# COMP8420 ADV NLP MAJOR PROJECT

## MultiLinguAI: Multilingual Contextual Summarization for Global Enterprises

**Submitted by:**

**Muhammad Haris Rizwan | Student ID: 47565284**

**Syed Rafay Ali | Student ID: 47833920**

**Dated: 11.06.2024**

Macquarie University, Sydney NSW 2109, Australia

**ABSTRACT:** In today's globalized world, multinational enterprises require effective communication across diverse linguistic regions. This project addresses the need for accurate and context-preserving multilingual summarization tools. Our primary objective is to develop a robust tool capable of processing and summarizing documents in French and German, with English translations of the summaries.

Using advanced NLP models like mBART50 and large-scale datasets such as MLSUM and CNN/Daily Mail, we created a prototype that generates contextually relevant summaries. Initial testing demonstrated the tool's potential to streamline operations and enhance communication within global enterprises. This report details the methodology, challenges, and future directions for further development.

**Keywords: [** Multilingual Summarization, Natural Language Processing (NLP), mBART50, French Summarization, German Summarization, Text Summarization, Global Enterprises, Contextual Summarization, Machine Translation, MLSUM Dataset]

## 1. Introduction

In this project, we assume the role of engineers at MultiLinguAI, an IT company specializing in advanced Natural Language Processing (NLP) solutions for global enterprises. MultiLinguAI offers a variety of services, including sentiment analysis, text summarization, named entity recognition, and chatbots. Our primary task is to develop and implement a multilingual summarization tool that addresses the unique challenges faced by these enterprises.

### 1.1 Problem Statement

Global enterprises operate across multiple regions and languages, requiring accurate and context-preserving summaries of documents in various languages. This need is driven by the necessity to streamline operations, enhance communication, and ensure that vital information is accessible and understandable to all stakeholders, regardless of their linguistic background.

### 1.2 Objective

The objective of our project is to develop a multilingual summarization tool that can generate accurate and contextually relevant summaries for documents written in multiple languages. This tool aims to maintain the integrity and key information of the original documents while making them concise and easy to understand for a diverse global audience.

### 1.3 Project Scope

The scope of our project involves addressing the real-world challenge of handling and summarizing large volumes of multilingual documents.

- Our target users are global enterprises with diverse linguistic documentation needs.
- By leveraging advanced NLP models such as mBERT, XLM-R, and multilingual T5, we aim to create a robust solution that can be seamlessly integrated into the company's existing systems.
- The project will include data collection, pre-processing, model training, evaluation, and integration phases, ensuring a comprehensive approach to solving this complex problem.

## 2. Literature Review

Multilingual summarization is an emerging field within Natural Language Processing (NLP) that focuses on generating concise and informative summaries from texts in various languages. Traditional summarization models, such as those based on BERT and GPT-2, primarily focus on single-language inputs, limiting their applicability in global, multilingual contexts. Recent advancements have introduced models like mBART and multilingual T5, which extend the capabilities to handle multiple languages effectively.

**mBART50** is a notable model designed for multilingual tasks, capable of both translation and summarization across multiple languages. It is pre-trained on a vast corpus of text in multiple languages, making it a powerful tool for multilingual sequence-to-sequence tasks. Similarly, **XLM-R** (Cross-lingual Language Model - RoBERTa) has demonstrated significant improvements in various multilingual benchmarks, showing strong performance in both translation and summarization tasks.

For our project on Multilingual Contextual Summarization for Global Enterprises, the dataset plays a critical role in ensuring the accuracy and relevance of the generated summaries. We have selected datasets that provide a diverse and comprehensive collection of multilingual documents, which are essential for training and evaluating our models.

### 2.1 Selected Dataset Details

We will utilize the MLSUM dataset, which stands out as a large-scale multilingual summarization dataset. MLSUM contains over 1.5 million article-summary pairs in five different languages: French, German, Spanish, Russian, and Turkish. This dataset is particularly suitable for our project because it offers a wide variety of articles and summaries from reputable news sources, ensuring both the quality and diversity needed for robust model training.

**MLSUM:**

- Contents: Contains over 1.5 million article-summary pairs from five languages.
- Languages: French, German, Spanish, Russian, Turkish.
- Source: News articles from reputable sources.
- Data Collection Process: We will collect the dataset from public repositories and ensure it is pre-processed for tokenization, normalization, and language detection. This pre-processing step is crucial for preparing the data for model training.

