MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Chatting About Data - Interacting with Voice Interfaces to Engage with Election Panel Data"

verfasst von / submitted by

Bernhard Jordan BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066935
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Medieninformatik
Betreut von / Supervisor:	Univ. Prof. Torsten Möller, PhD
Mitbetreut von / Co-Supervisor:	Laura Koesten, MSc PhD

Abstract

The importance of data is increasing and so is the use of voice-interface-chatbots. While tasks that can be handled by chatbots are still relatively simple, analysing data can be a difficult task for non-experts. The aim of the research was to show how an interaction with a voice-interface-chatbot to explore data can work. In a preliminary study we investigated what questions are asked by people when looking at tabular data. We then conducted a Wizard of Oz study with a voice-interface-chatbot that can answer questions about the National Council elections with text or graphs. In 15 interviews 1235 messages were exchanged, of which 159 were categorized as information requests. Our study shows that conversational interfaces should have the ability to include past conversation history in the evaluation of questions in order to encourage user engagement. We present a typology of questions and provide in-depth insights to inform the development of chatbots for human data interaction.

Kurzfassung

Die Bedeutung von Daten nimmt kontinuierlich zu und damit auch der Einsatz von Chatbots mit Sprache als Interaktionsmittel. Während die Aufgaben, die von Chatbots erledigt werden können, noch relativ einfach sind, kann die Analyse von Daten für Nicht-Experten eine schwierige Aufgabe sein. Ziel dieser Forschung war es, zu zeigen, wie die Interaktion mit einem Sprach-Eingabe-Chatbot zur Erforschung von Daten funktionieren kann. In einer Vorstudie untersuchten wir, welche Fragen an tabellarische Daten gestellt werden. Anschliessend führten wir eine Wizard-of-Oz-Studie mit einem Voice-Interface-Chatbot durch, der Fragen zu den Nationalratswahlen mithilfe von Text oder Grafiken beantworten kann. In 15 Interviews wurden 1235 Nachrichten ausgetauscht, von denen 159 als Informationsanfragen kategorisiert wurden. Unsere Studie zeigt, dass Sprachchatbots die Möglichkeit haben sollten, frühere Konversationen in die Bewertung von Fragen einzubeziehen, um die Motivation der Benutzer zu fördern. Wir stellen eine Typologie von Fragen vor und geben detaillierte Einblicke, um die Entwicklung von Chatbots für die Interaktion von Menschen mit Daten zu unterstützen.

Contents

List of Tables

List of Figures

1. Introduction

The importance of data is increasing as more and more data is being collected constantly. With a lot of data being available to the public on the internet (e.g. [CfRI18]) it is essential to understand and be able to make decisions based on data. For many, analysing data to find the information they need is a barrier to working in a data-informed way.

The aim of this work is to understand voice interaction with data better in order to make datasets more accessible and easier for people who are not data experts. For this purpose, we have conducted two studies: In the first step we conducted a preliminary study in which participants interacted with spreadsheets that were new to them. They were asked to think-aloud during a task based exploration session, focusing on the questions they might have about the data. Our aim was to collect the questions asked, identify what information from the dataset would interest them, how they interact with the spreadsheet, whether they think the data is trustworthy and why, and which type of presentation of the data would be considered helpful. Our aim was to gain insights into the sensemaking process of non-experts with data they are unfamiliar with.

Informed by the findings from the preliminary study we designed a Wizard-of-Oz study with a chatbot to investigate the interaction of non-experts (in this case first-time voters) with election panel data through spoken language. The thematic focus was chosen in order to make the task personally relevant and to achieve honest engagement.

We used the insights into question types and expected answers from the first study to develop a prototype of a chatbot. A conventional chat interface was chosen as the interface (see Figure 1.1), using voice as input from the participants and text, as well as images and data visualizations, as answers by the chatbot.

The chatbot was developed to answer questions about the national parliamentary elections in Austria. As research has shown that interest in data is related to personal interest and experience, we chose data on a political event, relevant to the participants [PAE19]. 16-25 year-olds were chosen as the target group for the study because the minimal voting age in this country is 16 and because they belong in the so-called "early voters" group meaning they are likely to have natural interest in deciding whether to vote and who for. In addition, as digital natives, this group of people are perhaps more open for novel interfaces. Hence, they are a particularly interesting group for the use of voice input. f

Conversational interfaces (CIs) have found their way into many households in recent years [LS16] (e.g. Apple's Siri [sir22], Google assistant [goo22a] or Amazon's Alexa [ale22]). Not only do they perform tasks such as: "Turn off the light" and "Set the heating to 21 degrees", they also answer questions about the weather, sporting events or the TV program. What these voice assistants have in common is that they all have speech as input. The output is also primarily spoken language with the additional option of displaying further information on a screen. Despite continuous advances, the questions

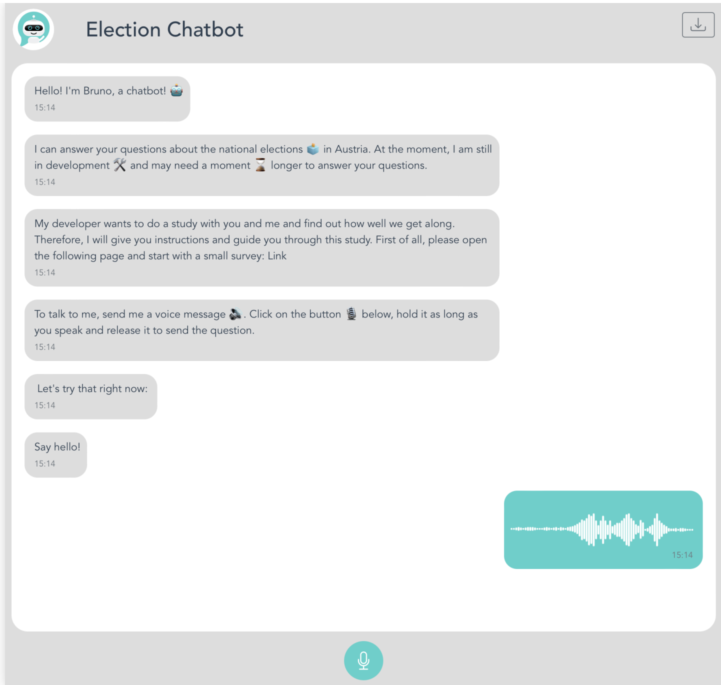


Figure 1.1.: Screenshot of the chatbot

that these CIs can answer are general and for their success "we need to make them capable of interacting with naive, uninformed humans in everyday situation" [KGKW05].

Our studies have shown different types of questions which are asked by non-experts, when exploring data. Furthermore, it is shown how the interaction with a conversational interface looks like that can answer data questions on the national parliamentary elections. In this study, visualisations were shown to play an important role in answering questions. It was shown that people refer back to individual elements from the visualisations in the course of the conversation and chatbots should be able to answer questions about these visual features. In order to encourage the engagement of the users of a chatbot, we have observed that a human communication style of the chatbot is helpful and that interactive charts are a great way to stimulate further engagement with the data.

1.1. Research questions

Developing a chatbot requires detailed knowledge about what users want to know about data and what questions they ask about it [DR20]. In the first step, we therefore looked at what types of questions are asked of data in spoken language when conversing with a person about the data and what answers users expect. In the second step, we looked at how we can build a chatbot that brings this knowledge together and encourages the users to use it. In more detail we took a closer look at the following:

- **RQ1** - What types of questions do young, lay people ask datasets and spreadsheets? What are the characteristics of these and what types of answers are expected?
- **RQ2** - If we want to engage non-experts to use natural language to explore datasets, what are the requirements for building such a tool?
- **RQ3** - What might a chatbot look like that can answer questions about National Parliamentary election panel data? What requirements are necessary in order to increase understanding and engagement of the users?

1.2. Contributions

This work presents two studies on people's interaction with data in which they interact with spreadsheets and chatbot prototypes respectively. The contributions of this paper are specifically:

- A typology of question types for voice input engagement with data
- In-depth insights into early voters engagement with voice interfaces about election data
- Detailed requirements for the design of a conversational interface to explore national election data.

2. Related Work

Voice interfaces have recently become more popular and accessible [LS16]. Many people already own home assistants [hom22, ama22, goo22b], but more and more devices such as computers, smartphones, smartwatches and cameras can also be interacted with by voice. Ewa et al. [LS16] have shown, that there is still a big gulf between users' expectation and their experience with conversational tools. This applies not only to speech-based interfaces, but also to text-based chatbots. The possibilities of the current voice interfaces to explore data are still limited [FCM+18]. How the interaction with voice interfaces to explore data can look like is not yet sufficiently researched.

2.1. Human-like Chatbot

Rapp et al. [RCB21] have shown that it is important for human trust in the chatbot that people have an idea of what the chatbot can do, that it answers quickly and is transparent in its responses. They and others [CG21, MSPV21] also show that the way the chatbot speaks is crucial for the user experience. In a Wizard-of-Oz study, Thies et al. [MMM+17] show which personalities young urban Indians find most compelling when using a chatbot. The personality of a "productivity oriented bot with nerd wit" was best rated by the participants. In a laboratory experiment, Rzepka et al. [RBH22] showed that people prefer speech-based interactions with voice assistants over text-based conversations with chatbots because of higher perceived efficiency, lower cognitive effort and greater satisfaction. What needs to be considered is not only how the chatbot needs to act, but also what patterns of data-centric sensemaking are prevalent, as Koesten et al. [KGGS21] demonstrated, and how exploration activities of data can be supported via such interfaces.

