Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	**739687**
Project Title	SMS SPAM DETECTION
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report:

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan:

Section	Description

Project Overview	The SMS Spam Detection project aims to develop a machine learning-based system to classify SMS messages as either spam or legitimate (ham). By leveraging natural language processing (NLP) techniques, the project involves cleaning and preprocessing text data, extracting meaningful features using methods like TF-IDF or word embeddings, and training models such as Naïve Bayes, Logistic Regression, and Support Vector Machines. The system will be evaluated using metrics like accuracy, precision, and recall to ensure reliable spam detection. Once optimized, the model can be deployed in a real-world application, providing users with an automated and efficient way to filter out unwanted messages and enhance their mobile communication experience.

Data Collection Plan	● Collect a diverse and representative dataset of SMS messages, labeled as either "spam" or "ham" (not spam), to train and evaluate a machine learning model for spam detection. ● Prioritize datasets with diverse demographic information.
Raw Data Sources Identified	Raw data for SMS spam detection can be sourced from public datasets, crowdsourcing, web scraping, and simulated message generation. Key variables include SMS content, spam/ham labels, message length, sender type, timestamp, and spam indicators like keywords. Data diversity and multilingual samples enhance model performance, while privacy is ensured by masking sensitive details.

Raw Data Sources Report:

Source Name	Description	Location/URL	Format	Size	Acces Permiss
Kaggle Dataset	The dataset comprises SMS detailes (labels, length), (sender type), and spam indicators outcomes.	https://drive.google.com/ /view file/d/1K4hMBJ3oMklTtxoyykgT7YPNdMh3ur1Q /view	CSV	5.2 ,MB	Public
UCI	This data concerns SMS SPAM applications; a good mix of attributes	https://archive.ics.uci.edu/dataset/228/sms+spam+collection	CSV	466.7 kB	Public