INFORMATICS INSTITUTE OF TECHNOLOGY
In Collaboration with
UNIVERSITY OF WESTMINSTER

Multilingual Dialogue Summary Generation System for Customer Services

A dissertation by

Mr. Tharindu De Silva

Supervised by

Ms. Dileeka Alwis

Submitted in partial fulfilment of the requirements for the BEng (Hons) Software Engineering degree at the University of Westminster.

April 2023

ABSTRACT

A written or spoken conversation between two or more people is known as a dialogue. In most modern-day applications where the conversation happens, it generates lots of data either in textual format or audio format. Summarizing these data or dialogues is known as the dialogue summarization where it only focusses on the relevant information. Dialogue summarization aims summarize only the necessary information where a reader can quickly capture the highlights of the dialogue without reviewing the whole.

Customer service is a domain where it generates a lot of data daily. Conversation between a support agent and the customer can take up a lengthy dialogue. At the end of the dialogue conversation, support agents should write a short description (a summary) of the dialogue which will help for later reference. This is time consuming and requires human resources.

'MultiDialogSum' is a dialogue summary generation system for customer services where it supports generating summaries for multiple languages. The use of recent cross-lingual transfer modals and machine translations techniques are utilized to achieve these capabilities. This will be developed and released as a web application for the users.

Keywords: Dialogue Summarization, Natural language processing, Cross-lingual transfer modals, Machine translation, Machine Learning

Project Descriptors:

Artificial Intelligence → Natural Language Processing → Language models, Machine translation

DECLARATION

I have personally written this dissertation and it has not been submitted for any academic certification or degree program from any university or institution, either currently or in the past. I have given proper credit to the authors of previous works by citing their information appropriately.

Name of the Student: Tharindu Niroshan De Silva

Registration Number: w1761890/2018367

Signature:  **Date**: 04/2023

ACKNOWLEDGMENTS

I'd want to thank and appreciate my supervisor, Ms. Dileeka Alwis, for her constant advice and support during my research journey. Her expertise, patience, and commitment have been invaluable in shaping the direction and quality of my thesis.

I am also grateful to Mr. Guhanathan Poravi, the module leader, for his insightful feedback and constructive suggestions that have significantly contributed to the development of my research work. His dedication and encouragement have been instrumental in helping me overcome numerous challenges.

I extend my heartfelt thanks to all the lecturers at the Informatics Institute of Technology, whose collective knowledge and expertise have laid the foundation for my academic growth. Their passion for teaching and research has been a source of inspiration, and their guidance has been integral to my success. Also, I’m thankful for the people who helped me in surveys, evaluations and recommendations during various stages in my research project for spending their valuable time.

My sincere appreciation goes to my friends from the university, who have been a constant source of camaraderie, motivation, and support throughout this journey. Their unwavering belief in me and willingness to lend a helping hand have made this experience more enriching and enjoyable.

Finally, I'd like to offer my heartfelt appreciation to my family members for their unwavering support and understanding. Their unwavering faith in my abilities and sacrifices made on my behalf have provided me with the strength and determination to persevere, even during the most challenging times.

To everyone who has contributed to my research journey, I sincerely appreciate your advice, support, and encouragement. Thank you very much.

TABLE OF CONTENTS

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Acronym	Description
AI	Artificial Intelligence
ML	Machine Learning
NLP	Natural Langue Processing
ROUGE	Recall-Oriented Understudy for Gisting Evaluation

CHAPTER 1: INTRODUCTION

1.1 Chapter Overview

This section of the thesis covers the summary of this research project which is a dialogue summary generation system that supports more languages. This chapter will include the information about the problem domain, research gap, research challenges and objectives that the author wishes to achieve by the end of project completion.

1.2 Problem Domain

A written or spoken conversation between two or more people is called a dialogue. Dialogue summarization is the technique of condensing a dialogue so that a reader can quickly understand the exchange. The dialogue summarization method involves extracting important information from the discourse to produce a summary highlighting the conversation's main points. With the constant development of communication technology in lately, dialogues have become an important way of information exchange.

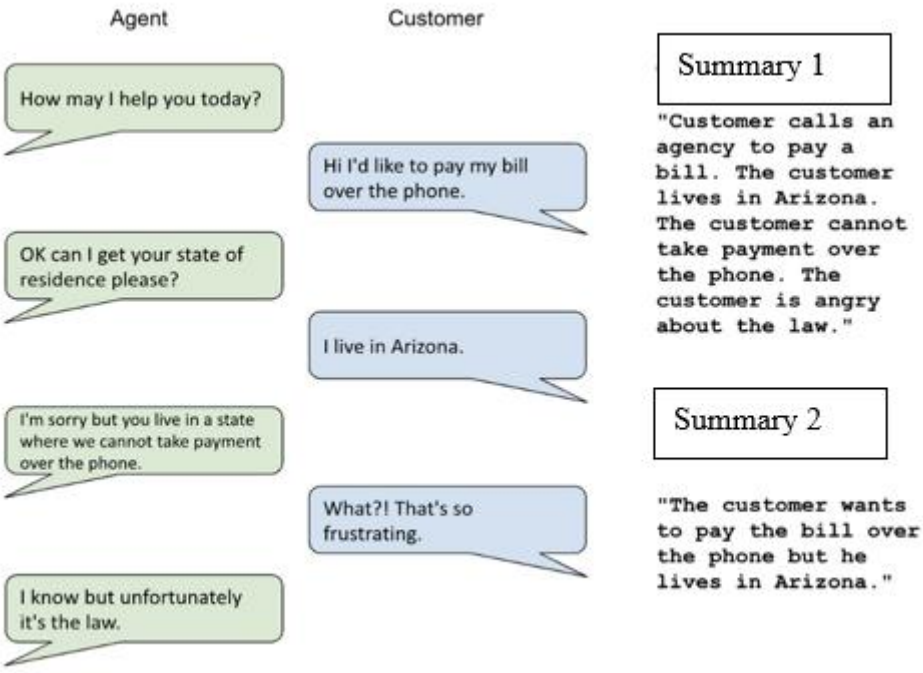


Figure 1 – Dialogue Summary

This proposal aims to provide the reader with an outline of dialogue summarization and how it has grown in recent years in different problem domains. The relevant study is discussed, which will clarify and justify the existing research gap and problems.

1.2.1 Impact of Dialogues on Communication Technology

Dialogues have been utilized within different domains as a communication technology for real-life applications. For example, some are in meetings, chats, interviews, and customer service. A dialogue can be represented in both textual and audible forms. Both are widely used in the real world. For example, chatbots are considered an intelligent way of communicating with minimal human interactions. These chatbots heavily use dialogues in a textual form as their primary source of communication. Another example is dialogues in the audible form in interviews. An interview between the two parties can be recorded and represented as a dialogue in an audible format.

1.2.2 The necessity of Dialogues Summarization

As previously mentioned, with these real-life applications' rapid growth, the dialogues' Summarization came into place because of various factors.

1.2.2.1 Lengthy Dialogues

Dialogues can be used as a way of storing information for later reference. But the vast amount of data generation and length of the dialogues practically caused problems. In the customer service domain, a conversation between a customer and a support agent can be as follows. A customer can raise a complaint, and a support agent tries to solve the issue. Support agents are asked to write a summary of the problem and possible solution at the end of their conversation. So, this summary can be used by other agents who may have to deal with similar or the same customer issues without going through the entire dialogue.

1.2.2.2 Time Consuming and Resource Cost

When it comes to massive data generation, analysing a longer conversation requires human interaction and can cost much time and resources. Sometimes it is overwhelming to go through lengthy dialogues to understand what happened within the conversation.

1.3 Problem Definition

To automate dialogue summarization without human interaction, many researchers have put their effort into building solutions using machine learning and natural language processing technologies for this challenging problem because of their unique application value (Gao and Wan, 2022). However, these solutions are heavily focused on the high resource languages such as English, Chinese etc.(Feng, Feng and Qin, 2022). High-resources languages are known as languages in which many data resources exist.

With globalization's acceleration, a domain like customer service can be involved with multinational participants. Yet the current solutions are only capable of primarily handling the dialogues which are in English languages. This can be a critical situation where global customer service is required to summarize the dialogues between an agent and a customer using human resources with linguistic fluency in multiple languages.

1.3.1 Problem Statement

Dialogue summarization in customer service is not developed to handle languages other than mostly English. Customer services can always be engaged with multiple languages, yet there is no solution for multilingual dialogue summarization.

1.4 Research Motivation

The problem stated in this proposal mainly applicable to the customer service domain and researchers who have extensive knowledge in Dialogue Summarization. Since dialogue summarization can be utilized in many fields, such as meetings, healthcare, and email threads, this potentially impacts many businesses' use cases. This work is expected to add a valuable contribution to the advancements of the dialogue summarization globally by expanding the possibilities of multilingual Summarization.

1.5 Research Gap

Based on the previous work on dialogue summarization, researchers have focused on the problem of low resources, which can be further divided into low linguistic resources and domain-specific data sets. Yet the approaches are heavily focused on English languages. This is due to the scarcity of the datasets, and the investment in such datasets can be costly (Zou et al., 2021). With globalization need of the dialogue summary generation solution with multilanguage capabilities required (Feng, Feng and Qin, 2022). This can be addressed as a theoretical gap in the Dialogue summarization domain.

Dialogue summarization in customer service is where multiple languages get involved. But the currently available solutions can only perform an English dialogue summarization. This can be identified as an empirical gap in the customer service domain. This project focuses on the empirical gap in customer service and the theoretical gap in the Dialogue summarization domain.

1.6 Contribution to the Body of the Knowledge

Upon the completion of this research project, contributions can be summarized as follow:

- A platform to generate dialogue summary for customer service: Data Science [Machine Learning]
- A novel approach to use existing datasets for low resource languages: [Cross-Lingual Transfer, Pre-trained Language Models]

1.6.1 Contribution to Problem Domain

A platform that supports dialogue summary generation for multiple languages for the customer service domain will be explored. Considering current solutions' availability, a platform capable of generating summaries for more languages can be valuable for the growth of the customer service domain globally. The proposed solution can save both time and cost when it comes to preparing summarize for large quantity of dialogues without human resources.

1.6.2 Contribution to the Research Domain

A novel approach to utilize the existing datasets for dialogue summary generation and overcome the multilingual and scarcity of the datasets in low-resource languages. The most recent developments in pre-trained language models and their capabilities of zero-short cross-lingual transfer will be explored to develop a multilingual dialogue summarizer. Later it is a hypothesis that this proposed approach can be used as a baseline method for low-resource natural langue tasks.

1.7 Research Challenge

Dialogue summarization is a trendy domain on which both large-scale enterprise companies and small companies are focused. Recently Microsoft cognitive azure team has started developing solutions for dialogue summarization.

During the past years, a wide variety of research projects have focused on text summarization, document summarization, news, etc. Yet the limited resources currently available and the complexity reduce the expansion of technology in this domain. Previous studies have explained how real-world dialogues challenge current summarization models (Zhang et al., 2021). Unlike text or document summarization, dialogues carry more complex attributes such as general knowledge, intentions, and informal sentences. Therefore, existing text summarization techniques cannot directly apply to dialogue summarization.

With the limited resources of datasets available in high-resource languages, existing work and future work is more focused on a few languages. This restricts the development of practical solutions that can summarize dialogues other than English. Considering these identified factors, a system to support a multilingual dialogue summarizer is needed.

1.8 Research question/s

RQ1: What are the current limitations in dialogue Summarization?

RQ2: How can existing resources be utilized to develop a multilingual dialogue summarizer using cross-lingual transfer techniques?

RQ3: What are the most recent advances in pre-trained language models that can be leveraged to construct a solution for multilingual dialogue summary generation?

1.8.1 Aims

The aim of this research is to design, develop and evaluate a multilingual dialogue summary generation system for customer services using dialogue summarization model with the help of the cross-lingual transfer method.

To explain the research aim, this research project will focus on developing a system which is capable of utilizing low linguistic resources to build a multilingual dialogue summarizer. The recent development in pre-trained language models and the capabilities of cross-lingual transfer learning will be applied to achieve this. Cross-lingual transfer learning is the mechanism of learning and transferring knowledge from one natural language to another. The number of supporting languages of the multilingual dialogue summarizer can depend on the selected pre-trained model and its capabilities.

The relevant knowledge will be investigated and explored throughout the project timeframe to confirm or reject the selected hypothesis, components will be built, and performance will be evaluated. This multilingual dialogue summary generation platform will be released on a hosted server for public use.

1.8.2 Research Objectives

Research Objectives	Description	Learning Outcomes	Research Questions

Literature Review	Gather required material on previous work and critically evaluate the findings. • **RO1:** Study on existing dialogue summarization techniques in the customer service domain. • **RO2:** Study on existing methods of text summarization techniques. • **RO3:** Conduct a preliminary study on pre-trained language models. • **RO4:** Analyze the recent advancements in multilingual models.	LO1, LO4, LO8	**RQ1, RQ2**
Requirement Elicitation	Determine the project's needs using the proper methods to give a solution for the research problems and gaps should be handled based on relevant prior research knowledge. • **RO1:** Gather information about requirements and resources related to dialogue summarization. • **RO2:** Gather requirements related to the pre-trained language model and understand the capabilities of recent advancements. • **RO3:** Gather information related to cross-lingual transfer learning.	LO2, LO6, LO8	**RQ2, RQ3**
Design	Designing a system capable of generating a dialogue summary with multiple languages involved. • **RO1**: To design a cross-lingual transfer learning model which can be implemented from the existing resources. • **RO2:** To identify a suitable algorithm for the proposed methodology and design to summarize the dialogues.	LO1, LO3, LO5, LO8	**RQ2, RQ3**

Implementation	Implementing a multilingual dialogue summarization platform. • **RO1:** To train the cross-lingual model using the existing resources. • **RO2:** Develop a model to summarize the dialogues. • **RO3:** Develop a user interface and which a user can interact with the application from a browser.	LO1, LO5, LO7, LO8	**RQ2**
Evaluation	Testing the implemented system and dialogue summary generation with evaluation metrics. • **RO1:** Test each component by creating unit tests. • **RO2:** Using performance metrics to evaluate the effectiveness of the summary generation. • **RO3:** Following the completion of the research, the functional and non-functional requirements should verify.	LO1, LO5, LO8	**RQ2**

Table 1 – Research Objectives

1.9 Chapter Summary

In this part of the thesis author has discussed about the problem with research gap, challenges and the expected contribution to each research and problem domain. Also the research objectives and the learning outcomes are mapped under the requirement of the University of Westminster, BEng(Hons) Software Engineering course final year research project module.

CHAPTER 2: LITERATURE REVIEW

2.1 Chapter Overview

This section of the thesis covers the literature review of the research project. This includes the problem domain which is the dialogue summarization, review of the existing systems and different approaches that the previous research projects are tried out. Furthermore, evaluation metrices that has been developed and used for dialogue summarization is also discussed.

2.2 Concept Map

Refer to **APPENDIX A**.

2.3 Problem Domain

2.3.1 Introduction to dialogue summarization

Dialogue summarization is a rapidly evolving field of research within natural language processing (NLP) and artificial intelligence (AI) that focuses on the automatic generation of concise and coherent summaries from spoken or written conversations (Gurevych and Strube, 2004). This area has gained significant attention recently due to its potential applications in various domains, including customer service, meetings and conferences, online forums, and conversational AI (Feng, Feng, and Qin, 2022). By providing accurate and informative summaries, dialogue summarization systems can help users digest critical information more efficiently, facilitate decision-making, and improve the overall effectiveness of communication.

Name	Domain	Language
ICSI [Janin *et al.*, 2003]	Meeting	English
AMI [Carletta *et al.*, 2005]		English
QMSum [Zhong *et al.*, 2021]		English
SAMSum [Gliwa *et al.*, 2019]	Chat	English
GupShup [Mehnaz *et al.*, 2021]		Code-Mix
CSDS [Lin *et al.*, 2021]	Customer Service	Chinese
TODSum [Zhao *et al.*, 2021]		English
TWEETSUMM [Feigenblat *et al.*, 2021]		English
CRD3 [Rameshkumar and Bailey, 2020]	TV Show	English
[Song *et al.*, 2020]	Medical	Chinese
SumTitles [Malykh *et al.*, 2020]	Movie	English
MEDIASUM [Zhu *et al.*, 2021]	Interview	English
DIALOGSUM [Chen *et al.*, 2021]	Spoken	English
EMAILSUM [Zhang *et al.*, 2021a]	Email	English
ForumSum [Khalman *et al.*, 2021]	Forum	English
ConvoSumm [Fabbri *et al.*, 2021]	Mix	English

Figure 2 – Dataset Availability

Meeting summarization has been developed recently with business digitalization. Meeting summaries can be very beneficial for grasping what happened within the meeting for participants and non-participants by referring to the generated summaries.

People have grown overwhelmed with large amounts of chat information as online chat applications have become vital tools for people to engage with one another. A new chat member may find it difficult to rapidly examine the primary topic of the dialogue in such a complex context. As a result, summarizing chats has emerged as a new norm.

Email threads can become lengthy and complex with multiple people as the conversation flow can change from one topic to another. This behaviour frequently causes the reader to find it hard to extract the information and follow the discussion flow. By giving concise summaries of the email conversations, a reader can rapidly grasp the important ideas and decisions made without going through the entire email thread, saving time.

The information exchange in customer service is direct between customer and the support agent. Automatic summarizing is one technique to improve customer service by providing the agent with immediate solutions based on the previous condensed summary. As a result, in past years, there has been an increase in research interest in customer service dialogue summarization.

2.3.2 Abstractive and Extractive Summarization

The summarization can be divided intro two paradigms. Those are extractive and abstractive summaries. Extractive approaches select important sentences as the summary, which is more accurate and truthful, but abstractive methods generate the summaries using new words, which increases the summary's conciseness and fluency (Chen et al., 2021). Recent advancements in neural networks, data-driven techniques have made great progress in these two paradigms. When it comes to abstractive approach, sequence-to-sequence learning integrated with at attention mechanism is developed as the backbone architecture for abstractive summarization (See, Liu and Manning, 2017)

2.3.4 Transformers

Transformers were introduced (Vaswani et al., 2017). It’s encoder-decoder based architecture modal has contributed a significant impact on the current foundation for many states of the art modals. Transformers have been used in recent research for improving the dialogue summarization task. Also for evaluating the dialogues, BERTScore was introduced by (T Zhang et al., 2020) .

2.4 Literature Review of the existing systems

2.4.1 Dialogue summarization and Taxonomy

Abstractive meeting summarization techniques have been more popular in recent years (Shang et al., 2018). Several studies have investigated the use of deep learning in addressing the summarization task since the creation of neural networks and have done so with notable success. While deep learning-based approaches have excellent modeling capabilities, relying just on literal data is insufficient. This is because meeting utterances include a variety of interacting signals, and the lengthy meeting transcripts further provide difficulties for conventional sequence-to-sequence models.

Instead, than summarizing the entire meeting, producing meeting summaries of a key feature, such as choices, efforts, ideas, and assumptions, may fit specific needs. (Zhong et al., 2021) recently suggested query-based meeting summarization, which attempts to describe a specific segment of a meeting based on the query provided. Meetings can include nonverbal information offered by participants such as sound, vision, and movement. These traits may be useful for identifying significant utterances during a meeting. As a result, the majority of studies look at both the extractive and abstractive multi-modal meeting summarization challenges, integrating both verbal and nonverbal data to better utterance representation. (Zhou et al., 2022)

Customer service interactions are inherently logical and focused on specific topics because participants typically have strong intent and clear motivations to address issues. As a result, some researchers have investigated topic modeling as a solution to this problem. (Liu et al., 2019) employ a coarse-to-fine generation framework that generates a sequence of key points (topics) to represent the logic of the dialogue before producing a detailed summary. A key point sequence might be question, solution, user approval, end, illustrating the progression of the dialogue. Instead of using pre-defined topics, (Zou, Zhao, et al., 2021) use neural topic modeling to explore implicit topics and propose a multi-role topic modeling mechanism. To address data scarcity, (Zou, Lin, et al., 2021) propose RankAE, an unsupervised framework that first selects topic utterances based on centrality and diversity before using a denoising auto-encoder to generate final summaries.

Customer service, on the other hand, is a type of goal-oriented conversation that includes informative elements, multiple domains, and two distinct participant roles. (Zhao et al., 2021) create a new dataset annotated with detailed dialogue state knowledge to aid in tracking the fine-grained dialogue information flow and producing accurate summaries in order to incorporate dialogue-specific information. Because customer service participants play different roles, (X Zhang et al., 2020) propose an unsupervised framework based on variational auto-encoders to generate separate summaries for the customer and the agent. (Lin et al., 2021) also present the CSDS dataset, which includes role-oriented summaries to capture the perspectives of different speakers.

2.4.2 Low-resource dialogue summarization

Producing extensive dialogue datasets with annotated summaries is expensive and time-consuming, which makes it challenging to create and train efficient summarization models, especially in new areas. As a result, it is crucial to develop dialogue summarization models that work well in settings with limited or no training examples. Techniques that involve domain adaptation and large-scale pretraining have become popular for low-resource summarization. These methods leverage external summary data from other domains, like the CNN/Dailymail news dataset, for initial model pretraining before refining it on dialogue summaries with scarce resources. Recent studies have shown that pretrained summarizers are effective in various dialogue contexts, such as chat logs and medical discussions (Zou, Zhu, et al., 2021).

By using domain adaptation for text summarization, this unique technique addresses domain differences. Domain adaptation has lately attracted significant study attention due to the fact that texts and their summaries from multiple domains may have commonalities and reciprocal benefits. An adversarial discriminator (critic) is used in the approach to learn the domain of each representation. It employs a gradient reversal method to create domain-invariant features by making feature distributions across domains as close as possible. This encourages the summarizer to prioritize content over domain-specific information.

The domain-agnostic multi-source pretraining approach for low-resource dialogue summarization method makes use of external large-scale corpora from a variety of sources to aid in dialogue modeling, summary language modeling, and abstractive summarization. To learn domain-agnostic summarization, adversarial signals are used in pretraining. The experimental results show that this approach is effective and adaptable in low-resource situations. Future work will look into how to keep token-level cross attention in the multi-

source pretraining strategy, how to integrate this approach into models with universal transformer architectures, such as BART, and how to capitalize on large-scale pretraining language models.

2.4.3 Cross lingual Summarization

The suggested method seeks to improve cross-lingual word-level representations by emphasizing the idea that words relevant to specific domains, such as dialogue domains (weather, alert, and reminder), are more essential than other terms. The approach accomplishes this by selecting 11 English terms linked with specific dialogue domains and translating them into other languages using bilingual lexicons, resulting in a limited number of parallel word pairings. These parallel word pairs are then used to improve the cross-lingual word embeddings inside (Artetxe et al., 2018). This method prioritizes domain-specific terms to refine the embeddings, resulting in more accurate cross-lingual transfer learning in related tasks. By modeling the distribution that encapsulates the differences in semantically similar sentences between languages, the latent variable model addresses the variation of similar sentences across languages. Even with very precise cross-lingual embeddings, noise remains due to variances in source and target languages. When combined with poor alignment, this noise becomes more severe, rendering point estimation subject to minor but significant variations among languages. The use of latent variables allows for the modeling of this distribution as well as the capture of variation in semantically comparable sentences across languages (Liu et al., 2019).

The paper (Bai, Gao and Huang, 2021) introduces MCLAS, a new multi-task learning framework that aims to perform cross-lingual abstractive summarization with few parallel resources. In other words, the goal is to generate summaries in a target language from original text in a source language, even when parallel data is scarce. The MCLAS model includes a shared decoder that generates both monolingual (single-language) and cross-lingual (source-to-target language) summaries sequentially. The model can leverage knowledge from monolingual summarization to help improve cross-lingual summarization performance by using this unified decoder for both tasks. The authors evaluated the performance of the MCLAS framework using two cross-lingual summarization datasets. The results show that the proposed model outperforms all baseline models (previous or competing approaches) in both low-resource and full-dataset scenarios (where more parallel data is available). This demonstrates that the MCLAS framework can produce high-quality cross-lingual summaries even when parallel resources are limited.

2.4.4 Pretrained language models

This study proposes utilizing DialogGPT, a state-of-the-art (SOTA) pre-trained model for conversational response generation, as an unsupervised dialogue annotator for dialogue summarization. The aim is to create a short summary while preserving valuable information. To do so, informativeness, redundancy, and relevance must all be considered.

DialogGPT is a highly effective conversational response generation model developed by (Y Zhang et al., 2020) that was pre-trained on 147 million conversation-like exchanges from Reddit comment chains. As a result, it has encoded a significant amount of dialogue background knowledge, which can be used for dialogue annotation tasks. By using DialogGPT as a annotator, it helps to determine the informativeness, relevance and redundancy of a particular dialogue (Feng et al., 2021).

- The extraction of keywords identifies the most important words or phrases in the dialogue, which aids in capturing the main points.
- Redundancy detection identifies and removes redundant or unnecessary information from the summary, making it more concise.
- Topic segmentation divides the dialogue into topically coherent segments, giving the summary a clear structure.