2.2. Chatbots for data exploration

Several chatbots that try to make data more explorable already exist. These are mostly based on text input and text output, and focus on specific topics. The Open Data Chatbot [KSV19] and Talking Open Data [NSV17] projects are attempting to make public datasets accessible, whereby the first is intended to recommend open data through dialogue, while the second also makes it possible to browse through the data structures via chatbots on popular networks such as Skype or Facebook. These systems help to find publicly available datasets on the searched topics, but they do not help to make those datasets explorable. Gamebot [ZM20] focuses on linking text and graphics as output in a sports context, while Bieliauskas and Schreiber [BS17] concentrate on the visualisation of software-architecture using conversational interfaces. Valletto [KR18] is a conversational interface as a prototype tablet app that can be used to build visualisations through voice input and multi-touch gestures. Iris [FCM$^+$18] deals with the relevance of memory, i.e. the reference to past messages in the chat process. This system can create visualisations and answer data science related questions. Unlike our approach, it does this with text-input and is geared towards data experts.

The previous chatbots all work with data, but none support exploration of datasets through voice input. The following works have looked at natural language interfaces, DataTone [GDA$^+$15] turns natural language queries into visualisations. Our aim was not to create visualisations with the help of voice interfaces, the focus of this work is the exploration of and engagement with data. The tools Eviza [SBT$^+$16] and Evizeon [HSTD18] support follow up questions in natural language conversations, which was extended with Orko [SS18] to include network diagrams.

2.3. Interaction with chatbots

The following studies show how a chatbot deals with the problems of natural language and connected user expectations. Based on the Valetto prototype [KR18], Kassel and Rohs [KR19] tested what expectations people with different statistical knowledge have of chat responses. They found out that the level of detail in the chatbot response influences user assessments of appropriateness. Hearst and Tory [HT19] used crowdsourced studies to show that there is a gap in people's preferences about visualisations. 40% prefer not to have charts as answers on questions about data. Those who prefer charts like to get additional context in addition to the requested data. Hearst et al. [HTS19] have looked at how to deal with underspecified terms such as "high" or "expensive" in natural conversations to make conversations meaningful, which we considered in building our prototype.

A study similar to the presented one was conducted by Setlur and Tory [ST22]. They conducted a comparative study with Wizard-of-Oz chatbots looking at 3 modalities (text interaction using Slack [sla22], voice interaction using a Bluetooth speaker, voice interaction using an iPad + Bluetooth speaker) and used a dataset about passengers of the Titanic. The aim was to evaluate user expectations about the conversation and the answers, whereas our study focuses more on the interaction with the chatbot. While they do not have a defined target group, we focus on young voters up to 25 years of age and use a dataset on the national elections, which might be closer to the participants information needs.

3. Method

3.1. Preliminary Study

In a first step, we tried to gain insights into how non-experts interact with data as spreadsheets. This should provide information on how a chatbot that works with data should function and which questions or which types of questions it must be able to answer. People who potentially come into contact with data but are not experts were chosen as the target group. The following questions will be answered with the help of the study:

- How do non-data experts interact with spreadsheets?
- What questions do they ask about the data?

For this, we conducted semi-structured interviews with 12 people, none of them are data experts. The participants were between 22 and 57 years old, eight were male, four female, all of them had an academic background and were recruited through personal contacts. For each interview, the participants were shown two spreadsheets, each with only a brief explanation of the topic. The participants were asked to come up with questions that spontaneously came to their mind while looking at the dataset. For each dataset, two tasks were given to the participants, and they had to formulate questions in order to gain further insights.

3.1.1. Datasets

When choosing the datasets, we ensured that the topic was not too specific and did not require a lot of prior knowledge. It was also important to map different types of data: format, location information, time data, etc. We used three datasets: *Global Temperature Time Series* [glo22], *CO_2 Emissions from Fossil Fuels* [CO222], *World Population Prospects* [wor22]. As can be seen in Table 3.1, the three datasets consist of time series data, some with location-dependent information, and some with aggregated values.

3.1.2. Design Decisions

The study setup included six individual interviews and two group interviews. The group interviews were conducted with three participants each. The idea was that in the conversation about the dataset users could ask each other questions and discuss uncertainties among themselves.

Title	**Global Temperature Time Series** [glo22]	**CO2 Emissions from Fossil Fuels** [CO222]	**World Population Prospects** [wor22]
Source	Carbon Dioxide Information Analysis Center (CDIAC) [cdi22]	GISTEMP [GIS22] & GCAG [gca22]	United Nations
Info	Global monthly mean and annual mean temperature anomalies in degrees Celsius	Per Country CO2 Emissions from fossil-fuels annually	Population estimates from 1950 to the present per country
Temporal Information	yearly/monthly	yearly	yearly
Spacial Information	no	Countries	Countries
Aggregated scores	no	yes	no

Table 3.1.: Description of the used datasets in the preliminary study

The individual interviews lasted 30 minutes each and due to the Covid-19 pandemic, five interviews were conducted remotely via Zoom and only one in person. Each participant was shown two of the three datasets. These were rotated for each participant so that all permutations occurred once and no interview was conducted in the same order and with the same datasets.

For the group interviews, the two datasets *World Population Prospects* and the *Global CO_2 emissions on fossil fuels* were used. The interviews lasted 45 minutes each, the first interview took place via Zoom and the second in person. The interviews were audio and video recorded with the consent of the participants and then transcribed using *Otter.ai* [ott22] and analysed with *Atlas.ti* [atl22].

3.2. Main Study

In the main study, we designed a chatbot that is informed by the findings from the preliminary study. The idea was to conduct a Wizard-of-Oz study to test a voice interface chatbot that can answer questions about a dataset on the national elections covering data from 2008 to 2019. This is publicly available data and gives information about the opinion of the people in the country about politics (see section 3.2.1 *Dataset*). The chatbot had to be able to display data as visualisations, use simple mathematical methods such as statistical methods (minimum, maximum, average) and comparisons, and also provide background information about the data. We conducted 15 interviews with potential first-time voters. We will now describe the components and setup of our study.

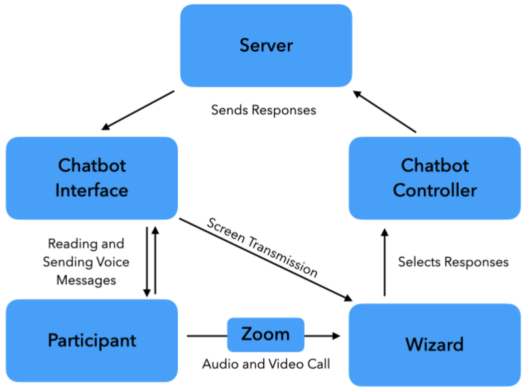


Figure 3.1.: The architecture of the system used for the Wizard of Oz study

3.2.1. Dataset

Research has shown that our personal experiences, educational backgrounds and political attitude strongly influence how we perceive and trust data visualisations [PAE19]. To increase this factor, we decided to use a new dataset about Austrias National Council elections. We believe that this data is of interest to many and that through this study we can promote interest in politics and democracy.

The data is provided by AUTNES [aut22], a social science analysis of the National Council elections. During four National Council elections between 2009 and 2016, mixed-mode panel surveys (via telephone and online) were conducted before and after the election. The voting behaviour of the population was examined, among other things with a focus on the topics of voting behaviour around 'voting at 16', populist attitudes and voting for right-wing populist parties, as well as whether and how media coverage influences attitudes and decisions.

3.2.2. Chatbot

The chatbot consists of a simple web interface with a design based on other well-known messenger services, as can be seen in Figure 1.1. Users can send voice messages with the microphone button on the bottom by holding the button pressed during voice input and releasing it to send the message. These messages are presented as sound waves and since audio is not really recorded (functionality is faked), it is not possible to listen to these messages again. Text input is also not possible. The chatbot's messages, on the other hand, can contain both text and images. The user can also click on pictures in messages to enlarge them. The entire conversation history is saved as a JSON-file [jso22] and can be downloaded by the user using the download button in the upper right corner. This file was the basis for the interview transcripts.

The developed system consists of three parts: Firstly the chatbot interface, through which the participant can read and send messages. Secondly a controller, that allows the wizard to send messages to the user and thirdly a server, which forwards the messages from the controller to the chatbot. As shown in Figure 3.1, the wizard and the participant were connected via a Zoom call [zoo22] during the study with the participant sharing the chatbot interface on their screen. The participant now has the ability to "send" messages, and these messages can be seen by the wizard though screen sharing. The wizard can then use the controller to send messages to the chatbot via the server.

The system is based on the web framework node.js [nod22], using Vue [vue22] and the open source front-end framework BootstrapVue [boo22b] based on Bootstrap [boo22a]. On the server-side, express.js [exp22] was used as a framework for sending messages.

