Detecting AI-Generated Text

The goal of this project was to develop a robust system for detecting AI-generated text. The approach involved acquiring knowledge about Natural Language Processing (NLP) fundamentals, exploring various language models, and implementing techniques to handle imbalanced datasets.

Research Phase (Week 1):

Explored basics of NLP to build a foundation for the project. Studied different language model architectures such as RNN, LSTM, Encoder-Decoder framework, Attention framework, and Transformer.

Data Preparation (Week 2):

Conducted Exploratory Data Analysis (EDA) on the competition dataset to understand its characteristics. Addressed dataset imbalance by adding more data. Created some ai generated text examples with instruction and prompt title and some with only using prompt title as an input to google generative ai.

Model Development (Week 3):

Leveraged pre-trained models like BERT and ROBERTA. Used bidirectional encoder framework specifically to capture the context of the text by looking both at left and right. Finetuned on the training set for adjusting model parameters for specific text classification task

Results and Analysis (Week 4):

Model	**BERT**	**ROBERTA**
validation accuracy	0.9853	1.00
ROC score on comp test set	0.61	0.59

In summary, the project successfully implemented a text detection system for AI-generated content in student essays. The Obtained ROC scores depicts the model is overfitted to some extent. Improvements can be done by training on more data with examples of text generated on different prompt titles.