If an utterance is unpredictable, the DialogGPT annotator inserts topic segmentation points before it, indicating a shift in the conversation topic. The paper's experimental results show that the proposed method by using the DialogGPT annotator achieves remarkable improvements on both datasets and establishes a new state-of-the-art performance on the SAMSum dataset, demonstrating its effectiveness in dialogue summarization.

2.5 Literature Review of the Technologies

2.5.1 Dataset Formats for Fine Tuning Transformers

TWEETSUMM consists of 1,100 dialogues recreated from tweets found in the Kaggle Customer Support on the Twitter dataset (Feigenblat et al., 2021). Each dialogue is accompanied by three extractive and three abstractive summaries created by human annotators. The Kaggle dataset, which is based on conversations between customers and support agents on Twitter.com (Hardalov, Koychev, and Nakov, 2018), covers an extensive array of subjects and services provided by various businesses, including airlines, retail, gaming, music, and more.

As a result, TWEETSUMM can function as a dataset for training and assessing summarization models across a broad spectrum of dialogue scenarios.

When it comes to fine-tuning transformers using the hugging face library, it requires the dataset to be in a specific format. The original paper by (Lewis et al., 2019) discusses the generalized explanation of the preprocessing steps, such as tokenization and masking, about fine-tuning and preprocessing the Bart modal. The Hugging Face Transformers library (Wolf et al., 2019) was used to fine-tune the model on the dataset. As recommended by the library's documentation and examples, the dataset was preprocessed and organized into source and target (Hugging Face, n.d.). Each line in the source column corresponds to an input sequence in this format, and the corresponding line in the target column contains the matching output sequence.

2.5.2 Modal Selection

After a thorough assessment of relevant research papers, the author has selected the bart modal over the Roberta model for the dialogue summarization task. The main reason for this is that BART is originally designed for text summarization task (Lewis et al., 2019). Bart, which stands for bidirectional and auto regressive transformer, denoising autoencoder framework that has shown state of the art performance in wide range of natural language task such as text generation, translation and comprehension (Lewis et al., 2019).

RoBERTa, a robustly optimized BERT pre-training approach, focused on the domain of text classification and sentiment analysis as outlined by (Liu et al., 2019). Although Roberta has achieved state of the art results in these natural language tasks, its capability of dialogue summarization task is not specifically aligned with the requirement. Bart;s denoising autoencoder framework makes it especially well-suited for abstractive summarization tasks, such as dialogue summarization, where the model must provide coherent summaries that accurately capture the core of the original text (Lewis et al., 2019). Based on these findings, the researcher concluded that BART is a better candidate for the dialogue summarizing task than RoBERTa.

The author decided to use the M2M-100 model for machine translation tasks instead of other available machine translation models after an extensive review of relevant research papers. The main reason behind this choice is the M2M-100's focus on many-to-many translation, which

supports translations between multiple languages without the need for intermediate pivoting through English (Fan et al., 2020).

If going through briefly MarianMT model by (Junczys-Dowmunt et al., 2018) model is a neural machine translation system that is built on the transformer architecture for machine translation. While the MarianMT model provides significant performance in a variety of machine translation tasks, main limitation is that it does not enable many to many translations. This means the MarianMT often requires the training of separate models for each translation direction (eg – English to French, French to English) rather than a single model that can handle many input and output languages at the same time.

While the M2M-100 modal supports many to many translations, is particularly advantageous in preserving the original meaning and nuances of the source text during translation, reducing the potential loss of information. M2M-100 is pretrained on a large-scale multilingual dataset, which comprises 7.5 billion sentences across 100 languages (Fan et al., 2020). This extensive pretraining enables the model to understand and generate translations for a wide range of languages, including low-resource languages, with higher quality compared to other machine translation models. Additionally, M2M-100 demonstrates strong zero-shot translation capabilities, which allow it to translate between language pairs on which it has not been explicitly trained (Fan et al., 2020). In contrast, other popular machine translation models like OpenNMT (Klein et al., 2017) and the original Transformer model (Vaswani et al., 2017) are limited in their ability to handle multiple languages or perform zero-shot translation tasks effectively. Furthermore, these models often rely on English as an intermediate language, which can lead to the loss of meaning and context during translation. Given these findings, the researcher concluded that M2M-100 is a more suitable choice for machine translation tasks compared to other available models.

2.5.4 Training Arguments for Fine Tuning

Seq2SeqTrainingArguments in the Hugging Face Transformers library is a class designed to configure the training process of sequence-to-sequence (seq2seq) models (Wolf et al., 2020). It is a subclass of the more general TrainingArguments class, which handles a variety of training configurations and hyperparameters for different types of Transformer models. The Seq2SeqTrainingArguments class extends TrainingArguments to include additional hyperparameters specifically tailored for seq2seq models, such as BART, T5, and others.

Seq2SeqTrainingArguments allows to fine-tune the model's performance on specific tasks by changing hyperparameters such as learning rate, batch size, weight decay, and warm-up steps. The selection of hyperparameters can have a significant impact on the training process, convergence time, and final model performance. As a result, selecting proper hyperparameters is critical for reaching optimal outcomes in tasks such as dialogue summarization.

In addition to hyperparameters, the Seq2SeqTrainingArguments class contains options such as evaluation strategy, logging, and checkpoint saving, which allows to monitor and analyse the process of training more effectively.

The following training arguments were selected for the training of the modal for dialogue summarization task.

- evaluation strategy ='steps. Evaluating the model's performance periodically after a fixed number of training steps enables the monitoring of model convergence and prevents overfitting.
- learning rate=3e-5: A learning rate of 3e-5 is a common value used in fine-tuning Transformer models, as recommended by (Devlin et al., 2019) for BERT and subsequently adopted for other models such as RoBERTa and BART.
- per_device_train_batch_size=4 and per_device_eval_batch_size=4: Smaller batch sizes can provide more stable gradient updates (Keskar et al., 2017) and are often necessary due to memory constraints when training large Transformer models (Wolf et al., 2020).
- gradient_accumulation_steps: Gradient accumulation mitigates the effects of smaller batch sizes by accumulating gradients from multiple mini batches before performing a weight update, effectively simulating a larger batch size.
- weight_decay: Weight decay is a regularization technique that prevents overfitting by adding a penalty term to the loss function.
- save_total_limit: Limiting the number of saved checkpoints helps manage disk space during training, while still providing multiple snapshots of the model's progress for evaluation and analysis (Wolf et al., 2020).
- num_train_epochs: Training for multiple epochs allows the model to iterate over the dataset multiple times, potentially improving performance (Wolf et al., 2020). However, too many epochs can lead to overfitting, so it is essential to monitor validation loss.

- predict_with_generate: This setting enables the use of the model's generate() method for evaluation, which is necessary for tasks like dialogue summarization that require text generation (Wolf et al., 2020).
- fp16=True: Using mixed-precision training (FP16) can reduce memory consumption and accelerate training without significantly affecting model performance (Micikevicius et al., 2018)
- warmup_steps and lr_scheduler_type='linear': A linear learning rate scheduler with a warmup phase can help stabilize training and prevent large weight updates early in the training process (Vaswani et al., 2017)

2.6 Evaluation

2.6.1 Metrices

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric was introduced by (Lin, 2004) as a family of automatic evaluation measures for summarization tasks. These metrics measure the overlap between n-grams in reference summaries and generated summaries. ROUGE has become widely used in the evaluation of summarization models, including dialogue summarization. The most common variants are ROUGE-1, ROUGE-2, and ROUGE-L, which measure unigram, bigram, and longest common subsequence overlaps, respectively (Lin, 2004).

Liu et al.'s research (Liu and Liu, 2008) investigates the relationship between ROUGE and human evaluation in the context of meeting summarization. They utilized the Spearman's rank coefficient to assess the correlation, finding that it is generally low. One limitation of the ROUGE evaluation method is its surface-level comparison, which doesn't account for lexical and compositional diversity. Yet, the researchers emphasize that, while ROUGE is an important evaluation technique, it may not be the best one.

The BLEU (Bilingual Evaluation Understudy) metric was proposed by (Papineni et al., 2002) for the evaluation of machine translation systems. BLEU measures the precision of n-gram overlap between generated texts and reference texts. Although primarily designed for translation tasks, BLEU has been adapted for dialogue summarization evaluation. BLEU's main strength lies in its simplicity and ease of computation, but it has limitations in capturing certain aspects of summarization quality, such as fluency and coherence.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) was introduced by (Banerjee and Lavie, 2005) as another evaluation metric for machine translation. It computes the harmonic mean of unigram precision and recall, considering exact matches, as well as approximate matches based on word stems, synonyms, and paraphrases. Like BLEU, METEOR has been adapted for dialogue summarization evaluation, offering a more nuanced measure of summary quality.

BERTScore – BERTScore is a text evaluation technique based on the BERT model created in by (Devlin et al., 2019). For comparative purposes, this technique computes weighted cosine similarity of embedded representations. In (Liu et al., 2019) demonstrated the utility of employing BERTScore as an evaluation metric for dialogue summarization by contrasting it with the well-known ROUGE metric. The researchers used the unweighted version of FBERT to fine-tune a pre-trained dialogue summarization model and BERTScore to evaluate the abstractive dialogue summarization.

BLEURT – by (Sellam, Das and Parikh, 2020) BLEURT addresses the challenge of evaluation metrics exhibiting weak correlation with human judgment in the context of dialogue summarization. Based on the BERT model, BLEURT is capable of modeling human judgment and utilizes a pre-training scheme with millions of examples. According to the researchers, unsupervised training and fine-tuning with human evaluation can result in an expressive evaluation score for dialogue summary.

Human judgment is still the best way for evaluating text summary. To evaluate their models, (Liu and Liu, 2008) used a combination of ROUGE and human evaluation. Human evaluation is required to guarantee that the underlying context is recorded, and the summary is computed correctly in order to acquire the most conclusive results for computer-generated summaries. However, this method is time-consuming and difficult to apply to huge datasets. Benchmarking gets more challenging with human review as a more qualitative tool.

2.6 Chapter Summary

In this part of the thesis offers an overview of the domain of dialogue summarization using relevant research papers and their methodologies to understand existing research gaps.

CHAPTER 3: METHODOLOGY

3.1 Chapter Overview

This section of the thesis covers the different methodologies of the project. The reasonings behind the selected methodologies and early identified risk and author's plan to mitigate them.

3.2 Research Methodology

Research Philosophy	From positivism and interpretivism, the author chooses **pragmatism** as the research philosophy. The selection was determined after considering evaluation methods and previously published research that included quantitative and qualitative data.
Research Approach	This research aims to evaluate and prove the hypothesis. Therefore, the development of cross-language transfer learning for low-resource settings can be a backbone for future research. A **deductive** approach was chosen as the research broadly applies existing theories to the domain of interest.
Research Strategy	The research strategy defines the methodology by answering the research questions. **Surveys** are considered for the research strategy as the primary option. Interviews will also be used as a secondary form of data collection, such as evaluation and feedback, during the final part of the research project.
Research Choice	The **mixed method** was chosen from among the mono, multi, and mixed research methods because this research will use both

	quantitative and qualitative data, such as surveys, performance values, and feedback.
Time Horizons	The time horizon determines the duration of the investigation. The **cross-sectional** time horizon was chosen from the available two since the data for this study will be collected at a single point in time.

Table 2 – Research Methodology

3.3 Development Methodology

3.3.1 What is the development methodology.

A **prototype** model was selected for the research project for the development model. To justify, this research project is developed on a prototype, then the prototype will be tested, and necessary modifications are applied until the desired result is achieved.

3.3.2 Requirement elicitation methodology

Among various requirement elicitation methodologies such as the survey, observation, literature review and brainstorming, for this research project, survey and literature review were selected. Literature reviews and surveys will be used to gather requirements for the development of the prototype.

3.3.3 Design methodology

In this research project, object-oriented analysis and design (OOAD) was selected as the design methodology to reuse the components to extend the system.

3.3.4 Evaluation methodology

Evaluation results are one of the outcomes of research. It can determine the effectiveness of the research findings. Evaluation metrics and benchmarking evaluation approaches will be used to evaluate the multilingual dialogue summarization.

3.3.4.1 Evaluation metrics

The following metrics based on previous research will be used to evaluate the quality of the generated dialogue summaries.

ROUGE(Lin, 2004) – most widely used evaluation metric in Summarization. Further, this can be expanded as ROUGE-1, ROUGE-2, ROUGE-L

BLEU(Papineni et al., 2002) – primary evaluation for machine translations. This can be used to evaluate generated text.

3.3.4.2 Benchmarking

Based on previous research benchmarking for dialogue summarization has been done using many benchmarking datasets. For this multilingual dialogue summarization, ClidSum benchmarking will be used. (Wang et al., 2022)

3.4 Project Management Methodology

Among different project management methodologies, **Prince2** was chosen for this research project as the project scope, and other stages have already been recognized.

3.4.1 Schedule

3.4.1.1 Gantt chart

Moved to Appendix.

3.4.1.2 Deliverables and dates

Deliverable	Date
Project Proposal Document **The initial proposal of the research project.**	7th November 2022
Ethics Form	7th November 2022
Literature Review Document **A critical review of existing work and solutions.**	27th October 2022
Project Specification Design and Prototype Document **A document specifying the approach of the multilingual dialogue summarizer.**	2nd February 2023
Prototype **A prototype of the research, which includes the multilingual dialogue summarizer**	2nd February 2023

Draft Project Report **A draft version of the final report to get feedback from the supervisor.**	20th March 2023
Final Thesis	27th April 2023

Table 3 – Deliverables and Dates

3.5 Resource requirements

3.5.1 Software resources

- Selection of operating system (Windows/macOS/Linux Distro) – Windows will be the default selection for the development because of overall software availability.
- Selection of base programming language (Python/R) – Python will be the main programming language for the project because of the availability of open-source packages and libraries.
- System Front-end programming language (Angular/React) – The angular framework will be used to develop the front-end application.
- IDE (Jetbrain PyCharm / Visual Studio) – Jetbrains PyCharm and Visual studio code will be used to support the development environment.
- Cloud-Based Development (Google CoLab/Azure Notebooks) – A cloud-based development environment to develop and train models. Google CoLab will be chosen as it has more RAM by default than the Azure Notebooks.
- Reference Manager (Mendeley/Zotero) – Research management tool to save and back up research resources in a cloud-based environment. Mendeley is chosen as it has both web and desktop versions.
- Cloud Storage to save files (Google Drive/One Drive/Dropbox) – To store necessary files related to this research project. One drive will be the primary selection as it can easily integrate with word applications. Also, google drive will be used as a backup.
- Source Code Management (Github/Gitlab/Bitbucket) – To manage and back up the code related to the research project. GitHub is chosen as it has more available tools than other source code management platforms.
- Continues Integration and Development Services (Amazon AWS, Microsoft Azure) – In order to continuously integrate changes in the implementation of the system a CI/CD pipeline service will required. Amazon AWS is selected as it has more options to choose.

3.5.2 Hardware resource

- Core i5 8th Gen processor – To execute high CPU resource-intense tasks.
- 50GB Disk space – To store datasets and models for local development.
- 16GB Ram or More – To perform high memory-intensive tasks.

3.5.3 Data requirements

- Available datasets on high-resource languages required for the project.

3.5.4 Skills requirements

- Skills for understanding mathematical equations.
- Knowledge of natural language processing.
- Knowledge of model evaluation methods.
- Creative report writing skills.

3.6 Risk and mitigations

These are the identified risk for this research project with plans for mitigations.

Risk	Probability of Occurrence	The magnitude of the loss	Mitigation Plan
Deep knowledge of language models and cross-lingual transfer learning theories.	5	5	Courses and online resources are available to improve knowledge in these areas.
Updating the resource requirements of the research project	4	2	Using the prototype development approach will be able to identify resource requirements.
Chance of exceeding the free tier in cloud services such as google CoLab and Amazon servers.	3	4	Analyse the available options and select the best plan for the development of the project.

File corruption or losing access to backup files	2	5	Share the backup folder with another personal account in case of losing access to one, so it will be accessible by the backup account.
Inability to conduct the research work due to sickness or illness	1	5	Have a record of project outcomes and maintain good documentation.

Table 4 – Risk and Mitigations

3.7 Chapter Summary

In this part of the thesis concludes the development and project management paradigms and the reasonings behind the selection for this research project. Also, the author's action plans for early identified risk and how to mitigate them.

CHAPTER 4: SOFTWARE REQUIREMENT SPECIFICATION

4.1 Chapter Overview

This section of the thesis focusses on determining various stakeholders and their interaction to the proposed system, a rich picture diagram and the different requirement gathering methodologies with results. Finally possible use cases with the system, functional and non-functional requirements of the prototype.

4.2 Rich Picture Diagram

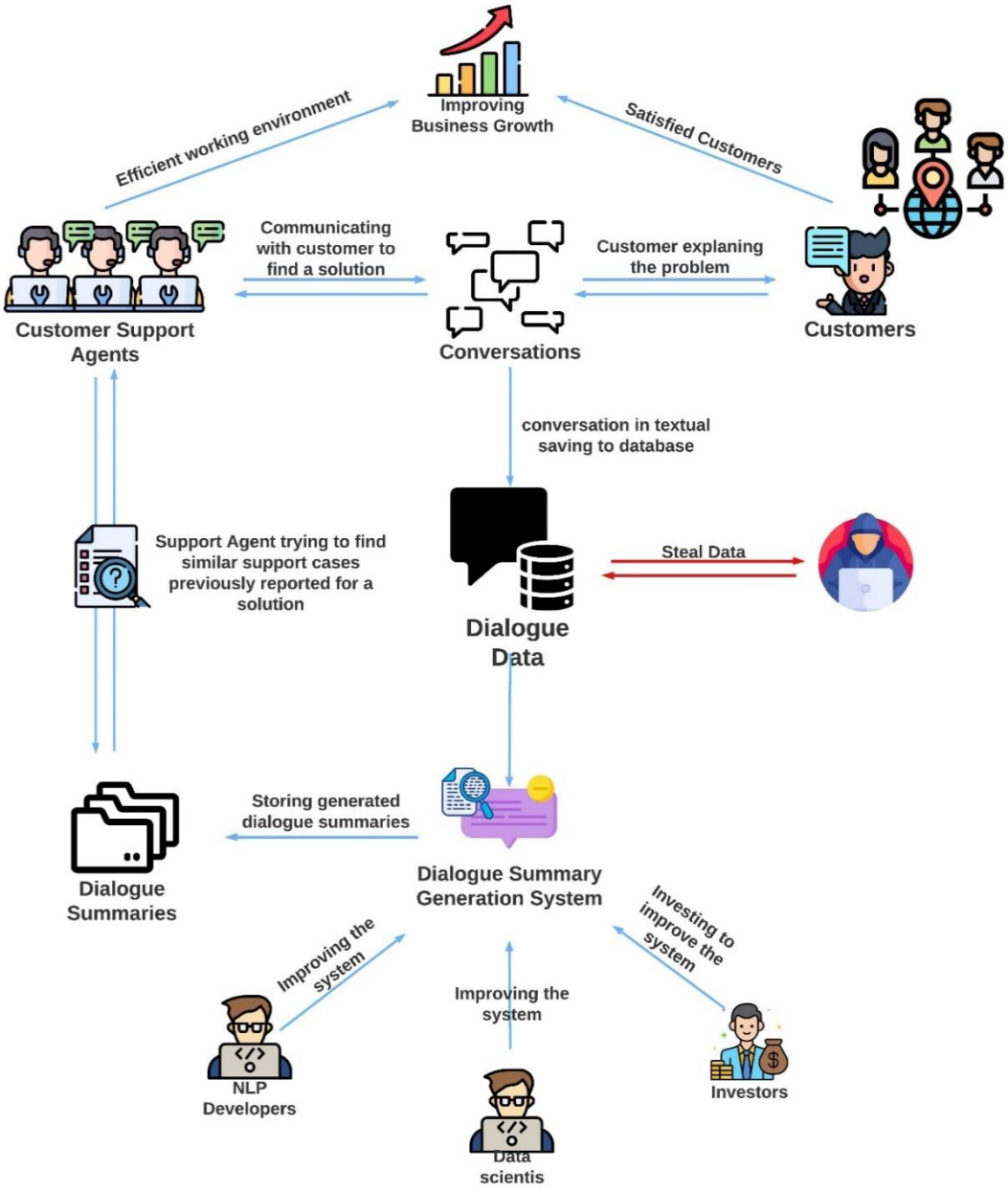


Figure 3 – Rich Picture Diagram

4.3 Stakeholder Analysis

4.3.1 Stakeholder Onion Model

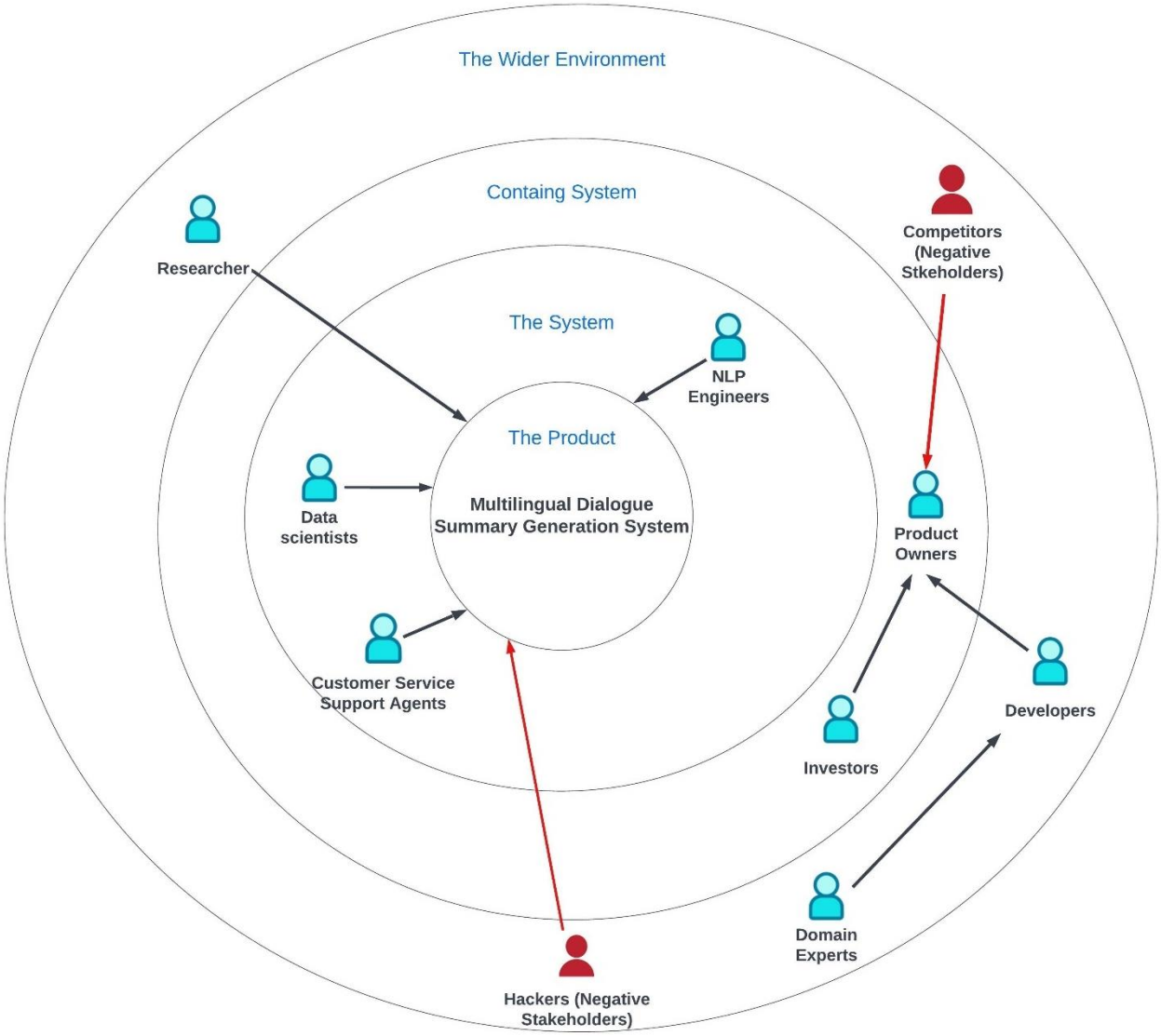


Figure 4 – Onion Model

4.3.2 Stakeholder Viewpoints

Stakeholder	Role	Description
Data scientist, NLP Engineers	Operational Maintenance	Develops the system using tools.
Customer Service Support Agent	Functional Beneficiary	Uses the developed system to summarize dialogues.
Product Owners	Functional Beneficiary	Owner of the dialogue summary generation system.

Investors	Financial Beneficiary	Invest to make profit and support the future developments and improvements.
Competitor	Negative Stakeholder	Creates a system that directly contest with the proposed system's features
Hacker	Negative Stakeholder	To disrupt the system and the data it holds.
Technical Experts	Expert	Decides whether the product is capable of handling set of requirements from a technical perspective.
Researcher	Advisor	Analyze the existing solutions for research purpose.

