Data Collection and Preprocessing Phase

Date	2 February 2026
Team ID	LTVIP2026TMIDS66217
Project Title	TransLingua – AI-Powered Multi-Language Translator
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report
Unlike traditional machine learning projects that rely on static datasets, TransLingua operates on **real-time user-provided textual data**. Data exploration and preprocessing focus on handling dynamic text input, validating user input, and preparing text prompts suitable for large language model processing. Python is used to preprocess and structure the text before sending it to the Gemini Pro model, ensuring accurate, context-aware translation results.

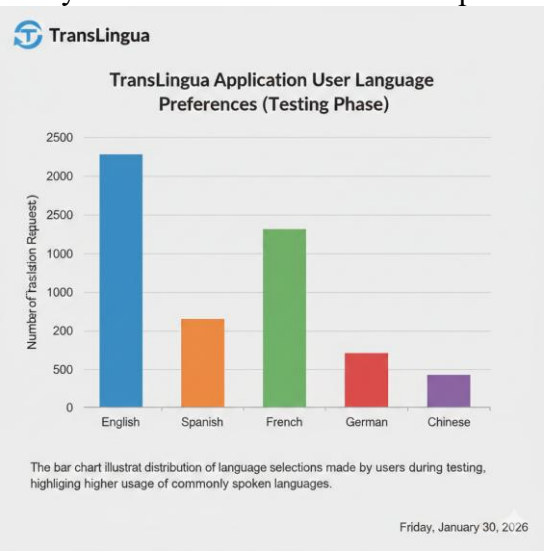
Data Overview

Section	Description
Data Type	Textual data (user-provided input text)
Data Source	Real-time user input via Streamlit interface
Data Format	Plain text
Data Size	Dynamic (varies per user input)
Nature of Data	Multilingual textual content

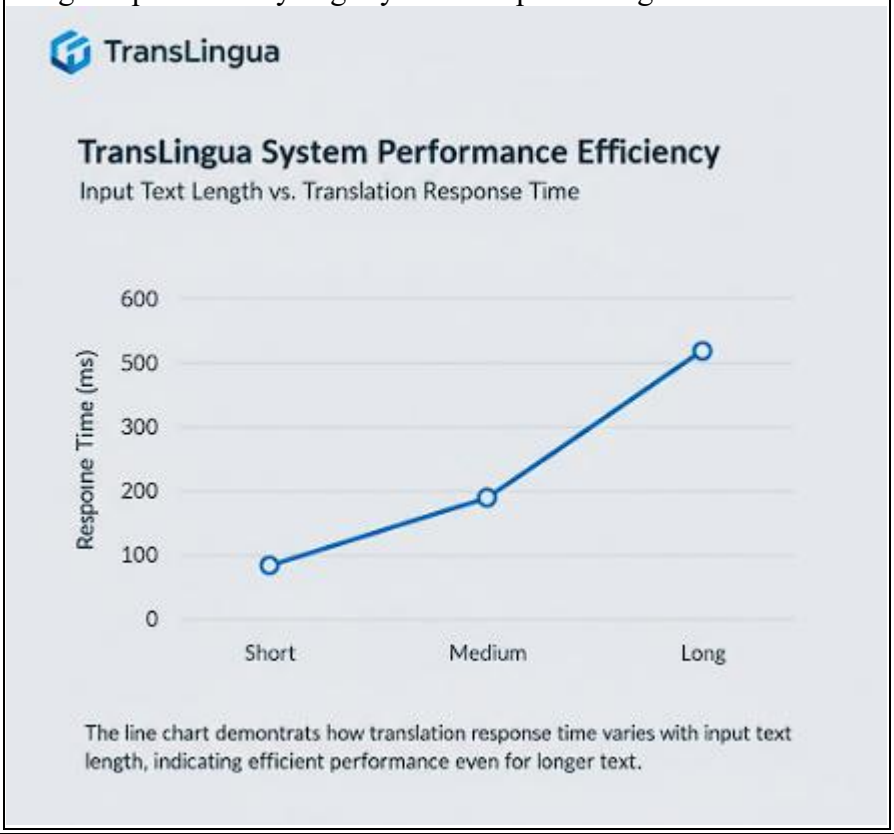
Descriptive Statistics

Aspect	Description
Statistical Measures	Not applicable, as the project does not use a fixed numerical dataset
Text Characteristics	Variable length sentences, multilingual content
Distribution	Depends on user input language and content

