualitative. Therefore, the case study method performed in this research project can be considered mixed (Robson & McCartan, 2016). Lastly, the case study can be considered fixed because the to-be collected data, the conversation itself and the respondent's opinion, is determined up front (Anastas, 1999).

2.4.3 Survey

As mentioned earlier, a survey element is added to the case study to include the capturing of the respondent's opinion on the chatbot conversation. The survey in this research project is conducted by means of a questionnaire because it is an effective and one of the most common approaches (Robson & McCartan, 2016). Including a survey that captures the respondent's opinion, allows for discovering patterns between the automatic metric scores and the perceived performance of the chatbot. This contributes to the main goal of this research, namely using these findings to propose a model that estimates a chatbot's perceived performance by automatically analyzing its conversations.

2.4.4 Data Analysis

Subsequently, the collected datasets are analyzed in order to answer the third and fourth sub-questions. According to Leek and Peng (2015), questions that can be answered by performing data analysis can be categorized into the following types: *descriptive, exploratory, inferential, predictive, causal* and *mechanistic*.

The third sub-question [SQ3] can initially be considered exploratory because first the data is analyzed to see whether there are trends, patterns, or relationships between variables (Peng & Matsui, 2016). Peng and Matsui also call this question type *hypothesis-generating* because the analysis is performed to look for patterns that could support the proposition of a hypothesis, rather than testing a predefined hypothesis. The result of the analysis is a set of hypotheses for which other chatbot conversations are analyzed to quantify whether they will hold beyond the initial datasets. Therefore, these newly proposed hypotheses can be classified as either inferential, predictive, causal or mechanistic. Furthermore, the analysis performed to answer the fourth sub-question [SQ4] is identical. First, an exploratory data analysis is performed to look for trends, patterns, or relationships between variables. Next, to fully answer this sub-question, the newly proposed hypotheses are analyzed by performing either inferential, predictive, causal or mechanistic data analyses.

In other words, many steps are carried out in order to find the answers to the sub-questions. To structure these activities, multiple data analysis methodologies can be followed. The most popular methodologies are KDD, SEMMA and CRISP-DM (KDnuggets, 2014). The choice was made to follow the CRISP-DM methodology (Chapman et al., 2000) because it is the most comprehensive process model for carrying out data mining projects (Azevedo & Santos, 2008; Wirth & Hipp, 2000). Recently, IBM released a refined and extended version of the methodology called ASUM-DM (IBM Corporation, 2016). However, the made changes are mainly focused on improving the infrastructure, operations and management side of implementing data mining projects (Haffar, 2016). Since these aspects are not relevant for this research project, the choice was made to stick to the original CRISP-DM methodology.

The CRISP-DM methodology, standing for CRoss-Industry Standard Process for Data Mining, was conceived during the late 90's by Chapman et al. (2000), to provide a standard process model for data mining projects. The methodology consists of sets of tasks described at four levels of abstraction, from general to specific: phase, generic task, specialized task, and process instance, see Figure 2.2. At the top, the model is divided into multiple phases, which each have a subset of generic tasks. These phases and tasks should be applicable to every data analysis situation and are therefore generic. At the bottom, the generic tasks are mapped to the situation at hand. The specialized tasks describe how the generic tasks are carried out for a specific situation. At last, the process instances are representing what happened in a particular engagement.

Furthermore, the methodology breaks the data mining process into six main phases, see Figure 2.3. The sequence of the phases is not fixed. While the project progresses, moving back and forth between the phases is required. As described above, multiple data analysis iterations are performed to reach the final goal. Therefore, the cyclical nature of data analysis is symbolized by the outer circle in Figure 2.3.

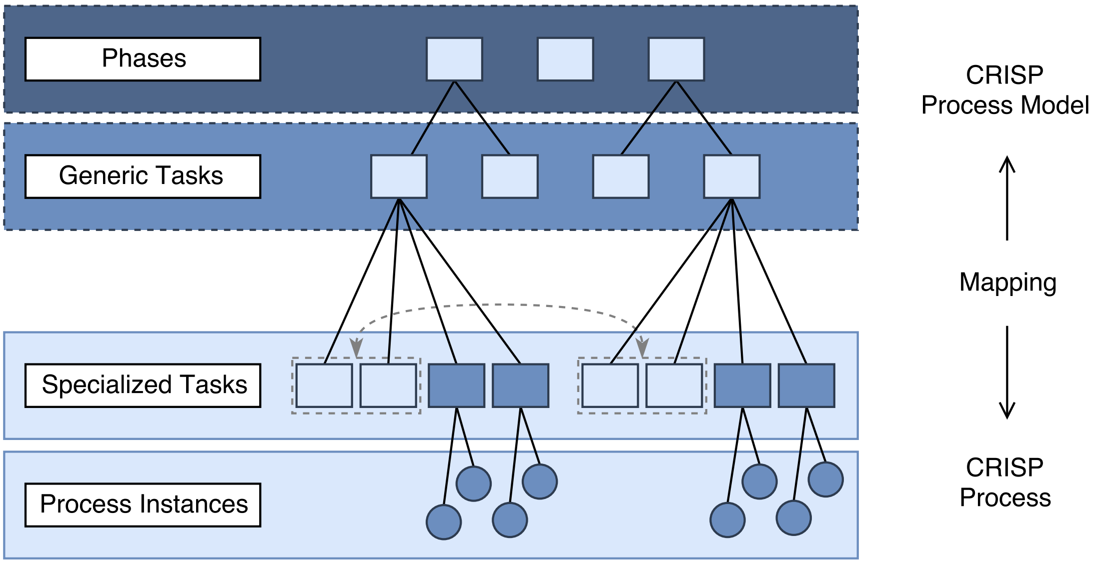


FIGURE 2.2: CRISP-DM breakdown

As mentioned earlier, the treatment implementation task of the engineering cycle is not performed in this study. Since the main goal of the CRISP-DM deployment phase is to implement the created model in day-to-day business activities, the phase is part of the last step in the engineering cycle and is therefore not performed. The other five phases are performed, however. Below, each phase is described including their corresponding generic tasks. The generic tasks are derived from the ones proposed by Chapman et al. (2000).

Business Understanding

The aim of the first phase is to get an understanding of the problem at hand, the potential solutions, and the goals. The business objectives determination task that is originally proposed for this phase, is replaced with a problem investigation task because it better fits the scope of a research project. This first phase is intertwined with the problem investigation task of the design cycle because the aims of the two are similar. Therefore, the systematic literature review is part of this phase.

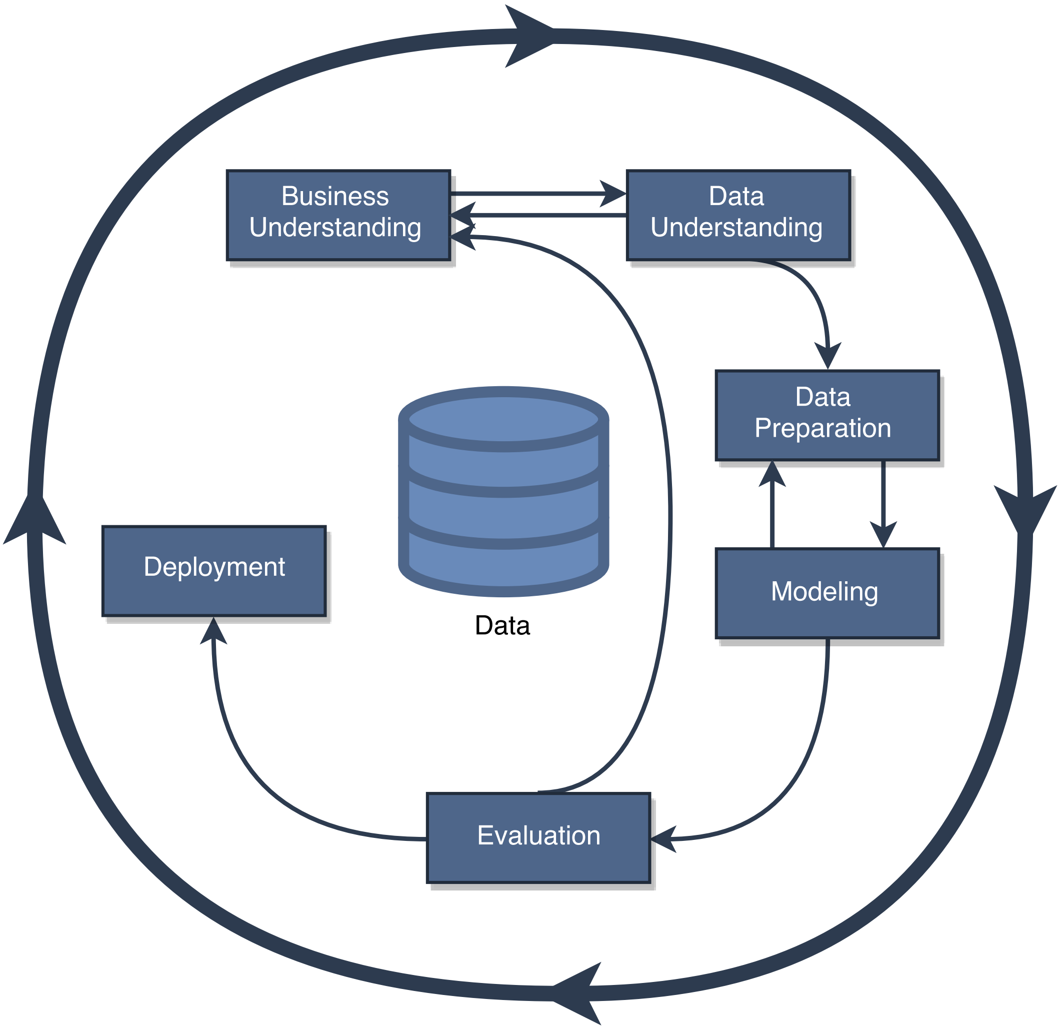


FIGURE 2.3: CRISP-DM process model lifecycle

Data Understanding

The second phase is performed to get a good understanding of the data that is analyzed. This phase focuses on collecting, describing, exploring and verifying the data. These tasks can all be considered as a part of the treatment design task of the design cycle. Therefore, the case study and the survey are part of this phase.

Data Preparation

Subsequent to understanding the data, preparation tasks are performed to prepare for the actual data analysis. During this phase, the case study and survey results are combined to form one integrated dataset.

Modeling

The actual data analysis is performed during the modeling phase. The modeling phase will be repeated multiple times to be able to answer the sub-questions.

Evaluation

After a model has been proposed, the final task is to evaluate the result. At the end of this phase, the decision is made whether the previous phases are to be re-executed or not.

2.5 Validity

To enforce that repeating this research project in an unaltered situation with different subjects has a similar outcome, the project needs to be reliable and valid (Yin, 2013). Although not identical, the two concepts do relate to one another. Reliable research does not imply it's valid, but unreliable research always implies that it is invalid (Baarda, 2014). Below, we elaborate on our efforts to ensure this research project is both reliable and valid.

Reliability

The reliability of research describes the extent to which the results are independent of chance. For research to be reliable, the used methodology should be stable to ensure that the measured results are stable. Therefore, the used research methodologies are specified in advance and described in great detail in section 2.4. Furthermore, the tools, R packages, and techniques that are used to transform the chatbot conversations into automatic metrics are described extensively in subsection 3.2.4 to ensure the reliability.

Validity

The validity refers to the credibility of the research project. We distinguish three validity types: *construct*, *internal* and *external*.

Construct validity relates to the establishment of correct operational measures for the studied concepts. In this research, the main construct to be measured is the performance of a chatbot. To ensure the measured chatbot performance scores are valid, we based the measuring technique on the results of numerous scientific research projects.

Internal validity refers to the degree to which confounding is avoided. This study is aimed at finding correlations between automatic metrics and chatbot performance. By including a large number of automatic metrics (dependent variables) during the data analysis, we aim to reduce the chance of missing alternative causes that influence the chatbot performance. Furthermore, to minimize bias in this study, the people that participated in the case study vary in age, background, educational attainment and experience with chatbots.

External validity relates to the generalizability of the research findings. However, this research project focuses on specific chatbots only, as is described in subsection 3.2.2. According to Lee and Baskerville (2003), a theory may never be generalized to a setting where it has not yet been empirically tested and confirmed. Therefore, the findings of this research are limited to the defined scope.

Finally, triangulation is applied to strengthen the overall validity of this research project. We distinguish three types: data source, theory, and methodological triangulation. The data analysis in this research project is performed on two datasets that originate from different sources. Moreover, multiple statistical and theoretical approaches are used to interpret the analysis results. Finally, multiple methodologies, such as the CRISP-DM, case study design, and design science are followed to gather and analyze the data.

3 Chatbot Evaluation Methods

In this section, previous research on chatbot evaluation methods are examined and a distinction is made between two different chatbot evaluation methods types. Subsequently, each method type is elaborated on by reviewing the current state of these evaluation methods in recent studies.

Ever since the first chatbot was developed, creators aim at measuring how well their chatbot is performing in order to improve it. Previous research on the topic of evaluating chatbots has led to multiple questionnaire-based evaluation methods. By examining existing literature on the topic of questionnaire-based evaluation methods, we distinguish two types: expert review and the user opinion.

Elaborating on the former, experts score an exhaustive list of chatbot evaluation criteria. According to Moller, Engelbrecht, Kuhnel, Wechsung, and Weiss (2009), an expert evaluation result in a measure that reflects the Quality of Service (QoS). Since expert evaluation methods that incorporate such criteria are created by academics who conducted valid scientific research, we consider these methods as most truthful and therefore assume these capture a chatbot's performance best.