Table 5 – Stakeholder Viewpoints

4.3.3 Stake Holder Grouping

Stakeholder Groups	**Stakeholder**	**What is the method that you will be used to gather requirements**
Group 1	Customer Service support agents (End-User)	Questionnaires
Group 2	Data scientists and NLP researchers	Interviews
Group 3	Competitors (Existing work)	LR

Table 6 – Stake Holder Grouping

4.3.4 Data gathering instruments.

Questionnaire for Group 1

Question	**Research/Relevance to the research**	**Research Question**

Do your organization uses dialogue summarizations?		RQ1
What is the current technique you are using for dialogue summarization?	Study the currently available solution in the industry.	RQ1
How often are summaries being used for later reference	Impact of dialogue summarization in customer services.	RQ1
Do you summarize all the dialogues?	Identifying the specific attributes that will be considered for dialogue summarization.	RQ1
If you selected "No, only if it is required" for previous questions, select the specific attributes considered to summarize dialogues.	Identifying the specific attributes that will be considered for dialogue summarization.	RQ1
How many languages do you support in customer service?	Overview of language support in customer service.	RQ1, RQ3
Do you find it challenging to use currently available solutions to summarize dialogues that are not in English?	Understanding limitations in current solutions from the end-user's perspective.	RQ1, RQ2
How much do you think it will be helpful to summarize dialogues which are not in English languages?	Identifying the end-user's requirement for the proposed system	RQ2, RQ3
Will you be interested in a solution that can summarize the dialogues between customer support and the customer?	Identifying the end-user's requirement for the proposed system	RQ2, RQ3
If you have any suggestions, please mention them below.	Identifying the end-users' suggestions based on prior experience with existing solutions.	RQ1

Table 7 – Questions for Group 1

Interview Questions for Group 2

Question	**Research / Relevance to the research**	**Research Question**
What are the current limitations of using text summarization techniques directly to dialogue summarization	Understanding the current limitations in the dialogue summarization domain.	RQ1, RQ2, RQ3

Recent advancements in cross-lingual transfer techniques in the NLP domain	An exploratory study on cross-lingual transfer techniques.	RQ2
How effective are cross-lingual transfer techniques for low-resource language use cases?	Identifying the practicality of using cross-lingual transfer techniques	RQ2, RQ3

Table 8 – Questions for Group 2

4.4 Selection of Requirement Elicitation Methodologies

To obtain requirements for this research project, different requirement elicitation approaches were used. This is accomplished through the use of a literature study, an interview, a survey, and prototyping.

Method 1: Literature Reviews
In the beginning of the project, the author conducted an analysis of the existing research projects related to the chosen domain to identify a solid research gap. The existing system and the technologies were thoroughly studied for this research project. These finding are mentioned in literature.
Method 2: Interviews
Collecting knowledge to the research body and to get qualitative feedback on the proposed solution, to get insights from the domain specific experts several interviews were conducted. By using this method, it supports the researcher to early identify the challenges of the proposed solution when it comes to development of the prototype.
Method 3: Survey
A questionnaire was used to gather requirements from the end user's perspective. This will help to identify the requirements and the problems with a existing solutions that the end user is facing. These data will be helpful to improve the proposed solution to become more practical.
Method 4: Prototyping
Selected software development life cycle for this research project is agile as it would help to improve the proposed system in a recursive development method.

Table 9 - Selection of Requirement Elicitation Methodologies

4.5 Discussion of Findings

4.5.1 Literature Reviews

Finding	Citation
The proposed method shows that latent variables cope with the variance of semantically similar sentences across different languages. Cross-lingual transfer between English to Spanish and English-to-Thai have demonstrated state-of-the-art results.	(Xiang et al., 2021)
The proposed framework can generate an abstractive summarization with limited amount of parallel resources by sharing a unified decoder that generates both monolingual and cross-lingual summaries	(Bai, Gao and Huang, 2021)
This proposed approach has shown the state-of-the-art models are very sensitive to language shift through automatic translation and combining training data for the two languages (English – Italian) is beneficial.	(Labruna and Magnini, 2021)
This proposed method uses a novel Leader-writer network with auxiliary key point sequences, ensuring the generated summary is logical and integral.	(Liu et al., 2019)

Table 10 0 Literature Reviews Findings

4.5.2 Interviews

To get opinion on the proposed solutions and the research area both research experts from dialogue summarization and domain experts from the customer service area were chosen. The interviews were conducted as open-ended questions therefore the output of these interviews was documented based on thematic analysis.

Codes	Theme	Conclusion

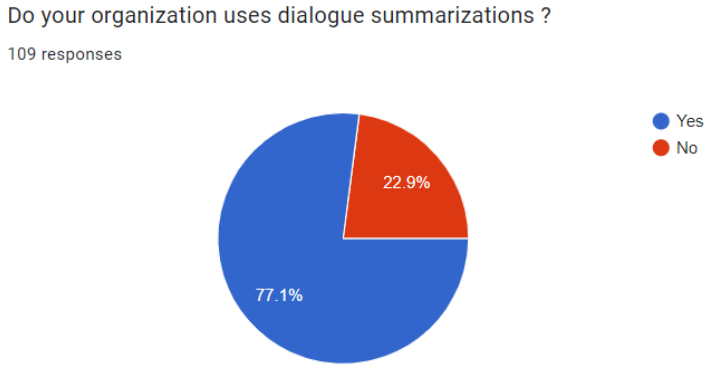
• Slow-moving • Utilized existing resources. • Missing context	Existing limitations in dialogue summarization domain.	All the participants were mentioned about the current text summarization techniques and why it cannot directly apply to the dialogue summarizations. Because the nature of the dialogues constructive and logical text data, keeping the context while summarizing is challenging compared to text summarization. According to the experts they suggest to fine tune text summarization techniques and use of pre-trained language models to build the base stage for dialogue summarizations rather than developing from the scratch.
• Not well explored. • Transfer learning	Integrating cross-lingual transfer techniques for low resources language.	Most of the cross-lingual transfer techniques are not yet explored in the dialogue summarization domain. Due to the minimum number of datasets that are currently available, use of cross-lingual transfer techniques will be very useful. Experts suggest to use of a cross-lingual transfer modal which can be used as a intermediate process for machine translation when developing the solution.

• Research gap • Contribution	Research gap and Scope	The opinion from the research domain experts were that the research gap and the proposed solution is an innovative use of existing methods to overcome the problem.
• Globalized customer base. • Human resources. • Practicality	Understanding the use of dialogue summarization in a practical use case.	As from the experts from the problem domain, they suggested that the use of supporting to more languages is very useful as one of the problems they are facing is the human resources cost that is required when it comes to manually summarizing the dialogues. Throughout the interviews the customer supports that have globally expanded have this as a major issue when comparing with services that are only available in English.
• Save work. • Later refer. • Generate Insights	Features of the proposed system.	The proposed system should be capable of saving the generated summaries and allow the users or the system administrators to refer later. As from the experts they mentioned that summaries can be used for various purposes later such as categorize and analyze to get new insights related to customers issues.

Table 11 – Interviews Thematic Analysis

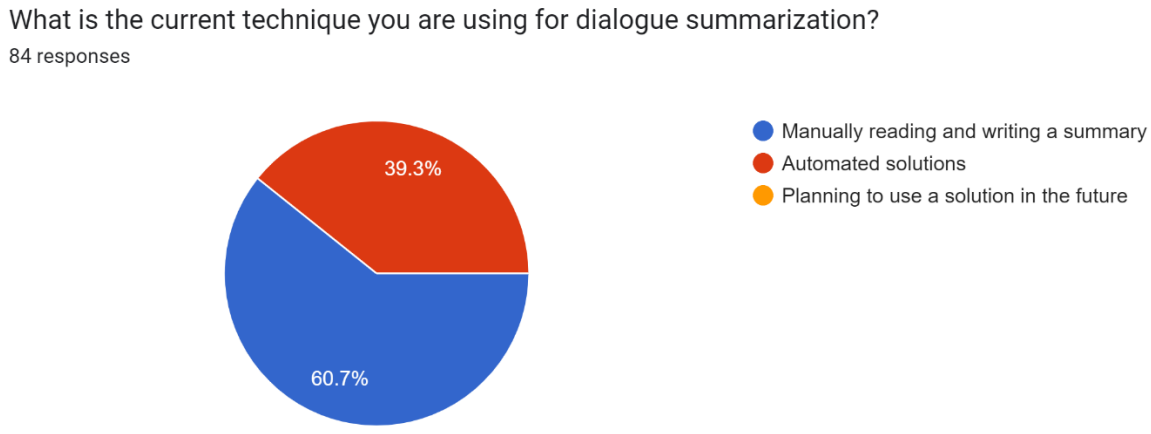
4.5.3 Survey

Question	Do your organization uses dialogue summarizations?
Aim of Question	To understand the awareness of these tools among the end users.



The majority of the participants are aware and currently using dialogue summarization within their organizations. Lesser number of responses shows that not all the customer services are not using dialogue summarizations.

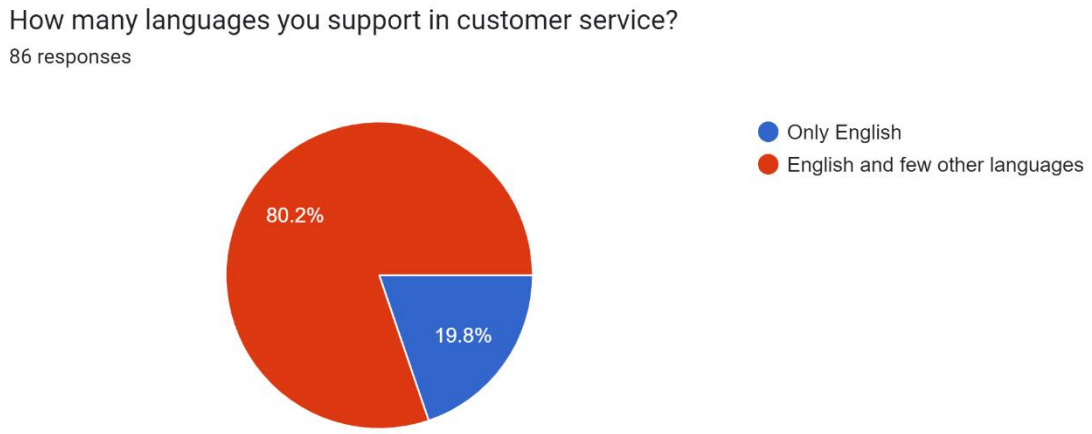
Question	What is the current technique you are using for dialogue summarization?
Aim of Question	Study the currently available solution in the industry.



From the gathered responses it clearly shows that the majority of the users are summarizing dialogues by manually. 39.3% of users are already using automated solutions for dialogues summarization. Observation from the this can be considered that the awareness of the

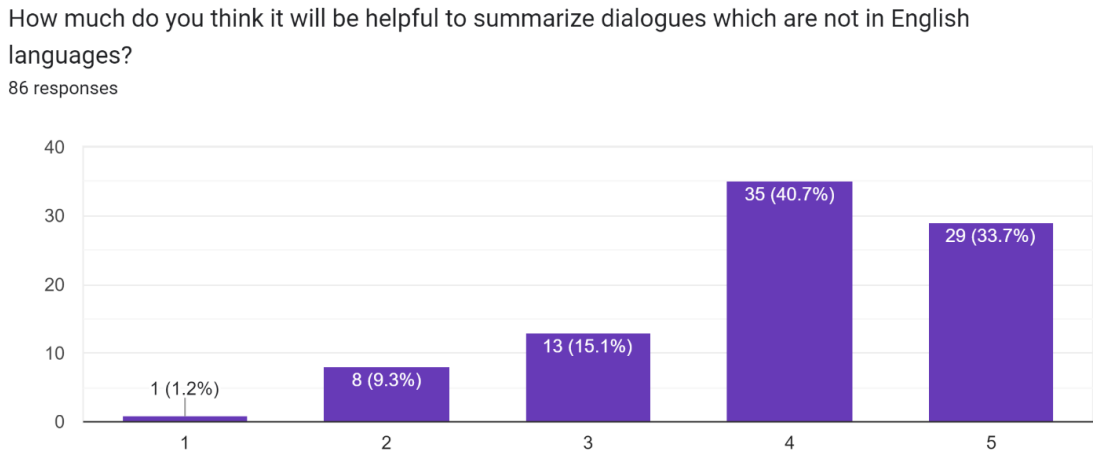
dialogue summarization solutions are not yet expanded in the industry. This can be a major reason that the tools are not yet well developed or in the early stages.

Question	How many languages you support in customer service?
Aim of Question	Overview of language support in customer service.



Majority of the customer services supports more than English language. This can be potential requirement as dialogue summarization that it should not be limited for English.

Question	How much do you think it will be helpful to summarize dialogues which are not in English languages?
Aim of Question	Identifying the end-user's requirement for the proposed system



From the responses that more than 60% of users are interested in the proposed solution which can be helping to solve the existing problems with the current solutions or to replace the current solutions.

Table 12 – Survey Results

4.5.4 Prototyping

Criteria	Findings
Developing core component	Use of transformers and their capabilities needs to be identified clearly. Different transformers have different capabilities
Applying cross-lingual transfer techniques	Final output of the summary must be clear. Context cannot be changed during the transfer process as it may lead to different meaning.

Table 13 - Prototyping

4.5.5 Summary of Findings

Finding	Literature Review	Interviews	Questionnaire	Prototyping
The proposed system will help to benefit users who are using existing solutions and users with no prior		✓	✓	✓

use of dialogue summarization solutions.				
The limitation in dialogue summarization systems can be pushed by using pre-trained language models and cross lingual transfer techniques.	✓	✓	✓	
Identified research gap would contribute to dialogue summarization research domain.	✓	✓		
Within the system summaries should be stored and allow user or administrators for later reference.		✓		

Table 14 – Summary of Findings

4.6 Context Diagram

Before beginning development, the scope of the proposed system and its interactions with internal and external components should be determined. The context of the proposed system is illustrated in the diagram below.

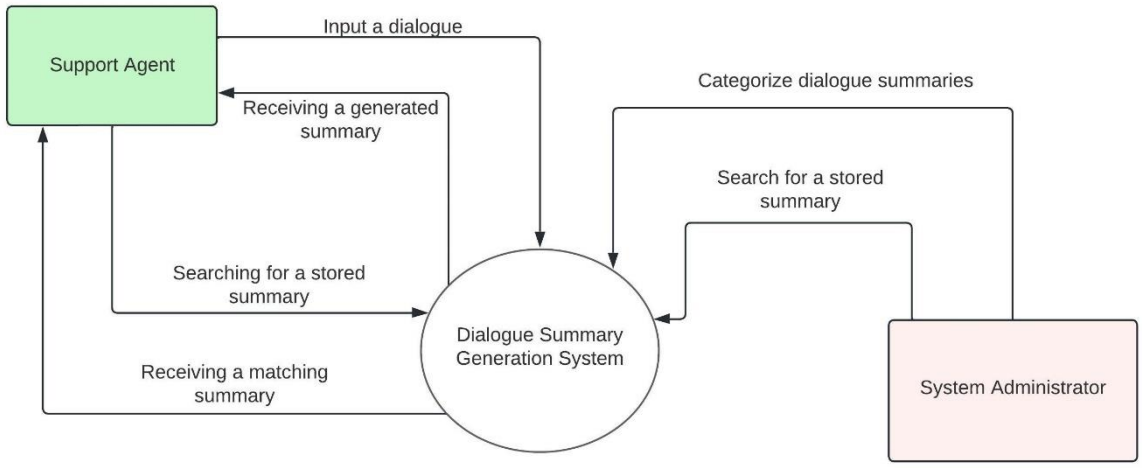


Figure 5 – Context Diagram

4.7 Use case Diagram

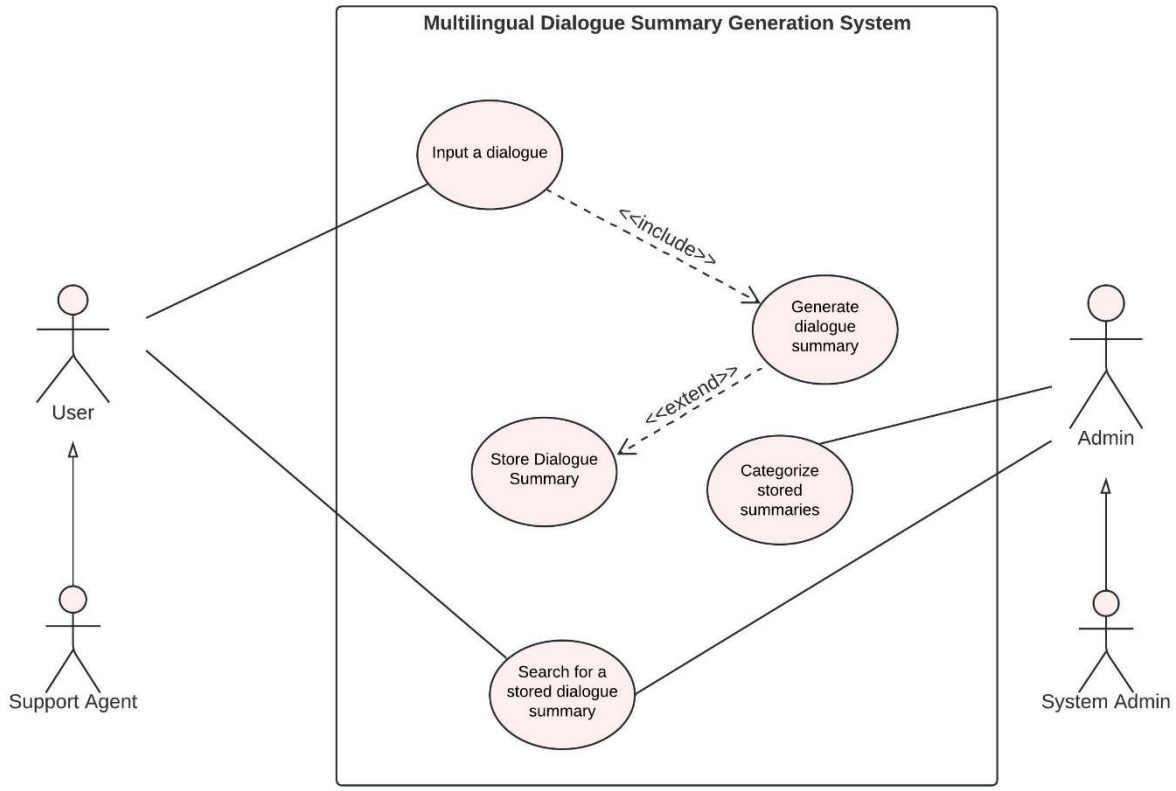


Figure 6 – Use case Diagram

4.8 Use case description

Use Case: UC1	Generate Dialogue Summary
Description	Generate a summary for the dialogue input
Primary Actor	Support Agent (User)
Pre-condition	Textual dialogue data should be in a valid input format.
Post condition	Generated summary should be displayed to the user.
Trigger	A support agent wishes to generate a summary for the dialogue which he/she had with the customer.
Success scenario	• A summary for the input data will be generated. • Generated summary will be stored in the system.

Table 15 – Use Case 1

Use Case: UC2	Search for a stored dialogue summary
Description	Find a previously generated summary with matching search results.
Primary Actor	Support Agent (User), System Admin
Pre-condition	Previously generated summaries should be saved on the system.
Post condition	Matching summaries should be displayed.
Trigger	A support agent / system admin wishes to search for previously generated summaries.
Success scenario	• Matching summaries will be displayed. • If there are no matching summaries will display a message as 0 results.

Table 16 – Use Case 2

Use Case: UC3	Categorize stored summaries
Description	Categorize summaries based on their similarities.
Primary Actor	System Admin
Pre-condition	Previously generated summaries should be saved on the system.
Post condition	Summaries should be categorized.
Trigger	A system admin wishes to categorize summaries based on their similarities to generate insights.
Success scenario	• Stored summaries will be categorize based on their similarities.

Table 17 – Use Case 3

4.9 Requirements

4.9.1 Functional Requirements

In order to prioritize the levels of the system requirements the MoSCoW principle was used.

Priority Level	Description
Must have (M)	This priority level indicates the core functional requirements and must be implemented.
Should have (S)	This priority level indicates the requirements that are not necessary but do add a great value.
Could have (C)	This priority level indicates the optional requirements
Will not have (W)	This priority level indicates the requirements that are out of the project scope.

Table 18 – MoSCoW Principle Priority Level

FR ID	Requirement	Priority Level	Use Case
1	User must be able to generate a summary for an inputted dialogue data at least for 1 language option.	M	UC1
2	Summary of the dialogue should be represented to the user.	M	UC1
3	User should be able to generate summary for a dialogue data with multiple language options.	S	UC1
4	The system should store the generated dialogue summaries and allow users to view the summaries	S	UC2
5	The system should allow the users to search stored summaries with matching keywords.	S	UC2
6	The system should allow the admins to categorize the stored summaries based on their similarities	C	UC3

Table 19 – Functional Requirements

4.9.2 Non-functional requirements

NFR ID	Requirement	Description	Priority Level
1	Quality of the Output	The quality of the generated summary should be clear and meaningful as much as possible.	M
2	Performance	Time to generate a summary should be acceptable.	S

3	Security	The system should prevent any data breaches from attackers to keep the information safe.	S
4	Usability	The system should be easy to use followed by good user interface and user experience principles.	M

Table 20 – Non Functional Requirements

4.10 Chapter Summary

In this part of the thesis covers the overview of the system and how it connects with the different parties have been displayed using the rich picture diagram. Sounder's onion model was used to display the stakeholders and how they connect with each other. In order to get opinions for the proposed solution different requirement elicitation methodologies were followed using the stakeholders. In the latter part the use cases of the system, functional and nonfunctional requirements were documented based on the input from the data gathering results.

CHAPTER 5: SOCIAL, LEGAL, ETHICAL AND PROFESSIONAL ISSUES (SLEP)

5.1 Chapter Overview

This section of the thesis includes the social, ethical, professional, and legal concerns while continuing this research project and how to reduce them to under the BSC code of conduct.

5.2 SLEP Issues and Mitigation

Social	Legal
• The interviewees' identities were not revealed in the thesis, and they were first informed that their contributions would be incorporated in the study work. • Information related to the interviewees were not used for political, religious other misbehaviors.	• Publicly available datasets were used for the development of the research project and have given credit for the original authors. • Tools and the developments platform were used for the implementation of the application is paid and no piracy of other illegal methods are used.
Ethical	**Professional**
• Participants were informed at the beginning of the survey, how their input will be considered for the research project.	• Academic standards and necessary guidelines were followed throughout the research project. • Interview process if conducted in a professional manner where interviewees are not manipulated to gather information.

Table 21 – SLEP Issues and Mitigations

5.3 Chapter Summary

In this part of the thesis describes the social, ethical, professional and legal concerns related to this research project and the action that are taken to mitigate them. The main objective of this analysis is to align the research project with the BCS Code of Conduct.

CHAPTER 6: DESIGN

6.1 Chapter Overview

This section of the thesis includes the all the design decision that were made to develop the proposed system. The authors' main design goals, overview of the architecture and both high-level and low-level design and wireframes will be discussed.

6.2 Design Goals

Design Goals	Description
Correctness	The output summary should be accurate and should not drop the context of the dialogues. The proposed system should be able to summarize the dialogues with similar to human annotated summaries.
Performance	The system should be capable of handling lengthy dialogues and optimized to generate summaries within a short period of time.
Usability	One of goal of the proposed system is to minimize the human resources that will take to summarize the dialogues manually, so it should be user friendly and be able to work with minimum effort.
Testability	The system should be divided into components within the development process to test and identify errors in the early stages.
Scalability	In the production environment there will be more users using the system and the workload should be handled to continue the process.

Table 22 – Design Goals

6.3 High-Level Design

6.3.1 Tiered Architecture

The tiered architecture diagram below provides an overview of the system architecture. The three-tier architecture organizes the presentation, logic, and data layers.

