Protocol for qualitative analysis

Categories Editable Document Protocol

1. Introduction

This document defines the protocol for conducting qualitative analysis for research into the application of the proposed software configuration management model in the context of developing an artificial intelligence model for software aimed at collecting data from legal documents. This document is essential for establishing the process to be used to perform qualitative analysis of the data obtained during the application of focus groups.

2. Qualitative Analysis Process

This section describes the process used to conduct the qualitative analysis. The process is based on the adaptation of the procedures by *Cruzes and Dyba and* **the article Investigating the Developer eXperience of LGBTQIAPN+ People in Agile Teams.** The process consists of 3 steps, described below:

- **(1) Data Extraction —** Extract the transcript of media files [audio and video recordings (Meet)] used to capture the primary study data;

- **(2) Data Coding —** Perform line-by-line coding of the data obtained through transcription. Identify the main relevant codes and relate them to the findings of the primary studies. This step will be done manually, that is, without the aid of tools;

- **(3) Grouping by Themes —** Group the set of codes obtained by related areas, forming inductive, unique themes that are coherent with the research questions. This step requires constant comparison and reciprocal translation between the data. This step will be carried out with the help of LLM.

2.1 Data Extraction

All recordings of the focus groups are stored in the drive folder.

Before performing the extraction, it is necessary to transcribe the recording.

- To perform the transcription, you need to:

1. Download the recording of the focus group in question.
2. Convert to MP3 using some online tool, for example: https://cloudconvert.com/mp4-to-mp3.
3. To transcribe audio to text, follow these steps:
   a. Access the Google Colab page
   b. Run the first cell, with the content:

```
%%shell
pip install openai-whisper pip
install python-docx
```

   c. After completing the dependency installation, ensure that the transcription audio has been completely uploaded to the Google Colab environment.

   d. Change the value of the **audio_file variable,** inside the conditional expression, with the name and path of the recording. In the second cell:

```
import whisper
from docx import Document

def transcribe_audio(audio_path):
    model = whisper.load_model("base")

    # Transcribe the audio
```

```
    result = model.transcribe(audio_path)

    return result["text"]

if __name__ == "__main__":
    audio_file = "p1.mp3"
    transcription = transcribe_audio(audio_file)

    doc = Document()
    doc.add_heading("Audio Transcript", level=1)
    doc.add_paragraph(transcription)
    doc.save("p1.docx")

    print("Transcript saved in transcript.docx")
```

e. Run the second cell, and wait until the transcript is finished.

f. With this, a docs file should have been generated in the directory Google Colab environment. It is extremely important that this file is saved in Drive, or in the local folder, as Google Colab discards the files as soon as the runtime environment is closed (it is volatile).

g. With the docs in hand, place them in the respective participant's folder on the drive, along with the recording and notes.

h. When you access the generated documents, you will see that it is a single continuous text with some spelling errors. The next step is to correct transcription errors and separate the statements of the researcher and the participants. In addition, you should add titles and subtitles that represent the parts of the script that were being discussed in the conversation, in order to facilitate later analysis.

i. With this, we have the transcription ready for the extraction of codes and themes.

2.2 Data Encoding

- This step is done manually, that is, without the aid of tools;

- A **code** is a summary of a relevant aspect (in relation to the research objective) contained in a transcription unit. It is a label or tag applied to data segments that serves to categorize and summarize relevant information. A transcription unit (quotation) is a part of the transcript that has a meaning and minimally provides context;

- Given a focus group transcript, the following steps should be taken to code the transcription units:
  - After the **individual** coding of the first **transcription block** (transcription of a focus group), an analysis will be carried out to synthesize the codes generated by each researcher;
  - With the aim of streamlining the coding process, the analysis will occur by dividing a block into 3 sections, related to the script applied during the focus group. Each researcher will be responsible for extracting the **essence** of what was said in the segment (code) belonging to the section to which they were assigned.

- Notes taken during the focus group should be used to help define the code;

- If necessary, the transcription excerpt can be analyzed by observing it in the original recording, to better support code extraction.
- Some guidelines:
  - A code must be atomic, that is, it cannot encompass other codes.
  - A transcription block may contain multiple codes. If this is the case, the transcript must be divided into smaller units, called transcription units, until the codes are completely separated.

