Text Semantic Analysis using BERT Transformer

Project Overview:

This project focused on text semantic analysis using the DistilBERT Transformer model. The objective was to classify text data into various semantic categories, including "positive," "negative," and "neutral." The report provides an overview of the project, data processing, model architecture, training, and results.

Data Collection and Preparation:

Total Data: The dataset comprised 27,480 samples.

Data Reduction: Due to resource constraints, the project proceeded with 5100 samples for analysis.

Data Processing:

Tokenization: Text data was tokenized to convert it into input format suitable for the BERT Transformer.

Conversion to PyTorch Tensors: The tokenized data was converted into PyTorch tensors for compatibility with the model.

Label Encoding: Sentiment labels ("positive," "negative," "neutral") were encoded into numerical values.

Model Architecture:

The BERT-based model used for text semantic analysis was created as follows:

```
model = DistilBertForSequenceClassification.from_pretrained("distilbert-base-uncased", num_labels=len(y.unique()))
```

DistilBERT Model: A pre-trained DistilBERT model was used, fine-tuned for sequence classification tasks.

Model Training:

The model was trained for 10 epochs. However, the training was interrupted at 7 epochs due to resource limitations. The results up to the seventh epoch were

obtained:

Epoch 6:

Training Loss: 0.267

Training Accuracy: **91.08%**

Validation Loss: 0.379

Validation Accuracy: **85.66%**

Epoch 7:

Training Loss: 0.214

Training Accuracy: **93.29%**

Validation Loss: 0.395

Validation Accuracy: **86.76%**

Challenges and Limitations:

Resource Constraints: The project faced resource limitations, with the training process crashing after the seventh epoch due to a lack of RAM of Collab.

Conclusion:

In conclusion, this project showcased the successful application of the DistilBERT Transformer model for text semantic analysis. Up to the seventh epoch, the model achieved promising results, with a validation accuracy of approximately **86.76%**. The limited training data and resource constraints necessitated the interruption of training, leaving further exploration of the model's potential for future work. Future steps could include fine-tuning the model on larger datasets and addressing resource limitations to complete the training process.