**AI ML Internship Log**

# Day 5 - Real Dataset Preprocessing & TF-IDF

# Date - 13 June 2025

# Team Role - Member

# Project Title - Personality Prediction from Social Media

# **1. What I Did Today**

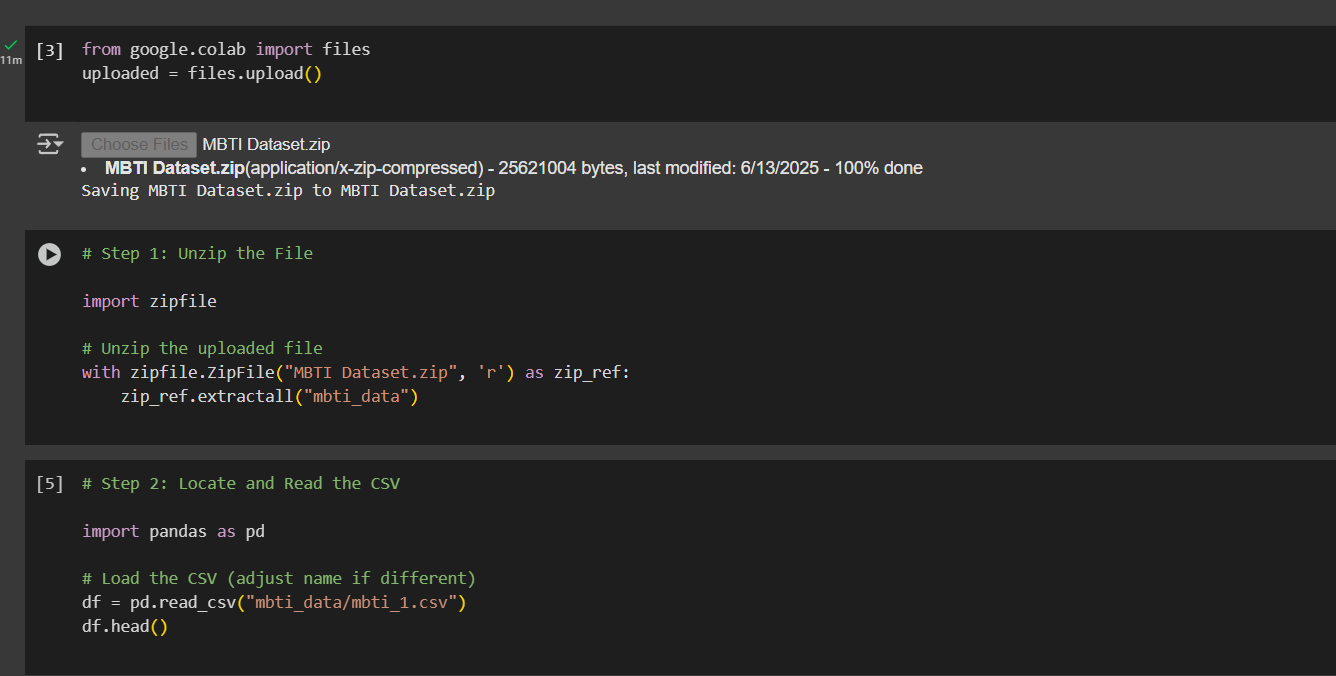
* Downloaded and understood the structure of the **MBTI Personality Dataset** (from Kaggle).
* Uploaded and unzipped the .zip file in Google Colab.
* Loaded the dataset (mbti\_1.csv) into a DataFrame using Pandas.
* Cleaned the text in the posts column (lowercased text, removed punctuation and URLs).
* Applied **TF-IDF Vectorization** using TfidfVectorizer from sklearn
* Generated a sparse matrix (X) of shape (8675, 1000).
* Stored the labels (y) as MBTI personality types.

