

## Data Collection and Preprocessing Phase

Date	23 September 2024
Team ID	LTVIP2024TMID25030
Project Title	FAKE NEWS ANALYSIS IN SOCIAL MEDIA
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 detect and classify fake news shared on social media platforms. Using datasets with features such as article text, user interactions, timestamps, and source reliability, the objective is to build a model that identifies misinformation effectively, promoting informed engagement on social networks.
Data Collection Plan	<ul style="list-style-type: none"> <li>• Search for datasets related to fake news, misinformation campaigns, and social media activity.</li> <li>• Prioritize datasets with labeled content (e.g.true/false) and diverse platforms(Twitter, Facebook, etc)</li> <li>• Include datasets that reflect patterns like user sentiment, engagement metrics (likes, shares), and publication timestamps.</li> </ul>

Raw Data Sources Identified	The raw data sources for this project may include datasets from kaggle, UCI Machine Learning Repository, and open-access academic datasets focused on misinformation. These datasets contain features like post text, user comments and article metadata.
-----------------------------	---

### Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	The dataset contains news articles, their labels (fake/real) and metadata like publication date and source. It also includes social media engagement metrics	<a href="https://www.kaggle.com/c/fake-news/data">https://www.kaggle.com/c/fake-news/data</a>	CSV	20 MB	Public
UCI Repository	This dataset focuses on misinformation spread covering user interactions and social media posts. It includes sentiment analysis and timestamps for news propagation	<a href="https://archive.ics.uci.edu/ml/datasets.html">https://archive.ics.uci.edu/ml/datasets.html</a>	CSV	10 MB	Public