Yet, the performance of a chatbot is not limited to the capturing of the QoS only. As Moller at al. argue in their paper, a QoS evaluation does not necessarily reflect the user satisfaction. By letting normal users review and evaluate chatbots with a collection of other evaluation criteria, a measure can be determined that reflects the Quality of Experience (QoE) of a chatbot. Users are generally subjective during system evaluations because they base their answers on how they perceive and experience the interaction with the chatbot. For this reason, the results of a user evaluation might not reflect the actual chatbot's performance but instead reflect the perceived performance of the chatbot. However, because the experience of the user is most relevant for many organizations that deploy chatbots, the perceived performance is of higher importance than the performance measured by experts. Therefore, we call the perceived performance the Golden Standard (GS) in this study.

However, the questionnaire-based evaluation methods are difficult to perform automatically with existing techniques because a vast majority represent qualitative values, such as measuring a chatbot's ability to respond to social cues, ethics and cultural knowledge of users and measuring a chatbot's trustworthiness. At this point in time, reliably measuring these attributes can only be done by a human and therefore is resource and time intensive. The goal of this project is to be able to quantify chatbot performances by solely assessing the automatically measurable metrics. Therefore, we distinguish another evaluation method type, called automatic metrics. At this point in time, very little research has been conducted that tried to associate metrics with the (perceived) performance of chatbots.

Concluding, we distinguish a total of two evaluation method types: the questionnaire-based methods, which include the expert review reflecting a chatbot's performance / QoS and the user opinion reflecting a chatbot's perceived performance / QoE, and secondly the automatic metrics that purely reflect the data, see Figure 3.1. The line in between the automatic metrics and the questionnaire-based methods represent a possible correlation between the two. The dotted lines indicate that the results of all evaluation methods are based on actual chatbot conversations. In the next sections, literature related to the defined types are discussed.

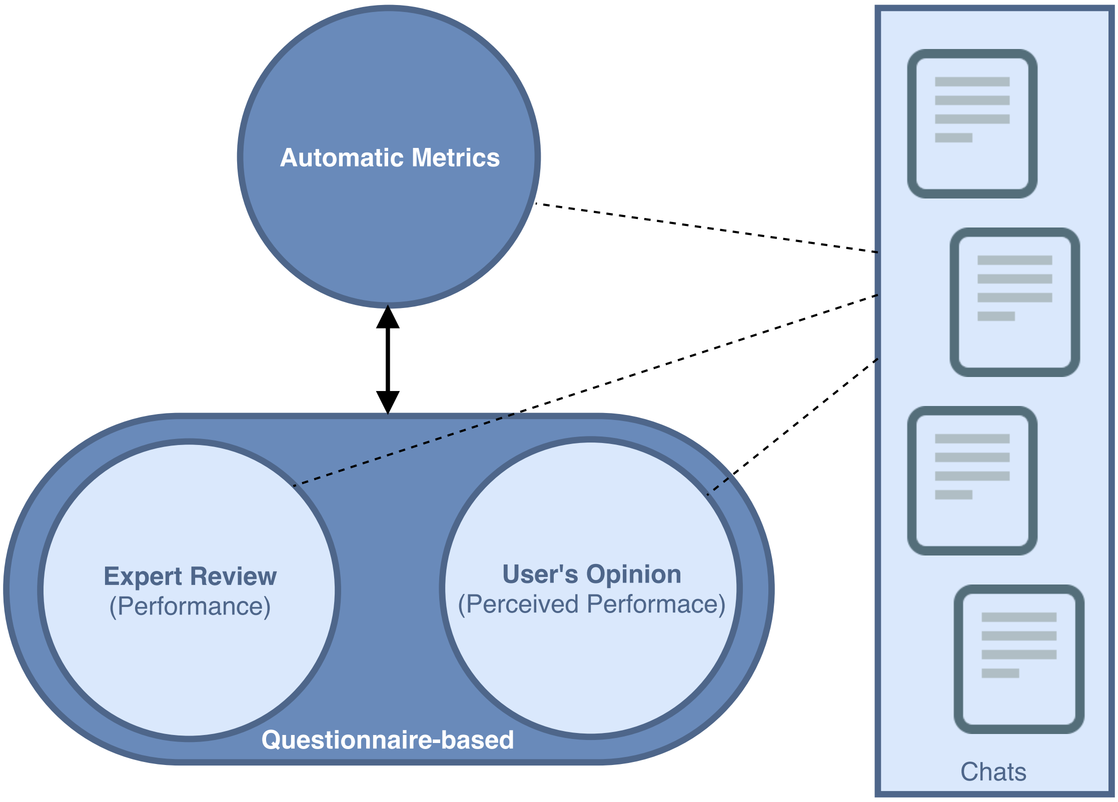


FIGURE 3.1: Chatbot evaluation methods

3.1 Questionnaire-based

We distinguish between two questionnaire-based evaluation methods: *expert review* and the *user opinion*.

3.1.1 Expert Review

As mentioned in the introduction of this study, evaluating the system at hand should be present in any system improvement cycle. A logical choice would be to let an expert perform this evaluation because a proper assessment requires an objective perspective on and a deeper understanding of the chatbot. The goal of an expert evaluation is to capture the chatbot performance, also called QoS.

In a recent paper, Radziwill and Benton (2017) review chatbot quality attributes and quality assessment approaches. They state that chatbot quality attributes proposed by researchers in the past are all greatly aligned with the ISO 9241 concept of usability: "The effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments" (Abran, Khelifi, Suryn, & Seffah, 2003). The chatbot's effectiveness relates to the completeness and accuracy with which users can achieve their goals during the conversation. Moreover, the efficiency refers to how well resources are applied to let the users achieve those goals. Lastly, the satisfaction refers to how the user feels about the interaction with the chatbot. For each of these three concepts, Radziwill and Benton defined underlying categories, see Table 3.1, which in turn consist of multiple chatbot quality attributes. Refer to their paper for a complete list of the quality attributes.

This evaluation method is performed by letting chatbot experts score each quality attribute on a 1-5 Likert scale to define a final score per category. By combining the category, a score can be calculated for each of the three concepts named above. Finally, the concept scores are combined to calculate a final score that reflects the performance of the evaluated chatbot.

TABLE 3.1: Chatbot evaluation concepts and underlying categories

Concept	Category
Efficiency	Performance
Effectiveness	Functionality Humanity
Satisfaction	Affect Ethics & Behavior Accessibility

3.1.2 User Opinion

Evaluating systems by asking its users for their opinion is a well-known approach to get insights about how the performance of the system is perceived. Unlike expert evaluations, user evaluations do not necessarily reflect the actual performance of the chatbot. Users often evaluate by radiating their opinion instead of objectively assessing the performance. Therefore, the results of a user evaluation are closer related to the QoE than to the QoS. A popular approach to capture the user opinion is to let them fill in a questionnaire after their interaction with a chatbot.

In the past, multiple researchers have created user evaluation methods to assess the perceived performance of a chatbot. For example, Semeraro, Andersen, Andersen, Lops, and Abbattista (2002) created a questionnaire using seven characteristics to score a chatbot's performance: impression, command, effectiveness, navigability, ability to learn, ability to aid and comprehension. A user was asked to score each characteristic with a 1-5 Likert scale, ranging from 'very unsatisfied' to 'very satisfied'. Moreover, in a similar study, Hung, Elvir, Gonzalez, and DeMara (2009) used the following characteristics in their questionnaire to measure a chatbot's perceived performance: ease of usage, clarity, naturalness, friendliness, robustness regarding misunderstandings and willingness to use the system again.

Furthermore, in a more recent paper, Kuligowska (2015) proposed another list of chatbot characteristics that aim at capturing its performance, usability and overall quality: visual look, form of implementation on the website, speech synthesis unit, built-in knowledge base (with general and specialized information), presentation of knowledge, and additional functionalities, conversational abilities and context sensitiveness, personality traits, personalization options, emergency responses in unexpected situations, possibility of rating chatbot, and the website by the user. Although named differently, many of these studies use characteristics that overlap. Therefore, we compare the different characteristics to form a combined list, see Table 3.2. A majority of the characteristics are based on the work of Semeraro et al. but are supplemented by two characteristics defined by Hung et al. and Kuligowska. Although the characteristic effectiveness and navigability occurred in two of the three studies, we have deliberately excluded them. The effectiveness is left out because this project does not focus on goal-oriented chatbots only. The lack of a conversation goal makes evaluating the effectiveness irrelevant. Moreover, the navigability is excluded because this project focuses on the messages of the chatbot only, and not its integration with other aspects of the website.

TABLE 3.2: Chatbot evaluation characteristics

Chatbot Characteristics	**Semeraro et al.**	**Hung et al.**	**Kuligowska**	**This study**
Impression	✓	✓		✓
Command / Robustness	✓	✓	✓	✓
Effectiveness	✓		✓	
Navigability	✓		✓	
Learnability / Ease of use	✓	✓	✓	✓
Aidability	✓			
Comprehension / Clarity	✓	✓		✓
Naturalness		✓	✓	✓
Friendliness / Personality		✓	✓	✓
Visual look / Form of implementation			✓	
Process feedback			✓	

3.1.3 Combined Approach

According to some researchers, in order to get a complete image of the system at hand, combining expert and user evaluations yields the best result (Desurvire, 1994; Desurvire, Kondziela, & Atwood, 1992; Karat, Campbell, & Fiegel, 1992). This is especially true for situations in which existing features of the chatbot are improved and new features are added. Since this study does not focus on the development of chatbots, we do not focus on this combined approach but instead decided to use the perceived performance as the Golden Standard during the data analysis.

3.2 Automatic Metrics

The automatic metrics evaluation method focuses purely on metrics that can be automatically measured by analyzing the conversations of chatbots. These simple metrics might relate to the (perceived) performance of chatbots. The aim of this research project is to form a collection of automatic metrics that can be measured by solely analyzing chatbot conversations. Since chatbot conversations are very similar to human-to-human chat conversations, previous research conducted in that field can also be relevant for this research.

However, only the metrics that fit the scope of this study are elaborated on in this. First, the scope of the project is limited to English metrics only because most metrics are based on research that has been performed by analyzing the English language. Moreover, only metrics that are automatically measurable are selected, because the collection of metrics should contribute towards the goal of this project, removing out the need for human evaluators. Lastly, the metrics for which previous research has shown that no correlation could be found are ignored. For example, recently a group of researchers studied the correlation of a number of metrics that are commonly used in the literature for the evaluation of unsupervised dialogue systems (C.-W. Liu et al., 2016), such as chatbots. In their study, two metric types are considered: word-overlap and embedding-based metrics. Word-overlap metrics assume that correct chatbot responses have significant word overlap with the ground truth responses. The word-overlap metrics considered in their study are: BLUE (Papineni, Roukos, Ward, & Zhu, 2002), METEOR (Banerjee & Lavie, 2005) and ROUGE (Lin, 2004). The embedding-based metrics are similar, but instead of looking at exact word-overlap, the meaning of words are compared by defining word embeddings. The researchers conclude that many of these metrics correlate very poorly with human judgment: the questionnaire-based methods. Therefore, the word-overlap and embedding-based metrics are not included in this study.

3.2.1 Metric Attributes

Before the existing automatic metrics and the corresponding measuring techniques are discussed, we elaborate on the different attributes a metric score has. The chosen attributes are partly based the ISO 24617-2:2012 dialogue act standard, which is designed according to the ISO Linguistic Annotation Framework and the ISO Principles for Semantic Annotation (Bunt, 2015; Ide & Romary, 2004). By specifying the different attributes and their possible values, a solid foundation is created for performing consistent data analyses.

Firstly, we define the level attribute. Metrics can be scored on three different levels: dialogue act level, conversation level, and entity level.

Dialogue act level

One single dialogue act is the lowest level for which a metric score can be determined. To define what a dialogue act is, we adhere to the following definition: "a dialogue act represents the meaning of an utterance at the level of illocutionary force" (Austin, 1962). Therefore, a dialogue act is roughly equivalent to the speech act as defined by Searle (1969). A dialogue act is not always the same as one chat message, because people could send multiple dialogue acts in one go or split it up over two or more messages. On this level, a metric score can be calculated by analyzing one dialogue act.

Conversation level

On this level, all the individual dialogue acts in a conversation are collected into one set. A metric score can be calculated by summarizing the metric scores of all the dialogue acts in the corresponding conversation.

Entity level

On the highest level, all the conversations of an entity are collected into one set. A metric score can be calculated by summarizing the metric scores of all the dialogue acts from the corresponding entity.

Moreover, another attribute we define is the source. The messages used to calculate a metric score can originate from three sources: the user, the chatbot or a combination of both. This attribute simply indicates which entity sent the message that is used to calculate the metric.

By combining the different attributes and translating it into a matrix, all available combinations become apparent, see Table 3.3. To illustrate, a C^U measurement contains information about all messages in a conversation (C) of a user (U). Note that one cell in the table is empty, meaning this attribute combination can never occur. A single dialogue act can never have two sources, simply because a unique dialogue act can't be sent by more than one entity.

TABLE 3.3: Automatic metric attributes

	User	**Chatbot**	**User & Chatbot**
Dialogue act level	D^U	D^C	
Conversation level	C^U	C^C	C^M
Entity level	E^U	E^C	E^M

3.2.2 Chatbot Attributes

Similarly, numerous attributes can be defined for a chatbot that might influence the metrics. These are attributes that belong to the chatbot, such as modality, device, style, and maturity (Cassell, Bickmore, Campbell, Vilhjálmsson, & Yan, 2000).

The modality of a chatbot is about the manner in which the chatbot communicates. A chatbot's modality can have multiple forms, such as text-based, voice-based or visual-based. Text-based chatbots communicate solely by chat, while voice-based chatbots have an extra layer that speaks the text out loud. Furthermore, visual-based chatbots make use of buttons, images and other widgets to communicate with its users. Advanced chatbots can make use of multiple modalities. The modality type could influence the chatbot's automatic metric scores or its perceived performance score.