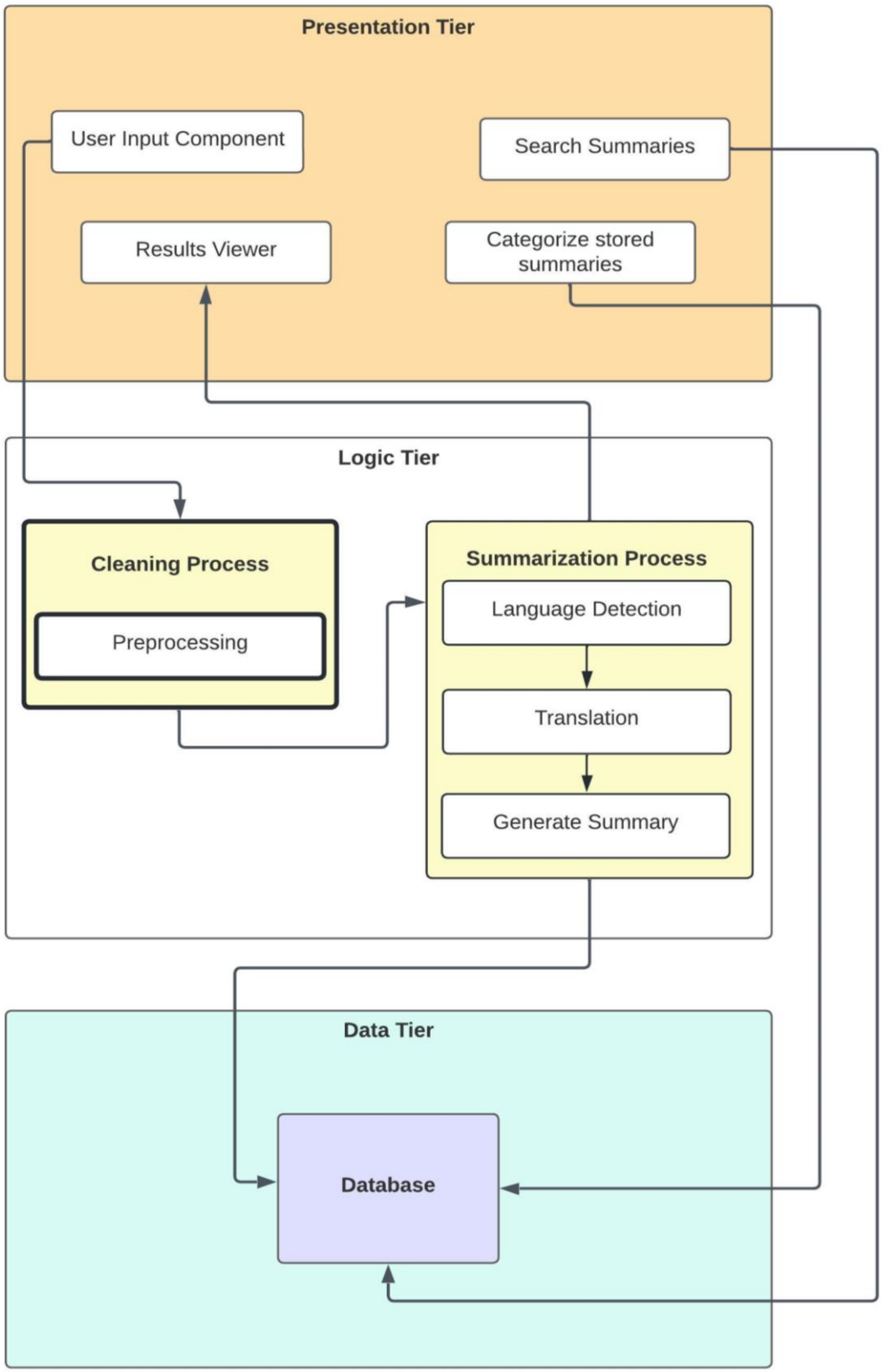


Figure 7 – Tiered Architecture

The above composed diagram is represented in a modular form which will be easy to understand, apart from the above diagram a backend services will be used to communicate among each layer.

6.3.2 Discussion of tiers

Data Tier

The data tier will consist of the database which will be used to store the generated dialogue summaries. Also, the user logins will use the database for authentication purposes to ensure system security.

Logic Tier

The logic tier will be responsible for processing the user input before putting through the summarization process. User input will be pre-processed by removing stop words etc. Then the processed input will be passed through the language detection modular. Machine translation will modular will go through process where it translates the input while minimizing the loss of context from the original user input. Then finally the summary will be generated.

Presentation Tier

User input component is where the end-user will input the dialogue in textual format. Results viewer will display the generated summary to the user. Search dialogue summaries will enable the user to search through previously generated summaries with matching keywords from the database.

6.4. System Design

6.4.1. Choice of the Design Paradigm

After the requirements gathering through multiple requirement elicitation methodologies Author has selected the **SSADM** (Structured Systems Analysis and Design Method) for the development of the proposed prototype because the requirement for the proposed system is well defined.

6.5 Design Diagrams

6.5.1 Data Flow Diagram

Level 01

Level 1 Data flow diagram represent the flow of data through different process within the system.

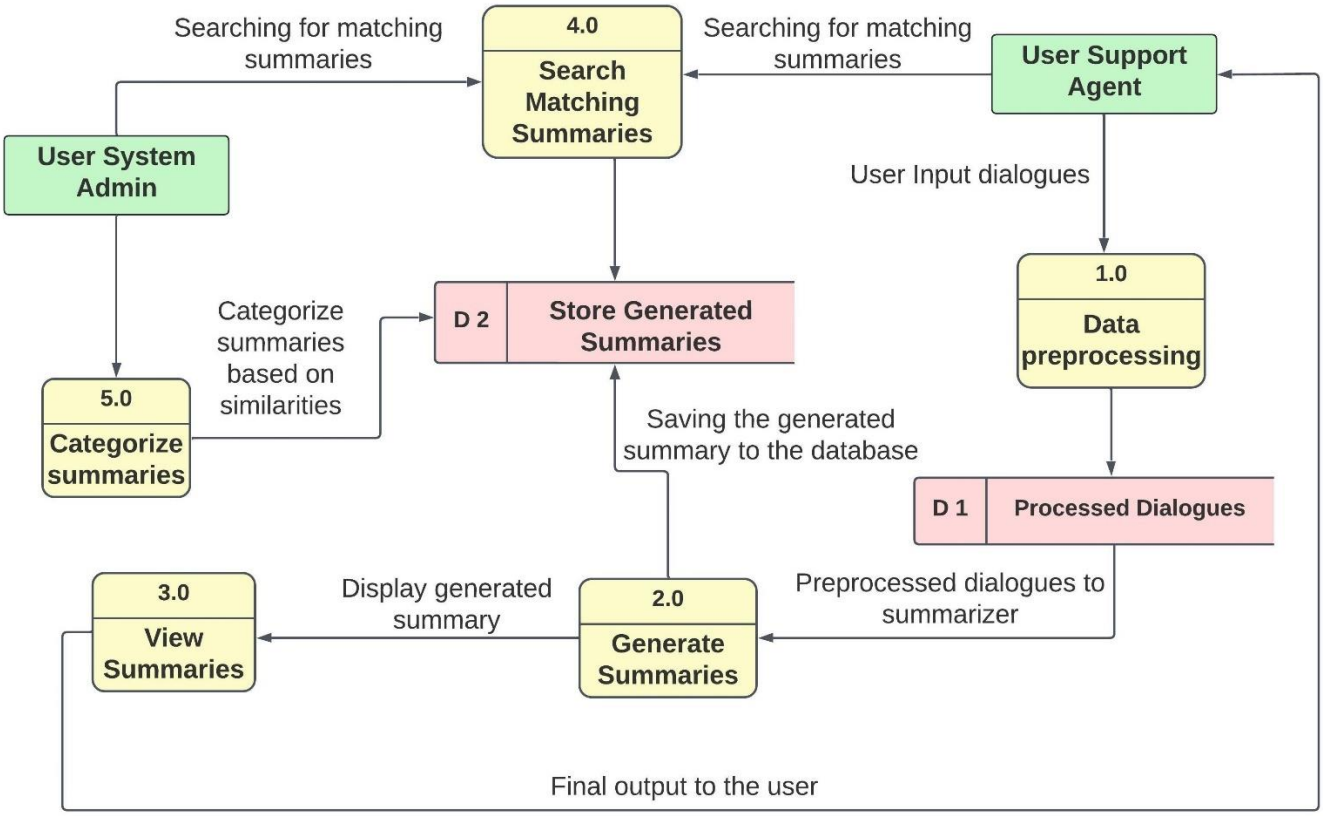


Figure 8 – Data Flow Diagram Level 01

Level 02

Level 2 Data flow diagram represent more expanded breakdown of the level 1 data flow diagram.

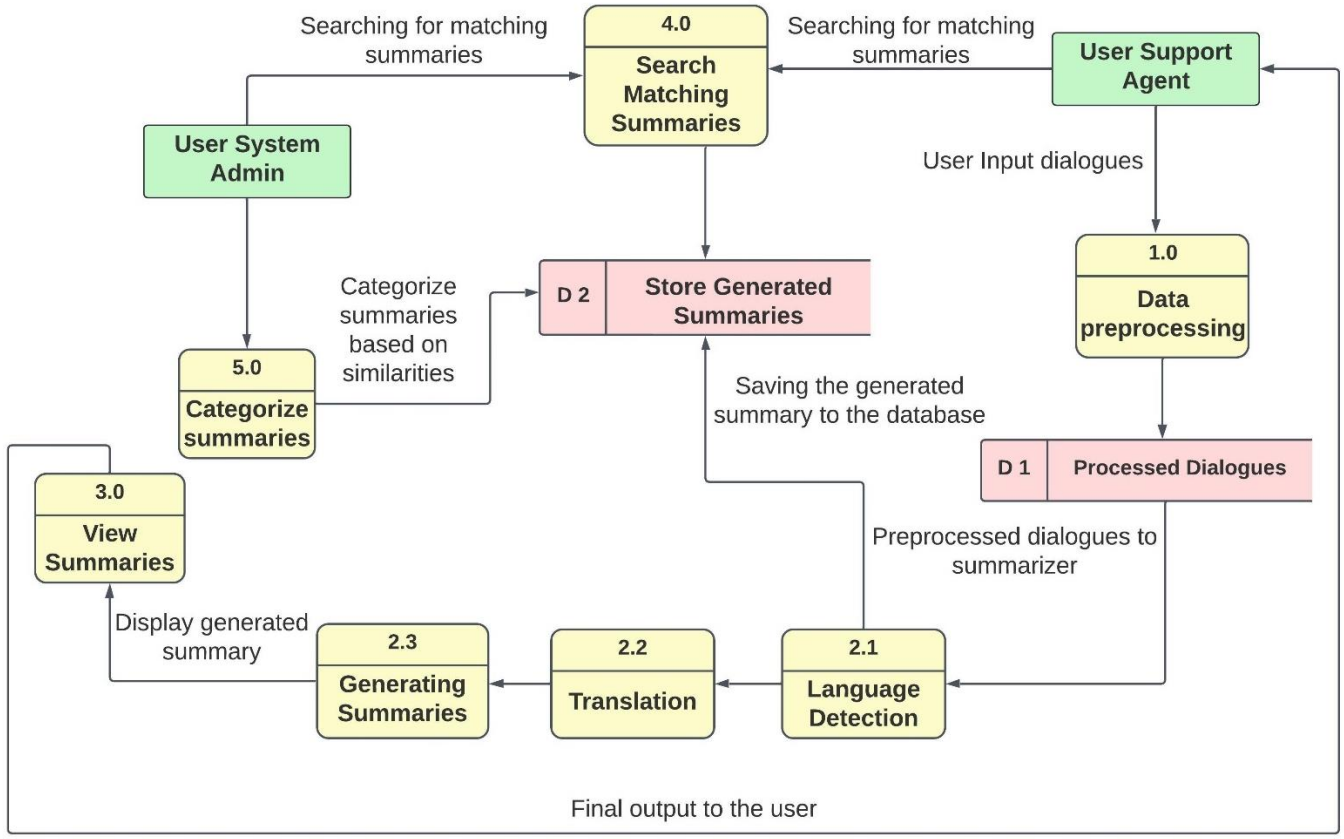


Figure 9 – Data Flow Diagram Level 02

6.5.2 System Process Flow Chart

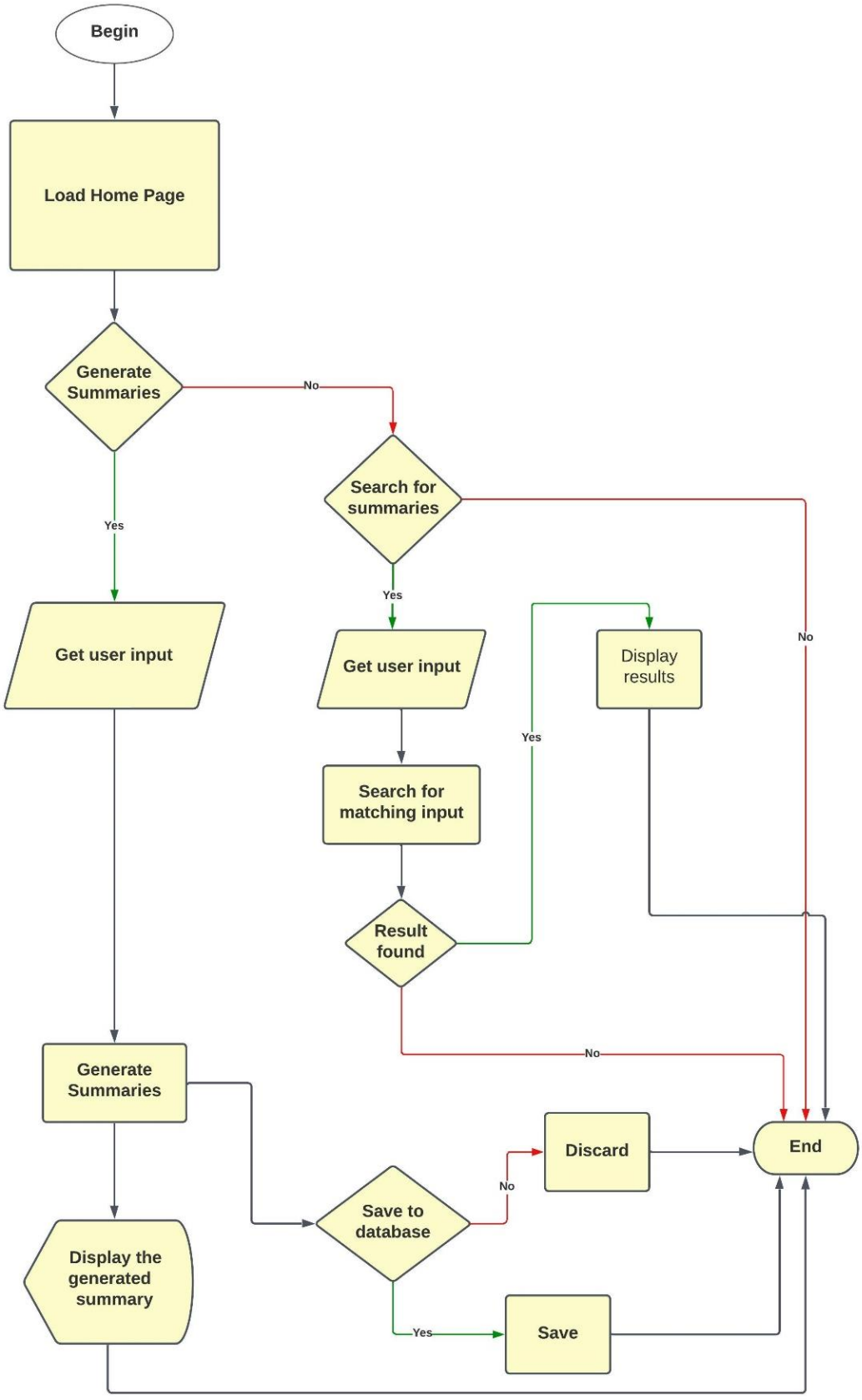


Figure 10 – System Process Flow Chart

6.5.3 User Interface Design

1. Low level fidelity wireframe diagram

Figure 11 – Low Fidelity Wireframes

2. High level fidelity prototype

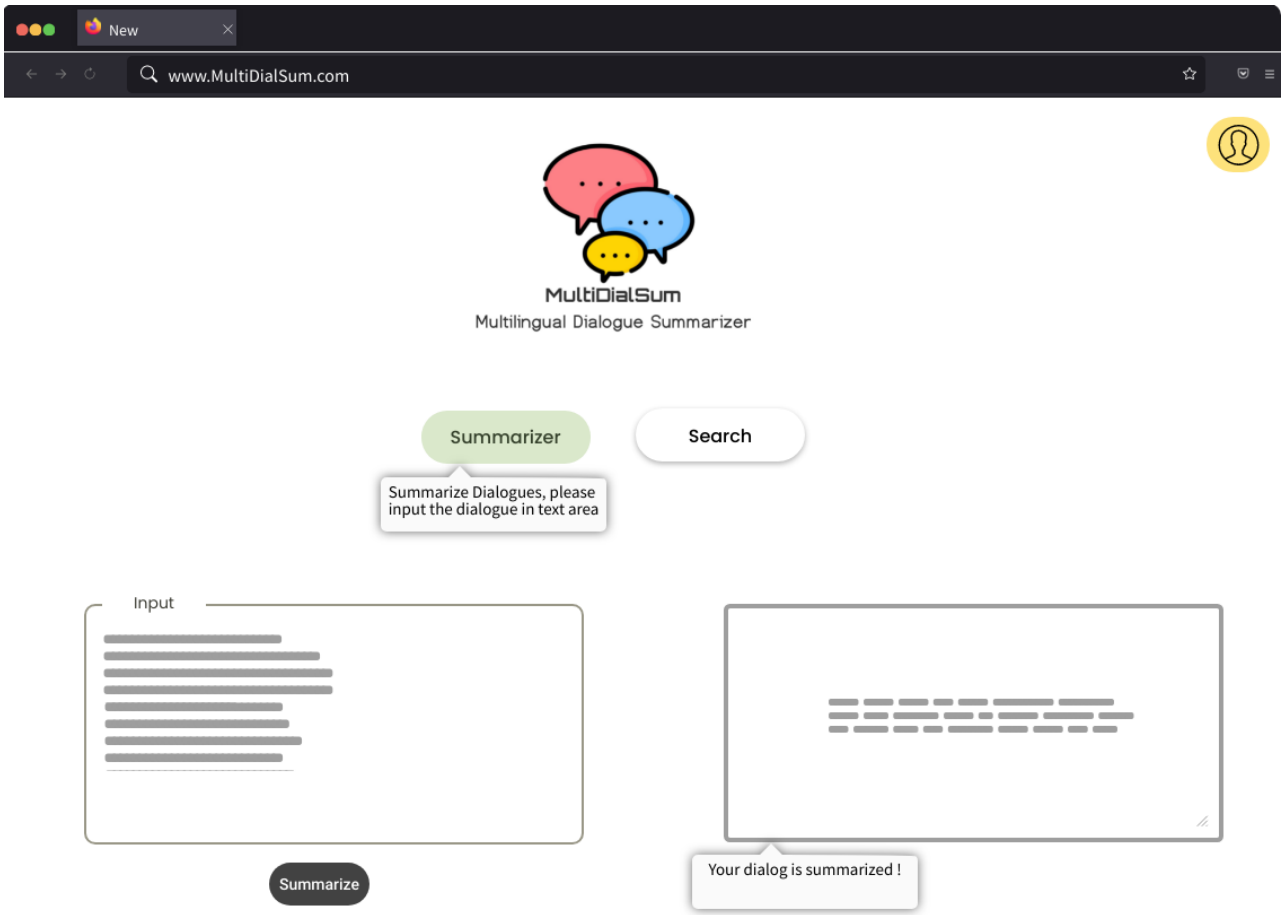


Figure 12 – High Fidelity Prototype

6.6 Chapter Summary

In this part of the thesis, design of the proposed system and the architecture of the system were discussed, also the overview of the expected UI for the system is explained using wireframe diagrams.

CHAPTER 7: IMPLEMENTATION

7.1 Chapter Overview

This section of the thesis consists of the technology stack, data set selection and other information related to the development of the prototype. Reasons for selecting those specific attributes will be discussed.

7.2 Technology Selection

7.2.1 Technology Stack

The selected technologies for the proposed system in every layer are represented in the diagram below.

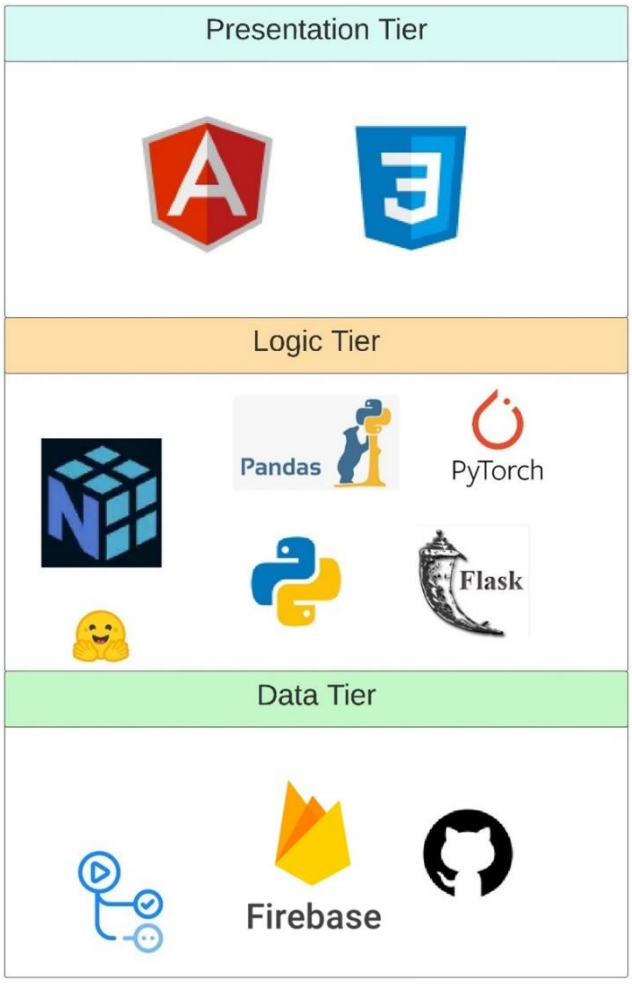


Figure 13 – Technology Stack

7.2.2 Data-set Selection

This research project requires a dataset with dialogues in customer service domain. As for the data set selection TWEETSUM dataset was chosen for the project. Tweetsum dataset is focused on twitter customer care conversation. It was made with 1,100 dialogs from tweets that is

available on the Kaggle customer support on twitter. Each dialogue is summarized with 3 extractive and 3 abstractive summaries by human annotators. Tweetsum dataset contains total summaries of 6500. (Feigenblat et al., 2021). Dataset is publicly available on the official GitHub repo under CDLA-sharing license.

7.2.3 Development Framework

Framework	Justification for selection
Flask	Since the proposed system is a web application, flask is chosen. Flask is a light weight and extensible web framework that will allow to develop web applications. Also, flask has integrated unit testing feature which will be useful for API test validation.
Material Design	Material design is used by many industrial applications. Material design is useful for rapid prototyping as it will allow the developer to focus on the logical components while not spending too much time on the css and other ui parts.
PyTorch	Main reason to choose pytorch over its competitive machine learning framework tensorflow is that pytorch well-designed comparing to tensorflows frequent api changes.

Table 23 – Selected Development Frameworks

7.2.4 Programming Languages

Python will be used as the main development language for the core component to preprocess the dataset and training modals. Python is one of the most chosen programming languages for data science projects because of its package's availability and support.

For the API development the flask framework is chosen, and the programming language will be python. Flask is a python based microframework which will allow to develop light weight web APIs.

For the front-end development, typescript was decided to be used. Even typescript will require more time to compile the code when comparing with the plain JavaScript, typescript allows to check the correctness at the compile time which will be useful while developing the prototype.

7.2.5 Libraries

Library	Justification for the selection
Pandas	Pandas dataframe library is capable of performing various operation on datasets such as cleaning, sorting, filtering etc.
Numpy	Numpy python library allow to perform mathematical operation on multidimensional arrays.
transformers	Transformers libraries provide API access to tools with large amount of pre-trained language transformer modals.
ROUGE	This rouge python package will allow to access the rouge metrices for evaluating the summarization and machine translations.
TensorBoard	TensorBoard is a package that will allow to visualize the machine learning related work such as loss and accuracu etc. TensorBoard can be easily integrate with pytorch.
Hugging Face	Hugging face is a library and a platform for accessing open source models for various fields.

Table 24 – Libraries Selection

7.2.6 IDE

IDE	Justification for the selection
Google Collab	More convenient cloud-based development environment and being able to work on multiple devices.
PyCharm	Fully fledged IDE which support lot of useful tools for local development.
VS code	Convenient IDE for development with lot of tools and extensions.

