# **1. Project Overview**

### **Project Name: Qualitative Data Analyzer**

This tool will assist researchers in analyzing qualitative data from academic papers. It will automate the data extraction, coding, and tone analysis of sentences mentioning user-specified keywords. Users upload PDF papers, provide keywords, and receive a structured summary report in spreadsheet format. The report consolidates relevant sentences, distinguishes between the author’s original findings and cited content, and analyzes the tone of the author’s statements regarding each keyword. This process reduces manual workload, improves consistency, and supports data-driven literature reviews.

# **2. Vignette Example:**

**Analyzing Gender-Related Discourse in Cybersickness Research**

Maya is conducting a literature review to explore how gender and other factors may influence cybersickness in virtual environments. She has collected a set of academic papers, but manually reading and summarizing each one is time-consuming and inconsistent. Therefore, she really wants a structured and time-efficient way to find out:

* From her set of papers, how many papers mention and discuss the specific factors? She also wants the name of the paper, the author, and its publication year.
* If one paper or a number of papers mention that factor, what does each paper say about those factors? She wants all the sentences that contain those factors and wants to know which ones are the paper's own concluding arguments and which ones are the arguments from other authors that the paper has cited to come to their own conclusion.
* For the sentences that are the author's own statement, do they use a supportive, uncertain, or opposing tone to discuss the factor’s impact on cybersickness?
* For the sentences that have references or in-text citations at the end, how many references are cited, and what are they?

To streamline this process, Maya uses the Qualitative Data Analyzer tool. She uploads PDFs of her selected academic papers and inputs a list of keywords representing the factors she’s investigating. The tool scans each document, extracts all sentences mentioning each keyword, and organizes them into a spreadsheet. For each author’s own statement about the factor (i.e., given keyword), the tool uses tone analysis to determine whether the claim supports, challenges, or expresses uncertainty about the factor’s role in cybersickness. In the final output, Maya receives a structured spreadsheet where each row represents a single keyword and includes:

* The title and author/year of the papers that mention the keyword
* All the sentences from the papers that include the keywords, separated by the ones that have an in-text citation at the end, and those that do not.
* The tone of the author’s own statements, where the sentences have no in-text citations or references (e.g., supporting, neutral, opposing)
* A list of references cited when discussing those factors.

With this structured summary, Maya is able to quickly identify which factors have a strong research consensus and which remain debated. Instead of manually combing through dozens of papers, she can focus her energy on synthesizing meaningful insights from the output of this tool.

# **3. Technical Details**

**User's Input**

The user will use a command line interface or a GUI (in a future version) to run the tool. The tool will require the following inputs to perform the qualitative analysis:

* PDF Files: One or more academic research papers in .pdf format.
* Keywords List: A list of factors or concepts (e.g., gender, VR experience, motion sensitivity) the user wants the tool to analyze.

**Keyword-Based Sentence Extraction**

The tool processes each PDF to extract the full text and tokenize it into individual sentences. It then identifies all sentences that contain one or more user-specified keywords.

Each sentence is analyzed for the following:

* Keyword Match: Whether it contains any of the specified factors.
* Citation Detection: Whether the sentence includes in-text citations using regular expressions (e.g., matches like Smith et al., 2020 or (Johnson, 2019)).

Content Source Classification:

Author’s Own Statements: Sentences without citations, assumed to reflect the paper’s own claims.

Cited Statements: Sentences with in-text citations, assumed to refer to or summarize prior research.

**Tone Analysis**

For each author’s own sentence, the tool analyse the tone of the sentences to detect the stance toward the keyword. The tones could be either supportive, neutral, or opposing. For example:

1. **Supporting** indicates that the paper’s authors assert or affirm the presence of an effect, relationship, or influence of the keyword on cybersickness. This tone often suggests that the factor matters, has a notable impact, or is a key contributor. Some phrases used by the authors that can help classify the supporting tone are: significant, critical, affects, influences, predicts, correlated with, plays a role, associated with, linked to, worsens, improves, highly sensitive to, strong relationship, results indicate, was found to contribute, etc.

Example sentences: Gender significantly influenced the severity of cybersickness symptoms.”

1. **Neutral** sentences are the ones that mention the factor but do not take a stance on whether it impacts cybersickness. Common indicators of this kind of tone are words like measured, included, collected, considered, recorded, examined, controlled for, and reported.

Example sentence: Gender was recorded as part of the demographic information.

1. **Opposing** indicates that the paper’s authors found no effect, no significant relationship, or even contradictory evidence between the keyword and cybersickness. The tone suggests the factor does not matter, is inconclusive, or shows inconsistent findings and uses negative phrases such as no significant, not related, was not found, did not affect, no difference, not associated, no clear pattern, mixed results, results were inconclusive, failed to show, inconsistent findings

Example sentences: Gender was not found to have a significant effect on cybersickness.

**Reference Mapping**

In cited statements, the tool captures all in-text citations associated with keyword discussions and aggregates them under the respective factor. It tracks:

• Number of references cited per factor in the current paper

• Which references are cited in relation to which factors

### **Output**

The output is an Excel file with the following fields for 1 single factor:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Paper 1: Paper title and author, and publication date** | **Paper 2: Paper title and author, and publication date** | **Paper 3:** |
| **Factor 1 (keyword)** | All the sentences that include the keyword | All the sentences that include the keyword | … |
| Statements that use references | Statements that use references | … |
| No of references used | No of references used |  |
| The list of references | The list of references |  |
| Authors own statements | Authors own statements | … |
| Tone of the author's own statements | Tone of the author's own statements | … |

|  |  |  |
| --- | --- | --- |
|  |  |  |

### **4. Major Components of the Tool**

* **PDF Input Handler**: Loads and extracts text from user-provided PDF papers.
* **Text Parser & Sentence Tokenizer**: Splits the extracted text into individual sentences using NLP libraries like NLTK or spaCy.
* **Keyword Extractor**: Searches for user-defined keywords within each sentence.
* **Citation Detector**: Identifies in-text citations using regular expressions for APA-style formats.
* **Content Classifier**: Categorizes sentences as the author’s own or cited content based on citation presence.
* **Tone Analyzer**: Determines the tone of the author’s own statements using a rule-based or ML-based classifier.
* **Reference Aggregator**: Counts and lists which references are used to support each keyword discussion.
* **Excel Exporter**: Compiles structured results into a table format for easy review, filtering, and synthesis.

### **Future Components**

The future component would focus on how to organize the data into one spreadsheet. Another component is to code the data using a scoring system.

# **6. Final Self-Assessment**

**1. What was the biggest or most unexpected change you had to make from your sketch?**

I initially wasn’t clear about the output I needed. As I developed the idea, I realized the importance of structuring the data, extracting keyword-related sentences, classifying tone, and linking references, all in a clear spreadsheet format.

**2. How confident do you feel that you can implement the spec as it's written right now?**

I feel more confident implementing this tool compared to my earlier project. The scope is clearer, and the tasks feel more manageable within my current skill level.

**3. What is the biggest potential problem that you NEED to solve (or you'll fail)?**  
Tone analysis and organizing the data clearly in a spreadsheet will be the biggest challenges. If the tone classification or output formatting fails, the tool won’t be useful. Another potential problem is to identify how to present the data in a single spreadsheet when there will be multiple keywords and each will have data from multiple papers.

**4. What parts are you least familiar with and might need my help with?**

I’m least familiar with building the tone analysis logic and structuring the output into a single spreadsheet. I may need help with both parts.