NLP Engineer Assignment report submission

**→ A brief introduction to the task and the dataset used**

The goal of this assignment is to build a text classification model using the Hugging Face library to classify a dataset of text into one of multiple categories. By using a pre-trained model such as BERT or GPT-2 as a starting point and fine-tune it on the classification task.

* Fine-tune A NLP Model: Fine-tuning a natural language processing (NLP) model involves adjusting the hyperparameters and architecture of the model, and often also involves adjusting the dataset, to improve the performance of the model on a specific task
* About Hugging Face : Hugging Face is a company that provides a platform for training and deploying natural language processing (NLP) models. The platform includes a library of pre-trained models that can be used for a variety of NLP tasks, such as language translation, text generation, and question answering.

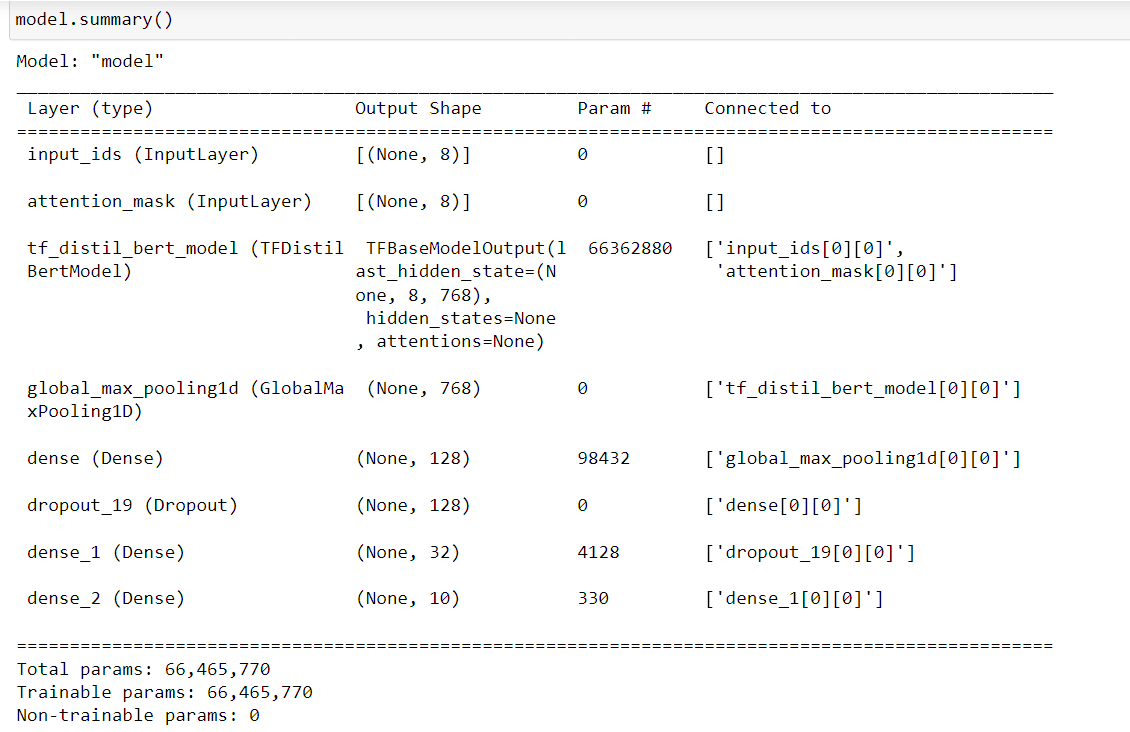
**Dataset contains:**

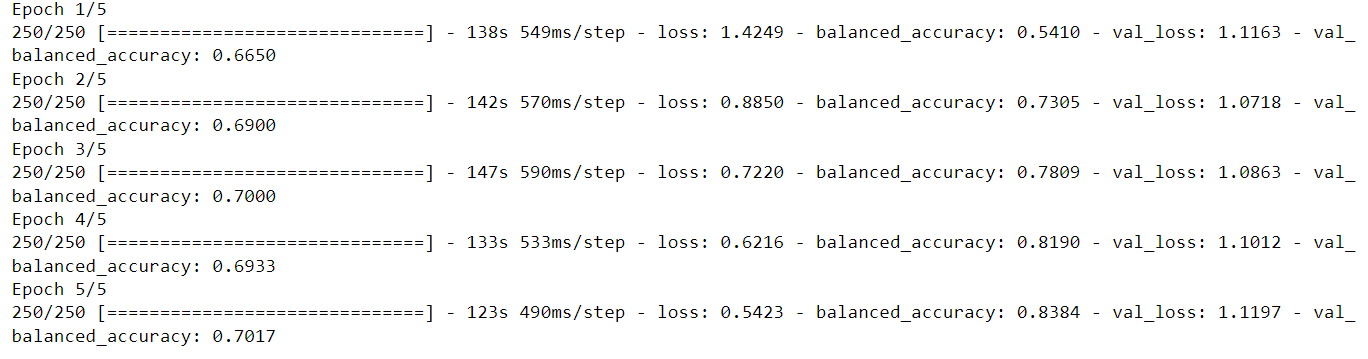
* 45500 rows and 5 columns
* Target column: Category ( Business , Politics, Food & Drink, TRAVEL ,Parenting, STYLE & BEAUTY ,Wellness, World news, Sports , Entertainment)  
