Sentiment Analysis and Clustering with DistilBERT

Introduction

In this project, I built a robust sentiment analysis pipeline using DistilBERT, an efficient and powerful transformer model. The project aimed to analyse text data, categorize sentiments (Positive, Negative, Neutral), perform clustering to identify text groupings, and summarize results effectively.

Using Hugging Face’s Transformers, Gradio for deployment, and essential data science libraries like Scikit-learn, I tackled the challenge of building, fine-tuning, and deploying the model successfully.

Project Overview

The project was divided into four main tasks:

1. Data Preprocessing

2. Model Evaluation

3. Clustering and Summarization

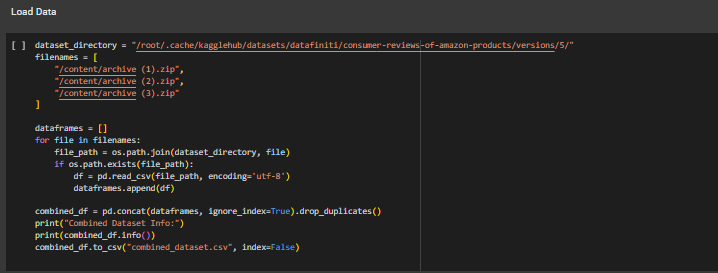
4. Model Deployment

Task 1: Data Preprocessing

The process began by loading and preprocessing the dataset.

Data Loading:

I loaded a dataset of reviews containing text and sentiment labels. I downloaded each dataset separately and later combined them into one dataset



A screenshot of a computer

Description automatically generated

Tokenization:

Text data was tokenized using the DistilBERTTokenizer.

A screen shot of a computer program

Description automatically generated

Splitting the Dataset:

I split the data into training and validation sets.

A screen shot of a computer program

Description automatically generated

Training the Model:

I loaded a pre-trained DistilBERTForSequenceClassification model and trained it using HuggingFace's `Trainer`.

A screenshot of a computer program

Description automatically generated

**Task 2: Model Evaluation**

After training, I evaluated the model performance using classification metrics:

A screenshot of a computer

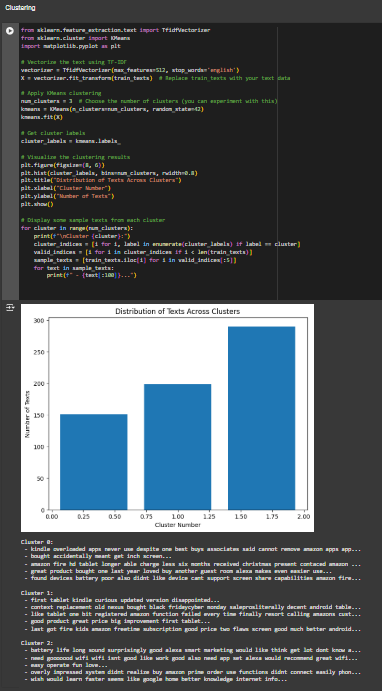
Description automatically generated

The evaluation showed that the model achieved average accuracy but correctly classified most texts.

**Task 3: Clustering and Summarization**

Clustering:

To group similar texts, I applied KMeans clustering on the hidden embeddings extracted from DistilBERT.



Summarization:

I used HuggingFace's pipeline to summarize representative texts from each cluster

A screenshot of a computer program

Description automatically generated

**Task 4: Model Deployment with Gradio**

Finally, I deployed the trained model using Gradio to create a user-friendly interface for sentiment prediction.

A screenshot of a computer

Description automatically generated

The Gradio app allowed users to input custom text and instantly receive sentiment predictions.

**Conclusion**

This project demonstrated the power of DistilBERT for sentiment analysis, clustering, and summarization tasks. The deployment with Gradio made the solution accessible and user-friendly.