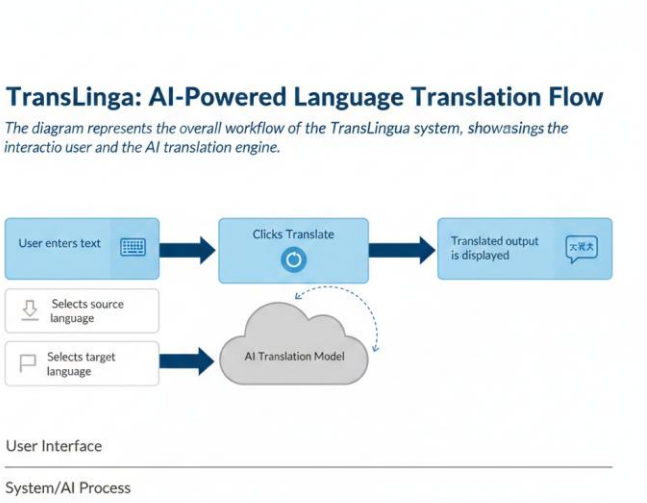
Univariate Analysis

Analysis	Description
Input Text Length	Analysis based on individual text input length 
Language Type	Single-language input per request
Character Distribution	Alphabetic, numeric, and special characters

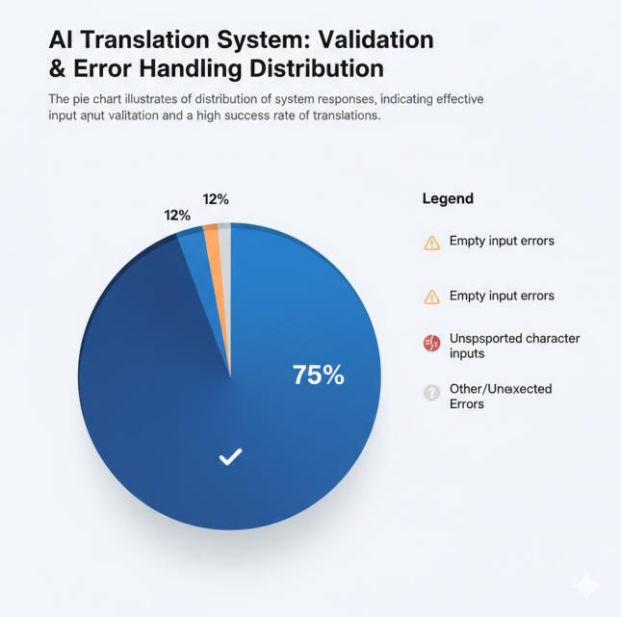
Bivariate Analysis

Analysis	Description
Source vs Target Language	Relationship between selected source and target languages
Text Length vs Response Time	Longer input text may slightly increase processing time 

Multivariate Analysis

Analysis	Description
Text + Source Language + Target Language	Combined influence on translation accuracy and response quality
User Input Patterns	Variation in usage across different language combinations 

Outliers and Anomalies

Category	Description
Outliers	Extremely long text inputs
Anomalies	Unsupported characters or empty inputs
Handling Method	Input validation and warning messages in UI 

Data Preprocessing Steps

Step	Description
Loading Data	Text is captured directly from the Streamlit text input field. No external dataset loading is required as the system operates on real-time user input.
Handling Missing Data	Empty or null input is handled through input validation. Users are prompted to provide valid text before the translation process begins.
Data Transformation	Text normalization is performed by trimming extra spaces and formatting the input into a structured prompt suitable for the Gemini Pro model. Encoding is handled internally by the model.
Feature Engineering	Source language and target language selections are treated as contextual parameters. Prompt engineering techniques are applied to improve translation accuracy and contextual relevance.
Save Processed Data	Processed data is not stored permanently. Translations are generated and displayed to the user in real time without persistent storage.