Data Collection and Preprocessing Phase

Date	12 July 2024
Team ID	SWTID1720161415
Project Title	JobSwift : Accelerating Careers with AI Powered Applications
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, 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 text data (e.g., inconsistent use of punctuation, capitalization)	Moderate	Normalize text data by converting to lowercase and removing unnecessary punctuation. Utilize text preprocessing libraries to standardize formatting.
Dataset 1	Presence of special characters and emojis that are not relevant to the training data	Low	Use regex or text preprocessing tools to filter out special characters and emojis. Ensure that only relevant text data is included.
Dataset 1	Missing values in some data 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 class 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 training data	High	Conduct a thorough review and manual verification of a subset of the dataset to ensure label accuracy. Correct any misclassifications found.
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 scenarios and contexts. Use data augmentation techniques to increase diversity.