The device attribute is about the device on which users interact with the chatbot, such as 'laptop > browser' or 'phone > app'. The device that is used by the user to interact with the chatbot could have an influence on the chatbot's automatic metric scores or its perceived performance score since the context is different per device.

The style attribute is about the chatbot's purpose. Chatbots can either be goal-oriented (GO) or non-GO. GO chatbots help users achieve a predefined goal, while non-GO chatbots are mainly informative or recreational. The user's purpose for interacting with a chatbot could have an influence on the chatbot's automatic metric scores or its perceived performance score.

Lastly, the maturity attribute is concerning the advancement of the chatbot. Often question & answer (Q&A) chatbots recognize keywords only, while more advanced chatbots make use of NLP. According to Snijder (2018), AI expert, the maturity of chatbots can be subdivided into five levels, see Figure 3.2. The more advanced the NLP and learning techniques, the higher the maturity level. The maturity level has a major influence on the chatbot's automatic metric scores and its perceived performance score. In this research, we focus mainly on, text-based, both GO and non-GO, web browser chatbots on maturity level 3. This choice was made because these chatbots are most commonly used and could profit most from the findings of this research.

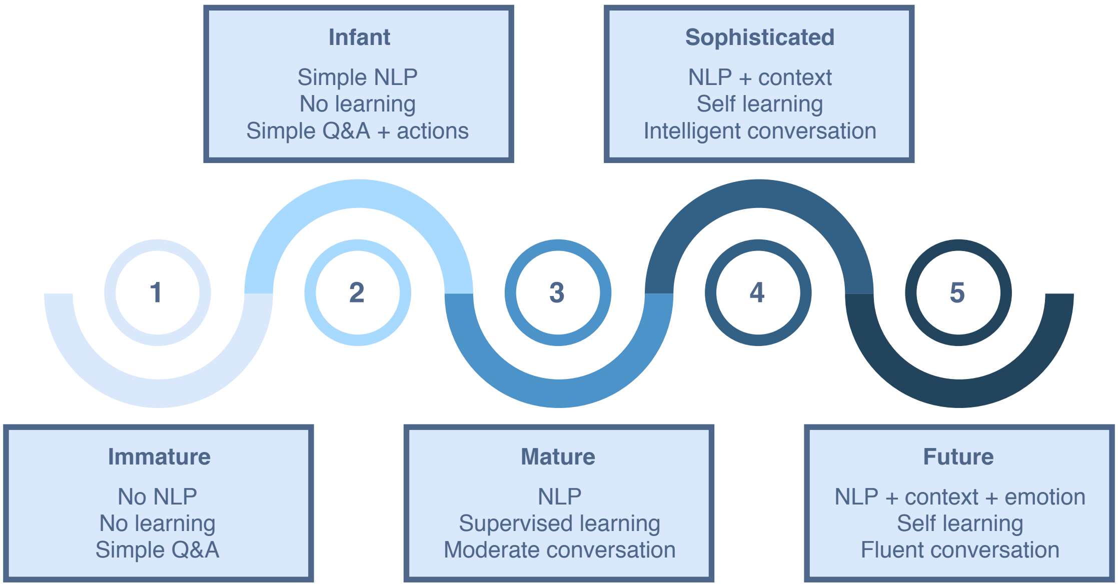


FIGURE 3.2: Chatbot maturity levels

3.2.3 Metric Analysis Approach

Besides these attributes, we distinguish two approaches to analyze the metric scores: 1) summarizing the metric scores of all the dialogue acts and 2) analyzing the trend of the metric scores of all dialogue acts to one another. By following the first approach, all individual dialogue acts in a set are combined to create a summarized portrayal of the entire set. A set can be one or more conversations. By following the latter approach, all the metric scores of the individual dialogue acts in a set are chained and compared to one another, in order to create a trend estimation of a metric. The trend can be interesting to analyze because it provides information on the variance/consistency of an entity throughout one or multiple conversations. For example, this type of analysis could be used to determine whether an entity is changing its dialogue acts based on the dialogue acts of the other entity.

3.2.4 The Automatic Metrics

Previous research has discovered that emotional contagion occurs in communications between two parties, where the positive emotions of one party influence the emotions of the other (Kiffin-Petersen, Murphy, & Soutar, 2012). Since chatbots engage in written text conversations with another party, these findings can be applied to the field of chatbots as well. For instance, it has been discovered that a chatbot that reads and responds to moods of human participants improve the user's satisfaction (Meira & Canuto, 2015). Moreover, Morrissey and Kirakowski (2013) argue that the user's satisfaction of a chatbot is higher when it responds to social cues.

However, in order for a chatbot to respond to these social cues or moods of human participants, they have to be measurable by analyzing the text only. In the previous century, multiple researchers asked themselves the question how language reflects the thinking styles, needs, behaviors, or other psychological states of a person Gottschalk (1997), Gottschalk and Gleser (1969), Stone, Dunphy, and Smith (1966). Over time, theories in this field of study matured which resulted in the insight that certain psychological states are more related to language than others. Following this reasoning, a piece of text could provide information about the emotions of the author that cannot be derived by solely looking at the meaning of the words.

The findings of these studies led to the creation of a text-analysis program called Linguistic Inquiry and Word Count (LIWC) to capture people's social and psychosocial states by analyzing text (LIWC, 2017). Numerous automatic metrics discussed in this section are inspired by previous research that has been conducted in collaboration with the LIWC program. More background information about the LIWC program is provided in subsection 3.2.5.

In the section below, a majority of the most important automatic metrics are elaborated upon. Each automatic metric is described together with its corresponding measuring technique.

Sentiment

Sentiment analysis, also known as opinion mining, is the discipline that analyzes evaluations, sentiments, opinions, emotions, and attitudes from written language (B. Liu, 2012). According to Liu, sentiment analysis is a core research area in the natural language processing (NLP) field. Just like in the real world, emotion plays a significant role in conversations (Kirange, Deshmukh, Jain, & Patil, 2011). Automatic techniques for sentiment mining are essential because manual extraction is very costly and inefficient (Hassan, Qazvinian, & Radev, 2010). Moreover, understanding emotions could help a conversational agent, like a chatbot or intelligent robot, to give more human-like responses based on the emotional state of a user (Kirange et al., 2011).

Initial opinion mining techniques started with identifying words that bear emotions or sentiment, such as 'happy', 'sad', 'amazing', 'bad', etc. Analyzing these type of words could classify sentences as either positive, neutral or negative. Over time, research on opinion mining led to the maturing of opinion mining techniques. However, correctly measuring the sentiment of a small piece of text, such as a chat message, can be a challenging task due to the limitation in length and the high concentration of misspellings, slang terms, and shortened forms of words (Kiritchenko, Zhu, & Mohammad, 2014). To overcome these challenges, Dos Santos and Gatti (2014) proposed a new deep convolutional neural network that exploits from character- to sentence-level information to perform sentiment analysis of short texts.

In this study, a sentiment score called emotional tone, which is included in the LIWC tool is used. A high score suggests a positive and upbeat mood, while a low score reveals greater anxiety, sadness or hostility. A score around 50 is associated with either a lack of emotionality or different levels of ambivalence. The emotional tone score can be calculated for all attribute combinations mentioned in Table 3.3. However, determining the sentiment of a single dialogue act can be less accurate due to the lower amount of words. Therefore, the values on the dialogue act level are probably less reliable than the values on the conversational and entity level.

Response Time

When comparing human-to-chatbot conversations with human-to-human conversations, one aspect greatly differs, namely the fact that a human is no longer needed to read the message, form an answer and write back that answer to the user. This change is accompanied by a couple of advantages related to the response time. Besides the advantage that chatbots are available 24/7 and generally support multiple concurrent conversations at once, whereas humans can only focus on one conversation at a time, the response time of chatbots is often a lot quicker than that of a human (IBM, 2017).

Although men could argue that faster response times improve the user satisfaction, the opposite could also be true. Providing inhumanly fast answers to questions could evoke a negative artificial feeling from the users. The response time can be measured for all attribute combinations mentioned in Table 3.3, with the exception of the initial dialogue act of a conversation.

Word Count

One of the most straightforward automatic metrics is the word count. This metric simply counts the number of words that occur in a message. In the previous century, Jakob Nielsen discovered that people tend to read less online than with conventional writings (Nielsen, 1997). He claims that shorter texts better fit the reading behavior of online users. However, at this point in time, there are no scientific research results that support the hypothesis that the length of chatbot sentences relate to the (perceived) performance of the chatbot. The word count score can be determined for all attribute combinations mentioned in Table 3.3.

Turn Count

Another simplistic automatic metric is the turn count. This metric represents the number of dialogue acts that were sent during the conversation. The turn count score can be determined for both the conversation and entity level attributes combinations mentioned in Table 3.3.

Composition Statistics

Beside simply counting the total number of words, it is also possible to count the number of words per word class. The word classes that exist are noun, verb, adjective, adverb, pronoun, preposition, conjunction, determiner, and exclamation. The process of automatically assigning words to one of these classes is called Part-Of-Speech (POS) tagging. The input to the tagging algorithm is a collection of words and a tag set. The output of the algorithm is a sequence of tags, a single best tag for each word (Daniel & Martin, 2017).

The main two challenges in POS tagging are dealing with ambiguous and unknown words (Güngör, 2010). The first is related to tagging words that can be tagged to multiple classes. Multiple methods exist that focus on tackling this challenge, such as the most frequent class baseline algorithm named by Daniel and Martin. This simplistic baseline algorithm decides between two or more tags for an ambiguous word by looking at the tag which is most frequent in the training corpus. Composition statistics can be measured for all attribute combinations mentioned in Table 3.3. However, determining the compositions of a single dialogue act can be less accurate due to the lower amount of words. Therefore, the values on the dialogue act level are probably less reliable than the values on the conversational and entity level.

Readability

In the past, multiple researchers have investigated the domain of readability, the ease with which a reader understands a piece of text. Multiple researchers have created algorithms that aim to capture a text's readability by automatically calculating a score (e.g. Flesch, 1951; George R. Klare, Rowe, John, & Stolurow, 1969; McLaughlin, 1969). Some of the well-known readability algorithms are: Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman-Liau Index, SMOG Index and the Automated

Readability Index. Although these formulas are critiqued by others, they do provide some insight into the complexity of a piece of text. A readability score can be determined for all attribute combinations mentioned in Table 3.3.

Analytical Thinking

In a recent study, scientist examined the influence of the smallest and most commonly used words in English, such as pronouns, articles, and other function words (Pennebaker, Chung, Frazee, Lavergne, & Beaver, 2014). They discovered that although these function words are almost invisible to the reader or writer, they can reveal ways people think and approach topics. The results of the study show that students with higher analytical thinking skills are associated with the use of categorical language (greater article and preposition use), while students scoring low on analytical thinking are associated with more dynamic language (greater use of auxiliary verbs, pronouns, adverbs, conjunctions, and negations).

By translating these research findings into an automatic metric, the use of functions words in a piece of text, regardless of whether this is an essay or a chat conversation, can be measured to calculate a score for analytical thinking. A high score is associated with logical, formal, and hierarchical thinking. A low score reflects informal, here-and-now, personal, and narrative thinking. An analytical thinking score can be determined for all attribute combinations mentioned in Table 3.3.

Confidence

In a similar study, a group of researchers investigated the effect of social hierarchy and confidence on the use of language (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2013). Through computerized text analyses, they discovered that the status and confidence of a person is reflected in the usage of pronouns. Their findings show that people with a higher status use more first-person plural and second-person singular pronouns and fewer first-person singular pronouns. They conclude that higher ranked people have more other-focus whereas lower ranked people have more self-focus.

These results can be translated into a metric that reflects the level of confidence of a speaker. A high score suggests that the author is speaking from the perspective of high expertise and is confident. On the other hand, a low score reflects a more humble, tentative, or even anxious style. A confidence score can be determined for all attribute combinations mentioned in Table 3.3.

Authenticity

In an earlier study, researchers tried to predict deception by analyzing linguistic styles (Newman, Pennebaker, Berry, & Richards, 2003). They discovered that liars use more negative emotion words and fewer self-references and other-references than truth-tellers.

The LIWC tool used these research findings to create an automatic metric, which intends to calculate a score for the authenticity of the author. A high score reflects a more personal, honest and disclosing form of discourse, while a low score is associated with a more distanced and guarded form of text. An authenticity score can be determined for all attribute combinations mentioned in Table 3.3.

3.2.5 LIWC

As mentioned earlier, previous studies in the field language usage led to the creation of a text-analysis program called Linguistic Inquiry and Word Count (LIWC) to capture people's social and psychosocial states by analyzing text (LIWC, 2017). The text-analysis is performed by determining scores for numerous variables by calculating the percentage of words in the text that belong to a particular category (Tausczik & Pennebaker, 2010). In order to perform these calculations, LIWC uses a library which refers to the collection of words that define a particular category. These dictionaries form the core of the LIWC program. The first LIWC dictionary was published in 2001, the second in 2007 and the latest in 2015. The latest library, LIWC2015, has over ninety (sub-)variables. This version is completely new compared to earlier ones. The new dictionaries accommodate punctuation, numbers and short phrases. Therefore, LIWC2015 is very suitable for analyzing chatbot conversations. The program can calculate a score for variables, such as analytical thinking, confidence, authenticity, emotional tone, total function words, affective processes, social processes, cognitive processes, drives and personal concerns. For a complete list of the variables, refer to the LIWC language paper (Pennebaker, Boyd, Jordan, & Blackburn, 2015).

4 Data Analysis

The findings of the literature review discussed in the previous chapter form the answers to the first two sub-questions. In order to answer [SQ3] and [SQ4] as well, a data analysis is performed on the two collected datasets by following the CRIPS-DM methodology, see subsection 2.4.4. In this chapter, the data analysis procedure and results are elaborated upon.

