Assignment Title: Data Collection & Cleaning

Project Title: Fake News Detection

1. Reading Summary:

This week, I studied how to collect and clean data using Pandas and NumPy.

I learned about removing duplicate rows, handling missing values, and detecting outliers or invalid entries.

These cleaning steps help improve data quality before training a fake-news-detection model.

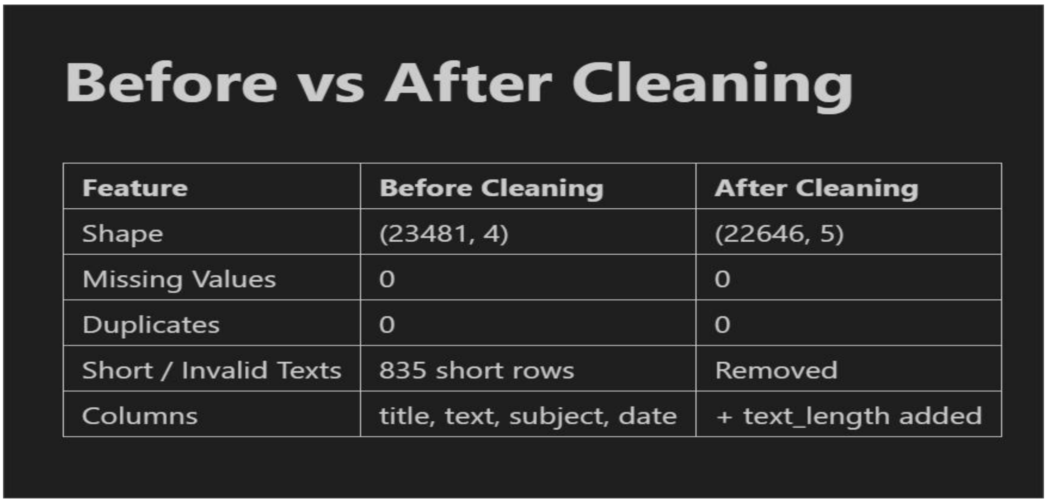
2. Task Performed:

- Loaded **Fake.csv** (23 481 rows, 4 columns) into Colab.
- Checked dataset info → no missing values found.
- Verified 0 duplicate rows.
- Calculated article text lengths and detected 835 short/incomplete articles.
- Removed short articles (`text_length < 50`).
- Saved the final cleaned dataset as **Fake_Cleaned.csv**.

3. Before Vs After Cleaning:

Description	Before Cleaning	After Cleaning
Rows	23 481	≈ 22 646
Columns	4	5 (text_length added)
Missing Values	0	0
Duplicates	0	0
Outliers (Short Texts)	835	Removed

4. Output:



Before vs After Cleaning

Feature	Before Cleaning	After Cleaning
Shape	(23481, 4)	(22646, 5)
Missing Values	0	0
Duplicates	0	0
Short / Invalid Texts	835 short rows	Removed
Columns	title, text, subject, date	+ text_length added

5. Learning Outcome:

From this task, I learned how to:

- Check dataset quality (missing, duplicates, outliers).

- Clean textual data using Pandas.

- Prepare datasets for machine-learning projects.

- Save and upload cleaned data to GitHub.

6. Challenges Faced:

- The dataset was **large (23k+ rows)**, which caused performance delays in Colab and upload issues on GitHub.

- Some news articles had **very short or incomplete text**, which could reduce model accuracy.

- **File download errors (“Failed to fetch”)** occurred due to large file size limits in Colab.

- The **date column contained inconsistent formats**, which may require further preprocessing.

- Ensuring the **cleaned file was properly saved and uploaded** to GitHub took multiple attempts.

7. GitHub Repository Link:

GitHub repository link here:
https://github.com/AlmasMalik66/DataScience-AI-Assignment