3.2.3. Participants

Participation in the interviews was limited by three factors: participants had to be eligible to vote in the national elections, be so-called "early voters" and not have participated in the preliminary study. In other words, they had to have Austrian citizenship and be between 16 and 25 years old. "Early Voters" refers to young people who vote for the first time. This is particularly interesting as this group is part of the "digital natives" and the use of chatbots and voice interfaces plays a greater role than in the rest of the population. The participants are further identified as (P#).

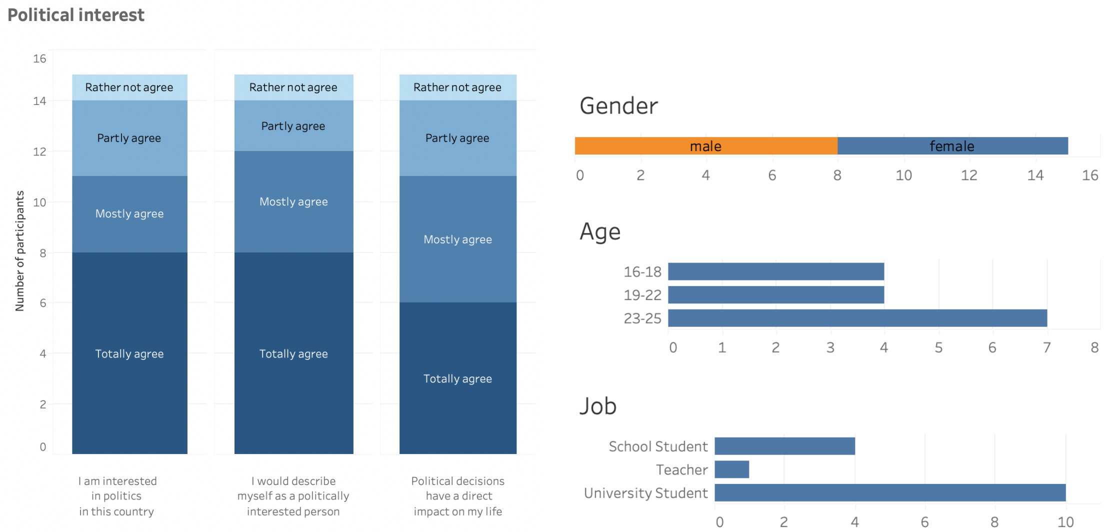
Interviews were conducted with a total of 15 people who were recruited through personal contacts in secondary schools, social media postings and personal conversations. Out of the 15 participants, eight identified themselves as male, seven as female. four people each belong to the age groups 16-18 and 19-22, and seven people were between 23 and 25 years old. Four participants attended school, ten were university students and one person was a teacher (see Figure 3.2b).

Of the 15 participants only two people have never used a voice assistant before. Only one study participant used it daily, one person said they used it once a day or once a week, and six said they had never used it or only used it once a month. In summary, it can be said that most participants had already used voice assistants, but did not use them frequently.

As can be seen in Figure 3.2a, the participants can be described as politically interested. Thus, 11 out of 15 state that they *mostly* or *totally agree* that they are interested in the country's politics, 12 people would *mostly* or *totally agree* with the statement that they see themselves as politically interested persons and 11 people *mostly* or *totally agree* that the decisions of politics have a direct impact on their lives. When asked whether they would vote in the next National Council election, all 15 participants answered with full agreement. The existing political interest of the participants and the selection of the data set resulted in a good symbiosis, resulting in a high level of engagement in the interaction with the chatbot.

3.2.4. Online Questionnaire

Besides the interaction with the chatbot, an online questionnaire was part of the study. Participants were asked whether they had already used voice assistants, which ones they had used and how often, as well as questions about their political interest. More details can



(a) Most participants describe themselves as politically interested people.

(b) The distribution by gender, age and employment of the participants

Figure 3.2.: Description of the participants in the main study.

be found in section *3.2.3 Participants*. Part of the survey was also the result of the tasks given by the chatbot. After completing the study with the chatbot, a System Usability Scale (SUS) [Bro95] was surveyed as well as feedback and demographic information.

3.2.5. Interview Procedure

The study took place in individual interviews using the video conferencing tool Zoom [zoo22]. 15 interviews were conducted, each lasting between 20 and 30 minutes. At the beginning of the interviews, the participants were informed about the study procedure and the voluntary nature of their participation. The interviews consisted of two parts, the interaction with the chatbot and an online survey. The interviews were video and audio recorded in order to prepare transcripts afterwards. The study participants were sent a link to the chatbot. The chatbot introduced itself briefly and explained that it was not yet fully developed, so it sometimes needed some time and could not yet answer all questions. From the chatbot, a link led to an online survey where participants had to provide information about their experience with voice interfaces and their interest in politics. The results of tasks also had to be entered into the survey, as well as feedback and the System Usability Scale at the end of the study.

The chatbot took over the moderation of the study. At the beginning, it asked the participants to say *Hello* and to ask for the *fact of the day* in order to test the voice input. Afterwards, the chatbot led the participants to two pre-selected topics: *Non-voters* and *media consumption and voting decisions*. For both topics, the participants could first ask questions that interested them before they were given tasks that they had to solve with the chatbot. Task 1 on the topic of *non-voters* was to find out which media young and first-time voters consume and whether that differed from the rest of the population. Task 2 on the topic *media consumption and voting decision* was to find out whether the type of media consumption had an effect on voting decisions. All participants were given task 1 to solve, after the initial six participants, we have decided to add a second task to collect even more questions and offer a broader range of topics. Task two was added because it turned out that it would be good to collect even more questions and offer a broader range of topics.

At the end, the participants had the opportunity to try out the chatbot and ask any questions they were interested in. Afterwards, the chatbot asked the following feedback questions:

- Did you enjoy this study?
- How did you like the interaction with me?
- What should I do better?
- Were the answers I gave you helpful?

4. Findings

4.1. Preliminary Study

4.1.1. Group interviews via Zoom

The first group interview did not work as expected for the following reasons: The idea was that in a group interview, the participants would exchange and discuss the tasks with each other and thus additional insights would arise. This was counteracted by the corona-based interview via Zoom. Due to the slight time delay and the absence of gestures, facial expressions and non-verbal communication, the interview felt like three individual interviews at the same time. The goal of gaining new insights through communication among the participants was not achieved. Due to reduced Corona measures it was possible to conduct the second group interview in person on-site.

Through gestures and eye contact, the second group interview could proceed without strict moderation. The participants did not know each other beforehand, but by briefly getting to know each other over a cup of coffee, the artificial group situation of the first interview could be avoided considerably. The difference can be seen in the number of words: In the first interview, 4439 words were spoken, while in the second interview 5210 words were spoken in the same timeframe.

4.1.2. Data Quality Uncertainty

The participants had little to some experience with datasets, but were mostly very critical of the information that was shown. Thus, at the end of the group interview, one participant asked whether the data had been prepared by us:

1. *„I maybe have one more question. Did you arrange these sheets, like this exercise for this project?“* (GI2, Study 1)

What often caused confusion were data entries of 0 or empty entries. Participants asked several times whether these entries meant 0, no information or missing information:

2. *„If it means that there's no information, wouldn't it be easier to understand if there'd be an X instead of a zero?“* (I3, Study 1)

4.1.3. Visualisations

The study participants found it difficult to use the spreadsheet and it's many numbers and to gather information from it. Two participants stated, that it would be easier to read if there was a visualisation. One participant justified this by saying that it is not necessarily the numbers themselves that are relevant, but the impact of the numbers:

3. *„Because I don't have to concentrate on numbers like 1000 or 1600. I see the impact with one view, and it's a little bit irritating the small numbers.“* (I6, Study 1)

When asked what other form of presentation could make the data easier to understand, the other participants uniformly answered that visualisation would be beneficial. One participant argued that he is used to reading graphs and that it is therefore easier for him:

4. *„I mean, I'm always preferring graphs, and some kinds of with graphical visualization. Because I'm just more used to it.“* (GI1, Study 1)

The majority of participants preferred graphs but also understood the relevance of data tables and the importance of individual data entries. A "nice combination" of both would be the best option according to one participant.

It is essential to look at the overwhelming effect of the amount of numbers on the participants. The study shows that users are primarily interested in getting a quick overview of the data and only need to obtain individual numerical values in certain cases. When it comes to developing automatic answers with the help of a chatbot, the possibility of charts as answers should definitely be considered.

4.1.4. Question Categorization

The interviews were analysed for uncertainties, questions, assumptions and interpretations. These passages will be referred to as "questions" in the following sections. Inspired by the categories of Kim, Hoque and Agrawala [KHA20], we then analyse the questions based on two parameters: Can the questions be answered by simply looking them up, calculated by simple statistical methods (minimum, maximum, average) or calculated by applying filters (*low-level*), or are several steps and calculations necessary (*high-level*)? These can be, for example, comparisons of one group with another, correlations of two variables or changes over time. Can the question be answered with the help of the dataset (*inside*) or does it require additional information (*outside*)? *Inside* refers to the data itself, while *outside* contains the metadata, explanations of the variables and additional knowledge questions that are relevant to the topic but not directly present in the dataset, e.g. *"Who is the queen?"* This classification creates 4 possible categories: *inside/low-level, inside/high-level, outside/low-level, outside/high-level.* You can find examples in Table 4.1.