4.1 Data Collection

Prior to the data analysis, two datasets containing chat conversations are collected. One dataset is collected by executing the case study and the other dataset is publicly available on the internet.

4.1.1 Open Dataset

To answer the third sub-question, patterns and relations in the automatic metrics are analyzed. For this analysis, a large and freely available dataset containing chat conversations is preferable. However, finding an appropriate open dataset is very difficult. Since the conversations of chatbots are often full of personal information, many datasets are held private. Furthermore, organizations that built their own chatbot often prefer to keep their chatbot datasets for themselves. Their data can be used to optimize chatbots, giving them a competitive advantage.

Nevertheless, we were determined to find an open dataset. In a recent study, an overview is provided containing all current freely available dialogue datasets (Serban, Lowe, Henderson, Charlin, & Pineau, 2015). Unfortunately, the fast majority of these datasets are human-human dialogues. Although these datasets seem similar, turn-taking for normal human-human conversations is richer than for human-chatbot dialogues (Doran, Aberdeen, Damianos, & Hirschman, 2001). Therefore, these datasets do not match the characteristics of this research project.

Filtering these out, very few suitable datasets remain. All of the remaining dialogue datasets are created using the Wizard-of-Oz (WoZ) approach. This approach let users believe they are communicating with a chatbot, while in fact a real human secretly plays the role of the chatbot (Medhi Thies, Menon, Magapu, Subramony, & O'Neill, 2017). Since users are let to believe that they are communicating with a chatbot, the turn-taking is also more similar. Although these WoZ datasets differ from real chatbot conversations, they are currently the most similar datasets that are publicly available.

The 'Frames' dataset (Asri et al., 2017), which was published by Maluuba - a Microsoft company, is such a dataset. It is designed to "help drive research that enables truly conversational agents that can support decision-making in complex settings" (Asri et al., 2017). This dataset was created by letting one human play the role of a chatbot travel agent (wizard) and the other the role of a customer. The wizard intended to act like a chatbot by sometimes providing wrong answers or bad behavior, making this dataset even more suitable. We decided to use the 'Frames' dataset, due to its similarities, size and because it contains a survey rating from the users. Additional metadata of the dataset is shown in Table 4.1.

4.1.2 Case Study Dataset

To answer the fourth sub-question, a dataset enriched with the perceived performance of each conversation is required. Due to the lack of online available datasets that include scores for the perceived performance components, we decided to set up a case study to create a suitable dataset. For this case study, we requested thirty participants to have a conversation with a chatbot, see section A.1. All participants signed an informed consent, see section A.2. We decided to make use of Mitsuku, a publicly available chatbot from AIML technology by Steve Worswick that is driven by AI. We chose to use Mitsuku because it won the Loebner Prize, which is awarded to the most "human-like" chatbot, three times (2013, 2016, and 2017). During the case study, the participants were requested to chat with Mitsuku for five minutes, the same number of minutes the jury of the Loebner Prize take to score all participating chatbots. The participants in this case study were free to talk about any subject, since Mitsuku is a recreational chatbot, see section A.3. Recreational chatbots are not goal-oriented but instead, are meant for 'small talk'. After the conversation, the participants were requested to fill in a survey to aims at capturing the perceived performance, see section A.4. The results of this experiment form the case study dataset. Metadata of the dataset is shown in Table 4.1.

TABLE 4.1: Metadata of the datasets

Name	Type	Topic	Avg. # of turns	Total # of dialogues
Maluuba Frames	Chat (WoZ)	Hotel and flights	15	1369
Mitsuku	Chat	Recreational	44	30

4.2 Data Preparation

Subsequent to the data collection, the data is prepared for the data analysis. The preparation steps that are taken are described below.

Extracting

At first, the dialogues are extracted from the raw data into a data table. For the frames dataset, the dialogues are extracted from JSON format and for the case study dataset, the information is extracted from plain text (chat log). Every message is given a turn identifier, a source (chatbot or user) and linked to a dialogue identifier.

Summarizing

Next, the messages of each dialogue are summarized and bundled per source: one summarization for the chatbot's messages, one for the user's messages, and one for all messages in the dialogue. Due to this summarization step, the conversation can be analyzed as a whole instead of each message separate.

Transforming

Subsequently, the content of the dialogues in the data table is transformed into automatic metrics. The LIWC tool, see subsection 3.2.5, is used for measuring the majority of the automatic metrics. Furthermore, the R package 'readability' is used to transform messages to multiple different readability scores. Moreover, timestamps of the messages in the dialogue are transformed into response time and total duration. Lastly, each dialogue gets a total turn count.

Appending

Next, the answers to the case study survey questions are bundled and used to calculate the scores for the perceived performance categories. These category scores, the average score and the overall rating by the user are appended to the corresponding dialogue in the case study data table. Since the frames dataset does only contain an overall rating of the user and no additional information regarding the perceived performance, only this overall rating is appended to the dialogues in the data table.

Cleaning

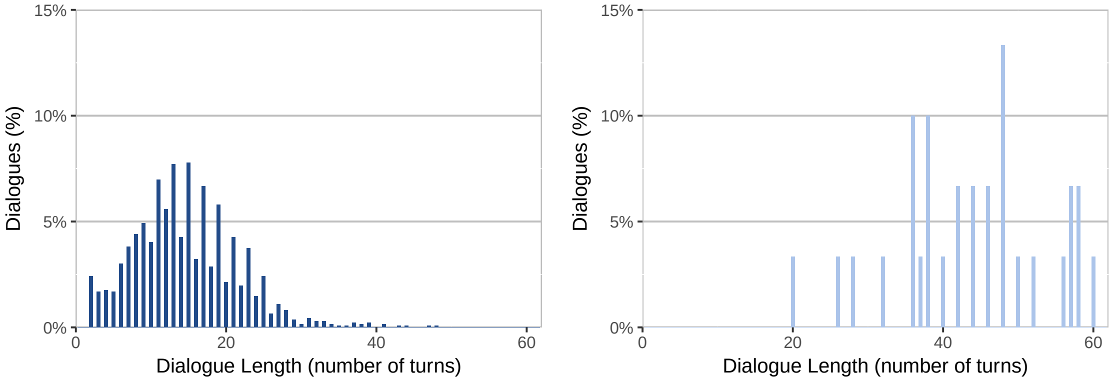
During the last step in the data preparation phase, the rows with missing values are removed. Fortunately, the case study dataset has no missing values and only very few values are missing in the frames dataset. Moreover, the columns with a standard deviation of zero are removed, because those variables are irrelevant for the data analysis and could obstruct certain analyses. A complete list of the remaining variables can be found in Appendix B.

4.3 Descriptive Statistics

In this phase, the datasets are described by their descriptive statistics. The frames dataset has a total of 1369 dialogues with an average of 15 turns per dialogue (sd = 7). The case study dataset has a total of 30 dialogues with an average of 44 turns per dialogue (sd = 10). The dialogue length distribution of both datasets is shown in Figure 4.1. Moreover, both datasets contain overall user ratings on a 1 to 5 Likert scale that represents the conversation satisfaction of the users. The average user rating is a 4.6 (sd = .83) for the frames dataset and a 3.6 (sd = .74) for the case study dataset. The user rating distribution of both datasets is shown in Figure 4.2.
A brief comparison of these descriptive statistics shows that the datasets are divergent. The frames dataset has a homogeneous character. This is visible in the user ratings, for which a vast majority of the users gave a five out of five. Moreover, the sentiment of the messages varies highly, as the box sizes in Figure 4.2b show. The use case dataset, on the other hand, is more equally distributed, meaning the dataset is more heterogeneous. The user's rating and the number of turns in a dialogue are both more normally distributed. Furthermore, the box sizes in Figure 4.3b are a lot smaller, meaning the sentiment of the messages is quite similar.

4.4 Hypothesis Testing

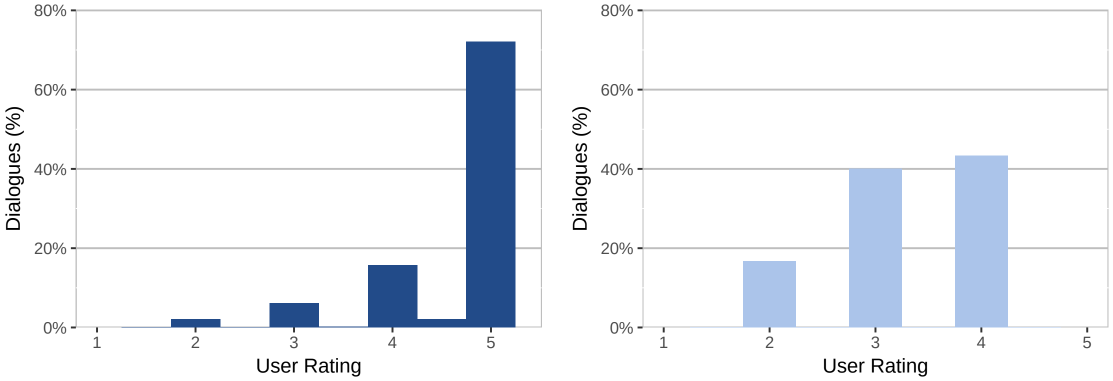
The analysis is extended to discover patterns and relationships in the variables. Correlations in the dataset are explored through visualization using correlation heatmaps.



(A) Frames dataset

(B) Case study dataset

FIGURE 4.1: Dialogue length distribution



(A) Frames dataset

(B) Case study dataset

FIGURE 4.2: User rating distribution

Strong correlations are further explored through visualization by scatter plots containing the linear regression line.

Moreover, multiple hypotheses are formulated in order to answer the last two subquestions. The hypotheses are either concerning patterns within the automatic metrics [SQ3] or concerning the relationship between the automatic metrics and the perceived performance [SQ4]. For the effect size we follow the guidelines of Cohen (1988), for which a correlation coefficient of .1 implies a weak correlation, .3 a moderate correlation and .5 a strong correlation.

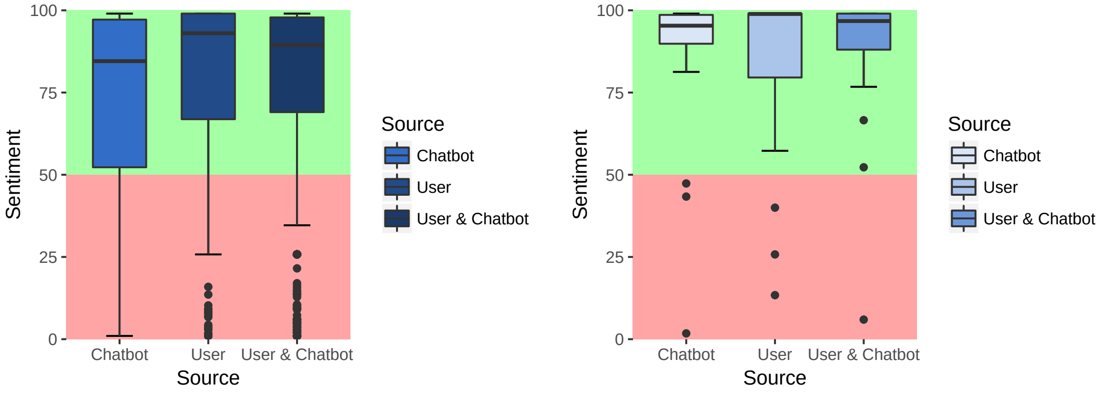
4.4.1 Patterns in Automatic Metrics

To assess whether expected patterns occur between automatic metrics, multiple hypotheses pairs are formulated and tested. For each expected pattern, an alternative and a null hypothesis are formulated.

H1$_0$: *There is no relationship between analytical thinking and readability.*

H1$_A$: *There is a relationship between analytical thinking and readability.*

To find the answer to the hypothesis, we compare the analytical thinking variable with one of the readability scores. The average grade level variable correlates strongest with all of the other readability scores, as can be seen in Figure 4.4a, therefore this variable will be used in the statistical test.

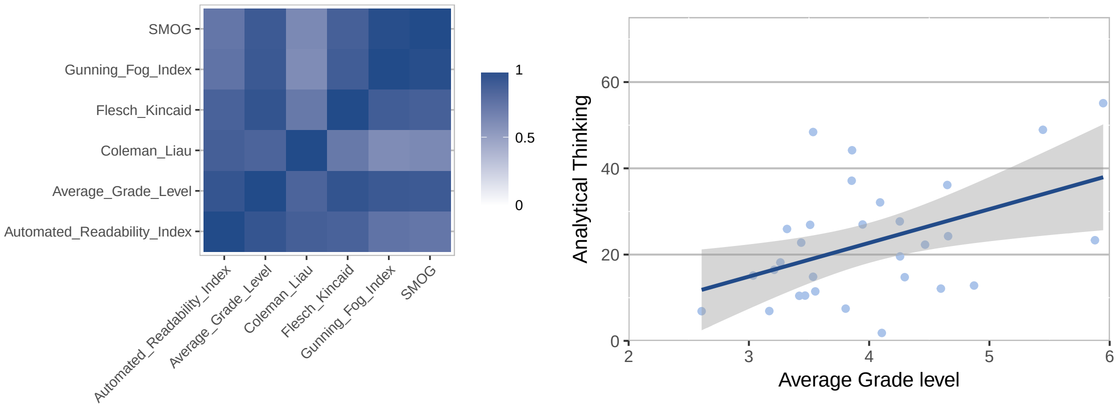


(A) Frames dataset

(B) Case study dataset

FIGURE 4.3: Sentiment of messages

Performing a Pearson correlation test on the case study dataset results in r(28) = .75, p < .001, which indicates a strong correlation. The regression line with a confidence interval is shown in Figure 4.4b. The same correlation is assessed for the frames dataset resulting in r(1361) = .35, p < .001, which indicates a moderate correlation. Although the correlation is weaker for the larger dataset, the correlation is present and significant. Therefore, we reject $H1_0$ and retain $H1_A$. Concluding, the more complex the sentences (higher readability score), the higher the analytical thinking score.



