Question One: Hugging Face

Part 0: Project Explanation

Project Idea:

a project that aligns with one of the Sustainable Development Goals (SDGs). SDG 4 (Quality Education).

Project Title: "Automated Assessment of Educational Content Readability Using NLP"

- Objective: The goal is to analyze educational materials and determine their readability levels to ensure they are accessible to a wide audience, including those with different learning abilities.

- Detailed Explanation:

- This project aims to utilize a pre-trained NLP model to assess the readability of various educational texts. By analyzing the language complexity, sentence structure, and vocabulary, the model will classify texts into different readability levels. This could help educators ensure that their materials are suitable for their intended audience.

Part 1: Research and Setup

I. Research and Setup:

- Hugging Face provides a wide range of pre-trained models for various NLP tasks such as text classification, sentiment analysis, translation, and more.

II. Model Selection:

- Model: bert-base-uncased for text classification.

Part 2: Implementation

I. Loading the Model:

- I load the bert-base-uncased model, which is pre-trained on a large corpus of text data. Specified that num_labels=3 because of classifying the text into three readability levels (easy, medium, difficult).

II. Data Preparation:

- Dataset: manually created a small dataset of texts with corresponding readability labels.

III. Test the Model:

- Splited the data into training and testing sets.

Training:

Defined training arguments like batch size, number of epochs, etc., and use the Trainer class from Hugging Face to train the model.

Part 3: Evaluation

I. Evaluation:

-After training, we evaluate the model on the test set using accuracy and F1-Score as the metric.

1. Accuracy

Accuracy is the ratio of correctly predicted instances to the total number of instances.

Formula: Accuracy = (Number of Correct Predictions) / (Total Predictions)

Calculation:

- Correct Predictions: 3 (positions 1, 2, and 3 where predicted labels match true labels)

- Total Predictions: 5

- Accuracy: 3 / 5 = 0.6 (or 60%)

2. F1-Score

F1-Score is the harmonic mean of precision and recall, providing a balance between the two.

Formula: F1-Score = 2 * (Precision * Recall) / (Precision + Recall)

Calculation for class 0:

- Precision: ≈ 0.67

- Recall: 1

- F1-Score for class 0: 2 * (0.67 * 1) / (0.67 + 1) ≈ 0.8 (or 80%)

Question Two: Finetuning Large Language Models

Project Idea:

a project that aligns with one of the Sustainable Development Goals (SDGs). Good Health and Well-being (SDG 3).

Project Title: " Fine-tuning a BERT Model for Classifying Medical Research Articles "

- Objective:

The primary goal of this project is to fine-tune a pre-trained BERT model to classify medical research articles into different categories such as "Prevention," "Treatment," and "Diagnosis." This classification task is critical in the healthcare industry, as it can significantly streamline the process of literature review, allowing healthcare professionals to quickly identify relevant research based on specific focus areas.

- Project Explanation:

What We Are Trying to Achieve:

- Model Fine-tuning: We aim to adapt a pre-trained BERT model, which has been trained on general text data, to perform well on a specific task related to healthcare. The task is to categorize medical articles based on their content, focusing on whether the article discusses prevention, treatment, or diagnosis.

- Improvement of Task-specific Performance: By fine-tuning the BERT model on a relevant dataset, the project seeks to improve its accuracy and reliability in classifying medical articles. This process involves adjusting the model's parameters to better capture the nuances of medical terminology and content.

- Evaluation and Comparison: The project will evaluate the model's performance before and after fine-tuning, providing insights into how fine-tuning improves the model's ability to understand and categorize specialized content. We expect to see significant performance gains after the fine-tuning process, demonstrating the effectiveness of this approach.

- What is Expected:

1. Install Necessary Libraries and Tools

pip install torch transformers datasets sklearn matplotlib

2. Exploratory Data Analysis (EDA)

- Load the Dataset: Load a dataset of health-related tweets.
- Perform Basic EDA: Analyze the distribution of sentiment classes, check for missing values, and understand the length of tweets.
- Visualization: Use libraries like matplotlib or seaborn to visualize the sentiment distribution and the tweet lengths.

3. Dataset Preparation

- Preprocessing: Clean the text data by removing special characters, converting to lowercase, and tokenizing the text using the BERT tokenizer.
- Label Encoding: Convert sentiment labels into numerical format (e.g., positive = 0, neutral = 1, negative = 2).

4. Model Selection

- Choosing the Model: We'll use the pre-trained bert-base-uncased model from Hugging Face.
- Loading the Model: Load the BERT model for sequence classification.

5. Fine-tuning Process

- Define Training Arguments: The hyperparameters for fine-tuning the model. Important parameters include the learning rate, number of training epochs, and batch size.
- Train the Model: Use the Trainer class from Hugging Face to fine-tune the model.

6. Evaluation

- Test the Model: This project demonstrates the process of fine-tuning a large language model (BERT) for a specific NLP task, in this case, sentiment analysis of health-related tweets.
- Compare Before and After Fine-tuning: By evaluating the model before and after fine-tuning, you will be able to see the impact of fine-tuning on task performance, which is expected to improve after the process.