Table 25 – IDE Selection

7.2.7 Summary of Technology Selection

Component	Tools
Programming language	Python, typescript
Development framework	Flask
UI framework	Material design of Angular
Libraries	Pandas, Numpy, transformers, ROUGE, TensorBoard

IDE	Google Collab (core functionality), PyCharm (API development), VS code (frontend)
Version control	Git, GitHub, Hugging Face
Web app hosting	Netlify (frontend-host), AWS (backend-host)
CI/CD	Github Actions

Table 26 – Technology Selection

7.3 Implementation of the Core Functionality

7.3.1 Modal Training

For the implementation, the format of the tweetsum data set needs to be changed. Summaries in the tweetsum dataset is annotated by humans. But the tweetsum dataset doesn't have original conversations, instead it has tweet Ids. So, for this research project will be using the twitter customer support data and tweetsum data set to map the conversations and their respective summaries into a single data set for the modal training purposes. The following code is to extract those conversations and summaries from both datasets to finally produce a single complete dataset.

```
import pandas as pd
import json
import csv
import numpy as np
import re

# Reading main customer support dataset from kaggle
df_twcs = pd.read_csv('../original_datasets/twcs.csv')
print('twcs dataset loaded')

# opening jsonl files to map with main customer support data
with open('../original_datasets/tweet_sum_data_files/final_train_tweetsum.jsonl') as f:
    lines = f.read().splitlines()
    df_tweetsum_json = pd.DataFrame(lines)
    df_tweetsum_json.columns = ['json_element']

# preparing the TWEETSUM dataset files to a csv files with mapped coloumns
df_tweetsum_json['json_element'].apply(json.loads)
df_tweetsum_csv = pd.json_normalize(df_tweetsum_json['json_element'].apply(json.loads))
df_tweetsum_csv.to_csv('tweetsum_formated.csv')  # tweetsum json to csv
print('tweetsum_formated.csv done!')
```

Figure 14 – Pre-processing

```
for tweet_id in tweet_id_seq_arr:
    concat_tweet = ''
    for tweet in tweet_id:
        text = df_twcs['text'].loc[df_twcs['tweet_id'] == tweet].values
        bounding = df_twcs['inbound'].loc[
            df_twcs['tweet_id'] == tweet].values  # bounding true = customer tweet, bounding false = agent
        text = boundingCheck(bounding, text[0])
        concat_tweet = concat_tweet + ' ' + text
        cleaned_contact_tweet = removeLinks(concat_tweet)
        cleaned_special_char_tweet = removeSpecialChar(cleaned_contact_tweet)
    review_array_cleaned.append(cleaned_special_char_tweet)

print('Cleaning process for tweet conversations done.....')

print('Creating final dataset.....')
with open('tweetsum_formated.csv', mode='w', newline='', encoding="utf-8") as file:
    writer = csv.writer(file)
    writer.writerow(['id', 'dialogue', 'summary'])
    for i, (a, b) in enumerate(zip(review_array_cleaned, summary_array_cleaned), start=1):
        writer.writerow([i, a, b])

print('Done.....')
```

Figure 15 – Pre-Processing and Creating New Format

After preparing the dataset, the next step is to fine tune the bart-large-xsum transformer from the hugging face library for dialogue summarization task. Finalized and prepared dataset is divided into 70% of training data, 20% testing data and 10% validation data.

```
import transformers
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM, Seq2SeqTrainingArguments, Seq2SeqTrainer
from datasets import load_dataset, load_from_disk
import numpy as np
import nltk

nltk.download('punkt')

# Hyper Parameters for modal tuning
max_input = 512
max_target = 128
batch_size = 3
model_checkpoints = "facebook/bart-large-xsum"
train_dataset = load_dataset("csv", data_files="../datasets/tweetsum_train.csv")["train"]
eval_dataset = load_dataset("csv", data_files="../datasets/tweetsum_test.csv")["train"]

# Tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_checkpoints)
```

Figure 16 – Loading Train and Test Datasets

```python
# Processing dialogues with input_ids, attention_mask and labels
mr-desilva
def preprocess_data(data_to_process):
    # get all the dialogues
    inputs = [dialogue for dialogue in data_to_process['dialogue']]
    # tokenize the dialogues
    model_inputs = tokenizer(inputs, max_length=max_input, padding='max_length', truncation=True)
    # tokenize the summaries
    with tokenizer.as_target_tokenizer():
        targets = tokenizer(data_to_process['summary'], max_length=max_target, padding='max_length', truncation=True)

    # set labels
    model_inputs['labels'] = targets['input_ids']
    return model_inputs


# Setting training and evaluation datasets through tokenizer
train_dataset = train_dataset.map(preprocess_data, batched=True)
eval_dataset = eval_dataset.map(preprocess_data, batched=True)

# Loading the modal
model = AutoModelForSeq2SeqLM.from_pretrained(model_checkpoints)
```

Figure 17 – Pre-Processing Input Data

The preprocess_data function in the code preprocesses the input dataset for training and evaluation. It extracts dialogues and tokenizes them using the pre-trained tokenizer, ensuring they have the same length by setting padding and truncation options. The summaries are also tokenized with the maximum target length and treated as target sequences. Finally, the tokenized summaries' input IDs are assigned as labels to the model inputs, which serve as the ground truth during training, allowing the model to learn generating summaries based on input dialogues.

The Seq2SeqTrainingArguments class in the provided code defines the training arguments that control various aspects of the training process. The output directory is set to 'dialogue_summarization_bart' for saving output files, and the evaluation strategy is set to 'steps', meaning evaluation occurs after a fixed number of steps. The learning rate is 3e-5, and the batch sizes for training and evaluation are both set to 4. Gradient accumulation steps are set to 4, and weight decay is set to 0.01. The total number of saved checkpoints is limited to 3, and the model is trained for 5 epochs. Prediction with generation is enabled, and mixed-precision training (FP16) is used if available. The learning rate scheduler is set to linear with 500 warm-up steps. Sine the google collab pro is using for modal training process, will be utilizing cuda cores.

```
args = Seq2SeqTrainingArguments(
    output_dir='dialogue_summarization_bart',
    evaluation_strategy='steps',
    eval_steps=500,
    learning_rate=3e-5,
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    gradient_accumulation_steps=4,
    weight_decay=0.01,
    save_total_limit=3,
    num_train_epochs=5,
    predict_with_generate=True,
    fp16=True,  # available only with CUDA
    warmup_steps=500,
    lr_scheduler_type='linear'
)

trainer = Seq2SeqTrainer(
    model,
    args,
    train_dataset=train_dataset,
    eval_dataset=eval_dataset,
    tokenizer=tokenizer
)
```

```
# Training the modal
print('Modal training started.......')
trainer.train()
print('Modal training done!')
```

Figure 18 – Training Args and Fine Tuning

7.3.1 Generating dialogue summaries

In order to generate a summary for input dialogue, it should consist of a context that is worth generating a summary. For example, a small dialogue practically does not need a summary as by reading it, a reader can capture the meaning. To set a threshold value for the input dialogue a distribution of the dialogues length from the dataset was considered. The below graph displays the dialogue length in words in the training set of the dataset.

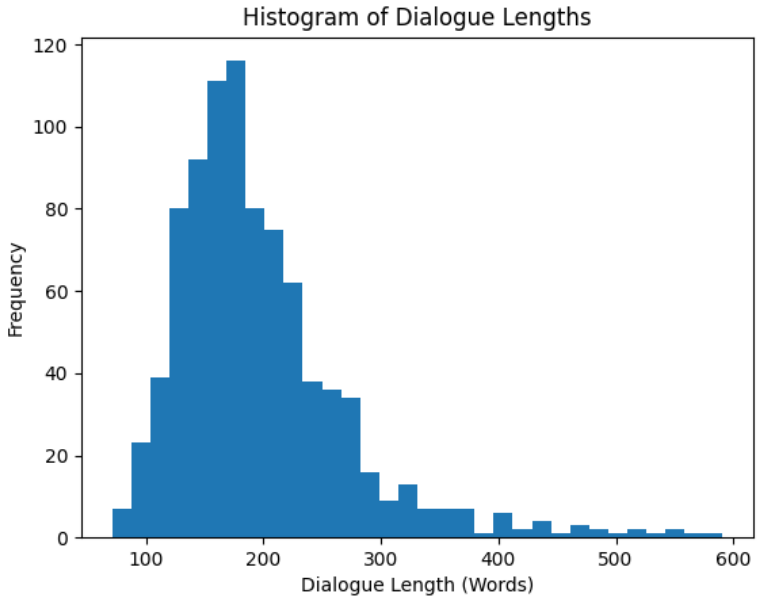


Figure 19 – Distribution of Word Count

Mean: 197.10921501706486, Median: 180.0, Standard Deviation: 75.17978084993548, Min: 71, Max: 591.

```
def thresholdCheck(dialogue):
    min_length_threshold = 60
    if input_length(dialogue) > min_length_threshold:
        return True
    else:
        return False
```

Figure 20 – Threshold Checker

A minimum of 60 words were considered for a input dialogues to generate summary.

```
def generate_dialogue_summary(dialogue):
    dialogue = remove_emojis(dialogue)
    if thresholdCheck(dialogue):
        max_new_tokens = 50
        input_ids = tokenizer_fine_tuned(dialogue, return_tensors='pt').input_ids
        summary_ids = fine_tuned_modal.generate(input_ids, max_new_tokens=max_new_tokens)
        summary_text = tokenizer_fine_tuned.decode(summary_ids[0], skip_special_tokens=True)
        return summary_text
    else:
        return 'Entered dialogue input is too short for generating a summary!'
```

Figure 21 – Generate Summary Method

Above method will generate summary for English dialogue input.

7.3.2 Integrating many-to-many translation model

To detect the language code of the input, will be using langid python package which is built top on the naïve bayes classifer. Then the source lang code will be set and target lang code will be set to English. Afterwards the fine-tuned dialogue summarizing modal will generate summary for English. Then the generated English summary will be back translated to it's original input language format.

```
import langid
from transformers import M2M100ForConditionalGeneration, M2M100Tokenizer
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM

# Dialogue summarizer modal
fine_tuned_modal_name = '../modal'
tokenizer_fine_tuned = AutoTokenizer.from_pretrained(fine_tuned_modal_name)
fine_tuned_modal = AutoModelForSeq2SeqLM.from_pretrained(fine_tuned_modal_name)

# m2m machine translation modal
model_name = "facebook/m2m100_418M"
tokenizer = M2M100Tokenizer.from_pretrained(model_name)
model = M2M100ForConditionalGeneration.from_pretrained(model_name)


mr-desilva
def generate_mult_dialogue_summary(dialogue):
    # Identify the source lang code
    src_lang_code = identify_src_lang(dialogue)
    print('Identified lang code for the input: ' + src_lang_code)
    # Translate dialogue to English
    translated_dialogue = m2m_translation(dialogue, src_lang_code, "en")

    # Summarize dialogue in English
    max_new_tokens = 50
    input_ids = tokenizer_fine_tuned(translated_dialogue, return_tensors='pt').input_ids
    summary_ids = fine_tuned_modal.generate(input_ids, max_new_tokens=max_new_tokens)
    summary_text = tokenizer_fine_tuned.decode(summary_ids[0], skip_special_tokens=True)

    print('summary text ' + summary_text)
mr-desilva
def m2m_translation(text, src_lang_code, tgt_lang_code):
    tokenizer.src_lang = src_lang_code
    input_tokens = tokenizer(text, return_tensors="pt")  # tokenizing source sentence
    translation_output = model.generate(**input_tokens, forced_bos_token_id=tokenizer.get_lang_id(
        tgt_lang_code))  # setting the model to generate the output based on the target lang code.
    translated_sentence = tokenizer.decode(translation_output[0], skip_special_tokens=True)
    print('translated sentence in english - ' + translated_sentence)
    return translated_sentence
```

Figure 22 – Machine Translation Modal

7.4 User Interface

Refer Appendix for more UI snapshots.

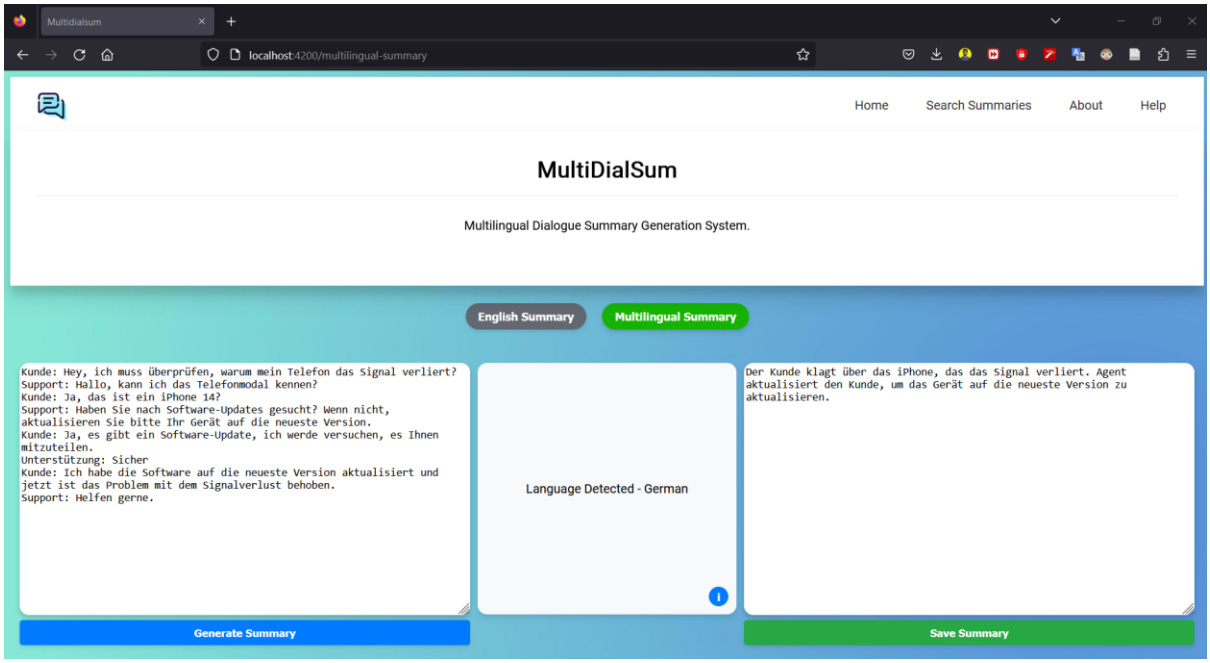


Figure 23 – Sample UI

7.5 Chapter Summary

In this part of the thesis consists of the technologies and tools that were selected to the development of the proposed system with the reasonings. Also, the discussion on the development of the core component.

CHAPTER 8: TESTING

8.1 Chapter Overview

This section of the thesis focused on the testing phase of this research project and both functional and non-functional testing and integration testing will be discussed in detail.

8.2 Objective and Goals of Testing

The main intent of software testing is to ensure that the system performs according to the requirements.

- Ensure the implemented system is working as expected with all the requirements.
- Make certain that the implemented system fulfills the essential "Must have" and "Should have" functional requirements defined by the MoSCoW method.
- Verify both required and important non-functional requirements are fulfilled.
- Identify improvements or bug fixes that can improve the overall performance of the system.

8.3 Testing Criteria

To evaluate whether the implemented system meets the requirements, two types of testing approaches were selected.

Testing Approach	Description
Functional Testing	The functional requirements are tested to ensure that the expectations are met.
Structural Testing	Testing the non-functional requirements to verify that the expectation are satisfied.

Table 27 – Testing Criteria

8.4 Modal Testing

8.4.1 ROUGE Score

Under the literature review chapter, the metrics for evaluation was identified. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation.

ROUGE-1: This metric measures the overlap of unigrams (single words) between the generated summary and the reference summary. It focuses on individual word matches, providing a sense of the system's ability to capture important keywords and concepts in the summarized text. Higher ROUGE-1 scores indicate better performance in matching unigrams.

ROUGE-2: This metric measures the overlap of bigrams (word pairs) between the generated summary and the reference summary. It focuses on word pair matches, providing an

understanding of the system's ability to preserve local word order and short phrases in the summarized text. Higher ROUGE-2 scores indicate better performance in capturing word pair relations.

ROUGE-L: This metric measures the Longest Common Subsequence (LCS) between the generated summary and the reference summary. It focuses on the longest sequence of words that appear in the same order in both summaries. This metric provides insights into the system's ability to maintain the overall structure and coherence of the summarized text. Higher ROUGE-L scores indicate better performance in preserving the structure and coherence of the summarized text.

ROUGE-Lsum: This metric is an extension of ROUGE-L, and it is specifically designed for evaluating summaries with multiple sentences. It computes ROUGE-L scores for each sentence in the generated summary against all sentences in the reference summary and aggregates the scores. This metric provides a more nuanced understanding of the system's ability to generate coherent multi-sentence summaries. Higher ROUGE-Lsum scores indicate better performance in generating coherent multi-sentence summaries.

8.4.2 Testing Generated Summaries with ROUGE score.

In order to get the results, rouge score requires generated summaries by the modal and referring ground truth summaries for the comparison evaluation. The TWEETSUM dataset summaries are annotated by humans which in this case can help to get a more accurate rouge score when comparing with the modal generated summaries.

Initially the TWEETSUM dataset is divided into three parts for training, testing and validation. The validation part of the dataset will be used for this purpose. Bellow diagram represents the achieved results by the fine-tuned modal.

Metric	Precision	Recall	F-Measure
ROUGE-1	44.79	49.39	45.51
ROUGE-2	21.92	23.48	21.97
ROUGE-L	38.64	42.74	39.30
ROUGE-L-SUM	38.71	42.78	39.37

Table 28 – ROUGE Score

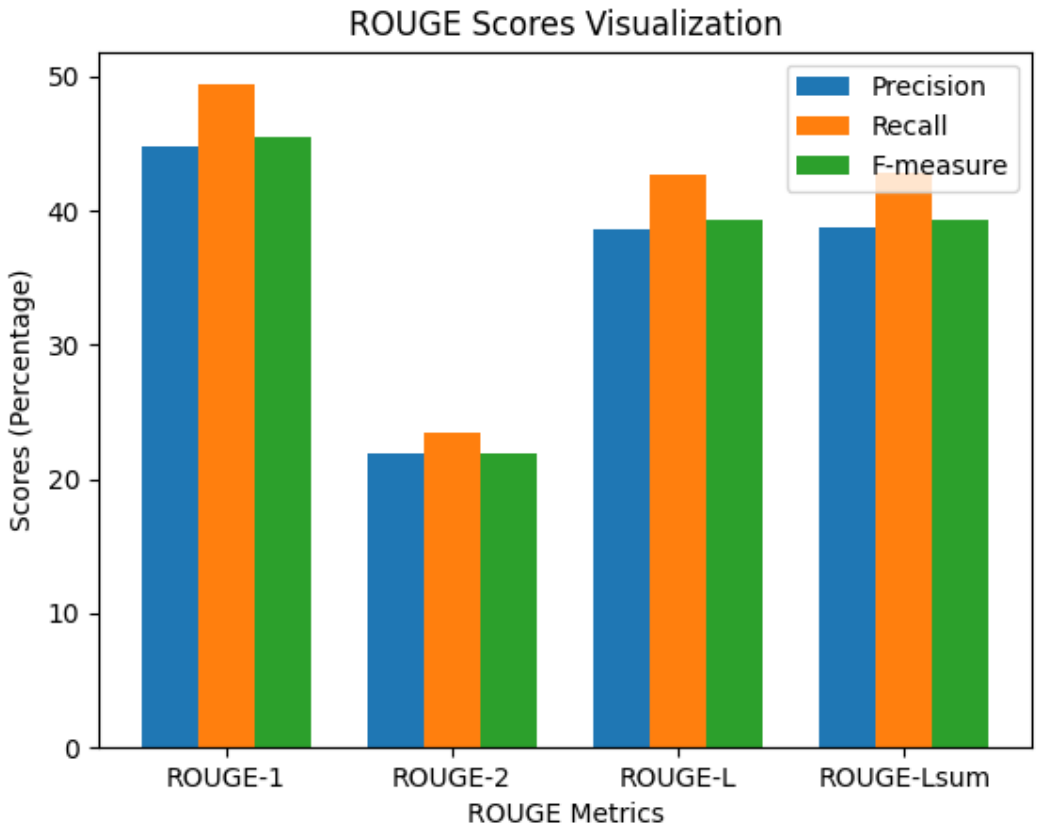


Figure 24 – ROUGE Score Graph

8.4.2 BERTScore

BERTScore is a metric for evaluating the quality of text generated by machine learning models, such as neural machine translation (NMT) or abstractive summarization systems (T Zhang et al., 2020). BERTScore measures the similarity of the two texts in the continuous vector space of BERT embeddings. It computes the cosine similarity between the generated text and reference text's token embeddings at different layers, resulting in a score that quantifies their similarity. BERTScore has been shown to correlate better with human judgments of text quality compared to other metrics, as it is able to capture more nuanced aspects of language, such as syntax, semantics, and context.

English dialogue summaries generated by the model was tested using BERTScore. The results for Mean BERTScore achieved 0.6798.

8.5 Functional Testing

Functional requirements that are specified in chapter 4 are tested against the application using black box testing methodology.

FR ID	User Action	Expected Result	Actual Result	Result Status
1	User generating a summary for an inputted dialogue data at least for 1 language option.	Summary should be generated in input language.	Summary is generated in input language.	Passed
2	User tries to view the generated summary.	Summary should be visible to the user	Summary is visible to the user	Passed
3	User tries to generate summaries for different languages.	Input language should be automatically detected and generate the summary.	Input language is detected, and the summary is generated.	Passed
4	User tries to store the generated summary and view it later.	User should be able to save the generated summary and view it.	Saved summary can be viewed later.	Passed
5	Users try to search for a matching keyword among saved summaries.	User should be able to view matching summary results if there are any.	Saved summaries with matching keywords can be viewed.	Passed

Table 29 – Functional Testing

8.6 Module and Integration Testing

Module	Input	Expected Result	Actual Result	Result Status
English Summary Generator	Dialogue data in English	Generated summary in English.	Generated summary in English.	Passed
Multilingual Summary Generator	Dialogue data in any language	Detect language and generate summary.	Detect language and generate summary.	Passed
Save Summary	Select save summary	Summary saved to database.	Summary saved to database.	Passed
Search Summary	Keywords to search	Matching summaries with input keywords	Matching summaries with input keywords	Passed

Table 30 – Module and Integration Testing

8.7 Non-Functional Testing

NFR ID	Test Case	Expected Result	Priority	Result
1	Quality of the Output	The quality of the generated summary should be clear and meaningful as much as possible.	High	Passed
2	Performance	Time to generate a summary should be acceptable.	Medium	Passed
3	Security	The system should prevent any data breaches from attackers to keep the information safe	Low	Passed
4	Usability	The system should be easy to use followed by good user interface and user experience principles	Medium	Passed

Table 31 – Non Functional Testing

8.8 Limitations of the testing process

One limitation of the testing process for the dialogue summarization system in the customer service domain is the lack of available baseline models specifically designed for this task. As dialogue summarization in customer service is a specialized domain, it presents challenges in

comparing the performance of the system with other models. Another limitation is the absence of a multilingual validation dataset, which makes it difficult to thoroughly evaluate the quality of the multilingual summaries generated by the system. In the current scenario, human evaluation is the best approach for assessing the quality of multilingual summaries, as discussed in the evaluation chapter. However, this method can be time-consuming and costly, posing additional challenges in the testing process.

8.9 Chapter Summary

In this part of the thesis concludes the testing phase of this research project, covering various aspects including model testing, functional and non-functional testing, as well as the limitations of the testing process encountered throughout.

CHAPTER 9: EVALUATION

9.1 Chapter Overview

This section of the thesis discusses the evaluation of the research project which includes the evaluations opinion according to the experts from research domain and application domain. Also, the author's own evaluation on the project. Furthermore, evaluation of the functional and non-functional requirements that were early mentioned will be evaluated.

9.2 Evaluation Methodology and Approach

The aim of this research project is to develop a dialogue summary generation system which can handle more than one language, both quantitative and qualitative approaches were chosen. The testing chapter covers the quantitative results of the evaluation. To collect data for qualitative evaluation, the author carried out surveys and interviews.

9.3 Evaluation Criteria

In order to breakdown the qualitative evaluation, the following criteria have been considered by the author for thematic analysis.

Criteria	Evaluation Purpose
Selection of research conducted	To show the importance of the chosen topic, field, unanswered questions, and detail in this study.
Contribution to the research	To justify the contribution upon the completion of this research project for dialogue summarization in customer service domain.
Quality of research documentation	To verify that a sufficient amount of relevant literature has been examined and the complete research procedure has been recorded and showcased in an acceptable manner.
Approach to development	To ensure that a suitable development strategy was employed to address the issue as effectively as possible, including the creation of a prototype.

Selection of the quantitative evaluation methods	To verify the selected metrics that are used to evaluate and analyze the results of the research project.
Future Improvements	To identify the improvements which can be done as a part of future works.
Overall experience on the application	To verify that the web application of this research project is convenient for the targeted audience.

Table 32 – Evaluation Criteria

9.4 Self-Evaluation

Self-evaluation was documented by the author of this research to the above criteria's.

Criteria	Evaluation Purpose
Selection of research conducted	The selected research domain is both trending and challenging one because of it's high application and business value.
Contribution to the research	The technical contribution for this research project can be identified as fine tuning a existing text summarization modal for the dialogue summarization task and use of cross-lingual transfer modal to enable multilingual or multi language support. The application contribution can be identified as a single web application which can support more than one language for dialogue summarization task.
Quality of research documentation	The documentation maintains the utmost quality standards according to the standard guidelines.
Approach to development	Combining two main datasets into a new format where the text summarization modals can be fine tuned for the dialogue summarization task is presented. Also, the approach is fully open source and will be

	presented to the public. Selection of the transformer modals has been justified with reasonings.
Selection of the quantitative evaluation methods	Both quantitative and qualitative analysis were carried out to evaluate the research project as much as possible.
Future Improvements	Opinions from the evaluators were considered and possible improvements have been done within the project timeline.
Overall experience on the application	The user experience on the web application is to be focused on more simplified and a readable manner.