(A) Correlation heatmap readability scores - frames dataset

(B) Scatter plot average grade level and analytical thinking - case study dataset

FIGURE 4.4: Hypothesis 1

$H2_0$: *There is no relationship between confidence and readability.*

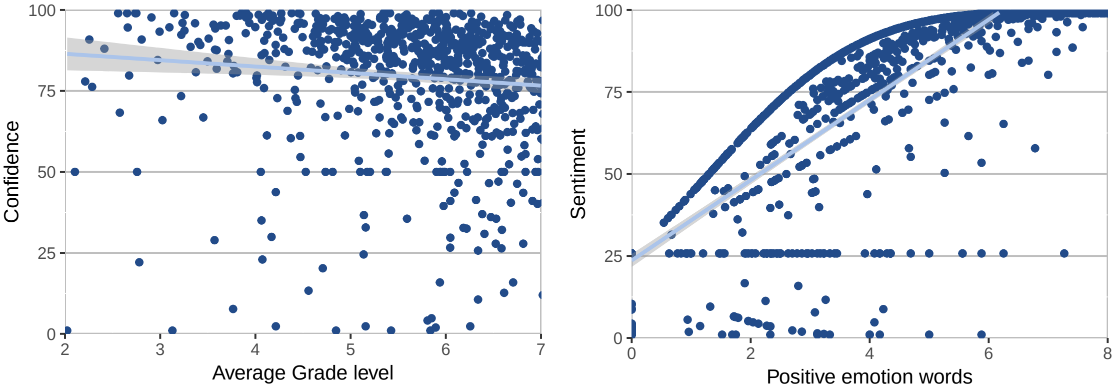
$H2_A$: *There is a relationship between confidence and readability.*

Performing a Pearson correlation test on the case study dataset results in r(28) = -.15, p > .05, which indicates a weak correlation, not significantly different from $H2_0$. The same correlation is assessed for the frames dataset resulting in r(1361) = -.04, p > .05, which indicates no correlation. The regression line with a confidence interval is shown in Figure 4.5a. Therefore, we reject $H2_A$ and retain $H2_0$. Concluding, there is no statistical relationship between confidence and readability.

H3$_0$: *There is no relationship between positive emotion words and sentiment.*

H3$_A$: *There is a relationship between positive emotion words and sentiment.*

Performing a Pearson correlation test on the case study dataset results in r(28) = .71, p < .001, which indicates a strong correlation. The same correlation is also assessed for the frames dataset resulting in r(1361) = .75, p < .001, which also indicates a strong correlation. The regression line with a confidence interval is shown in Figure 4.5b. The smooth line of data points perfectly show the strong correlation between the two variables. Therefore, we reject H3$_0$ and retain H3$_A$. Concluding, the more positive emotion words in the sentences, the higher the sentiment score. This outcome is logical because positive words play an important role in the sentiment algorithm, as described in section 3.2.4.



(A) Scatter plot average grade level and confidence - frames dataset

(B) Scatter plot sentiment and positive emotion words - frames dataset

FIGURE 4.5: Hypothesis 2 & 3

4.4.2 Relations to Perceived Performance

To assess how the automatic metrics relate to the perceived performance, multiple hypotheses pairs are formulated and tested.

H4$_0$: *There is no relationship between sentiment and user rating for the user's messages.*

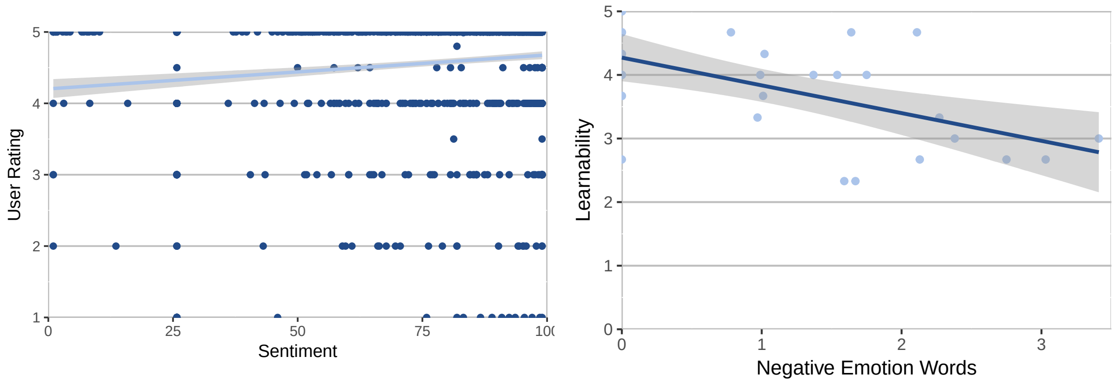
H4$_A$: *There is a relationship between sentiment and user rating for the user's messages.*

Performing a Pearson correlation test on the case study dataset results in r(28) = .38, p < .05, which indicates a weak correlation. The same correlation is also assessed for the frames dataset resulting in r(1361) = .16, p < .001, which indicates a weak correlation. The regression line with a confidence interval is shown in Figure 4.6a. Although the correlation is weaker for the frames dataset, it is present both are significant. Therefore, we reject H4$_0$ and retain H4$_A$. However, extending the hypothesis to make a statement about the perceived performance is invalid, since the user rating is only a truncated version of the perceived performance. Therefore, we perform another Pearson correlation test using the average rating because it represents the perceived performance score. However, the average rating is only available for the case study dataset. The results show r(28) = .50, p < .01, which indicates a strong correlation. Therefore, we conclude that a higher sentiment in the user's messages implies a higher user rating.

H5$_0$: *There is no relationship between negative emotion words and learnability for user's messages.*

H5$_A$: *There is a relationship between negative emotion words and learnability for user's messages.*

Performing a Pearson correlation test on the case study dataset results in r(28) = -.57, p < .001, which indicates a strong negative correlation. The regression line with a confidence interval is shown in Figure 4.6b. Because the frames dataset does not contain a measure for the learnability, the analysis can not be performed on that dataset. Concluding, we reject H5$_0$ and retain H5$_A$.



(A) Scatter plot sentiment and user rating - frames dataset

(B) Scatter plot negative emotion words and learnability - case study dataset

FIGURE 4.6: Hypothesis 4 & 5

H6$_0$: *There is no relationship between authenticity and friendliness for chatbot messages.*

H6$_A$: *There is a relationship between authenticity and friendliness for chatbot messages.*

Performing a Pearson correlation test on the case study dataset results in r(28) = -.06, p < .05, which indicates no correlation. Because the frames dataset does not contain a measure for the friendliness, the analysis can not be performed on that dataset. Concluding, we reject H6$_A$ and retain H6$_0$.

H7$_0$: *There is no relationship between readability and naturalness for the chatbot's messages.*

H7$_A$: *There is a relationship between readability and naturalness for the chatbot's messages.*

Performing a Pearson correlation test on the case study dataset results in r(28) = -.14, p > .05, which indicates a weak correlation, not significantly different from H7$_0$. Because the frames dataset does not contain a measure for the naturalness, the analysis can not be performed on that dataset. Concluding, we reject H7$_A$ and retain H7$_0$.

4.4.3 Hypotheses Insights

In this section, the outcomes of the hypotheses are combined to come to new insights. The outcomes of the Pearson correlation tests are summarized Table 4.2.

First, both H1$_A$ and H3$_A$ show that some automatic metrics correlate with one another. These inter-metric correlations could provide insights about the algorithmic operation of the automatic metrics, which is required for further refinements. Moreover, this insight can also be used to remove correlating metrics to reduce the number of variables, which reduces noise and increases the efficiency of the analysis. Secondly, both $H4_A$ and $H5_A$ show that some automatic metrics correlate with the perceived performance of the chatbot, especially with the performance characteristics. E.g. users that express negativity are more likely to rate the chatbot's learnability lower. Not understanding how to use the chatbot could be the cause. However, due to the lack of information, these insights could not be tested on the frames dataset.

TABLE 4.2: Pearson correlation results per hypothesis

	$\mathbf{H1_A}$	$\mathbf{H2_A}$	$\mathbf{H3_A}$	$\mathbf{H4_A}$	$\mathbf{H5_A}$	$\mathbf{H6_A}$	$\mathbf{H7_A}$
Case Study	.75***	-.15$^{n.s.}$	.71***	.38* & .50**	-.57***	-.06$^{n.s.}$	-.14$^{n.s.}$
Frames	.31***	-.04$^{n.s.}$	.75***	.16***	-	-	-

n.s. = $p > .05$ * = $p < .05$ ** = $p < .01$ *** = $p < .001$

4.5 Predictive Analysis

Taking the analysis a step further, multiple Prediction Models (PMs) are created in an attempt to predict the user ratings of the dialogues by solely using the automatic metrics. For the frames dataset, the to be predicted rating is the 'User rating' variable and for the case study dataset, this rating is the 'User average' variable. The following techniques are used to create the prediction models: Decision Tree, Principle Component Analysis (PCA) and Random Forest. Since the user ratings are on a scale of 1 to 5, the most logical choice is to create regression prediction models. For each data analysis technique, multiple regression prediction models are created with the following dataset splits (DS):

[DS_1] Case study training set (67%) & case study test set (33%)

[DS_2] Frames training set (67%) & frames test set (33%)

[DS_3] Frames training set (100%) & case study test set (100%)

For each dataset split, the user rating's 1) standard deviation (SD) of the training set 2) the mean of the training set 3) and the mean squared prediction error (MSPE) of the test set, using the training set's mean as the predictor, are provided, see Table 4.3. Furthermore, each dataset split is used to create three PMs on the conversation level, one per source type, see Table 3.3.

TABLE 4.3: User rating statistics per dataset split

	$\mathbf{DS_1}$	$\mathbf{DS_2}$	$\mathbf{DS_3}$
Mean *(training set)*	3.67	4.63	4.58
SD *(training set)*	.60	.73	.83
$\mathbf{MSPE_{mean}}$ *(test set)*	.42	.68	.94

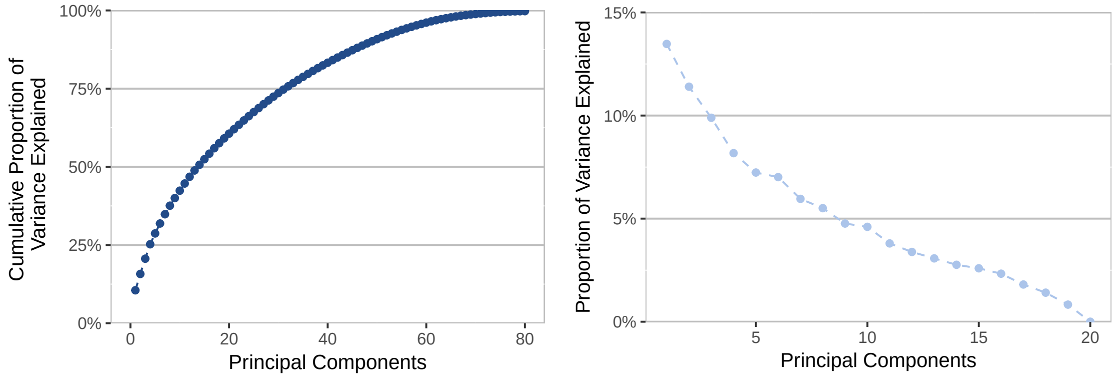
Although the user ratings are on a scale from 1 to 5, the analysis is extended by including classification PMs as well. Prior to these analyses, the user ratings are classified into three groups: low rating (1 – 2.3), medium rating (2.4 – 3.6), high rating (3.7 – 5). However, for each technique the classification PMs yielded worse results than the regressions PMs. Therefore, the classification PMs are emitted from the results.

4.5.1 Decision Tree

Firstly, the decision tree technique is applied to create PMs. However, the correlations in the dataset are too weak to come remotely close to a meaningful regression decision tree. As mentioned earlier, classification does not yield better results. Concluding, a single decision tree is not good enough of a predictor for the two datasets.

4.5.2 Principle Component Analysis

Subsequently, a PCA is performed for each DS to create PMs by reducing the dimensionality of the variables in the datasets. A set of observations, which possibly correlate, are converted into multiple sets of linearly uncorrelated variables, also called principal components. A PCA is performed for each DS and source type, but the results are unsatisfactory. For example, the results of the PCA for DS_2 with the user as the source are shown in Figure 4.7a. The figure shows a small distance in between the principal components, which is unfavorable. Ideally the cumulative variance explained would rapidly increase. Moreover, the results of the PCA for DS_1 with the user as the source are shown in Figure 4.7b. This time the variance explained by each principal component is shown. As can be seen, the best principal component explains just 14% of the variance. Ideally, this would be a lot higher. Concluding, the PCA yields very moderate results.



(A) DS_2 - User

(B) DS_1 - User

FIGURE 4.7: Variance explained by principal components

Nevertheless, the analysis is extended to assess the prediction capabilities. The results of the analysis are displayed in Table 4.4. The results show that for all dataset splits using the user's messages are best to predict the user's rating. However, comparing these MSPE values with the $MSPE_{mean}$ in Table 4.3 shows that only for DS_2 the prediction is slightly better. In other words, using the PCA as a prediction model is not yielding better results than using a static value. Since the predictions are not accurate, we do not discuss the loadings, which are the most predictive variables of the principal components.

