

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.

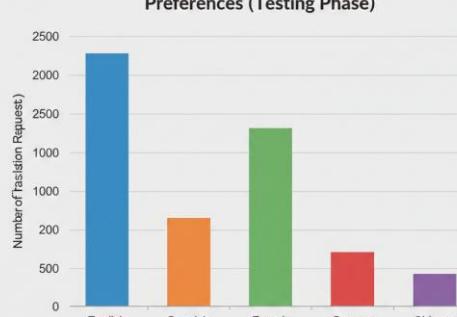
Data Overview

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

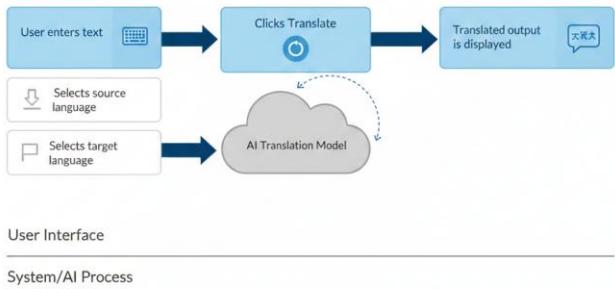
Univariate Analysis

Analysis	Description												
Input Text Length	<p>Analysis based on individual text input length</p> <p> TransLingua</p> <p style="text-align: center;">TransLingua Application User Language Preferences (Testing Phase)</p>  <table border="1"> <caption>Data from TransLingua Application User Language Preferences (Testing Phase)</caption> <thead> <tr> <th>Language</th> <th>Number of Translation Requests</th> </tr> </thead> <tbody> <tr> <td>English</td> <td>~2200</td> </tr> <tr> <td>Spanish</td> <td>~220</td> </tr> <tr> <td>French</td> <td>~1400</td> </tr> <tr> <td>German</td> <td>~150</td> </tr> <tr> <td>Chinese</td> <td>~50</td> </tr> </tbody> </table> <p>The bar chart illustrates the distribution of language selections made by users during testing, highlighting higher usage of commonly spoken languages.</p> <p>Friday, January 30, 2026</p>	Language	Number of Translation Requests	English	~2200	Spanish	~220	French	~1400	German	~150	Chinese	~50
Language	Number of Translation Requests												
English	~2200												
Spanish	~220												
French	~1400												
German	~150												
Chinese	~50												
Language Type	Single-language input per request												
Character Distribution	Alphabetic, numeric, and special characters												

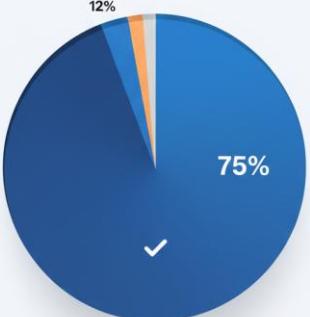
Bivariate Analysis

Analysis	Description								
Source vs Target Language	Relationship between selected source and target languages								
Text Length vs Response Time	<p>Longer input text may slightly increase processing time</p>  <p>TransLingua System Performance Efficiency Input Text Length vs. Translation Response Time</p>  <table border="1"> <caption>Data points estimated from the chart</caption> <thead> <tr> <th>Input Text Length</th> <th>Response Time (ms)</th> </tr> </thead> <tbody> <tr> <td>Short</td> <td>~80</td> </tr> <tr> <td>Medium</td> <td>~180</td> </tr> <tr> <td>Long</td> <td>~520</td> </tr> </tbody> </table> <p>The line chart demonstrates how translation response time varies with input text length, indicating efficient performance even for longer text.</p>	Input Text Length	Response Time (ms)	Short	~80	Medium	~180	Long	~520
Input Text Length	Response Time (ms)								
Short	~80								
Medium	~180								
Long	~520								

Multivariate Analysis

Analysis	Description
Text + Source Language + Target Language	Combined influence on translation accuracy and response quality
User Input Patterns	<p>Variation in usage across different language combinations</p> <p>TransLinga: AI-Powered Language Translation Flow <i>The diagram represents the overall workflow of the TransLingua system, showing the interaction user and the AI translation engine.</i></p>  <pre> graph LR A[User enters text] --> B[Clicks Translate] B --> C[Translated output is displayed] B --> D[AI Translation Model] D <--> E[Selects source language] D <--> F[Selects target language] </pre> <p>The diagram illustrates the workflow of the TransLingua system. It is divided into two main sections: User Interface and System/AI Process. The User Interface section shows a sequence of actions: 'User enters text' (with a keyboard icon), followed by 'Clicks Translate' (with a circular arrow icon). This leads to the 'Translated output is displayed' (with a speech bubble icon). The System/AI Process section shows the 'AI Translation Model' (represented by a cloud icon) receiving input from 'Selects source language' (with a download icon) and 'Selects target language' (with a copy icon). There are also feedback loops from the AI model back to the source and target selection steps.</p>

Outliers and Anomalies

Category	Description										
Outliers	Extremely long text inputs										
Anomalies	Unsupported characters or empty inputs										
Handling Method	<p>Input validation and warning messages in UI</p> <div style="background-color: #f0f0f0; padding: 10px;"> <p>AI Translation System: Validation & Error Handling Distribution</p> <p>The pie chart illustrates distribution of system responses, indicating effective input validation and a high success rate of translations.</p>  <table border="1"> <thead> <tr> <th>Error Type</th> <th>Percentage</th> </tr> </thead> <tbody> <tr> <td>Empty input errors</td> <td>1%</td> </tr> <tr> <td>Unsupported character inputs</td> <td>1%</td> </tr> <tr> <td>Other/Unexpected Errors</td> <td>12%</td> </tr> <tr> <td>Success (Valid Input)</td> <td>75%</td> </tr> </tbody> </table> <p>Legend</p> <ul style="list-style-type: none"> Empty input errors Unsupported character inputs Other/Unexpected Errors </div>	Error Type	Percentage	Empty input errors	1%	Unsupported character inputs	1%	Other/Unexpected Errors	12%	Success (Valid Input)	75%
Error Type	Percentage										
Empty input errors	1%										
Unsupported character inputs	1%										
Other/Unexpected Errors	12%										
Success (Valid Input)	75%										

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.