Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	SWTID1727180793
Project Title	SMS- Spam Detection Using NLP
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

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 Template

Section	Description
Project Overview	The machine learning project aims to predict whether an SMS is spam or not based on the content of the message. Using a dataset containing features such as the text of the message, sender information, and other metadata, the objective is to build a model that accurately classifies SMS messages as spam or not spam, assisting in filtering unwanted messages
Data Collection Plan	● Search for datasets related to SMS spam detection, text messages, and communication data. ● Prioritize datasets that contain diverse message content, languages, and sources to improve model robustness and generalizability. ● Gather datasets with a clear label for spam and non-spam messages.

Raw Data Sources Identified	The raw data sources for this project include datasets obtained from Kaggle, UCI, and other repositories such as SMS Spam Collection, Enron Email dataset, etc. The provided sample data represents a subset of collected information, containing SMS texts labeled as "spam" or "ham" (non-spam). The data includes various features such as message content, length, and sender details.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
spam_ham_dataset.	A collection of SMS messages labeled as "spam" or "ham" (non-spam), widely used for training spam detection models.	https://drive.google.com/file/d/1K4hMBJ3oMklTtxoyykgT7YPNdMh3ur1Q/view	CSV	150GB	Public