TABLE 4.4: PCA: MSPE user rating per dataset split

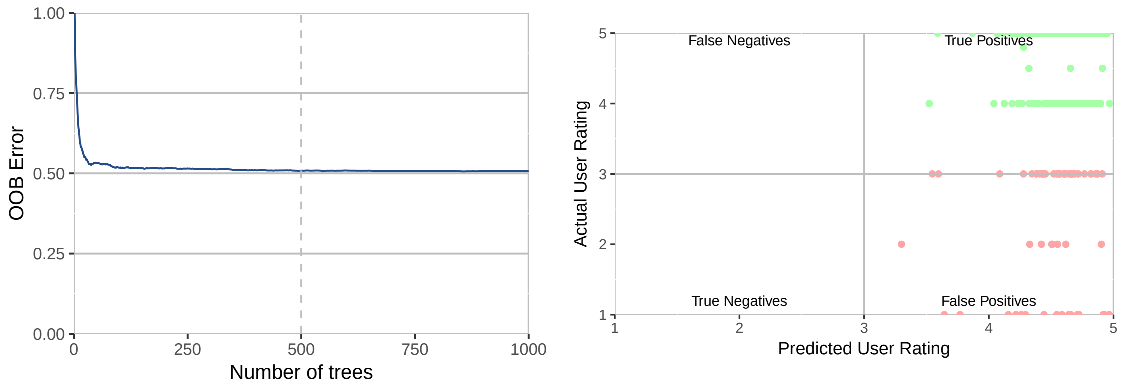
	DS_1	**DS_2**	**DS_3**
User	**.45**	**.66**	**.99**
Chatbot	.57	**.66**	1.10
User & Chatbot	.51	.70	1.10

4.5.3 Random Forest

In an attempt to perform better predictions on the user ratings, the random forest method is performed. The advantage of the random forest technique is that it tends not to overfit the data (Hastie, Tibshirani, & Friedman, 2008). For each of the dataset splits and source types, a random forest is grown. For each random forest, the bootstrap aggregating (bagging) method is applied to reduce variance in the PM (Breiman, 1996). Bagging is done by generating additional train data from the original dataset. The error on the left out data, out-of-bag (OOB) error, is used to estimate the accuracy of the random forest. Each random forest model is trained using 500 trees because analysis points out that more trees do not yield a lower OOB error, as can be seen in Figure 4.8a. For each tree, one-third of the variables are randomly sampled as candidates at each split. The results of the random forest PMs are shown in Table 4.5.

TABLE 4.5: Random forest: MSPE user rating per dataset split

	DS_1	**DS_2**	**DS_3**
User	.44	**.65**	.52
Chatbot	**.39**	.66	.52
User & Chatbot	**.39**	.67	**.50**



(A) Out-of-bag error per number of trees for DS_2 - User

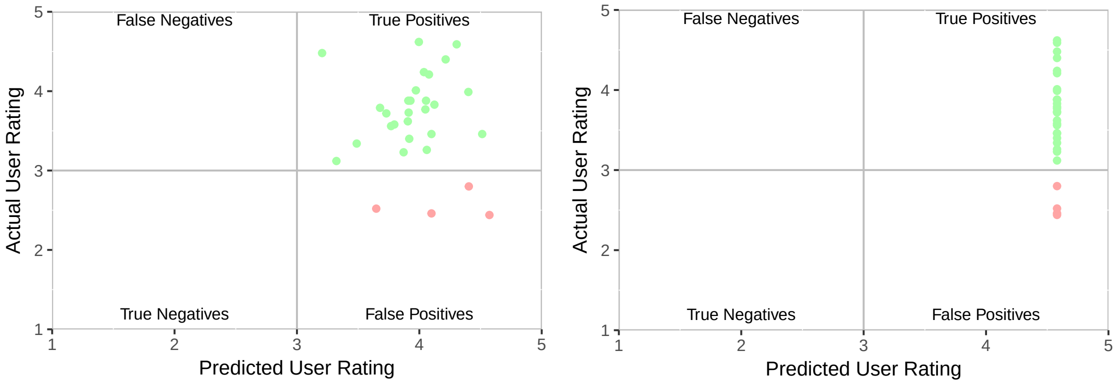
(B) Actual vs. predicted user ratings for DS_2 - User

FIGURE 4.8: Random forest

The results show that for DS_1 and DS_3 using both the user's and chatbot's messages are best to predict the user's rating. For DS_2 only using the user's messages is best. Comparing these results with the $MSPE_{mean}$ in Table 4.3 and the PCA MSPE in Table 4.4 shows that the random forest PM is more accurate for all dataset splits. Especially the random forest PC for DS_3 is substantially more accurate. However, the MSPE is still quite large. The importance of the variables in the models are assessed to determine which automatic metrics are the most predictive. The higher the Mean Square-Error (MSE) increase if a variable was to be randomly permuted, the more predictive the variable. The prediction variables with the highest importance are listed for each of the PMs, see Table C.1 in Appendix C.

Another way to look at the results is by dividing the ratings into two groups: negative ratings (1 ↔ 3) and positive ratings (3 ↔ 5). In practice, this is a logical step because a general impression of how the user experiences the conversation is probably enough. The predictions are assessed by comparing it to their actual values, labeling a prediction as correct (true) or incorrect (false). For DS_2, 382 out of 443 user rating's are true positives (TP) and 6 are false positives (FP) with no false negatives (FN) and true negatives (TN), resulting in a precision of 86%, see Figure 4.8b. Although this number seems high, this is mainly due to the homogeneity in the dataset. This becomes apparent when the same test set is predicted using the mean of the train set (4.63) as a static predictor. Since all predicted values fall in the positive rating group, this static predictor has the same precision of 86%.

Since the regression prediction for DS_3 is more accurate, the same classification is performed for this data split in an attempt to achieve a higher precision. In DS_3, 26 out of 30 user rating's are TP and 4 are FP with no FN and no TN, resulting in a similar precision of 86%, see Figure 4.9a. Again, since all predicted values fall in the positive rating group, the static predictor (4.58) achieves the same precision of 86%, see Figure 4.9b.



(A) Actual vs. predicted user ratings with random forest as predictor (B) Actual vs. predicted user ratings with training set's mean as predictor

FIGURE 4.9: Prediction results for DS_3 - User

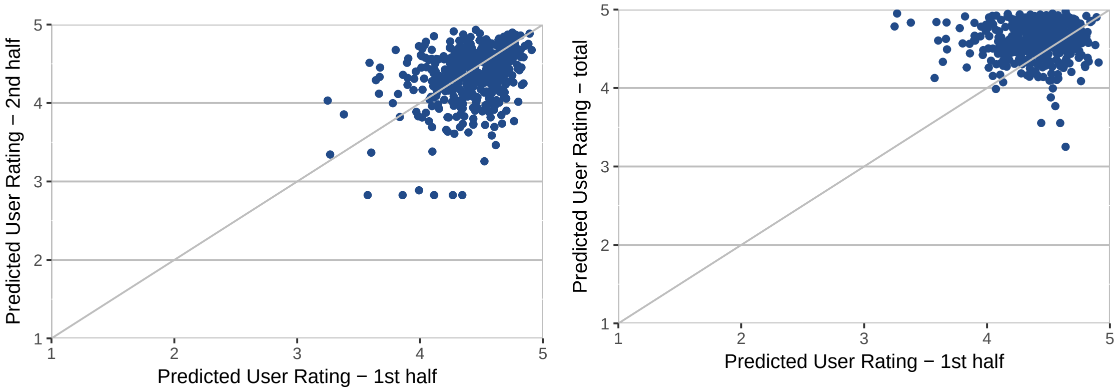
Lastly, the data analysis is extended to explore one additional approach that could add business value. All the analyses above are done on entire conversations. However, it could also be interesting to predict how the user experiences a chat conversation while it is still ongoing. With this approach, conversations could be flagged as negative halfway in to warn e.g. a real customer agent. For this additional analysis, the conversations are split into two halves. Subsequently, the random forest prediction process is repeated by using the first half of each conversation instead of the entire conversation, see Table 4.6. The results show that these predictions are very similar to the predictions that are based on entire conversations. This means that predicting using only the first half of the conversation is not much different from predicting using the entire conversation.

TABLE 4.6: Random forest: MSPE user rating per dataset split (1st half)

	DS_1	**DS_2**	**DS_3**
User	**.42**	.77	**.47**
Chatbot	.43	**.76**	.54
User & Chatbot	.82	.67	.51

To further investigate the influence of splitting the conversations, the predictions per half are compared. For DS_2, the predicted user rating based on the first half is compared to the second half in Figure 4.10a and to the entire conversation in Figure 4.10b. The same scatter plots are created for DS_3, see Figure 4.11. If both predictions would be the same, a point would be positioned somewhere on the gray diagonal line. The larger the distance from this line, the larger the difference between the two predictions. As can be seen in most of the figures, the points are centered around this gray line. Although the predictions per half vary a bit, the points scatter equally to both sides of the line meaning the average predicted user rating is similar.

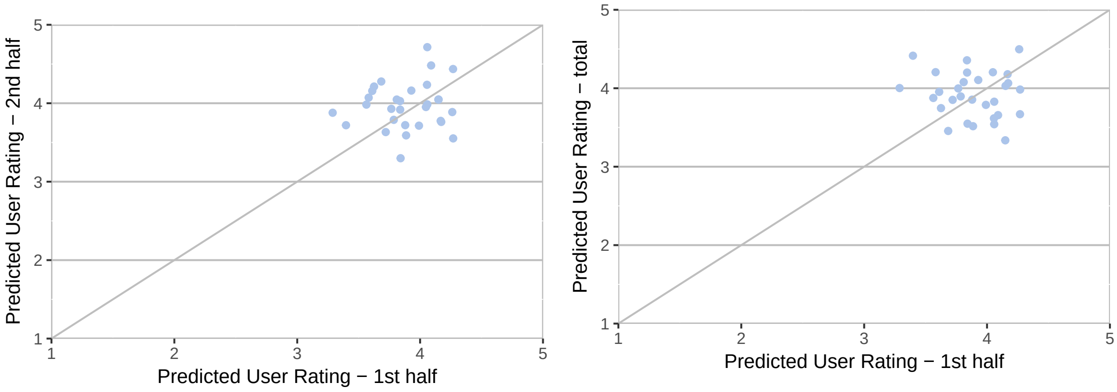
However, Figure 4.8b shows that the predicted user ratings based on the first half are generally lower than predictions based on the whole conversation. This is an interesting finding because it means that the prediction for the first half is generally lower than for the conversation as a whole. This could mean that the style of utterances of the chatbot in the frames dataset is different at the end of a conversation.



(A) Predicted user ratings 1st vs. 2nd half

(B) Predicted user ratings 1st half vs. entire conversation

FIGURE 4.10: Random forest prediction results for DS_2 (1st half) - Chatbot



(A) Predicted user ratings 1st vs. 2nd half

(B) Predicted user ratings 1st half vs. entire conversation

FIGURE 4.11: Random forest prediction results for DS_3 (1st half) - User

Concluding the random forest section, all the created models are not accurate enough. When a static predictor achieves similar results as a model, the model is not adding any value. This finding is strengthened by comparing the most predictive variables per model. The most predictive automatic metrics vary greatly per model, meaning the correlations are too weak to make accurate and consistent predictions. Moreover, splitting the conversations into halves does not improve the prediction precision. Therefore, no strong conclusions can be derived from this additional analysis.

5 Conclusion

In this chapter, the conclusion to the main research question is presented. First, the conclusions to the four sub-questions are discussed. Subsequently, the final conclusion to the main research question is provided.

5.1 Conclusion of Sub-Questions

To structure the research process, four sub-question were formulated. The first two sub-questions are based on theoretical foundations, which are derived from the body of literature on the topic of chatbot evaluation methods and natural language processing. The third and fourth sub-questions are based on the results of the performed data analysis.

[SQ1] *"Which approaches exist for evaluating chatbot performance?"*

Multiple existing chatbot performance evaluation methods are reviewed and compared. We distinguish two chatbot evaluation approaches: questionnaire-based and automatic metrics. The questionnaire-based research methods can be subdivided into two types: expert review and user opinion. We found that the expert review captures the chatbot's performance (quality of service), while the user's opinion captures the chatbot's perceived performance (quality of experience). Multiple methods were compared and merged, resulting in a list of six chatbot characteristics that together measure the perceived performance, see Table 3.2.

[SQ2] *"Which metrics can be automatically measured by analyzing chatbot conversations and with what techniques?"*

Research on analyzing natural language has been very active during the last couple of decades. The results of the literature review show that multiple metrics can be automatically measured, such as authenticity, confidence, readability, sentiment, etc. Linguistic researchers bundled many of these automatic metrics that led to the creation of the Linguistic Inquiry and Word Count (LIWC) tool, which is utilized in this research. Moreover, we found that each automatic metric can be measured on different levels and with different attributes, as discussed in subsection 3.2.1.

[SQ3] *"What patterns can be discovered in the automatic metrics and how are they related?"*

In order to answer the third sub-question, two datasets were gathered and analyzed. Three hypotheses focused on this sub-question were formulated and tested. The results of the data analysis show that few patterns exist between the automatic metrics. However, we found that analytical thinking is positively correlated with the readability in both datasets. Moreover, positive emotion words positively and strongly correlate with the sentiment in both datasets. Furthermore, some additional correlations that were found are not discussed due to their obviousness, such as correlations between the readability scores.

[SQ4] *"How can the automatic metrics be related to the perceived performance of chatbots?"*

The data analysis is extended to answer the fourth sub-question. Regarding this sub-question, four hypotheses were formulated and tested. The result to the first hypothesis shows that a positive correlation exists between sentiment and the perceived performance of both datasets. We conclude that a more positive tone in the messages reflects a more satisfied user. Furthermore, we found that negative emotion words are negatively and strongly correlated with the learnability. We conclude that users that express more negative emotions have more troubles understanding how to interact with the chatbot.

Finally, prediction models were created to investigate whether the automatic metrics could be used to predict the perceived performance. The findings show predicting with the automatic metrics is slightly more accurate than using a static predictor. Moreover, predicting by solely using the first half of each conversation yields similar results.

