LLM – Detect AI generated text

1. Introduction

The AI-Generated Text Classification project aimed to build a robust classifier using BERT-based models to discern between human-generated and AI-generated text. This report outlines the approach, challenges faced, insights gained, and summarizes the outcomes.

1. Approach
2. Data Preprocessing

1. Data Combination: Merged two datasets (`train\_essays.csv` and `final.csv`) to enrich the training set.

2. Text Preprocessing: Applied NLTK tokenization and removed stopwords to clean and enhance the quality of the text data.

3. Train-Test Split: Segregated the dataset into training and testing sets.

1. Tokenization and Dataset Creation

1. Transformer Models: Utilized BERT-based transformer models (`bert-base-uncased`) for tokenization.

2. Datasets: Created training and testing datasets using the Hugging Face `datasets` library.

1. Model Definition and Training

1. Custom Model: Developed a custom classification model, incorporating BERT for feature extraction and a dense layer for classification.

2. Compilation: Compiled the model using Adam optimizer, sparse categorical cross-entropy loss, and accuracy metrics.

3. Training: Trained the model with a batch size of 10, implementing early stopping and model checkpoint callbacks.

1. Challenges Faced

1. Data Enrichment: Merging datasets posed challenges in ensuring data integrity and alignment.

2. Model Training: Due to the substantial size of BERT-based models, training required significant computational resources and time.

3. Callback Compatibility: The project encountered issues with model saving using HDF5 format due to the subclassed nature of the model.

1. Insights Gained

1. BERT Features: Understanding the extraction of features from BERT embeddings and their role in downstream tasks.

2. Data Quality: Realizing the importance of data quality and preprocessing in enhancing model performance.

3. Callback Usage: Gained experience in using callbacks for training monitoring and model saving.

1. Summary

The AI-Generated Text Classification project successfully implemented a BERT-based model to classify text into human-generated and AI-generated categories. The challenges faced provided valuable insights into data management and model training. The project demonstrated the significance of preprocessing and the potential of transformer models in handling complex NLP tasks.

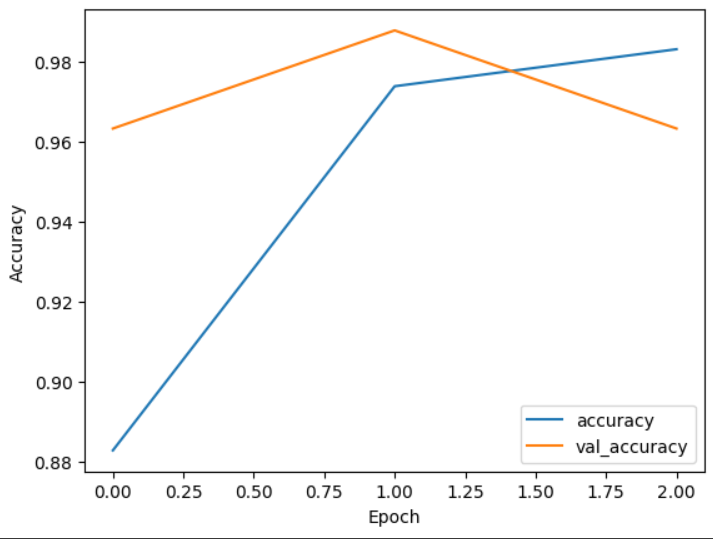
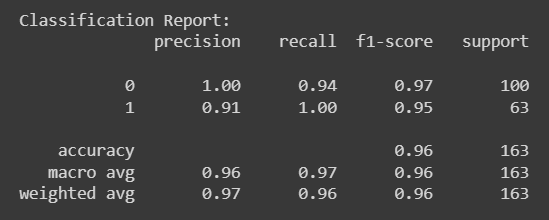


Figure 1 Accuracy 0.9835



SCOPE OF IMPROVEMENT

I have also tried to build a model containing Bert and BiLSTM layer because Bert process words in parallel and is weak in capturing sequential information of data while BiLSTM layer would have been useful to capture the whole context of data helping more in classification.

But I was not able to complete the model mainly because of mismatch of input data format in both BiLSTM and BERT and some tensorflow and pytorch dependencies.