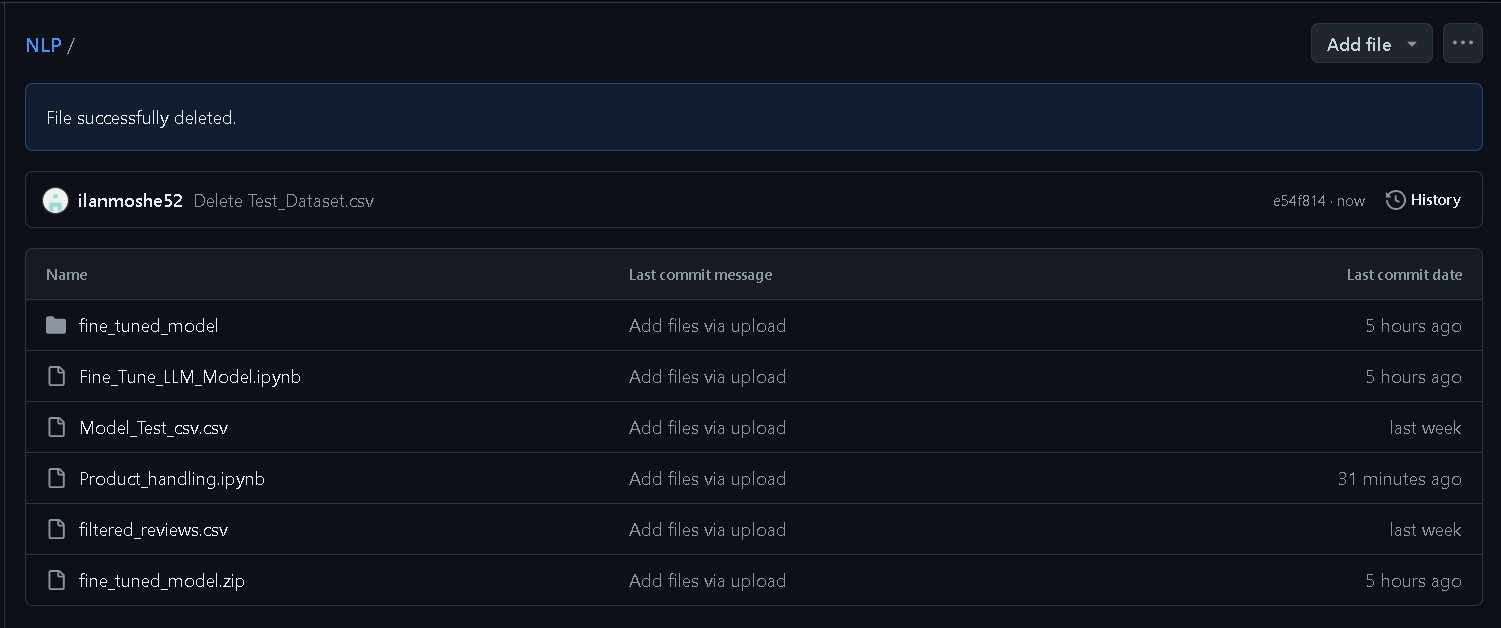
NLP Final Project Sentiment Analysis

Ilan Moshe (024176398)

1. File Description in the GIT:



**Data Files:**

* IMDB\_Dataset\_update.csv –

This file contains customer feedback from reviews, surveys, and social media about movies. It is used for fine tuning and training of the basic distilbert model. The original file link is below.

<https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews> ,

The file is in the google drive, because of the file size it can’t be stored on the GIT. (The link is in the notebook)

* Filtered\_reviews.csv – Products customer reviews csv file. This file contains customer feedback from reviews, surveys, and social media about consuming products. The link to the original Reviews file is (Reviews.csv):

<https://www.kaggle.com/code/mananhingorani/sentiment-analysis-from-customer-reviews/input?select=Reviews.csv>

I have done some processing and trimming to make it a more compact file. (Original file has more the 0.5 million records)

* Model\_Test\_csv – This dataset is used for testing the model performance after the fine tune process. It is a subset of the Reviews.csv mentioned above that is not part of the Filtered\_reviews.csv file.

**The Model Files:**

* Fine\_tune\_model directory (and Fine\_tune\_model.zip) – is the model based on distilbert model after LLM fine training using a new dataset of sentiment analysis.

Original mode before fine tuning process is here:

[distilbert/distilbert-base-uncased-finetuned-sst-2-english · Hugging Face](https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english)

**Code Files:**

* Fine\_Tune\_LLM\_Model.ipynb – The file is used to fine tune the basic model with new dataset (IMDB\_Dataset\_update.csv) and evaluate the updated model performance. More detailed information is available in the documentation in the file itself. The code can be can be run by "Run all cells", it will automatically fetch the data needed for it without any intervention.
* Product\_handling.ipynb – This file performs sentiment analysis of customers reviews per product and defines what are the products that are problematic from the customers point of view. More detailed information is available in the documentation in the file itself. The code can be can be run by "Run all cells", it will automatically fetch the data needed for it without any intervention.

(There are many graphs in these code files, that are not seen with the GIT viewer, therefore these code files must be viewed in the Google Colab)

**General Files**

NLP Final Project - Ilan Moshe 024176398.doc – General project explanations