5.2 Conclusion of Main Research Question

[MRQ] *"How can the performance of chatbots be automatically quantified by analyzing its conversations?"*

The goal of this research project was to investigate what metrics can be automatically measured by analyzing chatbot conversations and how they relate to the chatbot's performance. To realize this goal, existing chatbot evaluation methods that are intended to capture the performance were analyzed, which resulted in a newly constructed set of six chatbot performance characteristics. Next, various natural language processing techniques were investigated and the corresponding measuring techniques and algorithms were discussed. Furthermore, two datasets with conversations were gathered and transformed into the listed automatic metrics. A few correlations were found between the automatic metrics. Subsequently, the relation of the automatic metrics to the perceived performance was analyzed, showing that some automatic metrics relate to the perceived performance categories. Finally, the correlations were used to build a prediction model for the performance of a chatbot. We found that building a prediction model for the performance of a chatbot using solely automatic metrics is possible but it is not accurate enough to add much value.

6 Discussion

In this chapter, the research project is discussed, the limitations of the project are addressed and ideas for future work are proposed.

The aim of this research project is to contribute to existing chatbot evaluation methods. As Radziwill and Benton (2017) show in their summarizing paper, previous research on the topic of chatbots evaluation mainly focuses on qualitative, questionnaire-based methods. However, the results of this research project show that evaluating chatbots is not limited to using questionnaires only. Automatically assessing chatbot performance with metrics is an interesting but still challenging approach. The data analysis results show that multiple correlations exist between the automatic metrics and the perceived performance. This finding complements earlier research conducted by C.-W. Liu et al. (2016) on another type of automatic metric. Although the found correlations are too weak to create a generalizable prediction model, this research can be considered as a new step towards the automatic evaluation of chatbot conversations.

This new step is not only interesting for academics but also for businesses that make use of chatbots because it could add a lot of business value. For example, during the data analysis we investigated whether predictions could be made based on the first half of conversations only. Although the results were similar to the other findings, it does showcase the possibilities of this type of research and the added value it could have for businesses. Furthermore, it perfectly fits in the train of thought of the continuous improvement and learning curve theory (Zangwill & Kantor, 1998), as mentioned in the introduction.

Concluding, the research field of automatically evaluating chatbot performance is still in its early phase but, nevertheless, very relevant for both academics and businesses.

6.1 Limitations

The largest limitation of this project is the data. Short after starting the project we discovered that there were no online available datasets that suited this research. Either the datasets were generated, too small, contained no perceived performance per conversation or had nothing to do with chatbots. The best option was the frames dataset by Microsoft, which was still not perfect because although the users thought they were communicating with a chatbot, the conversations were actually human-human. Furthermore, although the frames dataset contained a user rating score, it could only be considered as a basic version of the perceived performance score. Therefore, it is debatable whether the data analysis results of the frames dataset are fully applicable for chatbots.

Furthermore, to create a better suitable dataset, we performed our own case study. However, this dataset turned out to be very small in size with only 30 samples. With over 100 variables, the sampling number was only one-third of the number of input variables, while this ratio should preferably be a lot higher (Osborne & Costello, 2004). Moreover, only one chatbot was used during the case study. Additionally, the chatbot that was used, Mitsuku, is not goal-oriented. Although users were given the task to converse with Mitsuku, there was no option to classify a conversation as successful or unsuccessful. These two points limit the ability to derive generalizable insights from the data.

Moreover, the LIWC tool that was used to quantify a vast majority of the automatic metrics. The libraries and algorithms used by this tool are mostly open-source, but not entirely. Therefore it is questionable whether the algorithms of this tool are actually measuring what they intend to measure.

6.2 Future Work

This research project opens up multiple possibilities for future studies in the field of automatically evaluating the performance of a chatbot.

First, additional research could be performed on the automatic metrics. A large number of automatic metrics were selected for this study, but also many were left out. Researching the effects of other automatic metrics could lead to new insights. Moreover, the techniques that were used for measuring the automatic metrics could be researched further. The LIWC tool has many built-in libraries and algorithms, but future research could investigate whether this tool is indeed measuring what it intends to measure. Finally, the LIWC tool is mainly based on research concerning natural language conversations. Future research could investigate whether these algorithms are also suitable for quantifying the conversations of chatbots.

Additionally, the research could be extended by including case studies of other chatbots to compare their performance with one another. Such a study could distinguish between two chatbot types, namely goal-oriented and non-goal-oriented chatbots. Since the reason for communicating is very different, it is very likely that the patterns between the automatic metrics and the perceived performance differ as well.

Finally, the results of the two datasets already showed promising patterns between the automatic metrics and the chatbot performance. Additional data analysis on a large scale dataset could further investigate these patterns. Subsequently, further research could lead to a prediction model that has a higher precision in predicting a chatbot's performance. Such a model could have a large business value when it is actively running during chat conversations. When the model flags a chat conversations as unsuccessful, a customer support employee could take over only those conversations that require the attention of a human.

Bibliography

Abran, A., Khelifi, A., Suryn, W., & Seffah, A. (2003). Consolidating the ISO usability models. In *Proceedings of 11th international software quality management conference* (pp. 23–25). doi:10.1.1.93.3969

Anastas, J. W. (1999). *Research Design for Social Work and the Human Services* (2nd). New York: Columbia University Press.

Asri, L. E., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., … Suleman, K. (2017). *Frames: A Corpus for Adding Memory to Goal-Oriented Dialogue Systems*.

Austin, J. L. (1962). *How to do Things with Words*. Oxford: Clarendon Press.

Azevedo, A. I. R. L. & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Baarda, B. (2014). *Dit is onderzoek! Handleiding voor kwantitatief en kwalitatief onderzoek.* Groningen/Houten: Noordhoff Uitgevers.

Banerjee, S. & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (June, pp. 65–72). Association for Computational Linguistics. doi:10.1007/s10590-009-9059-4

Baumeister, R. F. (2013). Writing a Literature Review. In M. J. Prinstein (Ed.), *The portable mentor: Expert guide to a successful career in psychology* (2nd, Chap. 8, pp. 119–132). New York: Springer Science+ Business Media. doi:10.1007/978-1-4614-3994-3

Baumeister, R. F. & Leary, M. R. (1997). Writing narrative literature reviews. *Review of General Psychology*, *1*(3), 311–320. doi:10.1037/1089-2680.1.3.311

Benbasat, I., Goldstein, D. K. D., & Mead, M. (1987). The Case Research Strategy in Studies of Information Case Research. doi:10.2307/248684

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. doi:10.1007/BF00058655

Budgen, D. & Brereton, P. (2006). Performing systematic literature reviews in software engineering. In *Icse '06 proceedings of the 28th international conference on software engineering* (p. 1051). doi:10.1145/1134285.1134500

Bunt, H. (2015). On the principles of interoperable semantic annotation. In *Proceedings of the 11th joint acl-iso workshop on interoperable semantic annotation* (pp. 1–13).

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., & Yan, H. (2000). Conversation as a system framework: Designing embodied conversational agents. *Embodied conversational agents*, 29–63.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0. *CRISP-DM Consortium*, 76. doi:10.1109/ICETET.2008.239

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd). Hillsdale, NJ: Erlbaum.

Daniel, J. & Martin, J. H. (2017). Part-of-Speech Tagging. In *Speech and language processing* (Chap. 10). Upper Saddle River, New Jersey: Pearson. Retrieved from https://web.stanford.edu/~jurafsky/slp3/10.pdf

Desurvire, H. W. (1994). Faster, cheaper!! Are usability inspection methods as effective as empirical testing? In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 173–202). New York: John Wiley & Sons, Inc.

Desurvire, H. W., Kondziela, J. M., & Atwood, M. E. (1992). What is gained and lost when using evaluation methods other than empirical testing. In A. Monk, D. Diaper, & M. Harrison (Eds.), *People and computers vii* (pp. 89–102). Cambridge: Cambridge University Press. doi:10.1145/1125021.1125115

Doran, C., Aberdeen, J., Damianos, L., & Hirschman, L. (2001). Comparing several aspects of human-computer and human-human dialogues. In *Proceedings of the second sigdial workshop on discourse and dialogue -* (Vol. 16, pp. 1–10). Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1118078.1118085

Dos Santos, C. N. & Gatti, M. (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. *COLING*, 69–78.

Flesch, R. F. (1951). *How to test readability*. New York: Harper.

George R. Klare, Rowe, P. P., John, M. G. S., & Stolurow, L. M. (1969). Automation of the Flesch Reading Ease Readability Formula , with Various Options. *Reading Research Quarterly*, *4*(4), 550–559.

Gottschalk, L. A. (1997). The unobtrusive measurement of psychological states and traits. *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*, 117–129.

Gottschalk, L. A. & Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior* (Doctoral dissertation, CA: University of California Press).

Güngör, T. (2010). Part-of-Speech Tagging. In R. Herbrich & T. Graepel (Eds.), *Handbook of natural language processing* (2nd, Chap. 10, pp. 205–235). Cambridge: Chapman & Hall/CRC.

Haffar, J. (2016). Have you seen ASUM-DM? Retrieved from https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/

Hassan, A., Qazvinian, V., & Radev, D. (2010). What 's with the Attitude?: Identifying Sentences with Attitude in Online Discussions. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1245–1255). Association for Computational Linguistics. Retrieved from http://portal.acm.org/citation.cfm?id=1870779

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning*. doi:10.1198/jasa.2004.s339

HubSpot. (2017). Artificial Intelligence Is Here - People Just Don't Realize It. Retrieved from https://research.hubspot.com/artificial-intelligence-is-here?_ga=2.103202371.23388408.1520943910-535391447.1520943910

Hung, V., Elvir, M., Gonzalez, A., & DeMara, R. (2009). Towards a method for evaluating naturalness in conversational dialog systems. In *Conference proceedings - ieee international conference on systems, man and cybernetics* (pp. 1236–1241). doi:10.1109/ICSMC.2009.5345904

IBM. (2017). How chatbots can help reduce customer service costs by 30%. Retrieved from https://www.ibm.com/blogs/watson/2017/10/how-chatbots-reduce-customer-service-costs-by-30-percent/

IBM Corporation. (2016). *Analytics Solutions Unified Method*.

Ide, N. & Romary, L. (2004). International standard for a linguistic annotation framework. *Natural language engineering*, *10*(3-4), 211–225. doi:10.3115/1119226.1119230

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2013). Pronoun Use Reflects Standings in Social Hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125–143. doi:10.1177/0261927X13502654

Karat, C., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of chi'92* (pp. 397–404). New York: ACM.

KDnuggets. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved from https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Kiffin-Petersen, S., Murphy, S. A., & Soutar, G. (2012). The problem-solving service worker: Appraisal mechanisms and positive affective experiences during customer interactions. *Human Relations*, *65*(9), 1179–1206. doi:10.1177/0018726712451762

Kirange, D., Deshmukh, R., Jain, T., & Patil, S. D. (2011). Opinion Mining: A Study on Automated Text Based Emotion Prediction. In *Ieee international conference on knowledge engineering* (Chap. 5).

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, *50*, 723–762. doi:10.1613/jair.4272

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(TR/SE-0401), 28. doi:10.1.1.122.3308

Klein, H. H. K. & Myers, M. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS quarterly*, *23*(1), 67–93. doi:10.2307/249410

Kuligowska, K. (2015). Commercial chatbot: Performance evaluation, usability metrics and quality standards of embodied conversational agents. *Professionals Center for Business Research*, *2*, 1–16. doi:10.18483/PCBR.22

Lee, A. S. & Baskerville, R. L. (2003). Generalizing Generalizability in Information Systems Research. *14*(September), 221–243.

Leek, J. T. & Peng, R. D. (2015). What is the question? *347*(6228), 1314–1315. doi:10.1126/science.aaa6146

Lemon, K. N. & Verhoef, P. C. (2016). Understanding Customer Experience Throughout the Customer Journey. *Journal of Marketing*, *80*(6), 69–96. doi:10.1509/jm.15.0420

Lester, J., Branting, K., & Mott, B. (2004). Conversational agents. *The Practical Handbook of Internet Computing*, 1–17.

Levy, Y. & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science Journal*, *9*, 181–212. doi:10.1049/cp.2009.0961

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Proceedings of the workshop on text summarization branches out (WAS 2004)*, *8*.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis lectures on human language technologies*, *5*(1), 1–167. doi:10.2200/S00416ED1V01Y201204HLT016

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., & Pineau, J. (2016). How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation.

LIWC. (2017). LIWC | Linguistic Inquiry and Word Count. Retrieved from http://liwc.wpengine.com/

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., & Pineau, J. (2017). Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. doi:10.18653/v1/P17-1103

McLaughlin, G. (1969). SMOG grading: A new readability formula. *Journal of reading*, *12*(8), 639–646. doi:10.1039/b105878a

McTear, M., Callejas, Z., & Griol, D. (2016). Creating a Conversational Interface Using Chatbot Technology. In *The conversational interface* (pp. 125–159). Cham: Springer International Publishing. doi:10.1007/978-3-319-32967-3{_}7

Medhi Thies, I., Menon, N., Magapu, S., Subramony, M., & O'Neill, J. (2017). How do you want your chatbot? An exploratory Wizard-of-Oz study with young, Urban Indians. *Lecture Notes in Computer Science (including subseries Lecture Notes* *in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10513 LNCS*, 441–459. doi:10.1007/978-3-319-67744-6{_}28

Meira, M. & Canuto, A. M. P. (2015). Evaluation of Emotional Agents' Architectures: an Approach Based on Quality Metrics and the Influence of Emotions on Users. In *Proceedings of the world congress on engineering* (Vol. 1).