Table 33 – Self Evaluation

9.5 Selection of the Evaluators

The project evaluator's selection can be divided into the following categories.

ID	Evaluator Type	Category
1	Technical Domain Expert	Experts who are in the field of dialogue summarization or text summarization.
2	Application Domain Expert	Experts who are currently using dialogue summarization systems or in the domain of customer service.

Table 34 – Selection of The Evaluators

9.6 Evaluation Results

9.6.1 Thematic Analysis

The expert feedback obtained has been examined based on the following emerging themes.

Criteria	Evaluator Type ID	Theme	Summary
Selection of research conducted	1	Research gap in dialogue summarization	The technical experts recognized that the multilingual dialogue summarization system addresses an important research gap in the field. They acknowledged that

			providing efficient and accurate dialogue summarization across multiple languages is crucial for improving customer service experiences. While the solution makes significant strides in this area, the experts emphasized the need for continued research and development to further enhance the system's performance and overcome potential limitations in preserving context and nuances during the translation and summarization process.
Contribution to the research	1	Technical Contribution	The evaluators acknowledged the research contribution made by the multilingual dialogue summarization system in addressing a pressing need within the customer service domain. They appreciated the integration of state-of-the-art models, such as bart-large-xsum and m2m100, to enable accurate summarization and translation across multiple languages. The experts recognized the system as a valuable step forward in enhancing customer service interactions, highlighting its potential to significantly impact the field. However, they also encouraged further improvements and refinements to maximize the solution's effectiveness and overcome any remaining challenges.
	2	Domain Contribution	In summary, customer service domain experts see the multilingual dialogue summarization system as a promising and valuable tool for improving global customer support interactions, but they

			encourage ongoing development to ensure its optimal performance in real-world scenarios.
Quality of research documentation	1	Readability and content representation	In conclusion, the experts found the research documentation to be well-organized and informative.
Approach to development	1	Preprocessing	Technical experts appreciated the combination of the Twitter support dataset and TweetSum for fine-tuning the bart-large-xsum model. They recognized that blending these two datasets is a valuable contribution, as it helps the model to better understand and adapt to the nuances and context of customer service interactions on social media platforms.
	1	Selection of the modal	Evaluators acknowledged the choice of the bart-large-xsum model for dialogue summarization, recognizing its proven summarization capabilities as a valid reason for selection. They understood that the model's ability to generate abstractive summaries can be particularly useful in the context of customer service dialogues. Regarding the decision not to use models like RoBERTa, the evaluators respected this choice, as RoBERTa is primarily designed for tasks like classification and natural language inference rather than abstractive summarization. They appreciated the focus on selecting a model that aligns well with the specific requirements of the task at hand.

Selection of the evaluation methods	1,2	Qualitative evaluation	The evaluators appreciated the incorporation of qualitative evaluation through surveys for human assessment of the dialogues and generated summaries. They recognized that this approach provides valuable insights into the system's real-world performance and helps identify potential areas for improvement that may not be captured by quantitative evaluation methods alone. The experts commended the effort to gather feedback from users, as it can help assess the system's ability to generate coherent, contextually accurate, and meaningful summaries. Such feedback can be instrumental in understanding the user experience and the overall effectiveness of the system from an end-user perspective.
	1	Qualitative Evaluation	The evaluators acknowledged the selection of the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score as a widely used and popular metric for evaluating summarization systems. They appreciated its ability to provide an objective measure of the similarity between the generated summaries and reference summaries, making it a suitable choice for assessing the performance.
Future Improvements	1,2	Generating more accurate summaries.	The evaluators suggested several potential areas for future improvement, including incorporating more diverse domain-specific data, exploring alternative models or architectures, enhancing context

			preservation and nuance handling, and implementing a continuous feedback loop for model refinement. They also encouraged conducting ongoing performance evaluations using a combination of quantitative and qualitative assessment methods. Overall, the experts emphasized the importance of continuous development and optimization to ensure the system's effectiveness in the customer service domain.
Overall experience on the application	1,2	Usability	The evaluators appreciated the overall experience of the application, noting the valuable functionality of saving generated summaries along with their original dialogues for later reference. They acknowledged that this feature could be particularly useful for customer service agents when reviewing past interactions and improving their service quality. Additionally, the evaluators commended the simplicity of the user interface, as it can enhance usability and accessibility for users with varying levels of technical expertise. They recognized that a straightforward UI contributes to a more efficient and seamless experience for users, which is essential in a customer service context.

Table 35 – Thematic Analysis on Evaluators Feedback

Unstructured feedback is included in **APPENDIX F**.

9.6.2 Evaluation results from survey

A survey was carried out for English language dialogues and other main 3 languages respectively French, Spanish and German. Survey participants were selected from the reddit and telegram social media communities which are dedicated for language learners. The participants were selected to validate the quality of the generated dialogue summaries by comparing its original dialogue. For each language a minimum of 7 participants were selected to get a diverse result. Full survey results are included in the **APPENDIX E**.

Conversation 1/2

Original Dialogue	Summary generated from the tool
Customer: Hello, I'm having trouble logging into my account. Support Agent: Hi there! I'm sorry to hear that you're experiencing issues. Can you please tell me if you're receiving any error messages? Customer: Yes, I'm getting an "Invalid username or password" message. Support Agent: Thank you for the information. Have you tried resetting your password? Customer: No, I haven't tried that yet. Support Agent: I suggest you try resetting your password first. You can do this by clicking on the "Forgot password" link on the login page. You'll receive an email with instructions on how to create a new password. Customer: Alright, I'll give that a try. Thanks for your help! Support Agent: You're welcome! If you still have trouble logging in after resetting your password, please don't hesitate to reach out to us again. Have a great day! Customer: Thanks, you too! Goodbye! Support Agent: Goodbye! If you need any further assistance, feel free to contact us.	Customer complains that he is unable to login into his account. Agent suggests to reset the password before trying to login, and says that he will receive an email with instructions on how to create a new password.

Figure 25 – Survey Dialogue 1

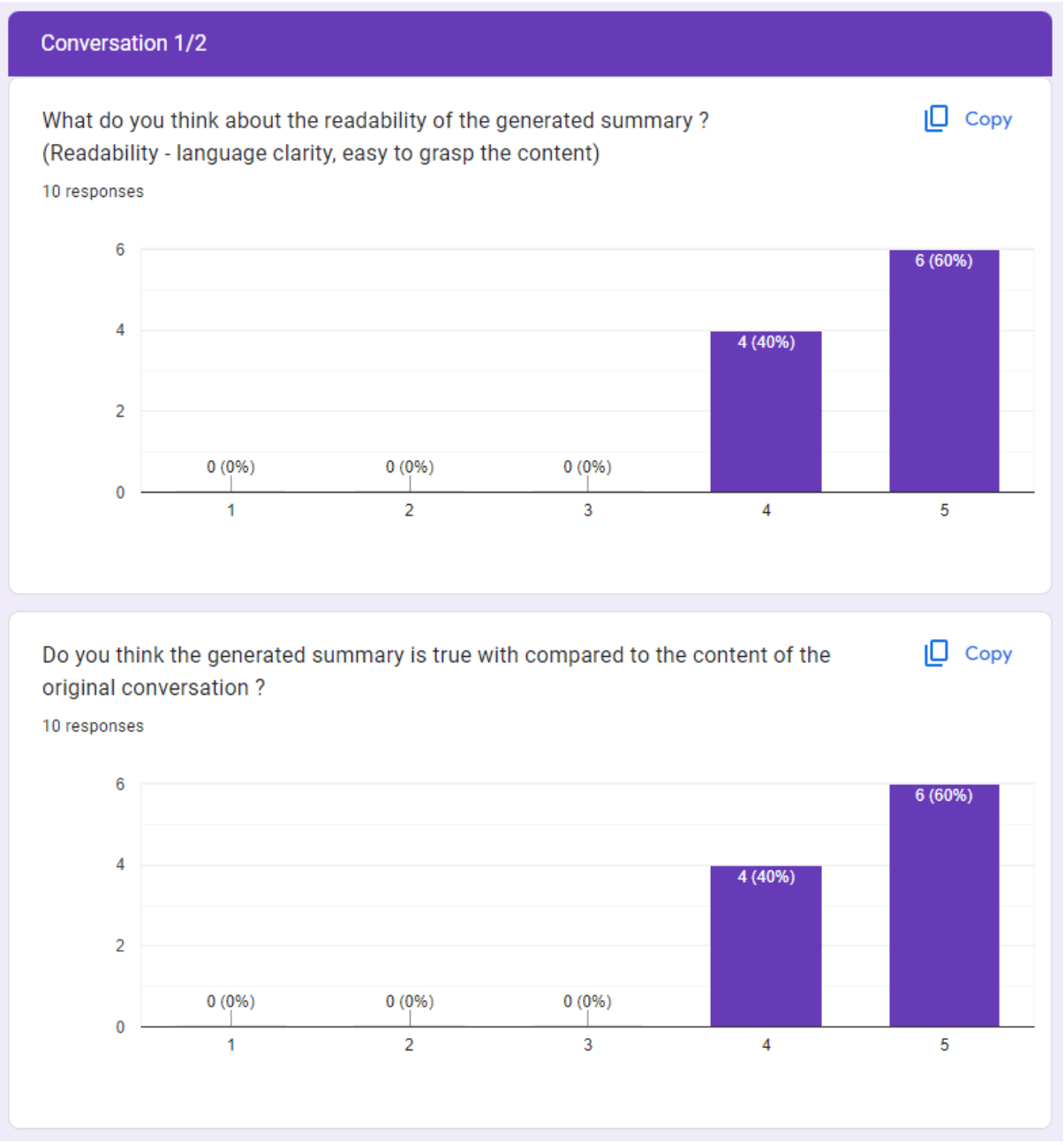


Figure 26 – Survey Results English

9.7 Limitation of Evaluation

A limitation of the evaluation process for the multilingual dialogue summarization system lies in the absence of well-established metrics specifically tailored for dialogue summarization. Currently, the best approach is human evaluation, which, while effective, can be costly and time-consuming.

Additionally, while quantitative evaluation metrics like ROUGE scores can provide an overall assessment of the dialogue summaries, they may not fully capture the nuances and context preservation aspects of the summarization process. As a result, relying solely on such metrics

might not offer a comprehensive understanding of the system's performance in real-world customer service scenarios.

9.8 Evaluation on Functional Requirements

FR ID	Requirement	Priority Level	Status
1	User must be able to generate a summary for an inputted dialogue data at least for 1 language option.	M	Done
2	Summary of the dialogue should be represented to the user.	M	Done
3	User should be able to generate summary for a dialogue data with multiple language options.	S	Done
4	The system should store the generated dialogue summaries and allow users to view the summaries	S	Done
5	The system should allow the users to search stored summaries with matching keywords.	S	Done
6	The system should allow the admins to categorize the stored summaries based on their similarities	C	Not done

Table 36 – Evaluation on Functional Requirements

9.9 Evaluation on Non-Functional Requirements

NFR ID	Requirement	Description	Priority Level	Status
1	Quality of the Output	The quality of the generated summary should be clear and meaningful as much as possible.	M	Done
2	Performance	Time to generate a summary should be acceptable.	S	Done
3	Security	The system should prevent any data breaches from attackers to keep the information safe.	S	Done
4	Usability	The system should be easy to use followed by good user interface and user experience principles.	M	Done

Table 37 - Evaluation on Non-Functional Requirements

9.10 Chapter Summary

In this part of the thesis concludes the evaluation methodologies and the considered criteria by the author and self-evaluation as well. The domain experts and other evaluators feedback also documented and discussed within the chapter.

CHAPTER 10: CONCLUSION

10.1 Chapter Overview

This section of the thesis discusses the various factors upon successfully completing the research project. How the author's existing skills and knowledge help to start the project and by the end of project completion what skills are newly acquired. Also the contribution of this research project is briefly discussed.

10.2 Achievements of Research Aims & Objectives

10.2.1 Research Aim

The aim of this research is to design, develop and evaluate a multilingual dialogue summary generation system for customer services using dialogue summarization model with the help of the cross-lingual transfer method.

Main objective of this project is achieved, and the results were presented within both Testing and Evaluation chapters.

10.2.2 Research Objectives

Research Objectives	Description	Learning Outcomes	Research Questions	Status
Literature Review	Gather required material on previous work and critically evaluate the findings.	LO1, LO4, LO8	**RQ1, RQ2**	Completed
Requirement Elicitation	Determine the project's requirements using the proper methods to give a solution for the research problems and gaps should be handled based on relevant prior research knowledge.	LO2, LO6, LO8	**RQ2, RQ3**	Completed
Design	Designing a system capable of generating a dialogue summary with multiple languages involved.	LO1, LO3, LO5, LO8	**RQ2, RQ3**	Completed

Implementation	Implementing a multilingual dialogue summarization platform.	LO1, LO5, LO7, LO8	**RQ2**	Completed
Evaluation	Testing the implemented system and dialogue summary generation with evaluation metrics.	LO1, LO5, LO8	**RQ2**	Completed

Table 38 – Conclusion of Research Objectives

10.3 Utilization of Knowledge from the Course

Module	Description
Software development group project	Upon completing this module, gave a very good background knowledge into machine learning and natural language processing domains. Also helped to improve the documentation skills.
Programming Principles modules	These two modules based on python and java helped to develop the programming skills and software testing skills.
Web Design and Development module	This module helped to improve the knowledge in web application development which is a must for this research project.

Table 39 – Utilization of Knowledge From The Course

10.4 Use of Existing Skills

- While doing the second year data science software development group project, the author had the opportunity of getting familiar with the NLP research domain as the project was based on the sentiment analysis for product reviews.

- Working as a software engineer intern at IFS helped to develop extensive skills in web application development which was well aided upon completing this research project application development.

10.5 Use of New Skills

At the beginning of this research project the author had no prior experience or knowledge in dialogue summarization domain. Moving forward, the author was to improve the knowledge

in the domain of dialogue summarization, cross-lingual transfer modals and machine translations.

Although the author had some prior experience in documentation, this project aided well to enhance not only the documentation skills but also to develop studying research papers and critically evaluate existing works.

10.6 Achievement of Learning Outcomes

Description	Learning Outcome
The solution for the chosen problem has been given after using the suitable approaches with valid justifications.	LO1
Requirement gatherings were done through literature review, conducting interviews and surveys. After those results represented in a readable manner.	LO2
Sufficient amount of literature reviews was studied and critically reviewed.	LO4
The defined tasks by the author were carried out with thorough study on possible options.	LO5, LO3
Project was continued with under the supervision of the supervisor and project activities were discussed at each stage.	LO6, LO3
Under the code of BCS Conduct, SLEP issues were considered and necessary actions were taken to mitigate them.	LO7
Everything in this research project is documented under the guidelines provided by the module leader and other standards were followed to produce a quality documentation.	LO8

Table 40 - Achievement of Learning Outcomes

10.7 Problems and Challenges Faced

Challenges	Mitigation
Running both summarization modal and the machine translation modal at once consumes a significant number of computes. Initially author's machine had 8GB of ram which was unable to run modals locally.	Expanded the machine ram by adding another ram stick. After that two modals can run parallelly without slowing down the machine.

Throughout the research project, a significant effort was dedicated to reviewing the previous research papers to understand the domain of dialogue summarization. It was challenging that the domain is relatively new and very small number of datasets were available under each sub domain.	Extensive literature review was carried out to get a better understand of existing works and available datasets under different sub domains in dialogue summarization.
Fine tuning modal to achieve the best accuracy was challenging.	Note down and conducted a analysis of all the hyperparameter changes to identify the best configurations for the modal.

Table 41 - Problems and Challenges Faced

10.8 Deviations

At the beginning of the project author's selection for the summarization modal is the ROBERTa modal which was developed by facebook. But after continues review on the literature, author identified that the ROBERTa modal was originally developed for classification task which may not the best choice for the dialogue summarization task. Because of that author decided to use the BART-LARGE-XSUM modal which is already had good performance in abstractive text summarization. All the reasonings are included in the literature review chapter.

10.9 Limitations of the Research

- The current system may not be able to handle very lengthy dialogues, within the given project timeline and the initial scope it will lead to another research direction.
- Even though the machine translation modal is capable of handling 97 languages, the amount of training data that it was originally trained can result in different performance in each language.
- Very small or small context dialogue summarization is not well performed by the modal. In a practical manner it is not necessary to generate a summary for a shorter dialogue. In order to prevent this threshold value is defined that it should require minimum amount of word count to generate a more meaningful summary.
- This system is strictly biased for the customer service domain, as it was trained on a such dataset. Generalizing this approach can be done by changing the dataset for any other sub domain.

10.10 Future Enhancements

- Able to categorize the summaries based on tags or topics.
- Fine tune the machine translation modal with domain specific words, this will help to improve the translation process.
- Enhance the dialog context perseveration while generating summaries.
- Explore alternative modals and architectures that can potentially improve the system's summarization and translation process.
- Optimize user interface and experience: Continuously refine the user interface and user experience based on feedback and user testing to ensure a seamless, efficient, and accessible experience for users of varying technical expertise.

10.11 Achievement of the contribution to body of knowledge

Upon the conclusion of this research project, the author has archived contribution in the domain of dialogue summarization and customer service.

10.11.1 Technical Contribution (Dialogue summarization)

1. Use of transformers and cross lingual transfer modals to develop a dialogue summarization solution which can support more than one language.
2. Use of existing text summarization modals for dialogue summarization tasks.

10.11.2 Domain Contribution (Customer Service)

A single tool capable of summarizing dialogues in more than one language. Currently there are no tools that can provide dialogue summarization for more than one language.

10.11.3 Additional Contribution

By combining TWEETSUM and Kaggle customer care dataset into a new format, the proposed dataset format can be utilized to fine tune text summarization modal into dialogue summarization task.

10.12 Concluding Remarks

In this part of the thesis summarizes the author's accomplishment on the research objectives and goals. How the author's prior experience throughout the course helped to achieve the success of the research project, limitations of the project, future enhancements and the contribution to each domain and technical areas. This research was carried out by following the guidelines of standards research project to achieve the best output quality.

REFERENCES

Artetxe, M. et al. (2018). Unsupervised Neural Machine Translation. Available from https://doi.org/10.48550/arXiv.1710.11041 [Accessed 13 April 2023].

Bai, Y., Gao, Y. and Huang, H. (2021). Cross-lingual abstractive summarization with limited parallel resources. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 6910–6924. Available from https://doi.org/10.18653/V1/2021.ACL-LONG.538.

Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. June 2005. Ann Arbor, Michigan: Association for Computational Linguistics, 65–72. Available from https://aclanthology.org/W05-0909 [Accessed 15 April 2023].

Chen, Y. et al. (2021). DIALOGSUM: A Real-Life Scenario Dialogue Summarization Dataset. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 5062–5074. Available from https://doi.org/10.18653/V1/2021.FINDINGS-ACL.449.

Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. June 2019. Minneapolis, Minnesota: Association for Computational Linguistics, 4171–4186. Available from https://doi.org/10.18653/v1/N19-1423 [Accessed 29 March 2023].

Fan, A. et al. (2020). Beyond English-Centric Multilingual Machine Translation. Available from https://doi.org/10.48550/arXiv.2010.11125 [Accessed 13 April 2023].

Feigenblat, G. et al. (2021). TWEETSUMM - A Large Scale Dialog Summarization Dataset for Customer Service. 7 November 2021. Available from https://research.ibm.com/publications/tweetsumm-a-large-scale-dialog-summarization-dataset-for-customer-service [Accessed 7 February 2023].

Feng, Xiachong et al. (2021). Language model as an annotator: Exploring DialoGPT for dialogue summarization. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 1479–1491. Available from https://doi.org/10.18653/V1/2021.ACL-LONG.117.

Feng, Xiachong, Feng, Xiaocheng and Qin, B. (2022). A Survey on Dialogue Summarization: Recent Advances and New Frontiers. Available from https://pypi.org/project/pyrouge/ [Accessed 29 October 2022].

Gurevych, I. and Strube, M. (2004). Semantic Similarity Applied to Spoken Dialogue Summarization. *COLING 2004: Proceedings of the 20th International Conference on*

Computational Linguistics. 23 August 2004. Geneva, Switzerland: COLING, 764–770. Available from https://aclanthology.org/C04-1110 [Accessed 29 March 2023].

Junczys-Dowmunt, M. et al. (2018). Marian: Fast Neural Machine Translation in C++. Available from http://arxiv.org/abs/1804.00344 [Accessed 13 April 2023].

Keskar, N.S. et al. (2017). On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. Available from https://doi.org/10.48550/arXiv.1609.04836 [Accessed 13 April 2023].

Klein, G. et al. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Proceedings of ACL 2017, System Demonstrations*. July 2017. Vancouver, Canada: Association for Computational Linguistics, 67–72. Available from https://aclanthology.org/P17-4012 [Accessed 13 April 2023].

Lewis, M. et al. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Available from http://arxiv.org/abs/1910.13461 [Accessed 11 April 2023].

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. 74–81. Available from https://aclanthology.org/W04-1013 [Accessed 2 November 2022].

Lin, H. et al. (2021). CSDS: A Fine-Grained Chinese Dataset for Customer Service Dialogue Summarization. Available from https://doi.org/10.48550/arXiv.2108.13139 [Accessed 13 April 2023].

Liu, C. et al. (2019). Automatic Dialogue Summary Generation for Customer Service. Available from https://doi.org/10.1145/3292500.3330683 [Accessed 17 October 2022].

Liu, F. and Liu, Y. (2008). Correlation between ROUGE and human evaluation of extractive meeting summaries. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers - HLT '08*. 2008. Columbus, Ohio: Association for Computational Linguistics, 201. Available from https://doi.org/10.3115/1557690.1557747 [Accessed 15 April 2023].

Liu, Z. et al. (no date). Zero-shot Cross-lingual Dialogue Systems with Transferable Latent Variables. Available from https://fasttext.cc [Accessed 29 October 2022].

Micikevicius, P. et al. (2018). Mixed Precision Training. Available from https://doi.org/10.48550/arXiv.1710.03740 [Accessed 13 April 2023].

Papineni, K. et al. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311–318. Available from https://doi.org/10.3115/1073083.1073135.

See, A., Liu, P.J. and Manning, C.D. (2017). Get To The Point: Summarization with Pointer-Generator Networks. Available from https://doi.org/10.48550/arXiv.1704.04368 [Accessed 29 March 2023].

Sellam, T., Das, D. and Parikh, A.P. (2020). BLEURT: Learning Robust Metrics for Text Generation. Available from https://doi.org/10.48550/arXiv.2004.04696 [Accessed 15 April 2023].

Shang, G. et al. (2018). Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. July 2018. Melbourne, Australia: Association for Computational Linguistics, 664–674. Available from https://doi.org/10.18653/v1/P18-1062 [Accessed 13 April 2023].

Vaswani, A. et al. (2017). Attention Is All You Need. Available from https://doi.org/10.48550/arXiv.1706.03762 [Accessed 29 March 2023].

Zhang, T. et al. (2020). BERTScore: Evaluating Text Generation with BERT. Available from https://doi.org/10.48550/arXiv.1904.09675 [Accessed 31 March 2023].

Zhang, X. et al. (2020). Unsupervised Abstractive Dialogue Summarization for Tete-a-Tetes. Available from https://doi.org/10.48550/arXiv.2009.06851 [Accessed 13 April 2023].

Zhang, Y. et al. (2020). DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. Available from https://doi.org/10.48550/arXiv.1911.00536 [Accessed 13 April 2023].

Zhao, L. et al. (2021). TODSum: Task-Oriented Dialogue Summarization with State Tracking. Available from https://doi.org/10.48550/arXiv.2110.12680 [Accessed 13 April 2023].

Zhong, M. et al. (2021). QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. Available from https://doi.org/10.48550/arXiv.2104.05938 [Accessed 13 April 2023].

Zhou, Z. et al. (2022). Learning Dialogue Representations from Consecutive Utterances. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022. Seattle, United States: Association for Computational Linguistics, 754–768. Available from https://doi.org/10.18653/v1/2022.naacl-main.55 [Accessed 13 April 2023].

Zou, Y., Zhu, B., et al. (2021). Low-Resource Dialogue Summarization with Domain-Agnostic Multi-Source Pretraining. Available from http://arxiv.org/abs/2109.04080 [Accessed 18 October 2022].

Zou, Y., Zhao, L., et al. (2021). Topic-Oriented Spoken Dialogue Summarization for Customer Service with Saliency-Aware Topic Modeling. Available from https://doi.org/10.48550/arXiv.2012.07311 [Accessed 13 April 2023].

Zou, Y., Lin, J., et al. (2021). Unsupervised Summarization for Chat Logs with Topic-Oriented Ranking and Context-Aware Auto-Encoders. Available from https://doi.org/10.48550/arXiv.2012.07300 [Accessed 13 April 2023].