**2. What I Learned Today**

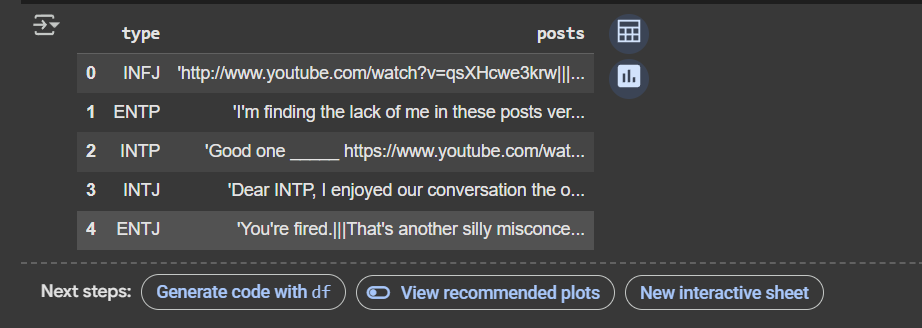
* MBTI categorizes people into 16 types based on personality traits (e.g., INTP, ENFP, etc.)
* The dataset contains real social media posts from Reddit users along with their MBTI type.
* Text cleaning and TF-IDF are essential to prepare messy text for ML.
* TF-IDF converts cleaned text into **numerical features** (importance scores of words).
* I now have a complete matrix (X) of features and corresponding labels (y), ready to be used for training a machine learning model.

**3. Code Screenshots:**

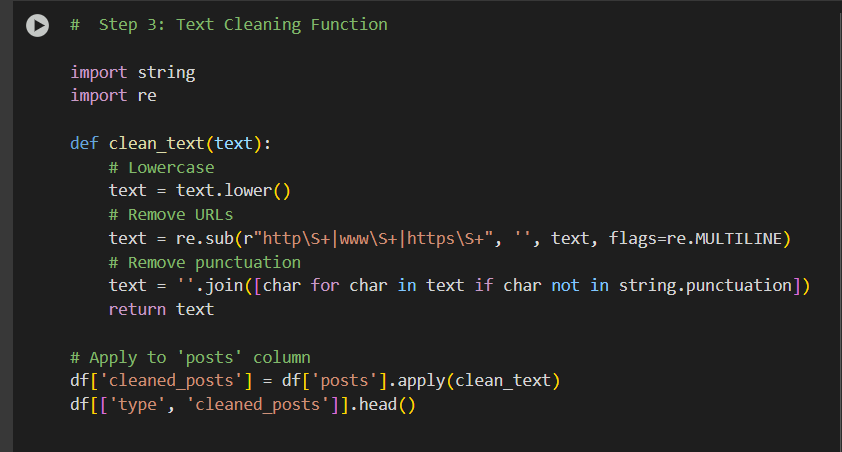
**Code:**

****

**Output:**

****

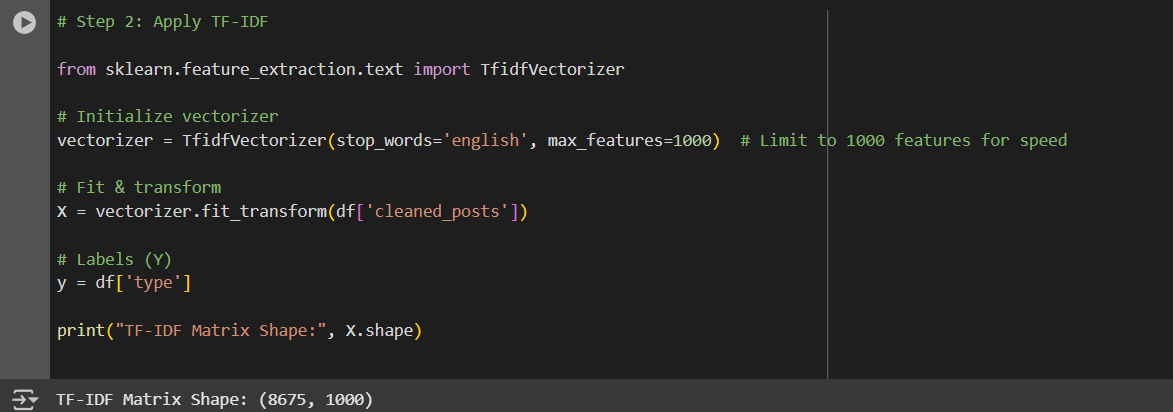
**Code:**

****

**Output:**

****

**Code:**

****

**4. How I Felt**

* I genuinely enjoyed today’s work.
* It felt like I was doing real project-level ML — not just running code, but truly understanding how raw social media text can be cleaned and turned into data for personality prediction.
* I’m also gaining confidence in handling real-world data, and I’m excited to move ahead with model building next week.

**5. Conclusion**

* Today I moved from sample text to actual project data. I learned how to upload, clean, and process a real dataset using TF-IDF**.**
* I now have both X (features) and y (labels) ready, which means the dataset is fully prepped for model training.
* This marks a turning point in the project — transitioning from preprocessing to model development.