Mihailidis, P. (2014). The civic-social media disconnect: exploring perceptions of social media for engagement in the daily life of college students. *Information Communication and Society*, *17*(9), 1059–1071. doi:10.1080/1369118X.2013.877054

Moller, S., Engelbrecht, K.-P., Kuhnel, C., Wechsung, I., & Weiss, B. (2009). A taxonomy of quality of service and Quality of Experience of multimodal human-machine interaction. *International Workshop on Quality of Multimedia Experience*, 7–12. doi:10.1109/QOMEX.2009.5246986

Morrissey, K. & Kirakowski, J. (2013). 'REALNESS' IN CHATBOTS: ESTABLISHING QUANTIFIABLE CRITERIA. In *International conference on human-computer interaction* (pp. 87–96). Berlin Heidelberg: Springer.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Personality and Social Psychology Bulletin. *Society for Personality and Social Psychology, Inc. 29*(5), 665–675. doi:10.1177/0146167203251529

Nielsen, J. (1997). How Users Read on the Web. Retrieved from https://www.nngroup.com/articles/how-users-read-on-the-web/

Osborne, J. W. & Costello, A. B. (2004). Sample size and subject to item ratio in principal components analysis. - Practical Assessment, Research & Evaluation. *9*(11).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *The 40th annual meeting on association for computational linguistics (acl).* (July, pp. 311–318). doi:10.3115/1073083.1073135

Peng, R. D. & Matsui, E. (2016). *The Art of Data Science*. lulu.com.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin. doi:10.1068/d010163

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. In *Plos one* (Vol. 9, *12*, pp. 1–10). doi:10.1371/journal.pone.0115844

Radziwill, N. M. & Benton, M. C. (2017). *Evaluating Quality of Chatbots and Intelligent Conversational Agents*.

Robson, C. & McCartan, K. (2016). *Real world research* (4th). John Wiley & Sons.

Runeson, P. & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, *14*(2), 131–164. doi:10.1007/s10664-008-9102-8

Searle, J. R. (1969). *Speech Acts*. London-New York: Cambridge University Press.

Semeraro, G., Andersen, H. H., Andersen, V., Lops, P., & Abbattista, F. (2002). Evaluation and Validation of a Conversational Agent Embodied in a Bookstore. In N. Carbonell & C. Stephanidis (Eds.), *Universal access*. Paris, France. doi:10.1007/978-3-642-23196-4

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., & Pineau, J. (2015). *A Survey of Available Corpora for Building Data-Driven Dialogue Systems*.

Snijder, J. (2018). *Chatbot Maturity Model*. Info Support.

Statista. (2017). Most famous social network sites 2017, by active users. Retrieved from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: MIT Press.

Tausczik, Y. R. & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. doi:10.1177/0261927X09351676

Turing, A. M. (1950). Mind Association, Oxford University Press. *Mind*, *59*(236), 433–460.

Van Eeuwen, M. (2017). *Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers* (Doctoral dissertation, University of Twente). Retrieved from http://essay.utwente.nl/71706/1/van%20Eeuwen_MA_BMS.pdf

Venturebeat. (2016). 3 stats that show chatbots are here to stay. Retrieved from https://venturebeat.com/2016/08/26/3-stats-that-show-chatbots-are-here-to-stay/

Webster, J. & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future : Writing a Literature Review. *MIS Quarterly*, *26*(2), 13–23.

Wieringa, R. J. (2014). *Design Science Methodology: for Information Systems and Software Engineering*. Berlin, Heidelberg: Springer. doi:10.1007/978-3-662-43839-8

Wirth, R. & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29–39).

Yin, R. K. (2013). *Case study research: Design and methods*. Sage publications.

Zangwill, W. I. & Kantor, P. B. (1998). Toward a Theory of Continuous Improvement and the Learning Curve. *Management Science*, *44*(7), 910–920. doi:10.1287/mnsc.44.7.910

A Appendix: Case Study

A.1 Information Sheet

Dear participant, thank you for participating in this master thesis research project that is being conducted by Cas Jongerius in collaboration with Utrecht University and Info Support. This research is positioned around the topic of automatically evaluating the performance of chatbots. The goal of this project is to find out whether we can make accurate estimations about how the user experiences a conversation with a chatbot, by solely analyzing the chat messages. This experiment serves the purpose of collecting a dataset containing chat conversations on which data analysis can be performed. In this experiment you'll be given the task to chat with a chatbot. After the conversation you're requested to fill in a short questionnaire to measure how you've experienced the interaction with the chatbot. This experiment will take 10 minutes approximately. Kindly read the following information regarding the confidentiality of the data:

- Your chat conversations and questionnaire results will be anonymized, meaning your name will not be linked to the conversations or questionnaire results. Moreover, personal information that is provided during the conversation will be replaced by pseudonyms.
- The anonymized chat conversations and questionnaire results will be archived by the researcher.
- The anonymized chat conversation and questionnaire results will be quantified for data analysis (meaning written text will be transformed into metrics) and might be shared with Info Support and other researchers or used in publications.

Before we start the experiment, feel free to ask any questions concerning the project or your participation. When ready, we kindly ask you to read and sign the informed consent and fill in the demographic questionnaire.

A.2 Informed Consent

I, the undersigned, confirm that (please tick box as appropriate):

1. I have read and understood the information about the project, as provided in the Information Sheet.
2. I have been given the opportunity to ask questions about the project and my participation.
3. I voluntarily agree to participate in the project.
4. I understand I can withdraw at any time without giving reasons and that I will not be penalized for withdrawing nor will I be questioned on why I have withdrawn.
5. The procedures regarding confidentiality have been clearly explained (e.g. use of names, pseudonyms, anonymization of data, etc.) to me.
6. The use of the data in research, publications, sharing and archiving has been explained to me.
7. I understand that other researchers and Info Support will have access to this data only if they agree to preserve the confidentiality of the data and if they agree to the terms I have specified in this form.
8. I, along with the Researcher, agree to sign and date this informed consent form.

Participant:

Name of Participant	Signature	Date

Researcher:

Name of Researcher	Signature	Date

A.3 Experiment Instructions

For this experiment we request you to have a conversation with the Mitsuku, a chatbot driven by artificial intelligence. The assignment is to have a short recreational conversation with Mitsuku, meaning the conversation has no goal other than 'small talk'. You're free to talk about anything that you like. Some example topics are:

- Food
- Sports
- Weather
- Traveling
- Jokes
- Animals
- Friends

When you are ready, start the conversation. Mitsuku can only communicate in English. After five minutes the researcher will ask you to end the conversation. Until then, try to keep the conversation going. After the conversation you're asked to fill in the questionnaire. Good luck!

A.4 Survey

TABLE A.1: Case study survey

Characteristic	**Question**
Impression	I would use the chatbot again in case I'm in for a fun conversation The chatbot was complex and not very comfortable to use Interacting with the chatbot was intimidating Interacting with the chatbot was frustrating I would recommend this chatbot to my friends / colleagues Interacting with the chatbot was very difficult Interacting with the chatbot was pleasing
Command	The chatbot responded simple and quickly to my messages I felt like I did not have control over the chatbot The chatbot did not respond fast enough to my messages It was easy to handle the chatbot exactly as I wanted
Learnability	It was clear from the start how I could communicate with the chatbot Using the chatbot was easy to learn I had trouble formulating messages that the chatbot would understand
Naturalness	The language skills of the chatbot were good The chatbot was capable of having a natural conversation The chatbot was not aware of its context
Comprehension	It felt like the chatbot understood what I was saying The answers provided by the chatbot were not consistent To get a proper answer to my questions, I had to ask multiple times Communicating with the chatbot was satisfying The answers of the chatbot were too complex I understood all the questions of the chatbot
Friendliness	It felt like the chatbot had a pleasant personality The utterances of the chatbot were indifferent The chatbot was unfriendly
User rating	How would you rate the chatbot?

B Appendix: Automatic Metrics

TABLE B.1: Automatic Metrics in datasets

Category	Abbreviation	Examples
Timing Variables		
Turn count	TC	
Duration in (sec)	Duration	
Average respond time (sec)	ART	
Readability Variables		
Flesch Kincaid	Flesch_Kincaid	
Gunnig Fog Index	Gun_Fog_Ind	
Coleman Liau	Coleman_Liau	
SMOG	SMOG	
Automated Readability Index	Auto_Readability_Ind	
Average Grade Level	Avg_Gr_Lvl	
Survey Variables		
User average*	Average_Rating	
Impression*	Impression	
Command*	Command	
Learnability*	Learnability	
Naturalness*	Naturalness	
Comprehension*	Comprehension	
Friendliness*	Friendliness	
User rating	User_Rating	
Summary Language Variables		
Analytical thinking	Analytic	
Confidence	Confidence	
Authentic	Authentic	
Sentiment	Sentiment	
Word count	WC	
Words per sentence	WPS	
Words > 6 letters	Sixltr	
Dictionary words	Dic	
Linguistic Dimension Variables		
Total function words	funct	it, to, no, very
Total pronouns	pronoun	I, them, itself
Personal pronouns	ppron	I, them, her
1st pers singular	i	I, me, mine

*Available in the case study dataset only

Table B.1 continued

Category	Abbreviation	Examples
1st pers plural	we	we, us, our
2nd person	you	you, your, thou
3rd pers singular	shehe	she, her, him
3rd pers plural	they	they, their, they'd
Impersonal pronouns	ipron	it, it's, those
Articles	article	a, an, the
Prepositions	prep	to, with, above
Auxiliary verbs	auxverb	am, will, have
Common Adverbs	adverb	very, really
Conjunctions	conj	and, but, whereas
Negations	negate	no, not, never
Other Grammar		
Common verbs	verb	eat, come, carry
Common adjectives	adj	free, happy, long
Comparisons	compare	greater, best, after
Interrogatives	interrog	how, when, what
Numbers	number	second, thousand
Quantifiers	quant	few, many, much
Psychological Processes		
Affective processes	affect	happy, cried
Positive emotion	posemo	love, nice, sweet
Negative emotion	negemo	hurt, ugly, nasty
Anxiety	anx	worried, fearful
Anger	anger	hate, kill, annoyed
Sadness	sad	crying, grief, sad
Social processes	social	mate, talk, they
Family	family	daughter, dad, aunt
Friends	friend	buddy, neighbor
Female references	female	girl, her, mom
Male references	male	boy, his, dad
Cognitive processes	cogproc	cause, know, ought
Insight	insight	think, know
Causation	cause	because, effect
Discrepancy	discrep	should, would
Tentative	tentat	maybe, perhaps
Certainty	certain	always, never
Differentiation	differ	hasn't, but, else
Perceptual processes	percept	look, heard, feeling
See	see	view, saw, seen
Hear	hear	listen, hearing
Feel	feel	feels, touch
Biological processes	bio	eat, blood, pain
Body	body	cheek, hands, spit
Health	health	clinic, flu, pill
Sexual	sexual	horny, love, incest
Ingestion	ingest	dish, eat, pizza

Table B.1 continued

Category	**Abbreviation**	**Examples**
Drives	drives	
Affiliation	affiliation	ally, friend, social
Achievement	achieve	win, success, better
Power	power	superior, bully
Reward	reward	take, prize, benefit
Risk	risk	danger, doubt
Time orientations	TimeOrient	
Past focus	focuspast	ago, did, talked
Present focus	focuspresent	today, is, now
Future focus	focusfuture	may, will, soon
Relativity	relativ	area, bend, exit
Motion	motion	arrive, car, go
Space	space	down, in, thin
Time	time	end, until, season
Personal concerns		
Work	work	job, majors, xerox
Leisure	leisure	cook, chat, movie
Home	home	kitchen, landlord
Money	money	audit, cash, owe
Religion	relig	altar, church
Death	death	bury, coffin, kill
Informal language	informal	
Swear words	swear	fuck, damn, shit
Netspeak	netspeak	btw, lol, thx
Assent	assent	agree, OK, yes
Nonfluencies	nonflu	er, hm, umm
Fillers	filler	Imean, youknow

C Appendix: Results

TABLE C.1: Top-9 most predictive variables in random forest

DS1					
User		Chatbot		User & Chatbot	
Variable	%IncMSE	Variable	%IncMSE	Variable	%IncMSE
relativ	5.85	bio	4.56	bio	4.44
Coleman_Liau	3.26	motion	2.30	assent	2.75
Gun_Fog_Ind	2.76	affiliation	2.15	Confidence	2.30
Avg_Gr_Lvl	2.06	i	1.93	certain	2.28
Duration	1.76	Duration	1.78	motion	2.15
Sixltr	1.73	affect	1.77	Gun_Fog_Ind	2.12
nonflu	1.59	leisure	1.64	percept	2.07
ipron	1.58	drives	1.53	see	1.63
Flesch_Kincaid	1.50	SMOG	1.48	informal	1.48
DS2					
User		Chatbot		User & Chatbot	
Variable	%IncMSE	Variable	%IncMSE	Variable	%IncMSE
prep	8.11	negate	8.13	Coleman_Liau	6.71
relativ	3.98	Duration	6.77	Sixltr	6.34
WC	3.61	QMark	5.41	AllPunc	5.54
verb	3.59	ppron	5.38	negate	5.53
TC	3.35	number	5.34	Duration	5.26
Confidence	3.24	article	5.08	number	5.20
posemo	3.15	WC	4.87	work	4.99
motion	3.13	you	4.66	Apostro	4.54
Analytic	3.06	WPS	4.56	Avg_Gr_Lvl	4.34
DS3					
User		Chatbot		User & Chatbot	
Variable	%IncMSE	Variable	%IncMSE	Variable	%IncMSE
Sixltr	6.02	negate	11.50	Coleman_Liau	8.46
see	5.44	TC	7.55	negate	7.16
affect	5.23	Duration	7.19	Duration	7.10
Duration	4.95	SMOG	6.67	Sixltr	6.76
prep	4.70	Sentiment	6.41	Apostro	6.60
posemo	4.62	conj	6.39	Avg_Gr_Lvl	6.33
Coleman_Liau	4.39	number	6.21	work	6.19
Exclam	4.25	WC	6.21	AllPunc	6.04
WC	4.13	function	6.01	Authentic	5.82