Data Collection and Pre-processing Phase

Date	17th July 2024
Team ID	SWTID1720025517
Project Title	CodeXchange: An AI-Powered Code Translator Tool Using PaLM's Chat-Bison-001
Maximum Marks	2 Marks

Data Quality Report

The Data Quality Report Template will summarize data quality issues from the selected sources, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
Dataset 1	Inconsistent formatting of code (e.g., inconsistent indentation, use of comments)	Moderate	Normalize code formatting using code beautifiers and linters for each language. Utilize text preprocessing libraries to standardize formatting.
Dataset 1	Presence of special characters and comments that are not relevant to the translation	Low	Use regex or text preprocessing tools to filter out special characters and irrelevant comments. Ensure that only relevant code data is included.
Dataset 1	Missing values in some code entries	High	Implement data imputation techniques or remove entries with missing values. Ensure that the dataset is complete before training the model.
Dataset 1	Duplicates in the dataset leading to biased model training	Moderate	Identify and remove duplicate entries to ensure that the dataset represents a diverse set of examples.
Dataset 1	Unbalanced classes leading to biased model performance	High	Use techniques like oversampling, undersampling, or class weighting to balance the dataset. Ensure that the model does not favor one language over others.
Dataset 1	Presence of noisy data and outliers	Moderate	Apply data cleaning techniques to identify and remove noisy data and outliers. Use statistical methods to detect anomalies.
Dataset 1	Inaccurate labels or misclassifications in the code pairs	High	Conduct a thorough review and manual verification of a subset of the dataset to ensure label accuracy. Correct any misclassifications found.

Data Source	Data Quality Issue	Severity	Resolution Plan
Dataset 1	Limited diversity in the training data, leading to poor generalization	Moderate	Augment the dataset with additional examples that cover a wider range of programming languages and scenarios. Use data augmentation techniques to increase diversity.