Data Collection and Preprocessing Phase

Date	10 April 2025
Team ID	259453
Project Title	SMS Spam Detection using NLP
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	Presence of duplicate messages	Moderate	Used `drop_duplicates()` function in pandas to remove repeated entries.
Dataset	Inconsistent casing (e.g., “Free” vs “free”)	Low	Converted all text to lowercase using `.lower()` during preprocessing.
Dataset	Presence of special characters, numbers, and punctuation	Low	Used regular expressions (`re.sub`) to remove unwanted characters.

Dataset	Class imbalance (spam messages are fewer than ham messages)	Moderate	Applied stratified train-test split and can use SMOTE if needed for balancing.
Dataset	Unnecessary words (stopwords like “the”, “is”, etc.)	Low	Removed using NLTK's stopword list during text preprocessing.
Dataset	Variants of same word (e.g., "loved", "loving", "love")	Low	Used stemming (`PorterStemmer`) to reduce words to their root form.