Inside/low-level

This category describes questions that relate to and can also be answered with the data, like the simple lookup question 5. Simple arithmetic operations such as average, minimum and maximum as well as filter operations find a place in this category, e.g. question 6 and 8.

	Inside the dataset	Outside the dataset
Low level	*„Which country has the highest (emissions) in total?“* (GI1, Study 1)	*„The questions that I have are about vocabulary.“* (I3, Study 1)
High level	*„Will Asia or Europe grow more in the next few years?“* (I4, Study 1)	*„I would be interested in how this is calculated?“* (GI1, Study 1)

Table 4.1.: Examples for inside/outside the dataset questions

5. *„I just wanted to note first, well, since when the documentation of this data is going on?“* (I3, Study 1)
6. *„So where does it start?“* (I4, Study 1)
7. *„Well down here it is divided by countries right? Up here, it's like sorted by income groups, geographic regions and development groups.“* (I2, Study 1)
8. *„Which country has the highest (emissions) in total? Which country has the highest per capita? or lowest?“* (GI1, Study 1)

In the beginning of each interview most questions were about the scope of the dataset (e.g. questions 5 and 6). Some questions (e.g. 7) dealt with the understanding of the data, i.e. whether the participants had read and understood the data correctly.

Inside/high-level

The high-level questions often deal with time-dependent trends, whether the data is complete, or how to interpret the data. The dataset on CO2 emissions from fossil fuels raised many questions. In the fossil fuels dataset, with only UK data available in the first few years and countries only being added gradually, many wondered what happened to the missing countries (e.g. question 9). Countries that had zero CO2 emissions from fossil fuels in a given year were not included in the dataset. This led to uncertainty about what was going on with the missing data, e.g. question 10. The most frequently asked question was about the evolution of the data over time (questions 11, 12 and 13). It was often not easy for the users to read these developments from the data, e.g. question 12.

9. *„Either they didn't use fuel at this time or there are no records of it.“* (I1, Study 1)
10. *„Are there all countries included?“* (GI2, Study 1)
11. *„Europe is getting smaller and smaller without migration.“* (I6, Study 1)
12. *"Over the years, over 140 years, it's like it's going down in the 1900s, and on middle 1900s, they just turned the curve upside down. It's going up.“* (I4, Study 1)
13. *„[...] then it says that temperatures are rising.“* (I4, Study 1)

Outside/low-level

Because the participants were confronted with the datasets for the first time, many questions arose about the individual columns, terms and calculations (see questions 14 and 15). Some questions also could not be answered with the metadata but needed additional information (e.g. question 16). Not only did terms come up that seemed unclear, but also concepts, in this case mathematical concepts, that were unclear and relevant to understanding the data (e.g. question 17). But also the way the spreadsheet program works was unclear to many and some wanted to know if there was a way to sort or filter the data (e.g. questions 18 and 19).

14. *„The questions that I have are about vocabulary.“* (I3, Study 1)
15. *„[. . .] what does zero migration mean?“* (I5, Study 1)
16. *„Could you first, like, give me a definition of the term datasets?“* (I3, Study 1)
17. *„The thing is, I don't really get what the mean is, like, what is the mean here?“* (I2, Study 1)
18. *„That's strange, How can you sort from A to Z if there are only numbers?“* (I4, Study 1)
19. *„OK, Can you sort it by land anyhow or by country?“* (I4, Study 1)

Outside/high-level

In the following, questions are addressed that concern uncertainties with the data and the categories, assumptions about the data and that can only be answered with the metadata or additional external information.

Further uncertainties arose with regard to the source, the process of data collection and processing. These do not only deal with definitions and general concepts, but go deeper and question the data, the categorisations, the measurement techniques, the processing steps and political questions (e.g. questions 20 & 21). Many questions also arose concerning categories. The participant who asked question 22, for example, what the categories meant, how the projections on which the categories are based are calculated or what a category contains. Or it does not seem clear what a category contains. A keyword describes the category, but without additional information one can only assume what exactly is meant by it (e.g. 23). Then there are uncertainties and political questions regarding the categorisations. The boundaries of categories can change. Political boundaries have often been redrawn, but the categorisation and development of individual countries in the world is also changing (e.g. questions 24 and 25).

20. *„Well, I would be interested in how this is calculated? So how the emissions are calculated.“* (GI2, Study 1)

	Messages Total	Messages User	Messages Chatbot	Questions	Duration
Average	82.33	26.27	56.07	10.60	15 min 19 s
Median	79.00	25.00	53.00	11.00	15 min 31 s
Sum	1235	394	841	159	229 min 58 s

Table 4.2.: Overview of the messages sent during the main study

21. *„I don't know how co2 emissions can be measured, which were 200 years before.“* (I6, Study 1)
22. *„I would be interested in how they get to these different variants for the prognosis, the prospects, because this is something I am just not familiar with.“* (I5, Study 1)
23. *„I wonder if the co2 emissions come from the use of the fuels, also the production maybe? Because production might be in a different country. So I wonder if it's included or not?“* (GI2, Study 1)
24. *„Obviously, borders change. So that's a complication.“* (GI2, Study 1)
25. *„I would want to know, how they divide more developed and less developed regions. So what is development in that context?“* (GI2, Study 1)

4.2. Main Study

The evaluation of the study is based on three parts: The transcript of the interviews, the data of the online survey and observations and notes of the study leader. The interview transcripts were analysed with Atlas.ti [atl22], the online survey with Tableau [tab22].

4.2.1. Questions

The total of 15 interactions with the chatbot lasted 3 hours and 49 minutes, with an average of 15 minutes and 19 seconds (see Table 4.2). A total of 1235 (82 on average) messages were sent, of which 394 (26 on average) were sent by the users and 841 (56 on average) by the chatbot. Of the 394 user messages, 159 (11 on average) were manually marked as questions or requests for information. These questions were then divided into the same categories as in section 4.1.4: *Inside/low-level, inside/high-level, outside/low-level, outside/high-level.* It should be noted that information goals were defined in the course of the semi-structured interview, which is why the number of questions in the categories has only limited significance.

Inside/Low Level

The questions in this category are primarily about the election results of the last elections. Common questions were the number of eligible voters, how many of them (did not) vote and what the election results were.

26. *"How many people voted in the last election?"* (P3, Study 2)
27. *"How many eligible voters did not vote in the last National Council election?"* (P3, Study 2)
28. *"What percentage did not vote in the last election?"* (P9, Study 2)
29. *"Who was elected in the last National Council election?"* (P4, Study 2)

These questions (26 - 29) require a filter operation (last election - election year 2019) and lookup. The answers could be numbers or charts, whereby the second question could be saved if the first question was given context in the form of a chart. Question 29 suggests answering with a bar chart of the election results because this is a familiar form of visualisation.

30. *"What is the turnout at the age of 16?"* (P13, Study 2)
31. *"Why do some people not vote?"* (P8, Study 2)
32. *"Which media do non-voters consume?"* (P0, Study 2)

Question 30 requires a further filter operation and implicitly refers to the previous question, which means that the year must also be correctly assigned by the chatbot, in this case 2019. Question 31 also does not define which year it is, which is why the question must either be set to all years, the last year or to one defined by the history of the conversation. Some questions are very imprecise and require interpretation. Question 32 does not define whether media means media types, specific magazines or social media.

Inside/High Level

A large part of the questions in this category are due to the semi-structured interview. In the task on *media consumption and voting decisions*, the users had to establish the connection between voting behaviour and the consumption of certain media. Accordingly, many questions are similar to the following:

33. *"What influence does social media consumption have on voting behaviour?"* (P0, Study 2)
34. *"How is social media and voting behaviour related?"* (P7, Study 2)
35. *"Is there a connection between media consumption and voting decisions, i.e. which medium is consumed and which party is voted for?"* (P1, Study 2)
36. *"Does media consumption influence voting decisions?"* (P9, Study 2)

The influence of one parameter on others was frequently queried. For example not only was the influence of media on voting decision queried, as is visible in questions 33 to 36, but also other correlations such as by income (37) and age (38):

37. *"How do non-voters differ according to income?* (P5, Study 2)
38. *"It would be interesting, if there is a specific age group, in which are more non-voters."* (P1, Study 2)

Questions about comparisons between different groups or disjunctive parameters dominate this category. For example, in question 39 and 40 the groups were divided by gender, in 41 a distinction is made by age, in 42 by participation in elections and in 43 by different years.

39. *"Do women vote differently than men?"* (P8, Study 2)
40. *"Do more women or more men vote?"* (P4, Study 2)
41. *"Do more old people vote than young people?"* (P8, Study 2)
42. *"Do non-voters inform themselves more often via the internet than voters?"* (P3, Study 2)
43. *"How have the results changed from 2008 to 2019?"* (P7, Study 2)

Questions 39 to 42 could be answered yes/no, which would be a correct answer but would not satisfy the user's need for information. The information interest here is more about the "*how*" and not the "*if*". The differences in these cases can best be shown with the help of charts, although it is not an easy task to show the difference. In the present study, therefore, multiple charts were chosen to compare the two groups.

