Data Collection and Preprocessing Phase

Date	01-10-2024
Team ID	LTVIP2024TMID25000
Project Title	SMS Spam Detection
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report outlines data quality issues from the selected dataset, including their severity levels and the proposed resolution plans. This report helps systematically identify and address any discrepancies or issues with the dataset to ensure high-quality input for the machine learning model.

Data Source	Data Quality Issue	Severity	Resolution Plan
Kaggle SMS Spam Collection Dataset	Presence of special characters, URLs, and numbers in text data	Moderate	Clean the text by removing special characters, links, and numbers using regular expressions.
Kaggle SMS Spam Collection Dataset	Imbalanced classes between spam and non-spam messages	High	Use class weighting or oversampling techniques like SMOTE to handle imbalance.
Kaggle SMS	Text data needs to	Modera	Use Tfidf Vectorizer to transform

Spam Collection Dataset	be converted into numerical format	te	text into numerical features for model input.
Kaggle SMS Spam Collection Dataset	Presence of duplicate messages in the dataset	Low	Remove duplicate rows to avoid model bias.