Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	SWTID1728136330
Project Title	Fake news analysis on social media using NLP
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification:

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	This project aims to identify and analyze fake news on social media platforms using Natural Language Processing (NLP). The primary goal is to build a model that accurately detects misinformation by evaluating textual patterns, sources, sentiment, and linguistic markers. The analysis will help mitigate the spread of misinformation on social media, contributing to a more informed digital society.
Data Collection Plan	Data will be gathered from multiple sources that contain social media posts, user comments, news articles, and labeled datasets focused on fake and real news. Key sources include publicly available datasets and web scraping from social media APIs, along with government and research datasets on misinformation.

Raw Data Sources Identified	Below is a detailed list of the raw data sources to be used in this project, with relevant descriptions for each dataset.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Fake and Real News Dataset	Verified news articles used for comparison.	https://timesofindia.indiatimes.com/readersblog/world-of-words/fake-news-and-social-media-33975/	CSV	XX GB	Public

Twitter API / Facebook Graph API	Data scraped from popular social media platforms including flagged content.	https://cris.maastrichtuniversity.nl/ws/portalfiles/portal/54095690/Spanakis_2020_Disinformation_in_Open_Online_Media.pdf	JSON	Varies	Private (API access required)
…	…	…	…	…	…