Data Collection and Preprocessing Phase

Date	28-09-2024
Team ID	LTVIP2024TMID25000
Project Title	SMS Spam Detection
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

This report outlines the data collection strategy and the identified raw data sources for the **SMS Spam Detection** project. It ensures a meticulous approach to data curation, integrity, and quality to support informed decision-making and accurate predictions.

Data Collection Plan Template

Section	Description
Project Overview	The machine learning project aims to classify SMS messages as **Spam** or **Not Spam** using a dataset of labeled SMS messages. The objective is to build a model that can accurately detect spam messages, enhancing the user experience and security. Features of the dataset include the raw SMS text and a label (spam or ham).

Data Collection Plan	Search for datasets related to **SMS messages** and **spam classification**. - Prioritize datasets with large sample sizes to cover a variety of spam message patterns. - Ensure datasets include labeled messages to facilitate supervised learning.

Raw Data Sources Template

Source Name	**Description**	**Location/ URL**	**Format**	**Size**	**Access Permissions**
Kaggle Dataset	The dataset contains labeled SMS messages, each classified as either **Spam** or **Not Spam (ham)**.	spam_ham_dataset.csv - Google Drive	CSV	5 MB	Public
UCI Dataset	A dataset of SMS messages labeled for spam detection,	https://archive.ics.uci.edu/dataset/228/	CSV	4.7 MB	Public

	providing a good balance of spam and ham messages.	sms+spam+collection			