Outside/Low Level

The questions in this category are roughly divided into three parts: Metadata questions, functionality questions and fact-finding questions.

44. *"From which years can the data be retrieved?"* (P7, Study 2)
45. *"What questions can you answer?"* (P11, Study 2)
46. *"It's a little small, can you make it bigger?"* (P6, Study 2)
47. *"Are Telegram chats a type of media?"* (P13, Study 2)
48. *"How many people live in Austria?"* (P4, Study 2)

In 44 to 46, the chatbot was asked questions about its functionality, in 47 a clarification of media types was requested and in 48 a simple fact-finding question was asked that is relevant in this context but cannot be answered using the available data alone.

Outside/High Level

The questions in this category are the most difficult when it comes to answering them automatically, but at the same time they occur the least.

49. *"What is the difference between Did-Not-Vote and Refused?"* (P0, Study 2)
50. *"Do the numbers in the survey match the actual numbers?"* (P13, Study 2)
51. *"I have a question about the chart: Why are there also people under the voting age of 16?"* (P14, Study 2)
52. *"Why then does the chart start at 10 years and not at 16 years?"* (P14, Study 2)

In question number 49 a user asks about the difference between two categories of voters. Although this can be answered by humans: *Did-Not-Vote* means a person that stated it did not vote, *Refused* means a person refused to give an answer to the question. It is not easy to automatically output the difference. The alternative would be to display both definitions side by side. Question 50 asks for a comparison of the poll data with the election results. This comparison can be implemented, but an automatic linking of the data is difficult. Questions 51 and 52 ask for a visual element in the chart and would like an interpretation from the chatbot. However, these are so specific that this can hardly be implemented automatically.

4.2.2. Referring to the past conversation

The frequency with which the participants referred to the conversation history was very noticeable. A total of 29 times they referred to a variable from the previous message with a question. Here, for example, P2 asks a question: *"Which media do non-voters consume?"* (P2, Study 2) and then in the following question implicitly refers to the group of non-voters with the word *"they"* and makes a comparison with another group: *"And do they differ from the rest of the population?"* (P2, Study 2)

Not only groups were referred to, but also other variables such as election years (*"And in the years before that?"* (P7, Study 2)), media type (*"Yes please show me these media types in detail too."* (P11, Study 2)), Electoral behaviour (*"How does it look with the newspaper Press?"* (P3, Study 2)) and visualisation type (*"I would like an example on social media."* (P9, Study 2)).

It also happened that a participant even referred to two variables explicitly, the party and the year:

53. Chatbot: *"The Liberals entered the National Council for the first time in 2013."*
    P11: *"How many votes did they get back then?"* (P11, Study 2)

A speciality among the questions referring to past messages are questions referring to features of previous charts. P14, for example, refers directly to the chart and asks the chatbot why, despite the minimum voting age of 16, the binned age groups start at ten years: *"I have a question about the chart: Why are there also people below the voting age of 16?"* (P14, Study 2). P11 on the other hand asks after receiving a message with a bar chart that shows the percentages of (non-)voters: *"How many people is that in absolute numbers?"* (P11, Study 2).

One participant tried out whether the chatbot is capable of "remembering" past messages and reacting appropriately to requests. Two stated that this is an important feature to make the chatbot easy to use, because constantly and precisely reformulating queries is unmotivating.

4.2.3. Answers

The chatbot's answers were largely pre-written and only had to be sent with one click in the dashboard. For unplanned questions, there was also the option of typing text quickly for the wizard. The duration of the answer is relevant for enthusiasm to continue while using the tool. The fact that the chatbot was not yet fully developed and was therefore still a little slow was pointed out to the participants in the welcome part of the chatbot. Nevertheless, the average response time was only 6.4 seconds, with a median of 4.4 seconds. As shown in section 4.2.5, the participants were satisfied with the interaction with the chatbot, only one participant would have liked to have been given feedback in case of longer loading times, whether an answer is still coming or whether the process has already been aborted.

In the evaluation of the chatbot at the end, all 15 stated that the answers given were helpful. During the interview, participants were frequently asked to rate answers, with only 3 times out of 42, the answers were rated as not applicable, in one case a 3 was given, in eight cases a 4, and in nine cases a 5 (on a scale from 0 not helpful to 5 very helpful). In the other cases, the answers were verbally assessed as good.

Overall, the participants probably found the answers authentic because 12 out of 15 participants believed that it was a functioning chatbot until the end. Due to technical problems one person was sure right at the beginning that the chatbot did not work completely automatically. Two participants said that during the interview they suspected that a human was behind it because of typing errors etc.

4.2.4. Objective and neutral answers

A big challenge in building a chatbot is to present the survey data as politically neutral as possible. This was also tested by a participant who asked: *"Bruno, do you have a favourite party?"* This was followed up with the sentence, *"That means you are politically neutral?"* and followed up one more time with, *"That means you would describe yourself as objective?"* (all P6, Study 2) This is a political science question that cannot be answered adequately here. If the goal of such a chatbot is to promote interest in politics and democracy, then the following question should definitely be able to be answered: *"What could be done to increase election participation?"* (P12, Study 2)

4.2.5. Voice Interface Chatbot

The interface and interaction with the chatbot was rated well by the users. On average, the bot achieved a System Usability Scale [Bro95] score of 84%, which is an excellent value. On a scale of 1-5, where 1 is the best and 5 the worst, the chatbot achieved a score of 1.8, with the worst score being a 4 only once.

All 15 participants said they had fun trying out the chatbot, with one person adding that it was *at least very interesting*. 13 people said that they liked the way they interacted, 2 said that they found this type of interaction *unusual* and did not like it so much. Two persons also stated that they did not like to speak in standard language and preferred to use dialect. It must be said that the chatbot instructor was able to understand dialect, but the participants assumed that they had to speak standard language without trying. This probably resulted from previous experiences with other existing language interfaces.

In developing the chatbot, we have tried to use informal language because we believe it encourages motivation to interact with the chatbot. This cannot be proven with the data, but one person stated that they found the chatbot very *friendly* and *easily accessible* and one person found it very *entertaining*.

The interaction with the chatbot was rated positively, but one user wished for a function similar to *Hey Siri* or *Ok Google* to initiate a question. Another person suggested activating and deactivating the record button by clicking it instead of holding it down. The different response times caused some irritation so users would have liked feedback if an answer took a little longer.

4.3. Comparison

As shown in the preliminary study it is difficult for non-data experts to become familiar with a dataset for the first time and it takes a while for people to get an overview. Accordingly, most questions refer to the understanding of the dataset and participants only ask questions about the data itself afterwards. On the one hand, this is logical, but also determined by the study design, as the time of the interviews was chosen to be relatively short.

This is also the reason why a lot of *outside questions* were asked in the first study, because the participants had to get to know the dataset first. In comparison, very few *outside questions* were asked in the main study. One reason for this is that the chatbot introduced itself at the beginning and gave a rough overview of its functionality. A second reason is that the selected dataset with the national elections is a topic that concerns everyone, as all participants are eligible to vote. In addition, data on elections is shown very often in the media, which means that people also have more of an idea of what kind of questions they can ask. In the preliminary study the *inside questions* refer on the *low-level side* mainly to mathematical key figures such as minimum, maximum and average. In the main study most of the questions were fact-finding questions about the election results and the eligible voters. Often these questions are connected with the filtering of individual variables. This can also be explained by the fact that in the first study all data were visible, whereas in the second study it was necessary to ask for them first.

On the *high-level* side, in the preliminary study many questions revolved around data completeness, data from one's own country and, above all, trends in time series. In contrast, in the main study, many questions were asked about the influence of one variable on the other or about the difference between different groups. Time trends were not as relevant here.

5. Discussion

5.1. Types of Questions

In this paper, we presented a categorisation of questions asked by users to voice interface chatbots about data from national elections. This is meant to be an aid towards the development of targeted answers and meaningful data visualisations to be used in such systems. Others such as Setlur and Tory [ST22] have developed a different taxonomy for questions that is more focused on the process of natural language processing. In their study these are listed separately for different prototypes testing different interaction modalities: Text and images (using Slack [sla22]), Voice and Images (Echo Show) and voice only (using Echo [ama22]).

In the following, we transfer the questions derived from our main study into the categorisation of Setlur and Tory and compare them with the results of the Echo Show, as it has the greatest similarities with our chatbot. The differences in the categories *Fact-finding* (Echo: 46 % vs. ours: 28 %), *Filtering* (Echo: 7 % vs. ours: 31 %) and *Grouping* (Echo: 18 % vs. ours: 0 %) are striking, while *Comparison* (Echo: 25 % vs. ours: 27 %) is very similar. However, it must be said here that the figures are only of limited significance as this likely depends on the study setup and the number of total questions is small. This also applies to the number of questions that make reference to previous questions (Echo: 38 % vs. ours: 20 %).