APPENDIX A – CONCEPT MAP

Figure 27 – Concept Map

APPENDIX B – GANTT CHART

	Task Name	Start Date	End Date
1	**Project Intiation**	08/15/22	11/07/22
2	Selection of research problem and a domain	08/15/22	08/23/22
3	Supervisor's Initial Idea discussion	08/22/22	08/23/22
4	Exploring the research domain	08/23/22	08/26/22
5	Identifying the research gap and the problem domain	08/23/22	09/09/22
6	Preparing draft project proposal	09/12/22	11/01/22
7	Supervisors feedback on draft project proposal	10/25/22	11/01/22
8	Self-evaluation on project proposal	11/01/22	11/02/22
9	Submission of the project proposal	11/03/22	11/07/22
10	**Literature Review**	08/23/22	11/28/22
11	Identifying components to cover in the literature review	10/28/22	11/07/22
12	Collecting related research papers	08/23/22	11/01/22
13	Reading and summarizing the papers	08/23/22	10/25/22
14	Prepare Draft LR	11/01/22	11/15/22
15	Supervisors feedback on draft LR	11/15/22	11/22/22
16	Finalizing LR based on feedback	11/22/22	11/25/22
18	**Requirement Gathering**	09/12/22	01/02/23
19	Explore requirement	09/12/22	11/11/22
20	Create questionnaires	11/16/22	11/18/22
21	Distribute questionnairs to target audience	11/18/22	11/29/22
22	Reach to research domain experts for insights	11/14/22	11/29/22
23	Analyze data from the questionnairs	12/02/22	12/05/22
24	Preparing SRS	11/11/22	12/23/22
25	Supervisors feedback on draft SRS	11/22/22	12/27/22
26	Self evaluation on SRS	12/28/22	12/29/22
27	Submission of SRS	01/02/23	01/02/23
28	**Designing the System**	11/01/22	02/02/23
29	Selecting component of the prototype	11/01/22	11/07/22
30	Creating High-level architecture	11/24/22	12/01/22
31	Supervisor feedback on High-level architecture	11/29/22	11/29/22
32	Creating system design	12/05/22	11/11/22
33	Preparing draft of project specification design (PSD)	11/24/22	01/13/23
34	Supervised feedback on draft PSD	12/27/22	01/10/23
35	Self evaluation on PSD	01/13/23	01/13/23
36	Submission of draft PSD	01/23/23	01/23/23
37	Submission of PSD	01/30/23	02/02/23
38	**Prototype Implementation**	12/07/22	02/02/23
39	Define prototype deliverables	12/07/22	12/13/22
40	Implementing core features	12/19/22	01/10/23
41	Implementing application components	01/10/23	01/13/23
42	Demo prototype to Supervisor and get feedback	01/24/23	01/24/23
43	Refining the prototype	01/25/23	02/01/23
44	Submission of the Prototype	01/30/23	02/02/23
45	**Testing and Evaluation**	01/12/23	03/17/23
46	Identifying functional and non-functional requirements	01/12/23	01/17/23
47	Creating test plans	02/06/23	02/10/23
48	Writing unit test	02/13/23	02/16/23
49	Conducting functionality tests	02/17/23	02/21/23
50	Conducting usability tests	02/22/23	02/24/23
51	Submission of Test and Evaluation Report	03/20/23	03/20/23
52	Final evaluation of prototype	02/27/23	03/17/23
53	**Final Documentation and Conclusion**	02/01/23	05/15/23
54	Drafting the dissertation	02/01/23	04/13/23
55	Supervisor feedback on draft dissertation	03/14/23	03/28/23
56	Revisit dissertation	03/28/23	03/29/23
57	Submission of draft dissertation	03/27/23	03/30/23
58	Submission of final dissertation	04/24/23	04/27/23
59	Writing documentation for the product	04/28/23	05/04/23
60	Project presentation and Viva	05/15/23	05/15/23

Figure 28 – Gantt Chart

APPENDIX C – IMPLEMENTATION

Appendix C1 – Threshold Distribution

```python
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# Read the CSV file
# Final decided threshold value is 60
df = pd.read_csv("../datasets/tweetsum_train.csv")
# Extract the 'dialogue' column and convert it to a list
dialogues = df["dialogue"].tolist()


# Calculate the number of words in each dialogue
 mr-desilva
def input_length(text):
    return len(text.split())


dialogue_lengths = [input_length(dialogue) for dialogue in dialogues]

# Compute summary statistics
mean_length = np.mean(dialogue_lengths)
median_length = np.median(dialogue_lengths)
std_length = np.std(dialogue_lengths)
min_length = np.min(dialogue_lengths)
max_length = np.max(dialogue_lengths)

print(
    f"Mean: {mean_length}, Median: {median_length}, "
    f"Standard Deviation: {std_length}, Min: {min_length}, Max: {max_length}")

# Visualize the distribution using a histogram
plt.hist(dialogue_lengths, bins="auto")
plt.xlabel("Dialogue Length (Words)")
plt.ylabel("Frequency")
plt.title("Histogram of Dialogue Lengths")
plt.show()
```

Figure 29 – Threshold Distribution Method

Appendix C2 – Modal Evaluation

```
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM
from datasets import load_metric
from datasets import load_dataset

fine_tuned_modal_name = '../modal'
tokenizer_fine_tuned = AutoTokenizer.from_pretrained(fine_tuned_modal_name)
fine_tuned_modal = AutoModelForSeq2SeqLM.from_pretrained(fine_tuned_modal_name)

print('modal loaded!')


mr-desilva
def generate_dialogue_summary(dialogue):
    max_new_tokens = 50
    input_ids = tokenizer_fine_tuned(dialogue, return_tensors='pt').input_ids
    summary_ids = fine_tuned_modal.generate(input_ids, max_new_tokens=max_new_tokens)
    summary_text = tokenizer_fine_tuned.decode(summary_ids[0], skip_special_tokens=True)
    return summary_text

rouge = load_metric("rouge")
dataseteval = load_dataset("csv", data_files="../datasets/tweetsum_valid.csv")["train"]


generated_summaries = [generate_dialogue_summary(example["dialogue"]) for example in dataseteval]
ground_truths = [example["summary"] for example in dataseteval]

rouge_score = rouge.compute(predictions=generated_summaries, references=ground_truths)
rouge1_f1 = rouge_score["rouge1"].mid.fmeasure
rouge2_f1 = rouge_score["rouge2"].mid.fmeasure
rougeL_f1 = rouge_score["rougeL"].mid.fmeasure



print(rouge_score)
print("ROUGE-1:", rouge_score["rouge1"])
print("ROUGE-2:", rouge_score["rouge2"])
print("ROUGE-L:", rouge_score["rougeL"])
```

Figure 30 – Evaluating Model Using ROUGE Score

APPENDIX D – UI

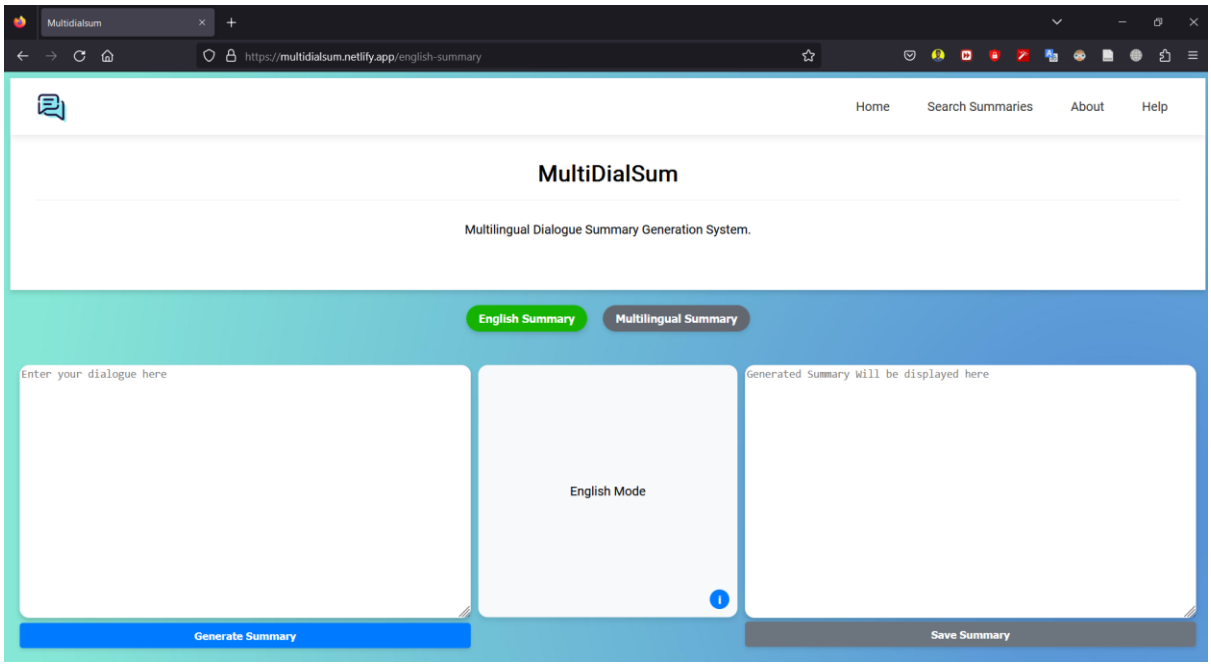


Figure 31 – Home Page

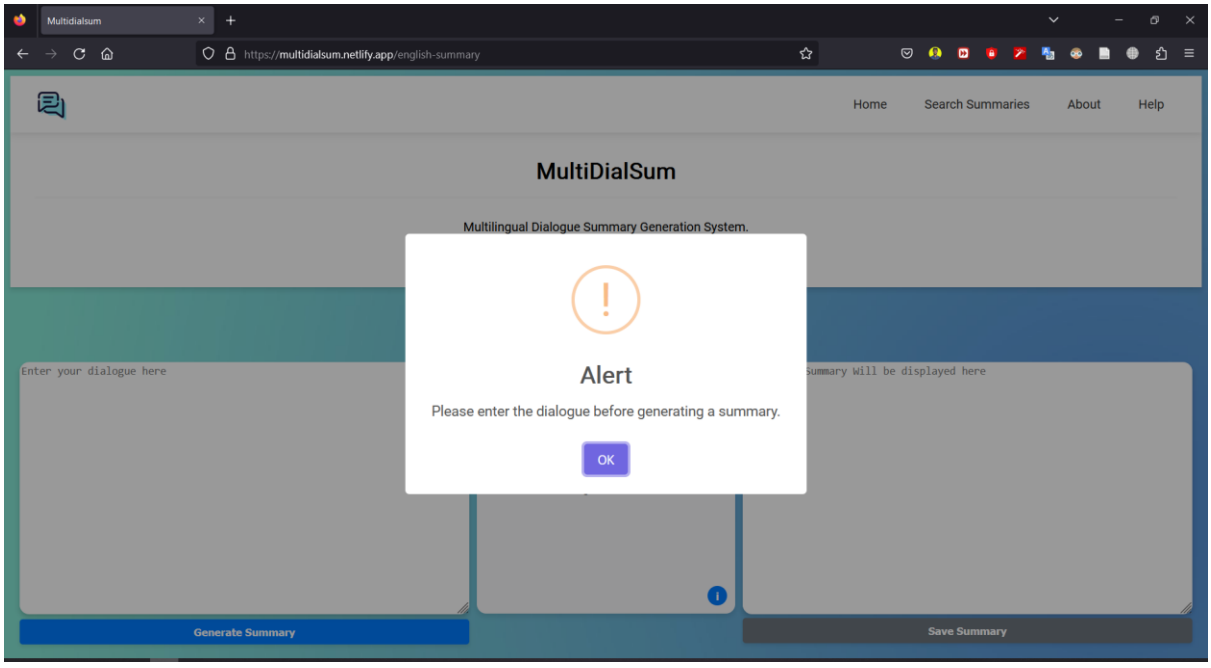


Figure 32 – Alert for Empty Input

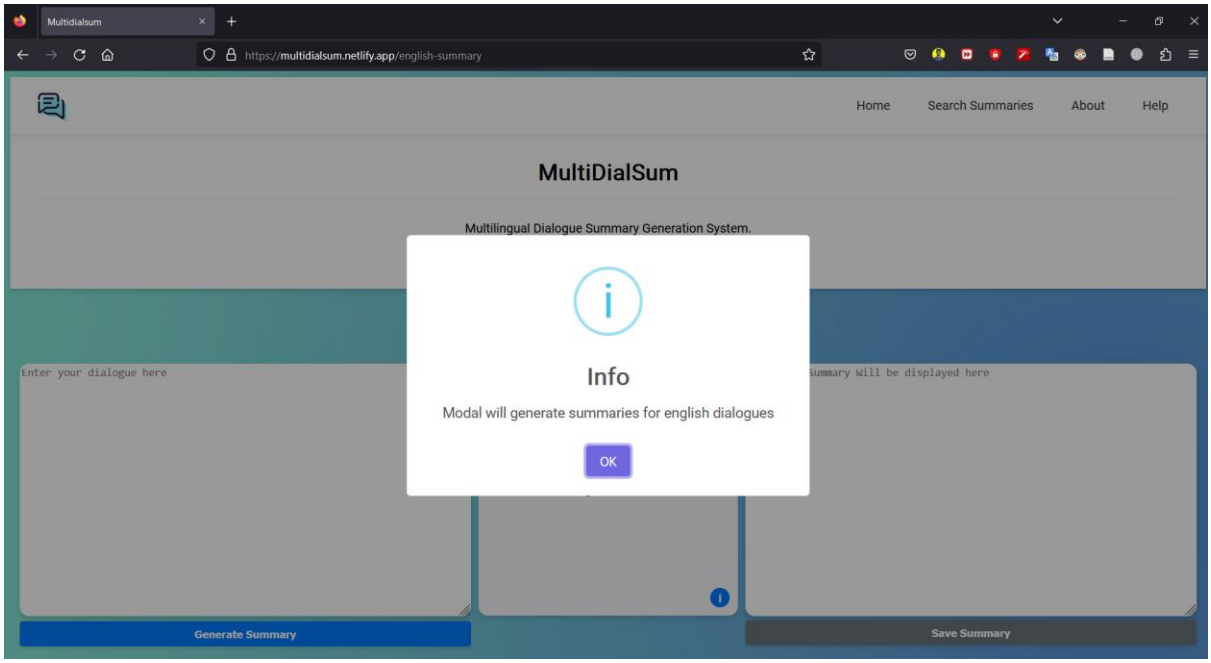


Figure 33 – English Summary Mode Info Card

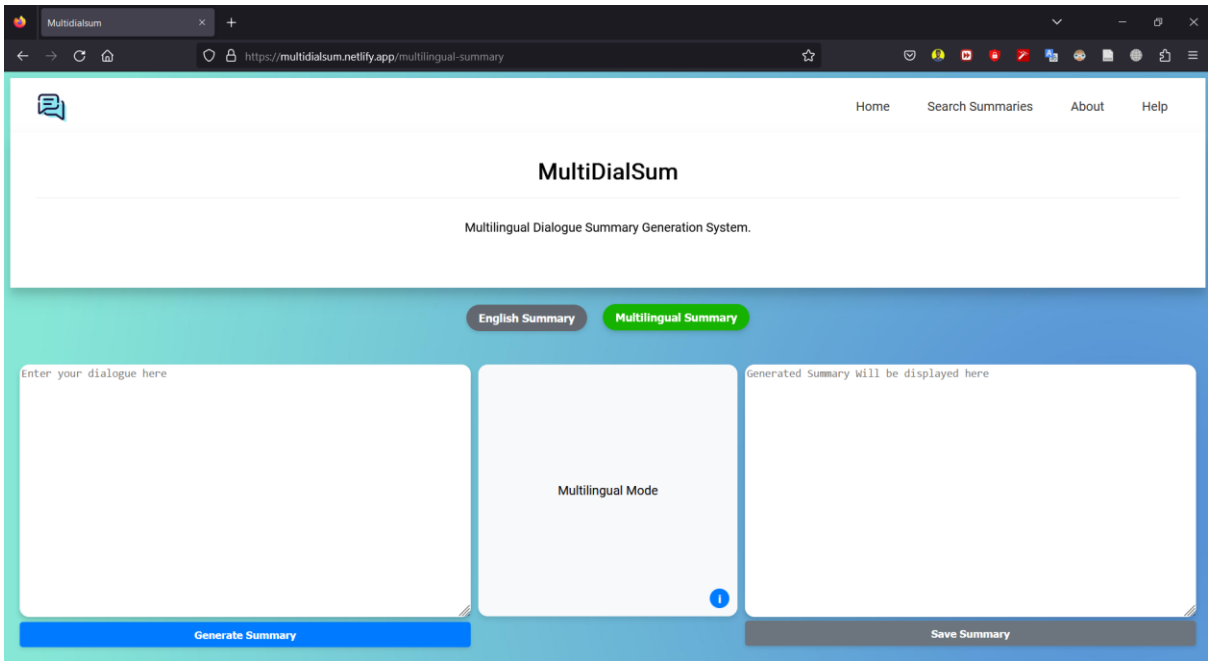


Figure 34 – Multilingual Mode

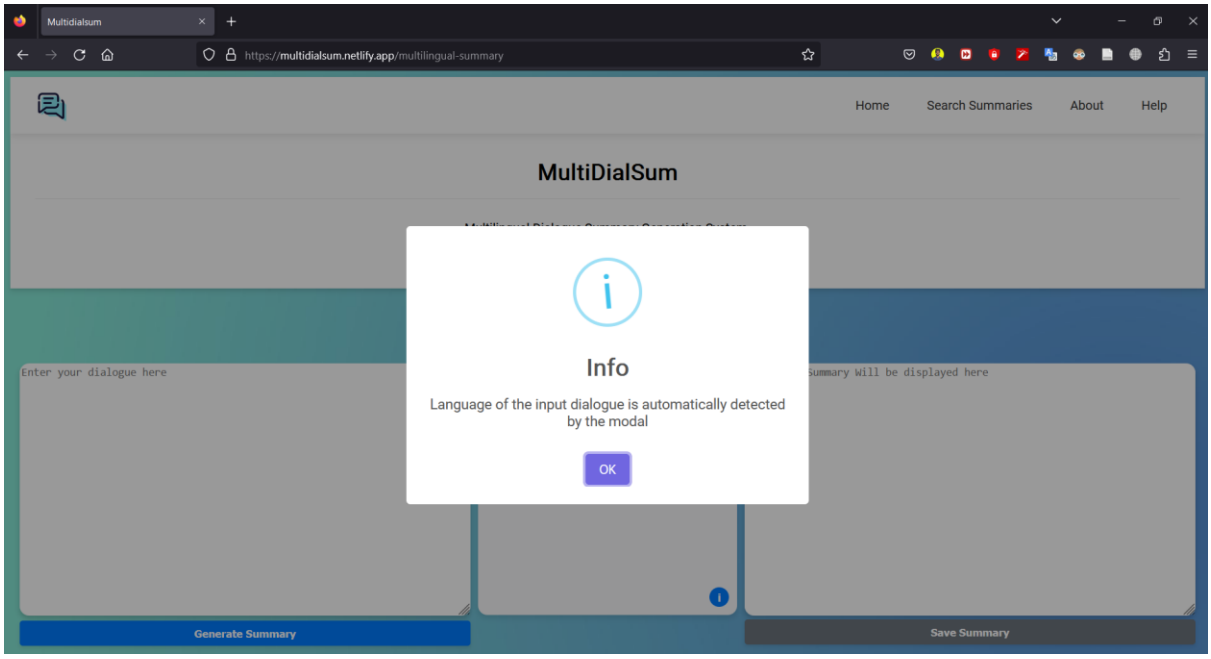


Figure 35 – Multilingual Summary Mode Info Card

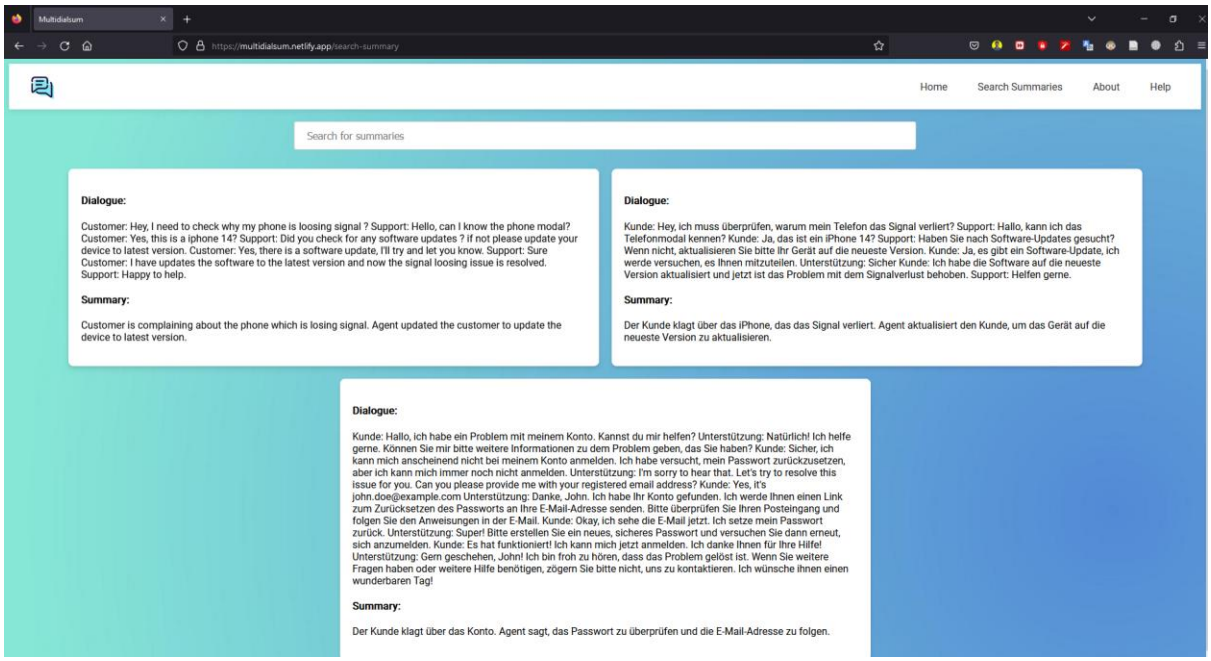


Figure 36 – Search Summaries

APPENDIX E – HUMAN EVALUATION SURVEY RESULTS

Appendix E1 - French Summary Human Evaluation Survey Results

Original Dialogue	Summary generated from the tool
Cliente: Hey, j'ai besoin de vérifier pourquoi mon téléphone perd le signal ? soutien: Bonjour, puis-je connaître le téléphone modal? Cliente: Oui, c'est un iphone 14 ? soutien: Avez-vous vérifié les mises à jour logicielles ? sinon, veuillez mettre à jour votre appareil vers la dernière version. Cliente: Oui, il y a une mise à jour logicielle, je vais essayer de vous tenir au courant. soutien: Bien sûr Cliente: J'ai mis à jour le logiciel vers la dernière version et maintenant le problème de perte de signal est résolu. soutien: Heureux de pouvoir vous aider.	Le client se plaint de l'iPhone qui perd le signal. l'agent a mis à jour le téléphone à la dernière version et maintenant le problème de la perte du signal est résolu.

