Key Differences Between This Project and Previous Ones

1. Live Data Collection vs. Preloaded Datasets

In our previous projects, we worked with pre-existing datasets (e.g., CSVs from Kaggle).
For this project, we are **building our own dataset** by collecting live data from APIs like Reddit, Twitter/X, and NewsAPI. This requires writing scripts to pull, store, and process the data ourselves.

2. Shared Scripts and Personal Notebooks

Everyone can continue using their personal `.ipynb` notebooks (e.g., `notebooks/leo.ipynb`) for experimentation.
However, to avoid duplication, we'll write **shared and reusable code** in the `scripts/` folder (e.g., for data collection, cleaning, sentiment analysis). This keeps our code clean and modular.

3. Project Folder Structure and Collaboration

We're using a more **structured folder layout** to mimic professional workflows.
Everyone will use the **same dataset** stored in the `data/` folder, rather than downloading or generating their own. This ensures consistency and easier collaboration.

4. Focus on Analysis, Not Prediction

This is not a predictive modeling project like those we've done before (e.g., recommending books or predicting likelihood of readmission in diabetes patients).
Instead, it's focused on **sentiment and topic analysis** of public discourse. While we may use models like VADER or BERT, we're not predicting future outcomes.

5. Version Control is Critical

Because the dataset is being built over time, version control via GitHub is essential.
We'll all be working off the same data and scripts, and changes should be pushed and pulled regularly to stay in sync.

Additional Notes

Why do we have a `models/` directory?
Even though we're not doing classic prediction, we may still use or fine-tune sentiment or topic

models (e.g., BERT, BERTopic). If so, those models will be saved in the `models/` directory for reuse.

Is this project deployable?
Primarily, it’s an analysis project. However, we could deploy it as a dashboard or app (e.g., using Streamlit or Flask) that visualizes public sentiment trends in real-time or accepts user input. Deployment could add value to the portfolio.