The impact of the study set-up on the questions can also be seen when comparing our findings with those of Amar, Eagan and Stasko [AES05]. They describe ten low-level analysis tasks that largely capture people's activities while employing information visualisation tools. These were extracted from nearly 200 data analysis questions provided by students to commercial visualisation tools. These tasks are as follows: *retrieve value, filter, compute derived value, find extremum, sort, determine range, characterize distribution, find anomalies, cluster and correlate.* Since one of the tasks of the Wizard of Oz study was to find out the relationship between media consumption and party preference, *correlation* and *filter* tasks occurred very frequently, others like *cluster, find anomalies, range determination* did not occur at all in our case. If one assigns the tasks to the question categories we described, low-level questions primarily use the tasks: *retrieve value, filter, compute derived value, find extremum* and *sort.* For high-level questions, further tasks are needed: *characterize distribution, find anomalies, cluster* and *correlate.*

5.2. Requirements for a voice interface chatbot for data exploration of national election data

The following discussion points are intended to provide an overview of what a voice interface chatbot could look like based on our findings, that can answer questions about national elections. We discuss what features are likely to increase user engagement amongst first and young voters, according to our data.

5.2.1. Answers

The choice of a classic interface with voice input and text/chart output proved to be reasonable. The possibility to look up past queries via the chat history reduces burden on mental memory and makes it easier to compare several pieces of information. This is a big advantage over interfaces that have speech as input and output, especially for data exploration scenarios. Other interfaces have used this setup in a similar way [FCM$^+$18, KR18, ZM20, BS17].

This chosen setup also makes it possible to present data visualisations to the user. While these were appreciated by the majority of participants, some were overwhelmed with the presentation, underlining the *split in people's preference for visualisations* described by Hearst and Tory [HT19] who found out that 40 % prefer not to see charts and graphs in the context of a conversational interface, while those who prefer to see charts wanted to see additional supporting context. While well-known graphs such as bar charts showing election results did not raise any complaints, stacked bar charts on sources of political information were sometimes seen as too complicated. In the feedback one person asked for help with interpretation of the graphs. This is also in line with the research of Kassel and Rohs [KR19], who showed that people with different statistical knowledge have different expectations of chatbot responses. They have shown that preference in the communication of facts is based on both the characteristics of the user and the desired analysis. This also influences the assessment of the appropriateness of the tool. Furthermore, we have shown that it is important not only to answer the question, but also to give context to it. This subsequently influences the engagement and motivation to use the tool. For example, when asked about the number of non-voters, the chatbot showed a bar chart with the number of non-voters and voters, which put the number in context and makes it understandable. The combination of a short text response with the result and an additional visualisation for context has also been found to work well. Hearst and Tory [HT19] and Zhi [MSZJ17] have also presented similar findings.

5.2.2. Humanlike Chatbot

The participants in our study enjoyed the interaction with the chatbot. One person said that although they had a little difficulty answering the two tasks, they still had fun trying out the chatbot. A second stated: "The system is very entertaining and also informative." With a few exceptions, of people who find it strange to talk to computers or who would rather speak dialect, the "personality" of the chatbot was perceived as positive and as "nice". It was described as easy-going in its wording and as humorous. To test the interaction participants were prompted to ask for the fact of the day, which was very informal and potentially "nerdy". When asked if they would like to hear another fact, 14 out of 15 answered affirmatively. The nerdy answers packed with a bit of wit are in line with the observations of [MMM+17], who recommend a "productivity oriented nerd with wit". That human traits are important for a chatbot is also shown by Chaves and Gerosa [CG21], and Mygland et al. [MSPV21].

5.2.3. Referring back

The importance of a memory, in other words the relevance of reference to past messages, has already been shown by the tools Iris [FCM+18], Eviza [SBT+16] and Evizeon [HSTD18]. Chaves and Gerosa [CG21] have also described this as a relevant human characteristic that a chatbot needs. However, our study has shown that for chatbots tailored to data engagement it is not only important to refer to text, but that it is also necessary to be able to refer to visual elements of charts. Therefore, it is necessary not only to be able to produce charts, but also to be able to discuss their elements, axes, colours etc. An automatic chart-question-answering pipeline has already been developed by Kim, Hoque and Agrawala [KHA20]. Their answers reach an overall accuracy of 51 % on a corpus of real-world charts with human-generated questions. We can imagine that with the specific, interactive charts generated by the chatbot it could be even easier to answer questions and achieve a higher level of accuracy. How this can be embedded in the context of chatbots and data-specific user needs is subject to further research.

5.2.4. Interactive visualisation

In section 5.2.1 *Answers* it was already mentioned that a short textual description of visualisations is relevant for a fast understanding of the answer. That a coupling between text and visualisation is helpful as an interactive support for reading has been shown by Metoyer et al. [MSZJ17].

Setlur and Tory [ST22] demonstrated the importance of interaction with data visualisations in the context of chatbots with voice interface prototypes. However, in their prototype the interaction is limited to buttons that can be used to filter the data by different groups. It would be relevant that one can easily swap variables, set filters, get tooltips and perform highlighting. Questions of which visualisation requires which type of interaction in the context of data chatbots must first be investigated in further studies.

5.3. Future work

Voice interfaces are a promising medium for human-data interaction, for which several interesting research directions would be interesting:

5.3.1. Collaborative exploration of data via voice interface

In this paper, the collaborative exploration of data via voice was tested in the first part, but not investigated any further. Other messengers offer group chats. This concept could be transferred to voice interface chatbots. It would be interesting to see how interaction patterns differ when using voice interfaces in group settings.

5.3.2. Influence of informal language on user satisfaction

Our chatbot used very informal language and also emojis to communicate with the users. The idea was that this would improve engagement and people would have more fun using such a tool. An interesting research direction would be to investigate whether and how the user experience changes when using formal or informal language in voice interface chatbots. Possible side effects here would be that trust in the answers could suffer under too informal language.

5.3.3. Interactive visualisations in the voice interface chatbot context

As described in section 4.2.2, it is relevant for voice interface chatbots to include conversation history in the evaluation of requests. Often, these references refer to elements or variables in visualisations. Since the recognition of implicit references is not easy, it would be interesting to see if interactive charts are a way to reduce such requests. If you can easily read individual values from a chart, swap variables and get the desired results with just a few clicks, this could further improve engagement.

6. Conclusion

As the relevance of data is increasing and the use of voice interfaces is becoming more common, it becomes apparent that current commercial tools are still only capable of performing rather basic tasks. In the specific context of voice interfaces for exploring datasets, we have shown which questions young voters ask chatbots. The developed categorisation of questions provides a basis for future tools and gives hints for which capabilities should be implemented in the future to increase user satisfaction. We also showed that it is important for user engagement that the chatbot can answer references to text and visualisations from the past conversation. A human communication style with a bit of wit motivates users to continue asking questions. Interactive charts that provide even more context to the questions asked can encourage further interaction. Future tools need to take steps in this direction so that voice interfaces can achieve an edge over traditional data exploration tools and establish a satisfied user base.

Bibliography

[AES05] Robert A. Amar, James Eagan, and John T. Stasko. Low-level components of analytic activity in information visualization. In John T. Stasko and Matthew O. Ward, editors, *IEEE Symposium on Information Visualization (InfoVis 2005), 23-25 October 2005, Minneapolis, MN, USA*, pages 111–117. IEEE Computer Society, 2005.

[ale22] Amazon alexa. `https://developer.amazon.com/alexa`, 2022.

[ama22] Amazon echo. `https://www.amazon.com/smart-home-devices/b?ie=UTF8&node=9818047011`, 2022.

[atl22] Atlas.ti. `https://atlasti.com`, 2022.

[aut22] Autnes - the demand side. `https://www.autnes.at/home/demand-side/`, 2022.

[boo22a] Bootstrap. `https://getbootstrap.com`, 2022.

[boo22b] Bootstrapvue. `https://bootstrap-vue.org`, 2022.

[Bro95] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.

[BS17] Stefan Bieliauskas and Andreas Schreiber. A conversational user interface for software visualization. In *IEEE Working Conference on Software Visualization, VISSOFT 2017, Shanghai, China, September 18-19, 2017*, pages 139–143. IEEE, 2017.

[cdi22] Cdiac - carbon dioxide information analysis center. `https://cdiac.ess-dive.lbl.gov`, 2022.

[CfRI18] European Commission, Directorate-General for Research, and Innovation. *Turning FAIR into reality : final report and action plan from the European Commission expert group on FAIR data*. Publications Office, 2018.

[CG21] Ana Paula Chaves and Marco Aurélio Gerosa. How should my chatbot interact? A survey on social characteristics in human-chatbot interaction design. *Int. J. Hum. Comput. Interact.*, 37(8):729–758, 2021.

[CO222] Data hub - co2 emissions from fossil fuels since 1751, by nation. `https://datahub.io/core/co2-fossil-by-nation`, 2022.

[DR20] Ferliana Dwitama and Andre Rusli. User stories collection via interactive chatbot to support requirements gathering. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 18:870, 04 2020.

[exp22] Express. `https://expressjs.com`, 2022.

[FCM+18] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S. Bernstein. Iris: A conversational agent for complex tasks. In Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox, editors, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 473. ACM, 2018.

[gca22] Gcag - the global component of climate at a glance. `https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/global/background`, 2022.