Figure 37 – French Survey Dialogue 1

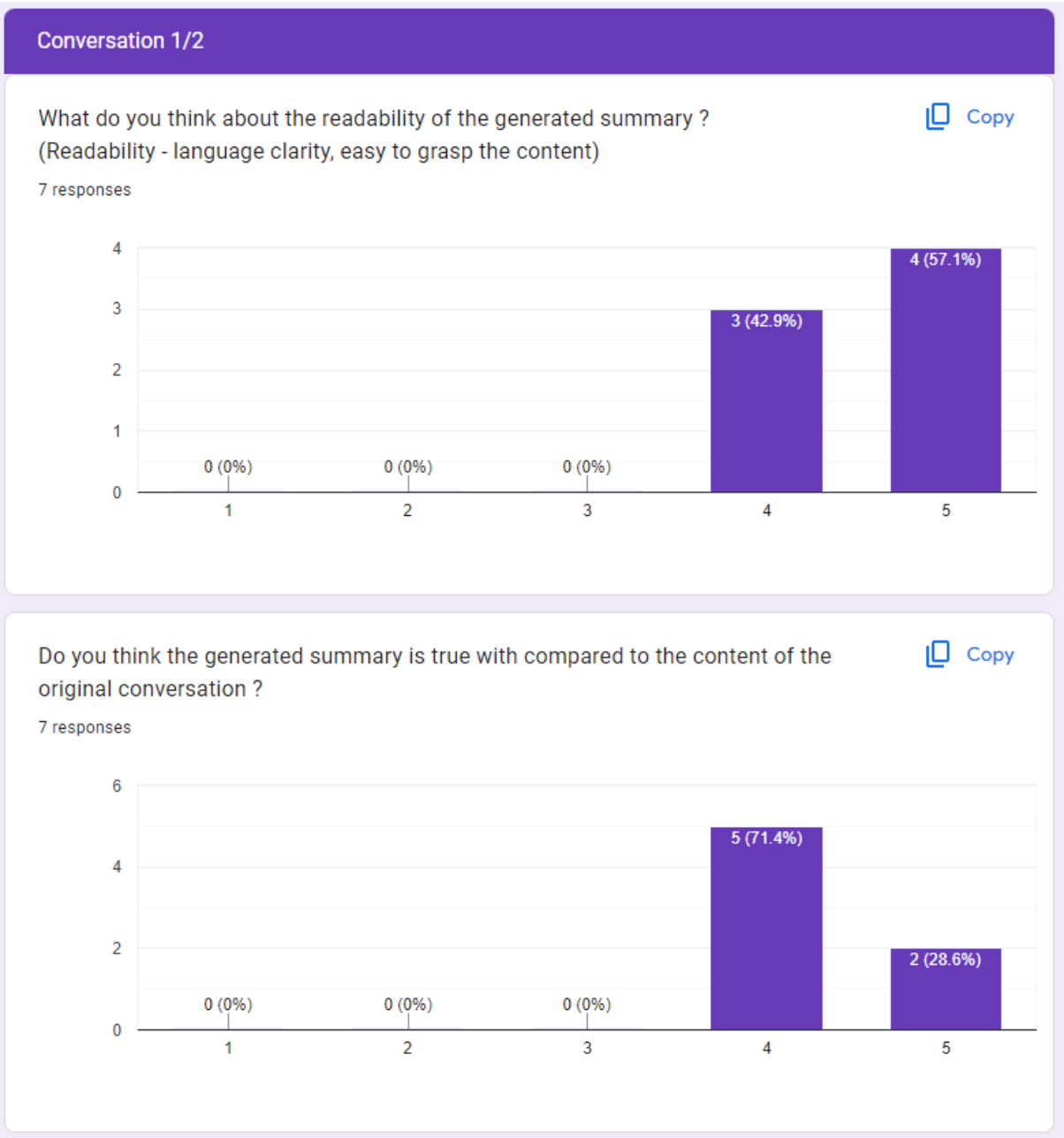


Figure 38 – French Survey Results Dialogue 1

Original Dialogue	Summary generated from the tool
Cliente: Bonjour, j'ai un problème avec mon compte. Pouvez-vous m'aider? soutien: Bien sûr! Pouvez-vous me donner plus d'informations sur le problème que vous rencontrez ? Cliente: Je n'arrive pas à me connecter à mon compte. J'ai essayé de réinitialiser mon mot de passe, mais je ne peux toujours pas entrer. soutien: Pouvez-vous s'il vous plaît me fournir votre adresse e-mail enregistrée? Cliente: Oui, c'est john.doe@example.com soutien: Je vais vous envoyer un lien de réinitialisation de mot de passe à votre adresse e-mail. Veuillez vérifier votre boîte de réception et suivre les instructions contenues dans l'e-mail. Cliente: D'accord, je vois l'e-mail maintenant. Je réinitialise mon mot de passe. soutien: Super! Veuillez créer un nouveau mot de passe sécurisé, puis réessayez de vous connecter. Cliente: Cela a fonctionné ! Merci pour votre aide ! soutien: Je t'en prie	Les clients se plaignent qu'ils ne peuvent pas se connecter à leur compte. l'agent dit qu'ils vont leur envoyer un lien de réinitiation de mot de passe à leur adresse e-mail.

Figure 39 – French Survey Dialogue 2

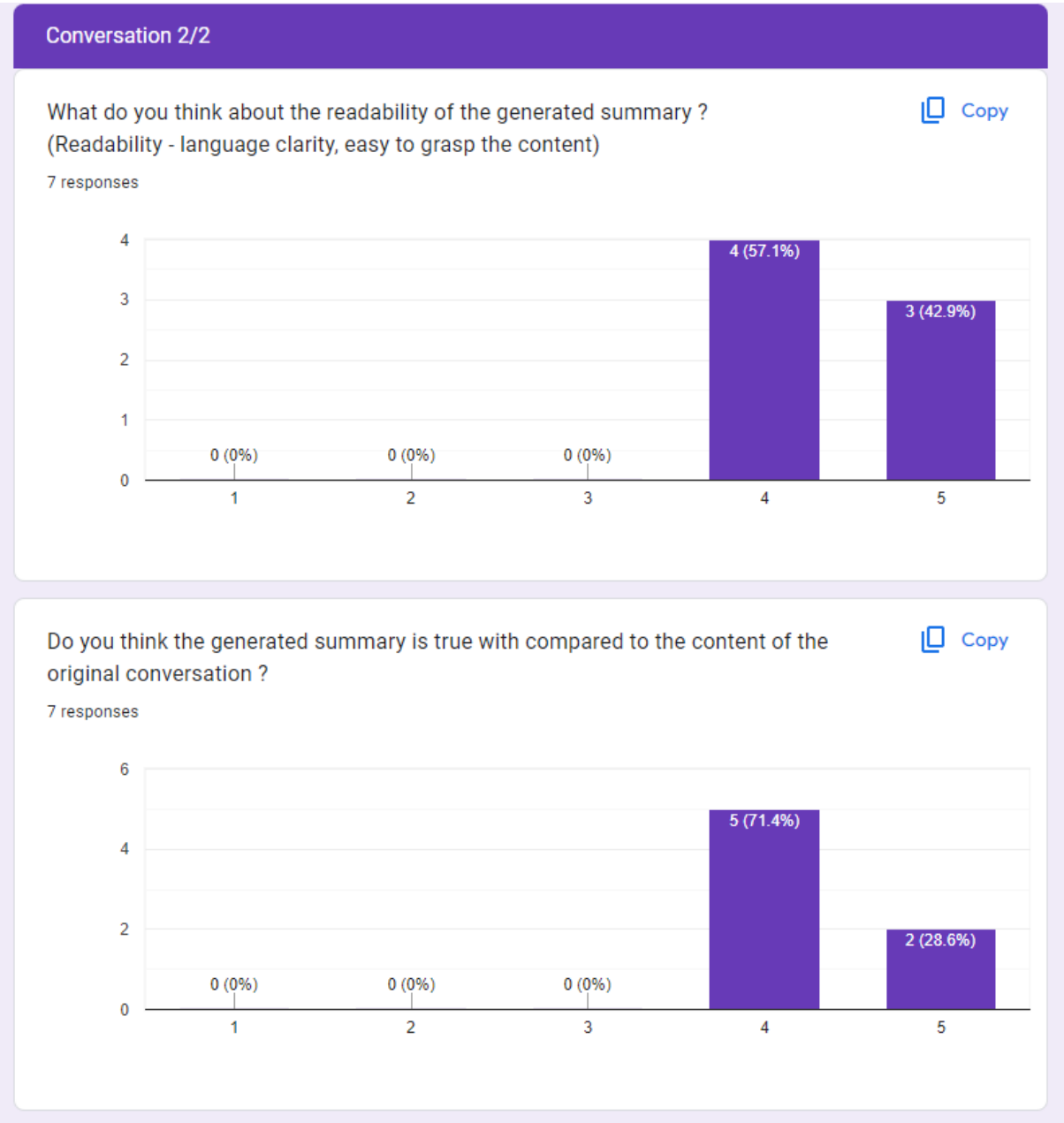


Figure 40 - French Survey Results Dialogue 2

Appendix E2 - Spanish Summary Human Evaluation Survey Results

Original Dialogue	Summary generated from the tool
Cliente: Hola, tengo algunos problemas con mi cuenta. ¿Me puedes ayudar? Soporte: ¡Por supuesto! ¿Puede por favor proporcionarme más información sobre el problema que está experimentando? Cliente: Parece que no puedo iniciar sesión en mi cuenta. He intentado restablecer mi contraseña, pero sigo sin poder entrar. Soporte: ¿Puede proporcionarme su dirección de correo electrónico registrada? Cliente: Sí, es john.doe@example.com. Soporte: Voy a enviarle un enlace de restablecimiento de contraseña a su dirección de correo electrónico. Por favor revise su bandeja de entrada y siga las instrucciones en el correo electrónico. Cliente: Bien, ahora veo el correo electrónico. Estoy restableciendo mi contraseña. Soporte: ¡Genial! Cree una nueva contraseña segura y luego intente iniciar sesión nuevamente. Cliente: ¡Funcionó! ¡Gracias por su ayuda! Soporte: De nada	El cliente se queja de que no puede acceder a su cuenta. El agente sugiere que revise su bandeja de entrada y siga las instrucciones en el correo electrónico.

Figure 41 - Spanish Survey Dialogue 1

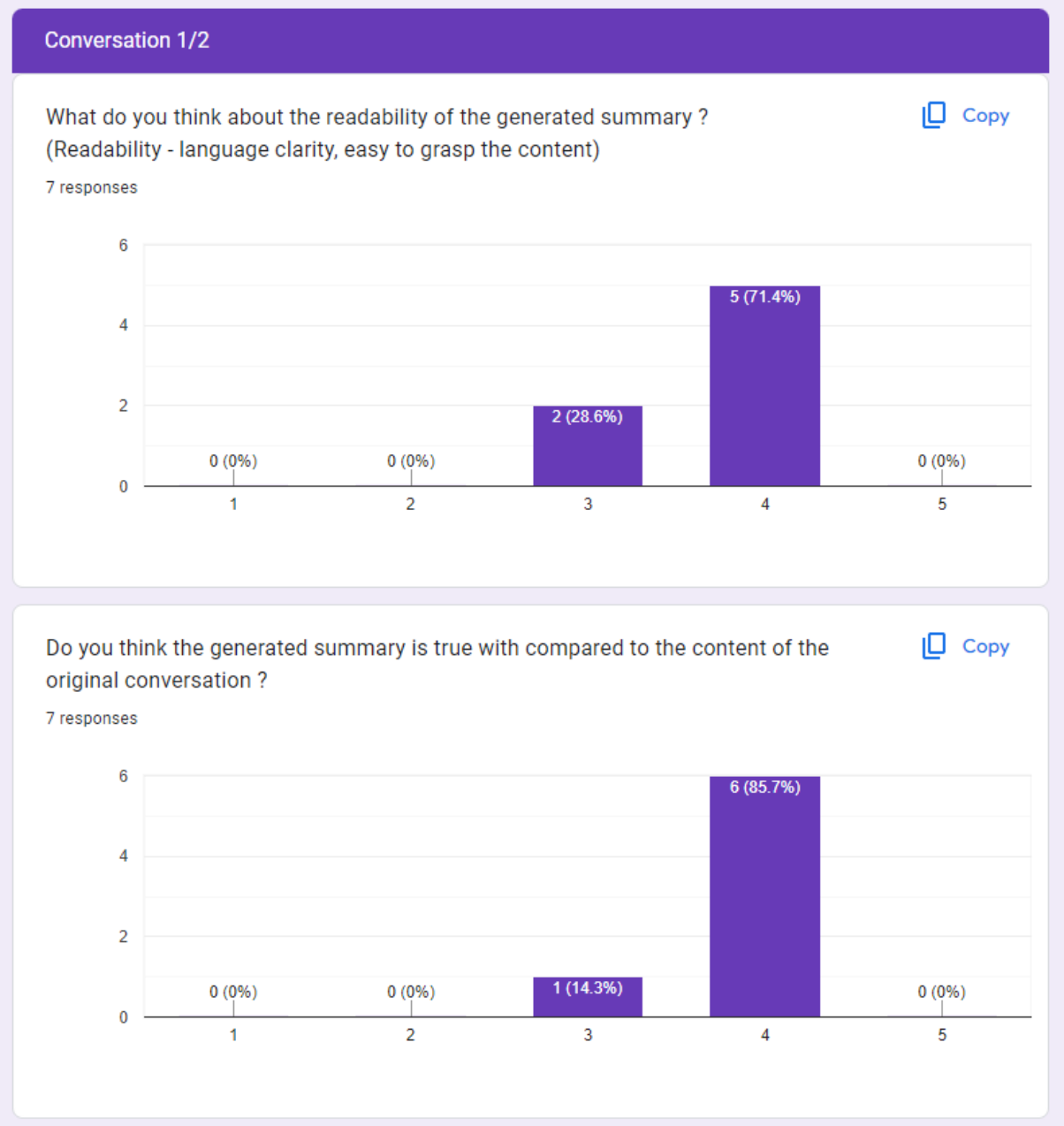


Figure 42 – Spanish Survey Results Dialogue 1

Original Dialogue	Summary generated from the tool
Cliente: Oye, necesito verificar por qué mi teléfono pierde señal. Soporte: Hola, ¿puedo saber el modal del teléfono? Cliente: Sí, ¿es un iphone 14? Soporte: ¿Buscaste alguna actualización de software? si no, actualice su dispositivo a la última versión. Cliente: Sí, hay una actualización de software, lo intentaré y te lo haré saber. Soporte: Claro Cliente: actualicé el software a la última versión y ahora se resolvió el problema de pérdida de señal. Soporte: Feliz de ayudar.	El cliente se queja de que el iPhone pierde señal. el agente sugiere actualizar el teléfono a la última versión y ahora se ha resuelto el problema de pérdida de señal

Figure 43 - Spanish Survey Dialogue 2

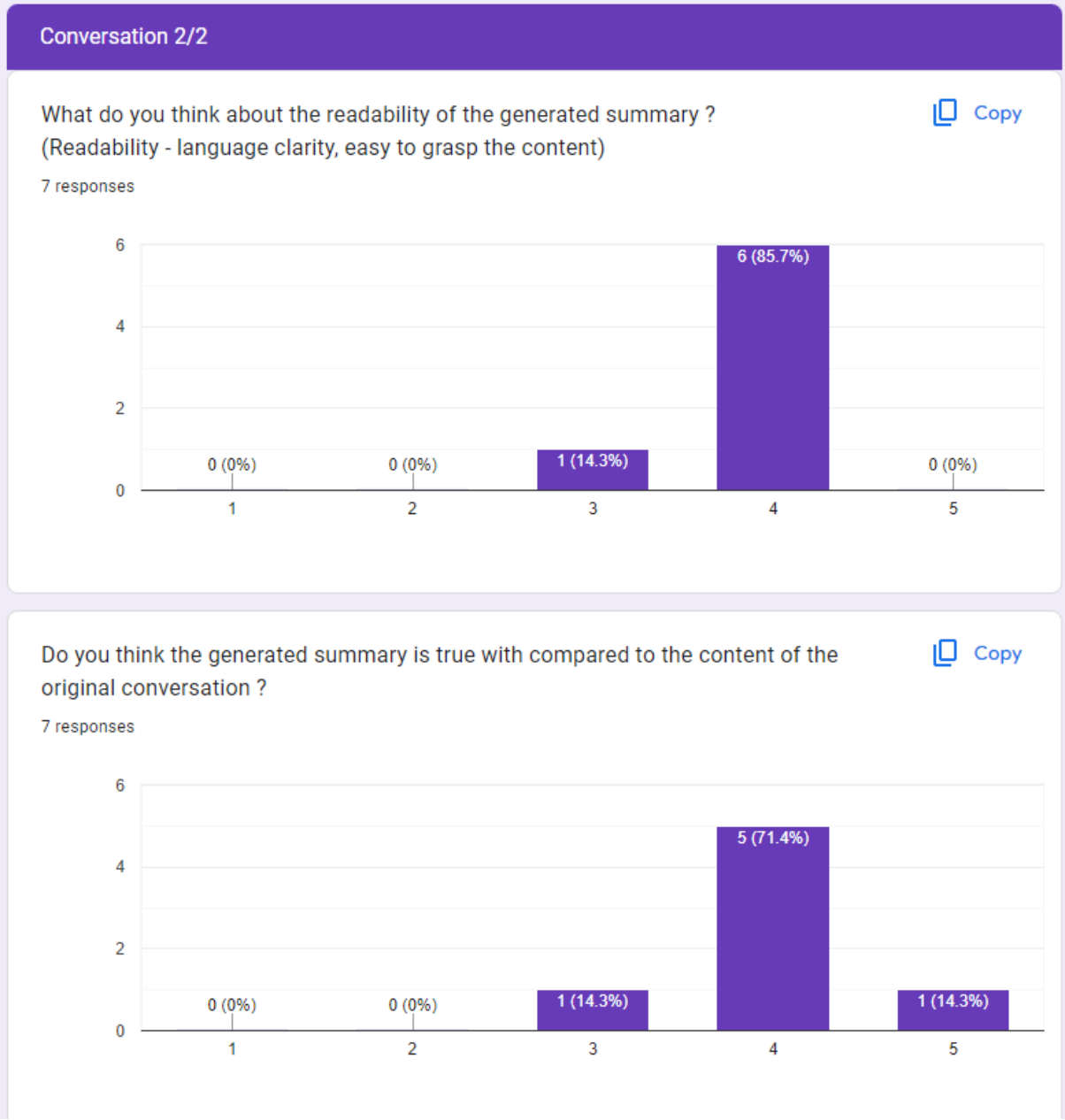


Figure 44 - Spanish Survey Results Dialogue 2

Appendix E3 - German Summary Human Evaluation Survey Results

Original Dialogue	Summary generated from the tool
Kunde: Hey, ich muss überprüfen, warum mein Telefon das Signal verliert? Agent: Hallo, kann ich das Telefonmodal kennen? Kunde: Ja, das ist ein iPhone 14? Agent: Haben Sie nach Software-Updates gesucht? Wenn nicht, aktualisieren Sie bitte Ihr Gerät auf die neueste Version. Kunde: Ja, es gibt ein Software-Update, ich werde versuchen, es Ihnen mitzuteilen. Agent: Sicher Kunde: Ich habe die Software auf die neueste Version aktualisiert und jetzt ist das Problem mit dem Signalverlust behoben. Agent: Helfen gerne.	Der Kunde klagt über das iPhone, das das Signal verliert. Agent aktualisiert den Kunde, um das Gerät auf die neueste Version zu aktualisieren.

Figure 45 - German Survey Dialogue 1

Figure 46 - German Survey Results Dialogue 1

Original Dialogue	Summary generated from the tool
Kunde: Hallo, ich habe ein Problem mit meinem Konto. Kannst du mir helfen? Unterstützung: Natürlich! Können Sie mir bitte weitere Informationen zu dem Problem geben, das Sie haben? Kunde: Ich kann mich anscheinend nicht bei meinem Konto anmelden. Ich habe versucht, mein Passwort zurückzusetzen, aber ich kann mich immer noch nicht anmelden. Support: Können Sie mir bitte Ihre registrierte E-Mail-Adresse mitteilen? Kunde: Ja, es ist john.doe@example.com. Support: Ich werde Ihnen einen Link zum Zurücksetzen des Passworts an Ihre E-Mail-Adresse senden. Bitte überprüfen Sie Ihren Posteingang und folgen Sie den Anweisungen in der E-Mail. Kunde: Okay, ich sehe die E-Mail jetzt. Ich setze mein Passwort zurück. Unterstützung: Super! Bitte erstellen Sie ein neues, sicheres Passwort und versuchen Sie dann erneut, sich anzumelden. Kunde: Es hat funktioniert! Vielen Dank für Ihre Hilfe! Unterstützung: Gerne	Der Kunde klagt, dass er nicht in der Lage ist, sich in sein Konto anzumelden. Agent sagt, dass er ihnen einen Link senden wird, um das Passwort auf ihre E-Mail-Adresse wiederherzustellen.

Figure 47 - German Survey Dialogue 1

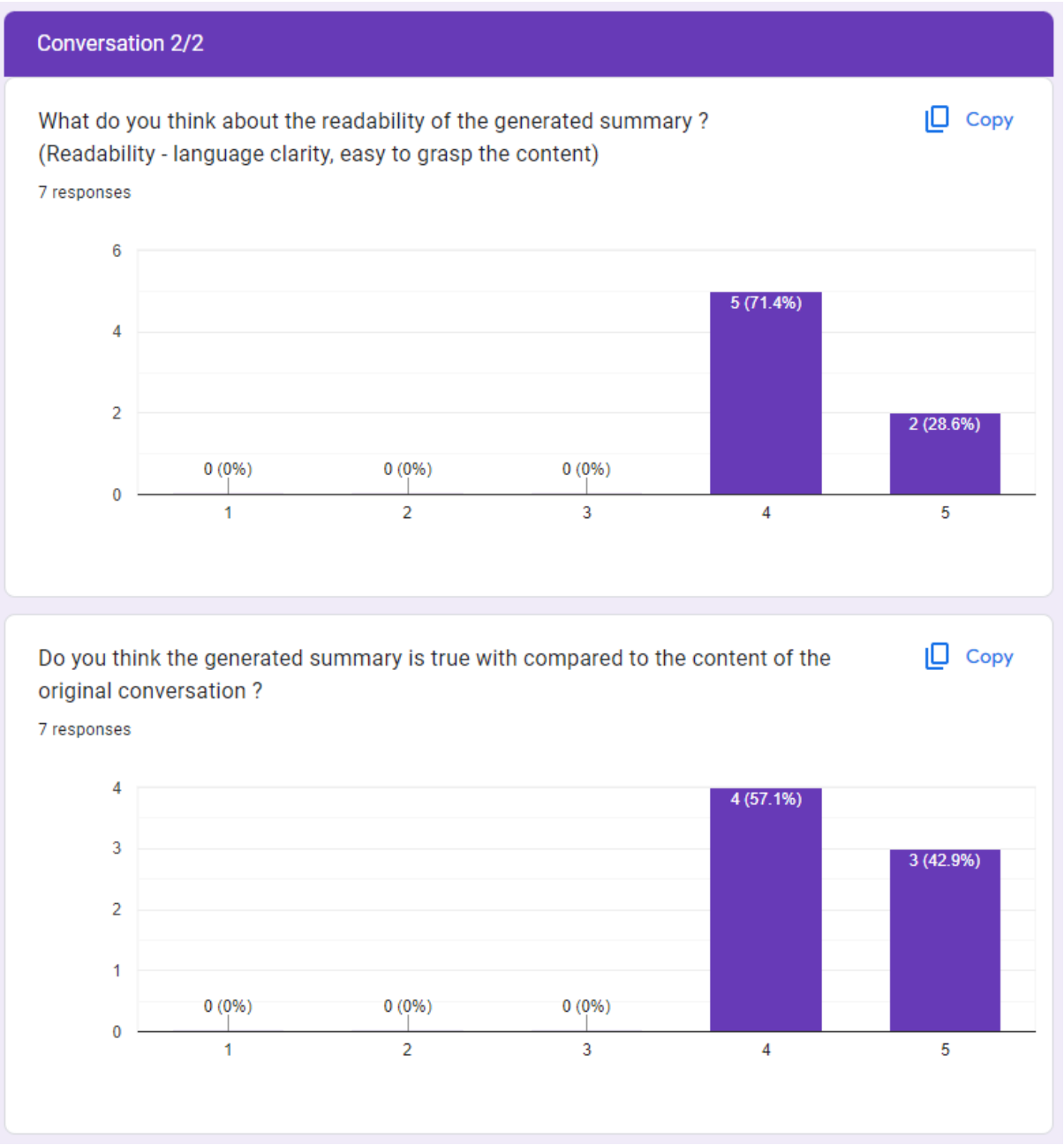


Figure 48 - German Survey Results Dialogue 2

APPENDIX F – EVALUATORS FEEDBACK

Evaluator ID	Feedback
1 (Ph.D. Student, Harbin Institute of Technology)	• Commendation for successful completion of innovative multilingual dialogue summarization system for customer service industry • Suggestions to improve the translation and the summary generation. • Addressing important aspects such as evaluation of summary quality, assessment of translation accuracy, tackling limitations, and exploration of domain adaptation techniques. • Encouragement to continue refining and expanding work.
1	• Commendation for successful completion of innovative multilingual dialogue summarization system. • Thoughtful approach to developing system with the integration of m2m100 model for multilingual support and fine-tuning of BART-large-xsum model on customer service data. • Encourage to explore alternative model architectures for further enhancing system performance.
2	• Appreciation for thoughtful and user-friendly system design and features that ensure practicality in customer service industry. • Inclusion of features such as the ability to search through saved summaries for efficient review and analysis of past interactions. • Valuable addition of option to generate summaries in multiple languages, catering to diverse linguistic needs of global audience and making application accessible to broader range of users. • Dedication to creating versatile and efficient system that addresses unique demands of customer service industry is commendable. • Overall design and features demonstrate thorough understanding of field and challenges faced by customer service professionals. • Project serves as good contribution to ongoing improvement of customer service practices and overall customer experiences.

2	• Congratulating on successful completion of innovative project focused on summarizing customer interactions across multiple languages in customer service industry. • Potential to transform way customer service teams analyze and understand customer conversations, leading to more effective support strategies and improved customer experiences. • Thorough consideration of important aspects such as generating valuable and relevant summaries for customer service professionals and providing accurate translations for a wide-ranging, international audience • Resulting in versatile and efficient system that is highly applicable in customer service industry
2	• Expressing admiration for work on developing multilingual dialogue summarization system for customer service field • Potential to simplify and improve way customer interactions are reviewed and understood, leading to better customer support experiences. • Appreciation for system's ability to handle multiple languages, making it useful for people worldwide.

Table 42 – Evaluators Feedback