Q2:Fine-Tuning Large Language Models for Violence Type Classification

Project Overview

The objective of this project was to fine-tune a pre-trained language model from Hugging Face for classifying tweets into different types of violence. This project aligns with the Sustainable Development Goal (SDG) 16: Peace, Justice, and Strong Institutions, focusing on reducing violence and ensuring justice.

Project Explanation

The primary aim was to enhance a pre-trained model's ability to accurately classify tweets based on the type of violence described, such as physical, verbal, or psychological violence. Fine-tuning a large language model (LLM) like BERT allowed us to adapt it to a specific task, improving its performance on a dataset that it was not originally trained on.

1. Install Necessary Libraries and Tools

I installed essential Python libraries, including **transformers**, **datasets**, **pandas**, **seaborn**, and **matplotlib**, to facilitate data handling, visualization, model training, and evaluation.

2. Exploratory Data Analysis (EDA)

- **Data Loading:** The dataset **violence.csv** was loaded into a pandas DataFrame.
- **EDA:** Basic exploratory analysis was performed to understand the distribution of violence types and tweet lengths. Visualizations included:
    - **Distribution of Violence Types:** Count plot to visualize the frequency of each violence type.
    - **Tweet Length Distribution:** Histogram to observe the variation in tweet lengths.

3. Dataset Preparation

- **Data Cleaning:** Unnecessary columns, such as **Tweet_ID**, were dropped from the dataset.
- **Label Encoding:** Violence types were mapped to numerical labels to prepare the data for model training.

- **Tokenization:** Tweets were tokenized using the **bert-base-uncased** tokenizer, converting the text data into a format suitable for the model.

4. Model Selection

- **Pre-trained Model:** The **bert-base-uncased** model was selected from the Hugging Face model hub for fine-tuning. This model is pre-trained on a large corpus of English text and is suitable for text classification tasks.

5. Fine-Tuning Process

- **Training Arguments:** Specific parameters were set for the training process, including the number of epochs, batch size, learning rate, and output directory for storing the results.
- **Fine-Tuning:** The pre-trained model was fine-tuned on the prepared dataset to adapt it to the task of violence type classification.

6. Evaluation

- **Model Evaluation:** The fine-tuned model was evaluated on a test set. The final model achieved an accuracy of 89.75%, demonstrating a significant improvement in performance after fine-tuning.
    - **Performance Comparison:** The fine-tuning process resulted in enhanced accuracy, proving the effectiveness of adapting a pre-trained model to a specific task.

Conclusion

This project successfully fine-tuned a pre-trained BERT model for the specific task of violence type classification in tweets. Through exploratory data analysis, careful data preparation, and model fine-tuning, we achieved a final accuracy of 89.75%. This underscores the power and flexibility of large language models when adapted to specific tasks. The model can now be effectively used as a tool for content moderation, aiding in the identification and classification of different forms of violence in social media posts.