Data Collection and Preprocessing Phase

Date	5 February 2026
Team ID	LTVIP2026TMIDS66217
Project Title	**TransLingua – AI-Powered Multi-Language Translator**
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report:
This report outlines the data collection strategy and identifies the raw data sources used in the TransLingua project. Since the system operates using a pre-trained large language model, the focus is on real-time text input collection and maintaining data quality rather than sourcing static datasets.

Data Collection Plan:

Section	Description
Project Overview	The TransLingua project aims to provide accurate and context-aware language translation using a pre-trained generative AI model. The system accepts real-time user input text and translates it into a selected target language through an interactive web interface.
Data Collection Plan	• Collect real-time textual input from users via the Streamlit application. • Allow users to select source and target languages dynamically. • Validate input text and language selections before processing. • Use prompt-based preprocessing to prepare text for AI model inference.
Raw Data Sources Identified	The raw data source consists of real-time user-provided text input entered through the web interface. No external datasets are required, as translations are generated dynamically using the Gemini Pro large language model.

Raw Data Sources Report:

Source Name	Description	Location / URL	Format	Size	Access Permissions
User Input (Streamlit UI)	Text entered by users for translation, including multilingual sentences and phrases	Streamlit Web Application	Text	Dynamic	User-controlled
Gemini Pro Model Output	AI-generated translated text produced in response to user input	Google Generative AI API	Text	Dynamic	API-based access