[GDA+15] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In Celine Latulipe, Bjoern Hartmann, and Tovi Grossman, editors, *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology, UIST 2015, Charlotte, NC, USA, November 8-11, 2015*, pages 489–500. ACM, 2015.

[GIS22] Gistemp - giss surface temperature analysis by nasa. `https://data.giss.nasa.gov/gistemp/`, 2022.

[glo22] Data hub - global temperature time series. `https://datahub.io/core/global-temp`, 2022.

[goo22a] Google assistant. `https://assistant.google.com/`, 2022.

[goo22b] Google nest mini. `https://store.google.com/at/product/google_nest_mini`, 2022.

[hom22] Apple homepod mini. `https://www.apple.com/homepod-mini/`, 2022.

[HSTD18] Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. Applying pragmatics principles for interaction with visual analytics. *IEEE Trans. Vis. Comput. Graph.*, 24(1):309–318, 2018.

[HT19] Marti Hearst and Melanie Tory. Would you like A chart with that? incorporating visualizations into conversational interfaces. In *30th IEEE Visualization Conference, IEEE VIS 2019 - Short Papers, Vancouver, BC, Canada, October 20-25, 2019*, pages 36–40. IEEE, 2019.

[HTS19] Marti Hearst, Melanie Tory, and Vidya Setlur. Toward interface defaults for vague modifiers in natural language interfaces for visual analysis. In *30th IEEE Visualization Conference, IEEE VIS 2019 - Short Papers, Vancouver, BC, Canada, October 20-25, 2019*, pages 21–25. IEEE, 2019.

[jso22] Json - javascript object notation. `https://www.json.org/`, 2022.

[KGGS21] Laura Koesten, Kathleen Gregory, Paul Groth, and Elena Simperl. Talking datasets - understanding data sensemaking behaviours. *Int. J. Hum. Comput. Stud.*, 146:102562, 2021.

[KGKW05] Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. A conversational agent as museum guide - design and evaluation of a real-world application. In Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist, editors, *Intelligent Virtual Agents, 5th International Working Conference, IVA 2005, Kos, Greece, September 12-14, 2005, Proceedings*, volume 3661 of *Lecture Notes in Computer Science*, pages 329–343. Springer, 2005.

[KHA20] Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. Answering questions about charts and generating visual explanations. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM, 2020.

[KR18] Jan-Frederik Kassel and Michael Rohs. Valletto: A multimodal interface for ubiquitous visual analytics. In Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox, editors, *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. ACM, 2018.

[KR19] Jan-Frederik Kassel and Michael Rohs. Talk to me intelligibly: Investigating an answer space to match the user's language in visual analysis. In Steve Harrison, Shaowen Bardzell, Carman Neustaedter, and Deborah G. Tatar, editors, *Proceedings of the 2019 on Designing Interactive Systems Conference, DIS 2019, San Diego, CA, USA, June 23-28, 2019*, pages 1517–1529. ACM, 2019.

[KSV19] Sophia Keyner, Vadim Savenkov, and Svitlana Vakulenko. Open data chatbot. In Pascal Hitzler, Sabrina Kirrane, Olaf Hartig, Victor de Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, Karl Hammar, Nelia Lasierra, Steffen Stadtmüller, Katja Hose, and Ruben Verborgh, editors, *The Semantic Web: ESWC 2019 Satellite Events - ESWC 2019 Satellite Events, Portorož, Slovenia, June 2-6, 2019, Revised Selected Papers*, volume 11762 of *Lecture Notes in Computer Science*, pages 111–115. Springer, 2019.

[LS16] Ewa Luger and Abigail Sellen. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade, editors, *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 5286–5297. ACM, 2016.

[MMM+17] Indrani Medhi-Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'Neill. How do you want your chatbot? an exploratory wizard-of-oz study with young, urban indians. In Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler, editors, *Human-Computer Interaction - INTERACT 2017 - 16th IFIP TC 13 International Conference, Mumbai, India, September 25-29, 2017, Proceedings, Part I*, volume 10513 of *Lecture Notes in Computer Science*, pages 441–459. Springer, 2017.

[MSPV21] Morten Johan Mygland, Morten Schibbye, Ilias O. Pappas, and Polyxeni Vassilakopoulou. Affordances in human-chatbot interaction: A review of the literature. In Denis Dennehy, Anastasia Griva, Nancy Pouloudi, Yogesh K. Dwivedi, Ilias O. Pappas, and Matti Mäntymäki, editors, *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society - 20th IFIP WG 6.11 Conference on e-Business, e-Services and e-Society, I3E 2021, Galway, Ireland, September 1-3, 2021, Proceedings*, volume 12896 of *Lecture Notes in Computer Science*, pages 3–17. Springer, 2021.

[MSZJ17] Ronald Metoyer, Walter Scheirer, Qiyu Zhi, and Bart Janczuk. Coupling story to visualization: Using textual analysis as a bridge between data and interpretation. 12 2017.

[nod22] Node.js®. `https://nodejs.org/en/`, 2022.

[NSV17] Sebastian Neumaier, Vadim Savenkov, and Svitlana Vakulenko. Talking open data. In Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Lawrynowicz, Fabio Ciravegna, and Olaf Hartig, editors, *The Semantic Web: ESWC 2017 Satellite Events - ESWC 2017 Satellite Events, Portorož, Slovenia, May 28 - June 1, 2017, Revised Selected Papers*, volume 10577 of *Lecture Notes in Computer Science*, pages 132–136. Springer, 2017.

[ott22] Otter.ai. `https://otter.ai`, 2022.

[PAE19] Evan M. Peck, Sofia E. Ayuso, and Omar El-Etr. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos, editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 244. ACM, 2019.

[RBH22] Christine Rzepka, Benedikt Berger, and Thomas Hess. Voice assistant vs. chatbot - examining the fit between conversational agents' interaction modalities and information search tasks. *Inf. Syst. Frontiers*, 24(3):839–856, 2022.

[RCB21] Amon Rapp, Lorenzo Curti, and Arianna Boldi. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *Int. J. Hum. Comput. Stud.*, 151:102630, 2021.

[SBT+16] Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. Eviza: A natural language interface for visual analysis. In Jun Rekimoto, Takeo Igarashi, Jacob O. Wobbrock, and Daniel Avrahami, editors, *Proceedings of the 29th Annual Symposium on User Interface Software and Technology, UIST 2016, Tokyo, Japan, October 16-19, 2016*, pages 365–377. ACM, 2016.

[sir22] Apple siri. `https://www.apple.com/siri/`, 2022.

[sla22] Slack. `https://slack.com/`, 2022.

[SS18] Arjun Srinivasan and John T. Stasko. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE Trans. Vis. Comput. Graph.*, 24(1):511–521, 2018.

[ST22] Vidya Setlur and Melanie Tory. How do you converse with an analytical chatbot? revisiting gricean maxims for designing analytical conversational behavior. *CoRR*, abs/2203.08420, 2022.

[tab22] Tableau®. `https://www.tableau.com`, 2022.

[vue22] Vue.js. `https://vuejs.org`, 2022.

[wor22] United nations - world population prospects 2021. `https://population.un.org/wpp/`, 2022.

[ZM20] Qiyu Zhi and Ronald A. Metoyer. Gamebot: A visualization-augmented chatbot for sports game. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, CHI 2020, Honolulu, HI, USA, April 25-30, 2020*, pages 1–7. ACM, 2020.

[zoo22] Zoom. `https://zoom.us`, 2022.

A. Appendix

A.1. Code Repository

The code for the chatbot can be found here:
`https://github.com/twentythreeme/ChattingAboutData`

A.2. Interview Scripts

On the following pages you will find the scripts for the preliminary study as well as for the main study.

Interview Scripts

Preliminary Study

Supplemental Material to:
CHATTING ABOUT DATA - INTERACTING WITH VOICE INTERFACES TO ENGAGE WITH ELECTION PANEL DATA

For the one person interviews and the group interviews slightly different scripts and tasks were used.

Single Interviews

REVIEW CONSENT FORM

Start Recording

INTRODUCTION

Thank you for participating in this study. It is about what questions people ask if they are getting confronted with datasets. It will last about 30 minutes, during which you will be shown two datasets and asked small questions and hypothetical scenarios. There is no right or wrong.
You can cancel the study at any time without giving a reason. If you have any questions or feel uncomfortable or unsure, please talk to me. If a question is too difficult or incomprehensible, please ask for clarification.

EXPERIENCE WITH DATA

Do you have some experiences with datasets? Have you ever worked with spreadsheets? If yes, what kind of data have you worked with?

PRESENTATION OF FIRST DATASET

In the study three datasets were used, but only two per interview. They were used in rotation, which means that each dataset was used 4 times in the 6 interviews. The following general questions are used for all datasets.

Ask for screen sharing and send link to dataset

Now i present to you the first dataset. At first get a bit familiar with the data, and please ask questions, if some things are not clear to you. After that, I will have some questions for you.

General Questions
- Are there some questions regarding the different columns/data entries?
- What would you like to know from/about this dataset?
- Would you like to have some information which isn't contained in this dataset?
- Would you like to extract some aggregated information from this data?
- If you could ask everything about this dataset, what would you like to know?
- Is there anything missing that you would like to know about the dataset in order to use it yourself?
- Data quality: How do you know that this data is „good“ data?

