Classification & Sentiment Analysis of Twitter Tweets

This project focuses on the classification and sentiment analysis of Twitter tweets using Machine Learning (ML) and Deep Learning (DL) techniques. The implementation includes preprocessing steps, exploratory data analysis (EDA), and model training using different algorithms.

Please follow the step-by-step instructions below to run the project successfully on **Google Colab** only.

Installation & Setup

Accessing Files and Setting Up Google Drive Folder Connection

a. Access the project submission via the Google Drive folder named **IT1244_Team1_Project**.

b. Download the project folder and upload it to your Google Drive under **My Drive**.

Preprocessing

a. Navigate to the folder:

   IT1244_Team1_Project > Code > Dataset EDA & Preprocessing

b. Open the notebook:

   Dataset Preprocessing - Removing Duplicates.ipynb

c. Run the entire notebook:

   Click on Runtime > Run all, or use the keyboard shortcut Ctrl + F9

   The processed dataset will be saved to:

   IT1244_Team1_Project > Model & Dataset > removed_dups_dataset.csv

Preparing for ML Model Training

d. Open the notebook:

   Data Preprocessing.ipynb

e. Run the entire notebook:

   Click on Runtime > Run all, or use the keyboard shortcut Ctrl + F9

   The processed datasets will be saved to:

   IT1244_Team1_Project > Model & Dataset > ml_cleaned.csv

   IT1244_Team1_Project > Model & Dataset > ml_cleaned_deduplicated.csv

Preparing for DL Model Training

f. Open the notebook:

   Dataset Preprocessing for DistilBERT.ipynb

g. Run the entire notebook:

   Click on Runtime > Run all, or use the keyboard shortcut Ctrl + F9

   The processed datasets will be saved to:

   IT1244_Team1_Project > Model & Dataset > bert_cleaned_dataset.csv

   IT1244_Team1_Project > Model & Dataset > bert_cleaned_removed_dups_dataset.csv

Data Visualisation

a. Navigate to the folder:

   IT1244_Team1_Project > Code > Dataset EDA & Preprocessing

b. Open the notebook:

   Dataset EDA (Original Dataset).ipynb

c. Run the entire notebook:

   Click on Runtime > Run all, or use the keyboard shortcut Ctrl + F9

   The visualisations will be generated within the notebook.

Running the Models

a. Navigate to the folder:

IT1244_Team1_Project > Code

The Code folder contains five model subfolders:

- Logistic Regression
- Maximum Entropy
- Random Forest
- Stochastic Gradient Descent
- DistilBERT

b. To execute each model:

- Open the respective model folder.
- Open the corresponding notebook(s).
- For Logistic Regression, Maximum Entropy, Random Forest and Stochastic Gradient Descent:
  - Run the entire notebook:
    Click on Runtime > Run all, or use the keyboard shortcut Ctrl + F9
- For DistilBERT:
  - Important Notes:
    - Ensure you are connected to a GPU runtime in Google Colab.
    - You are not recommended to run the entire notebook at once, as it may take more than 10 hours even on powerful GPUs.
    - It is strongly recommended to run the notebook chunk by chunk to manage resources and prevent timeouts.
    - Carefully read and follow any warnings or notes embedded in the notebook for a smoother execution.

The trained models will be saved to:

IT1244_Team1_Project > Model & Dataset > <*respective model folder*>

Note: The trained DistilBERT model is not stored in the folder directly. To access it, please refer to the file:

Trained Models Links in IT1244_Team1_Project > Model & Dataset > DistilBERT