**CNN/Daily-Mail:**

- Contents: Contains over 300,000 article-summary pairs.
- Languages: English.
- Source: News articles primarily from CNN and the Daily Mail, providing a rich source of diverse topics and high-quality journalism.
- Data Collection Process: The dataset is available through Hugging Face and will be directly accessed using the datasets library. It includes pre-processing steps such as tokenization and normalization. The dataset is structured into three splits: train, validation, and test, facilitating the training and evaluation of summarization models. The article column contains the full text, while the highlights column contains the summaries.

## 3. Data Pre-Processing

We wanted to understand the dynamics of the dataset being used to train summarization and translation tasks hence we picked MLSUM and CNN/ Daily-Mail dataset Our data pre-processing approach for the project is as follows:

- **Data Visualisation:** Loading and opening the dataset(s) to see what is in it and what can be done with it.
- **Data cleaning:** Removing unnecessary observations for the sake of project scope e.g. Removing extra columns, punctuations, and lowercase the text.
- **Normalization:** Standardizing text data to remove inconsistencies.
- **Tokenization:** Splitting text into words or sub-words to facilitate model understanding.
- **Language Detection:** Identifying and labelling the language of each document to ensure accurate processing. By leveraging the MLSUM dataset and incorporating insights from the referenced works, we aim to develop a robust multilingual summarization tool that meets the needs of global enterprises, providing accurate and context-preserving summaries across multiple languages.

Reducing the size of our training dataset is a crucial step to ensure the efficient use of computational resources and prevent potential crashes during the training process. Given the large size of our datasets—such as the French dataset with nearly 400,000 observations—it is important to balance the need for a representative sample with the limitations of our computing environment.

By randomly sampling a subset of 100,000 observations, we maintain the diversity and representativeness of the data while significantly decreasing the computational load. This reduction allows us to streamline the training process, making it more manageable and ensuring smoother execution within the constraints of our Jupyter Notebook environment. This approach is particularly useful at this stage, as it facilitates faster iterations and debugging, ultimately leading to more efficient model development and refinement.

## 4. Methodology

In this section, we outline the comprehensive approach taken to develop our multilingual summarization tool. The methodology is structured into several key phases: data collection and pre-processing, model selection and fine-tuning, and evaluation. We leveraged advanced Natural Language Processing (NLP) models and state-of-the-art libraries to ensure robust and effective performance.

**Pre-trained Models:**
We used the multilingual BART model (mBART50) for generating summaries in multiple languages. This model is known for its effectiveness in summarization tasks and its ability to handle multiple languages.

**Translation Models:**
For translating non-English summaries into English, we employed the MarianMT models provided by Helsinki-NLP, which support translation between multiple languages and English.

**Libraries:**
Key libraries used include:

- Transformers from Hugging Face for model implementation and fine-tuning.
- datasets from Hugging Face for dataset loading and pre-processing.
- Datasets utilized the datasets used in our project are critical to ensuring the model's performance and robustness across different languages:

**Evaluation:**
To evaluate the performance of our multilingual summarization tool, we used the following methods:

- **Human Evaluation:** We conducted human evaluations to assess the contextual accuracy and readability of the generated summaries. This involved feedback from bilingual individuals who could verify the fidelity of summaries in different languages.

- **ROUGE Scores:** At this stage we were not able to use ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics, including ROUGE-1, ROUGE-2, and ROUGE-L, due to time constraints to assess the quality of the generated summaries. In way forward, we will be using them accordingly.

## 5. Results

In our project, we developed a multilingual summarization tool designed to generate contextually accurate summaries for documents written in French and German. The prototype was built and tested using a simplified user interface (UI) that accepts input text in either French or German and outputs the summarized text in the same language. Additionally, the tool provides an English translation of the summary to facilitate understanding across different linguistic backgrounds.