  -Each category class contains 4500 rows  
  -It contains nan values only in keywords column
* Apart from that, the original dataset had lots of third person statements (like "This statement is irrelevant" says the officials)  
  -Keyword column has been added where main keywords in a url are extracted (urls were in the original dataset)

**→ The preprocessing steps taken :**

* Preprocessed the data by removing stopwords, punctuations and lemmatized using WordNetLemmatizer

**→ The architecture of the model used, and how it was fine-tuned**

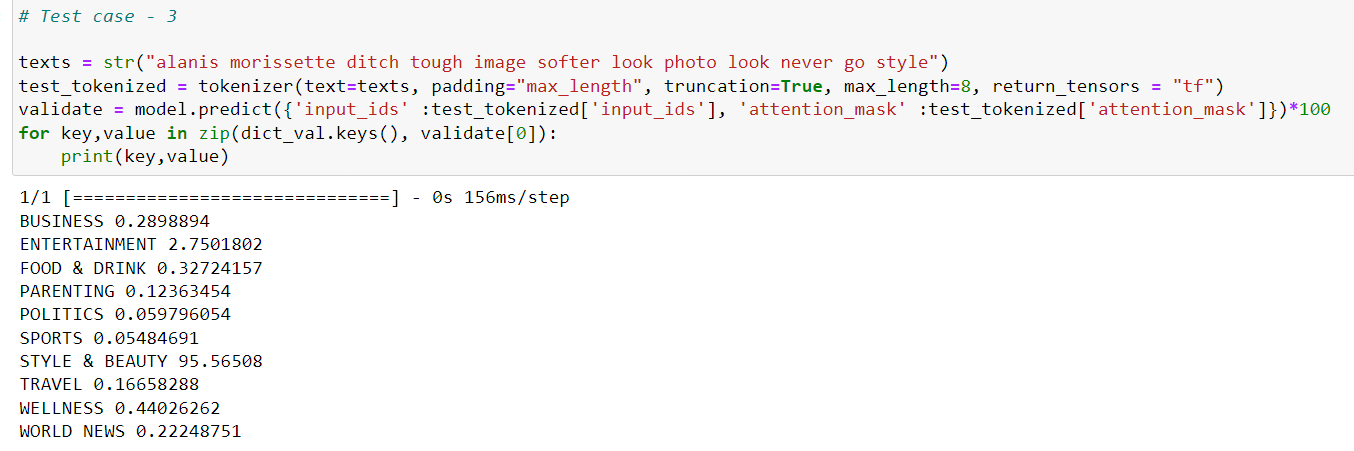




**→ A discussion of the performance of the model and possible ways to improve it**

* As epochs increased model performance also increased. And we got very less loss in both train and test
* By using a higher **Batch Size,** We can increase the model accuracy. I have used batch\_size=8. Due to my system GPU is running out of memory.

**→ Sample predictions and their explanations**



And you can see more in the code.

**DATASET LINK :** https://www.kaggle.com/datasets/setseries/news-category-dataset