- Coding will be individual, that is, to avoid biases among researchers, there will be a synthesis for the first group of codes generated. The synthesis should occur after the coding of each researcher on the team has finished;

- Questions should be recorded in individual coding environments (Spreadsheet pages) and should only be addressed during the team's code synthesis phase;

- Quotations should include a prefix that allows for the traceability of which focus group was run, section (letter “S” followed by the section number), topic and (if applicable) subtopic from which the quotation was taken. In addition, the participant’s identification code should also be added.

  The format should therefore be:

  `[RouteSectionCode]-[Topic Code].[Subtopic Code]`

  Coding example:

  `S2-1.a-P1: “Quotation”`

  (In Section 2 of the script, in topic 1 and subtopic “a”, participant “P1” answered “Quotation”)

- Once the coding stage is complete, these will be evaluated by at least one team member who did not participate in the coding process;
- The codes and transcription units (quotations) should be placed on the focus group spreadsheet.

2.3 Grouping by Themes

- To extract the themes, LLM (ChatGPT 4th) will be used, following the recommendations of:
  - J. Roberts, M. Baker, and J. Andrew, “Artificial intelligence and qualitative research: The promise and perils of large language model (llm)'assistance',” Critical Perspectives on Accounting, vol. 99, p. 102722, 2024.
  - Yuyi Yang, Charles Alba, Chenyu Wang, Xi Wang, Jami Anderson, and Ruopeng An. 2024. GPT Models Can Perform Thematic Analysis in Public Health Studies, Akin to Qualitative Researchers. Journal of Social Computing 5, 4 (2024), 293–312
- The following prompt should be applied in GPT Chat, based on the ”structured task description” and ”Input-ProcessOutput (IPO)” patterns:

> I am performing a thematic synthesis process based on the responses to a survey. I will provide the manually
> extracted codes, then I will ask you to generate possible themes

from the codes.

Associated codes - Sequence of manually extracted codes;

- GPT Chat will generate several **themes** and the final theme will be considered from the following process (references below the process):
  - **Who performs the analysis:** researchers, in groups (to reduce the bias of each member), manually.

ÿ **Criteria considered for theme selection:**

1. **Relevance to the RQs: The** topic must help answer the proposed RQ of the research and present relevance to the same;
2. **Uniqueness of theme:** According to Braun and Clarke's principles a theme must be coherent (self-contained) and essentially different from others;
3. **Theme compliance with codes: The** theme must cover all codes of that semantic;
4. **Iterative refinement:** Thematic synthesis is an iterative process, where themes are progressively reviewed, tested and refined until they become appropriate.

- **Criteria that will not be considered:**
  - **Make sure that the proposed themes are truly representative of the nuances and context present in the original data.** A theme may seem logical based on the codes, but it may miss important contextual information from the original responses.
    - It will not be considered due to the robust process used for code analysis: researchers constantly reviewed the transcription units, updating and rewriting them, building codes that represented them in the most objective way possible.
  - **Look for evidence in the raw data for each proposed theme.** The theme that is most strongly and consistently supported by the raw data is probably the most appropriate.
    - It will not be considered due to the way it was the analysis was carried out: the codes represent the transcription units (original data), and therefore, when the theme represents the codes, it also represents the units.
  - **Feedback and revision:** If proposed themes are not satisfactory, clear feedback should be provided (why they are not suitable, what is missing, what connections were missed) and the prompts, questions or groupings should be refined.
    - It will not be considered, as the selection of generated themes will be done manually by researchers.

References:

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology;*
- Braun, V., & Clarke, V. (2019). Reflecting on reflective thematic analysis. *Qualitative Research in Sport, Exercise and Health;*
- Nowell, L. S., Norris, J. M., White, DE, & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods;*
- Guest, G., MacQueen, K. M., & Namey, E. E. (2012). *Applied thematic analysis.*

    - **If there is still doubt between one or more themes,** these can be modified or merged, including the respective codes for each one.

- Once the codes have been grouped by theme, they will be evaluated by at least one team member who did not participate in the grouping process;

- The topics must be placed on the participant's spreadsheet.