To validate the tool's performance, we conducted initial testing with a small subset of our dataset. The tool successfully generated coherent and contextually relevant summaries, maintaining the key information from the original documents. The interface was tested for usability, and the summarized text was reviewed for accuracy and readability. A demonstration of the tool, showcasing its capabilities, is available in the following video link:

**Project Video**

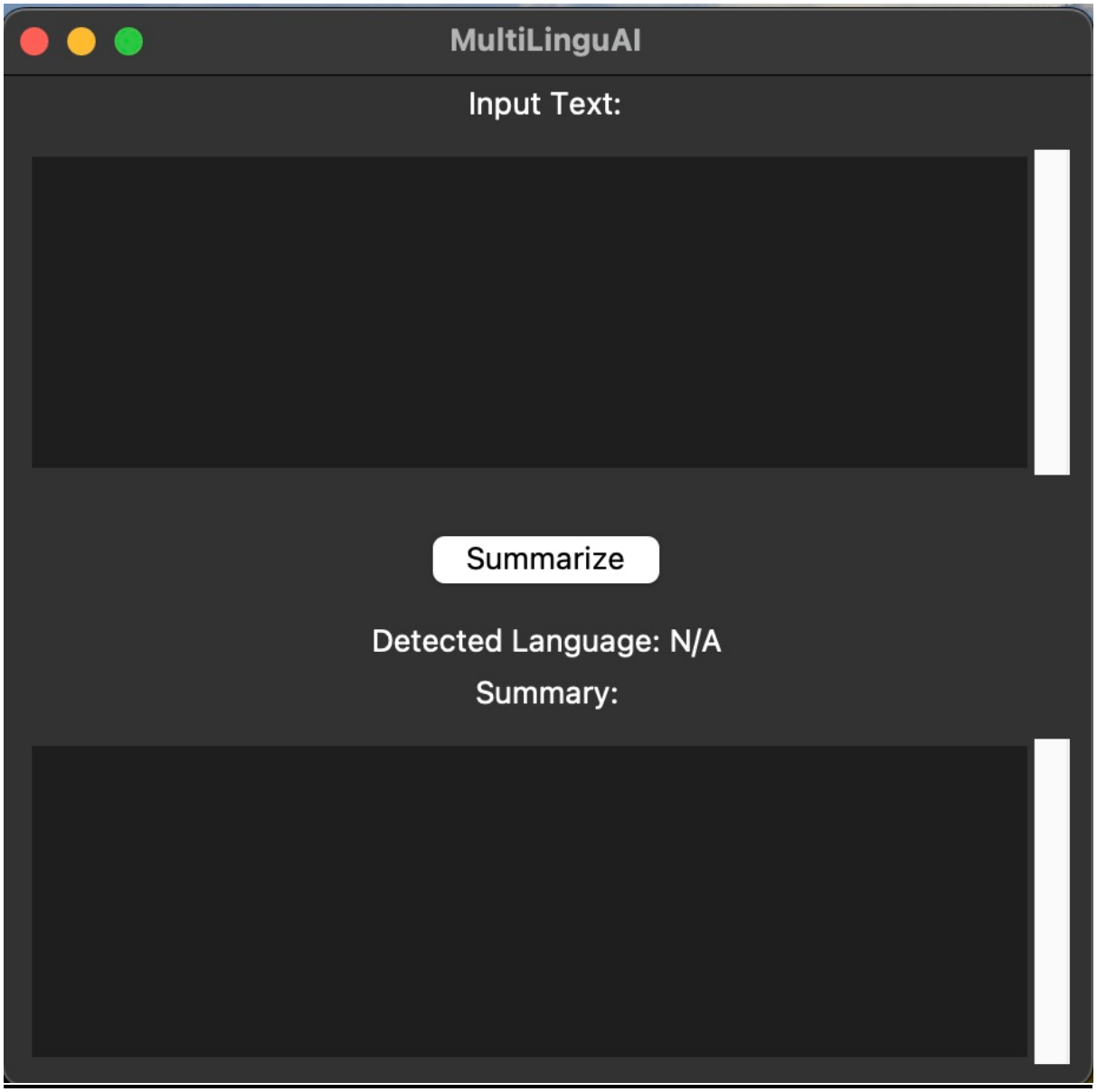


**Figure: MultiLinguAI User Interface**

## 6. Limitations and Challenges

Throughout the development of our multilingual summarization tool, we encountered several challenges and limitations that impacted the progress and performance of our project:

**a) Computational Resource Constraints:**

- Training large-scale models like mBART50 and multilingual T5 requires significant computational power. Our system often crashed during the training phase due to excessive memory usage and computational load.