General Tasks
- Imagine your are a journalist for a newspaper and you have to summarize the findings of this data into a headline. What would it be?

GLOBAL CO2 EMISSION - DATASET

Questions
- In this dataset are cells with 0 and cells with an empty body. What do you think is the difference between them?
- What do you notice when you look at the "Countries" column?

Tasks
- Imagine you are doing some research on this data and you want to know on which continent of the world the total CO2 emission/capita are the highest. What steps would you take to get to this information?
- Please look at column bunker fuels. Could you please describe this column and what information is represented in these numbers.

GLOBAL TEMPERATURE TIME SERIES - DATASET

Questions
- Why are there two entries per date?
- Don't look into detail, just say what you think: How is the global temperature developing in the last 20 years?
- Look at the column „Mean“. What does the data represent in this column?
- Look at the two data sources. Can you find some huge differences between those sources?

Tasks
- Don't look into detail, just say what you think: How is the global temperature developing in the last 20 years?
- If you look at the degree column. How would you summarize these data in 2-3 sentences?

WORLD POPULATION GROWTH - DATASET

Questions

- Did you find some countries, which population will decrease in the next years according to this prognosis?
- Which categorization of countries can you find in this dataset?
- If you had to make a one-sentence summary of this dataset, what would it be?

Tasks

- Imagine you are politician in the greek parliament and you want to know how the greek population will develop in the next 20-30 years. It is important for you to plan capacity for schools, hospitals and other public institutions. Which steps would you take to get information which could help you

REFLECTION

- How difficult was it for you to read information out of the datasets?
- Is there anything else what you like to know about this dataset?

DEBRIEF

Is there anything else you would like to add?

Thank you for taking the time. We very much appreciate it! If you would like to follow-up or know more about the results please feel free to send an email.

Group Interviews

REVIEW CONSENT FORM

Start Recording

INTRODUCTION

Hello and thank you all for being here and for participating in this study. It is about what questions people ask if they are getting confronted with datasets. It will last about 45 minutes, during which you will be shown two datasets and asked small questions and hypothetical scenarios. You are asked to talk and work with each other.
It is important that there is no right or wrong. You can cancel the study at any time without giving a reason. If you have any questions or feel uncomfortable or unsure, please talk to me. If a question is too difficult or incomprehensible, please ask for clarification.

INTRODUCTION OF PARTICIPANTS

But before we start I am asking you to briefly introduce yourselves. I suggest that we keep it short and follow the following pattern: Name, Place of residence, Occupation (Study, Job, etc.)

GLOBAL CO2 EMISSIONS - DATASET

Optional Questions
- Are there some questions regarding the different columns/data entries?
- If you could ask everything about this dataset, what would you like to know to make sense of it?
- Data Quality: How do you know that this data is „good“ data?
- Do you think this dataset is trustworthy? What would make you doubt?
- What was the first thing you looked at, when you saw the dataset?
- Is there something you found interesting?
- What would you like to know from/about this dataset?
- What additional information would you like to have/would you need?

Obligatory Questions
- From this data, could you think of how the CO2 emission will develop in the next 20 years? What factors could influence this?
- How should this data be presented so that it's easier for you to read it?

Tasks
- Imagine you as a group are part of an environmental protection group and got this spreadsheet. You are looking for some data that underlines your call for better environmental protection. What arguments could you extract from this source maybe?
- The other way around: What if you are a denier of climate change, what arguments could you extract from this data that support your ideas?
- Imagine you all are working at some environmental protection group and you are searching for answers of the development of the CO2 emissions in the last 20 years. What steps would you need to take to get such an information? Look at the dataset and find the most climate-friendly country. Decide for one and argue why you choose it.
- Discuss: Which of the fossil fuels does have the biggest impact on the CO2 emissions?

WORLD POPULATION GROWTH - DATASET

Optional Questions
- Are there some questions regarding the different columns/data entries?
- If you could ask everything about this dataset, what would you like to know?
- Data Quality: How do you know that this data is „good“ data?
- Do you think this dataset is trustworthy? What would make you doubt?
- What was the first thing you looked at, when you saw the dataset?
- Is there something you found interesting?

- What would you like to know from/about this dataset?

Obligatory Questions
- If you had to make a one-sentence summary of this dataset, what would it be?

Tasks
- Imagine you are a group of officials at the EU and you are planning the budget for the educational sector (like schools, university) for the EU-countries in the next 20 years. What information of this spreadsheet could help you planning?
- If somebody is asking you: Which is the country with the largest expected increase in population? How would you answer this question and which additional information would you need?
- If you look at Austria, can you find arguments that the population will grow/ decrease in the next 80 years? And why?
- Look at the dataset and try to find a continent, which will have the most surprisingly changes in the population in the next few years. Choose one and argue, why you choose it.
- When you look at the world bank income groups. What differences would you expect between high-income countries and low-income countries? After that, take a look at the data and see what the data says.
- If you look at „more developed regions“ and „high income countries“, do you believe the development of the population correlates?
- Could you think of any conflict, catastrophe or war in the last 70 years which maybe influenced the population in a country? Prove it with the data.

REFLECTION

- How difficult was it for you to read information out of the datasets?
- Is there anything else what you like to know about this dataset?

EXPERIENCE WITH DATA

Question to the round

Do you have some experiences with datasets? Have you ever worked with spreadsheets? If yes, what kind of data have you worked with?

DEBRIEF:
Is there anything else you would like to add?

Thank you all for taking the time. We very much appreciate it! If you would like to follow-up or know more about the results please feel free to write me!

Interview Script

Main Study

Supplemental Material to:
CHATTING ABOUT DATA - INTERACTING WITH VOICE INTERFACES TO ENGAGE WITH ELECTION PANEL DATA

Tool: Video conference via Zoom

WELCOMING

CHECKING THE SIGNED CONSENT FORM

SHORT DESCRIPTION OF THE STUDY

- It will take about 20-30 minutes
- It consists of 2 parts: Questionnaire and chatbot
- All instructions will come from the chatbot, the study leader will turn off the camera and microphone in the meantime
- All questions should be asked to the chatbot
- Participants are asked to share their screen
- The participant can stop the study at any time
- Let the participants know that the study is now being recorded

When the screen is split and recording is started, send the link to the chatbot, ask if everything is clear and turn off the camera and microphone and watch.

CHATBOT INTERVIEW

From now on, the chatbot "speaks" and the user answers.

Hello! I'm Bruno, a chatbot! 🤖

I can answer your questions about the national elections 🗳️ in this country. At the moment, I am still in development 🛠️ and may need a moment ⏳ longer to answer your questions.

My developer wants to do a study with you and me and find out how well we get along. Therefore, I will give you instructions and guide you through this study. First of all, please open

the following page and start with a small survey:
Link

To talk to me, send me a voice message 🔊.
Click on the button 🎙️ & below, hold it as long as you speak and release it to send the question.

Let's try that right now:

Say hello!

Hello Bruno!

Good day from me too! 👋 That has already worked great.

Now let's go one step further.

Now ask me for the fact of the day!

What is the Fact of the Day?

Satisfaction with democracy in this country increased from 61.8% in 2017 to 62.8% in 2019.

Super! We have learned the basics.

Let's get to the actual part: I can currently answer questions about the National Council elections of 2008, 2013, 2017 and 2019. I get my data from online and telephone surveys.

I would like to discuss the following topics with you today: 1. demographics and motivation of non-voters 👫 2. media consumption and voting 📰

Let's start with the topic of non-voters: Is there anything you would like to know about this topic?

TOPIC NON-VOTERS 👫

Is there anything you would like to know about non-voters?

Questions & Answers

Now I have a task for you: Find out which media non-voters consume and whether this differs from the rest of the population. Enter the result in the empty fields of the survey I sent you at the beginning. When you have finished, please let me know.

Questions & Answers

I'm finished now.

TOPIC MEDIA CONSUMPTION AND VOTING 📰

Is there anything you would like to know about young and first-time voters?

Questions & Answers

Now I have a task for you: Find out whether the way you consume media has an effect on your decision to vote. Enter the result in the empty fields of the survey I sent you at the beginning. When you have finished, please let me know.

Questions & Answers

I'm finished now.

FEEDBACK

Finally: Do you have any further questions about these topics?

No

You now have the opportunity to try out the chatbot and ask me all the questions that interest you.

Questions & Answers

I have a few questions for you about the evaluation:

Did you have fun with this study?

Yes

How did you like the interaction with me?

Good

What should I do better?

Suggestion for improvement

Were the answers I gave you helpful?

Yes

Thank you for the feedback!

Finally, click on the download button at the top right and send this file emailadress. Then go to the questionnaire one last time and fill it out to the end. Then you are done. Thanks, it was fun!

From this point on, the study leader joins in again, asks again about the JSON, and answers final questions.

Finally, the functionality of the chatbot is explained and the participants are told that they did not interact with a real chatbot, but that the study leader sent the answers. It is explained that this is a common method for developing an interface. Before taking big implementation steps, the interaction is tested.

Stop recording.