Data Documentation

Data Documentation / Storage Protocol - Chatbot 5

Ron L. Tabuchov
Jarro Teunissen
Stijn van der Pas
Vladislav Stoimenov

Data Science and Artificial Intelligence, Breda University of Applied Science

DISCOVER YOUR WORLD

Breda University OF APPLIED SCIENCES

Index

1 Documentation & Metadata

Our data is collected in 2 ways, one way is in a form of survey (Quantitative), and personal interviews or focus group (Qualitative). All data from the survey and interviews will be stored and protected in GitHub or Brightspace as Excel, voice records, transcript or CSV files in the right folder. Python coding will be done in the data analysis and interpretation stage to get insights and explore correlations.

The survey distributed online and will be answered by respondents globally, Interviews will be completed by each team member individually where each one is free to select the location (in the Netherlands), time and type of interviewer. Combining all the results will provide the team necessary insights and lead to the completion of the research.

Tools for data collection:

- Qualtrics – Publish and collect data from respondents.
- Teams and voice memo applications – to record and transcript interviews.
- GitHub, SharePoint and Brightspace – Storing and version controlling.

FAIR Principles:

- **Findable**: Clear metadata to locate data sets.
- **Accessible**: Stored in GitHub and Brightspace that can be easily accessed.
- **Interoperable**: Codebook with explanation and vocabularies will allow data to be integrated with other datasets and clear to the researcher.
- **Reusable**: Data is available for analysis and file format can be adjusted.

The team will be using machine-readable formats like **XLSX**, **JSON**, or **CSV** so that it can be used in different systems and databases.

Data Processing and GDPR Compliance

In this research, we will collect and handle personal data in line with the General Data Protection Regulation (GDPR). Below is an outline of how personal data will be managed.

1. Purpose of Data Processing

The data collected will be used specifically for this research project, which aims to study the satisfaction level of consumer with chatbots. The data will not be used for any other purposes.

2. Legal Basis for Processing

We will collect personal data based on **informed consent**. Respondents will be informed about how their data will be used, and they can withdraw their consent at any time. Additionally, this research serves a **legitimate interest** in contributing to chatbots for SMEs, while respecting individuals' privacy.

3. Risk Assessment

The risk level for processing personal data is considered **low** because the data will be anonymized and cannot be traced back to individuals. If there is uncertainty about the level of risk, we will consult with the Data Protection Officer (DPO) to ensure all precautions are taken.

4. Anonymization and Security

We will anonymize all personal data, meaning no identifying information will be linked to the respondents. The data will be securely stored using Github and BUas-approved systems within the EU. Only authorized personnel will have access to the data, and security measures like two-factor authentication will be used.

5. Data Sharing

If we work with external partners, we will make sure they follow GDPR guidelines too. A cooperation agreement will be in place to ensure all personal data is handled securely.

- **Informed Consent**: Document participants gave consent can be found here and in the DMP folder of the project.
- **Data Minimization**: Collection data which is relevant to the survey and interviews will be done by selecting and approving the questions with the mentor.
- **Anonymization**: By renaming the files, removing personal information before storing the data and avoiding asking or requesting personal information which can indicate or lead to personal identification will ensure anonymity.
- **Retention Period**: Information and data will be stored on GitHub and SharePoint for unlimited time.
- **Participant Rights**: participants can access, correct, or delete their data by contacting one of the researchers.

Contact information – Chatbot group 5:

Ron L. Tabuchov – 221846@buas.nl
Jarro Teunissen – 2331667@buas.nl
Stijn van der Pas – 202327@buas.nl
Vladislav Stoimenov – 235030@buas.nl

2 File naming & folder structure

Using a well-organized folder structure description is crucial part in the project, Folder Structure and file management in GitHub and SharePoint helps the researchers to navigate and access the data easily.

Folder Structure:

/2024-25a-fai2-adsai-group-chatbot-5/Data

- /Audio/
  - Interview_P001_23-09-24_Audio.mp3
  - FocusGroup_25-09-24_Audio.mp3
- /Video/
  - Interview_P001_23-09-24_Video.mp4
  - FocusGroup_25-09-24_Video.mp4
- /Transcripts/
  - Interview_P001_23-09-24_Transcript.docx
  - Interview_P002_24-09-24_Transcript.docx
  - FocusGroup_25-09-24_Transcript.docx
- /Consent_Forms/
  - Consent_P001.pdf
  - Consent_P002.pdf
  - Consent_FocusGroup.pdf

The main folder is /2024-25a-fai2-adsai-group-chatbot-5/DMP/, and all the subfolders are nested within this directory.

The date format is maintained as DD-MM-YY for consistency and readability.

Each file is named according to its Participant ID, date, and file type (e.g., Audio, Transcript, etc.).

File naming makes the data finable and clear; file naming formats will be used to ensure simple structure and consistency in the research.

Interviews file naming are saved in consistent format by each team member in the corresponding folder location, The file name includes what is the type of file or what it contains, date, ID and file type.

<Type>_<*Date in* DD-MM-YY>_<ParticipantID>.<Format>
Interview_23-09-24_P001.mp3

3 Version control

By uploading all the files to a GitHub repository, we can use the version control to automatically save different versions of each file. This allows us to track changes over time, see who uploaded or modified files, and access previous versions when needed. Using a version control system removes the need to rename files for each iteration, preventing confusion and providing a clear status of each file.

By labeling our file submissions in Brightspace and Trello as "Draft" or "Final," we can clearly communicate the status of each delivery to our teachers. This helps them understand whether a file is a initial version needing feedback or a completed assignment, ensuring clarity about which files still require their attention.

4 Readme file

This folder contains the following files related to data management and research preparation:

1. Codebook_Template.md

Description: A codebook describing all variables to be used in data collection. This is crucial for preparing data collection, as it defines the variable names and their descriptions.

2. NWO-DMP-Template-version-September-2020.docx

Description: The Data Management Plan based on the NWO template. It outlines how data will be managed throughout the research project.

3. BUas Research Ethics Review Application Form.pdf

Description: An application for a BUas Ethics Review.

4. Research_Information_Letter.docx

Description: A letter containing information about the research, to be presented to participants in interviews and included at the beginning of the questionnaire.

5. Informed Consent Form.docx

Description: A letter of informed consent to be presented to participants in interviews and included after the Research Information Letter in the questionnaire.

6. Data Storage Protocol.pdf

Description: A Data Storage Protocol detailing how data will be securely stored and managed.

7. Privacy and GDPR checklist.docx

Description: A checklist to ensure compliance with privacy laws and GDPR regulations.

8. FAIR Checklist.pdf

Description: A checklist to ensure that data follows the FAIR principles.