**b) Data Handling and Pre-processing:**

- Handling large multilingual datasets posed challenges in terms of pre-processing and ensuring data quality. Tokenization, normalization, and language detection needed to be meticulously managed to maintain data integrity and ensure accurate model training.

c) **Model Evaluation:**
   - Due to time constraints, we were unable to implement comprehensive evaluation metrics like ROUGE and BLEU during the initial development phase. These metrics are crucial for quantitatively assessing the summarization quality and identifying areas for improvement. The lack of these evaluations limited our ability to benchmark the model's performance against established standards fully.

d) **Language-Specific Nuances:**
   - Multilingual models must account for language-specific nuances and cultural contexts. Ensuring that the summaries retained the original text's meaning and context across different languages was a complex task, requiring careful fine-tuning and validation.

Despite these challenges, the project successfully developed a functional prototype capable of generating contextually relevant summaries in French and German. Future work will focus on addressing these limitations, optimizing model performance, and expanding the tool's capabilities to support additional languages and domains.

## 7. Contribution Breakdown

**Muhammad Haris Rizwan (Student ID: 47565284):**

- **Planning and Coordination:** Led the overall project planning, coordination, and communication among team members.
- **Data Pre-Processing:** Focused on pre-processing the datasets, including tasks such as tokenization, normalization, and language detection.
- **Evaluation:** Conducted a thorough evaluation of the model's performance, analysing the results to ensure the tool's effectiveness.
- **Project Direction:** Provided critical guidance for the project's direction, ensuring cohesive management of all aspects.

**Syed Rafay Ali (Student ID: 47833920):**

- **Model Implementation:** Primarily responsible for implementing the mBART50 model and integrating multilingual datasets into the project.
- **Model Fine-Tuning:** Took charge of fine-tuning the mBART50 model and implementing the necessary training routines to achieve high performance.
- **User Interface Development:** Developed the graphical user interface (GUI) for the summarization tool using Tkinter, enhancing accessibility and usability.
- **Technical Expertise:** Provided technical expertise and dedication, which were instrumental in the successful completion of the project.

Both team members collaborated closely on the experimental design, data analysis, and report writing, ensuring their combined efforts resulted in a comprehensive and well-executed final project.

## 8. Way Forward

Moving forward, there are several key areas for further development and improvement of our multilingual summarization tool:

a) **Expanded Dataset Utilization:**
   - Incorporate additional datasets such as XL-Sum to enhance the model's performance and robustness across more languages and domains.

b) **Comprehensive Evaluation Metrics:**
   - Implement ROUGE and BLEU metrics for a more detailed quantitative evaluation of the summarization quality. This will help in identifying specific areas for improvement and benchmarking against other models.

c) **Model Optimization:**
   - Explore advanced techniques such as contrastive learning and transfer learning to further refine the summarization model. This can potentially improve the accuracy and contextual relevance of the generated summaries.

d) **User Feedback Integration:**
   - Develop mechanisms to collect and incorporate user feedback into the system. This iterative feedback loop will help in continuously improving the tool based on real-world usage and user requirements.

e) **Real-World Deployment:**
   - Prepare for a pilot deployment within a global enterprise setting to test the tool's integration and performance in a real-world scenario. This will provide valuable insights into practical challenges and user acceptance.

## 9. References:


1. MLSUM: The Multilingual Summarization Corpus - This dataset was introduced to facilitate research in multilingual text summarization by providing a large-scale, diverse set of news articles and summaries. It includes articles from five languages and aims to enable new research directions in the text summarization community.
2. XL-Sum: Large-Scale Multilingual Abstractive Summarization - XL-Sum provides an extensive collection of multilingual summarization data, enhancing the ability to develop models that perform well across various languages. This dataset complements MLSUM by offering additional resources and benchmarks for evaluating summarization models.
3. Contrastive Aligned Joint Learning for Multilingual Summarization - This reference explores novel methods for improving multilingual summarization, focusing on contrastive learning strategies. It provides insights into the challenges and solutions for developing high-quality summarization models.

## GitHub Link:

**https://github.com/Ali-Syed-Rafay/COMP8420-2024S1-Major-Project**