RNN_HW1

1. Data Preprocessing

The original dataset contains 487,235 rows of text with binary labels. For computational efficiency, I sampled 100,000 instances (50,000 from each class) to ensure a balanced dataset.

1.1 Text Tokenization

Text samples were tokenized using the BERT tokenizer from the Hugging Face Transformers library. This choice provides several advantages:

- Handles out-of-vocabulary words through subword tokenization
- Preserves contextual information with special tokens
- Offers consistent preprocessing across training and inference

Texts were padded or truncated to a maximum length of 512 tokens to accommodate the BERT tokenizer's constraints and to ensure consistent input dimensions for the neural network.

1.2 Dataset Split

The dataset was divided into three sets:

- Training set (75%): Used for model training
- Validation set (10%): Used for hyperparameter tuning and early stopping
- Test set (15%): Used for final evaluation

2. Model Architecture

2.1 Bidirectional LSTM

The core of our model is a Bidirectional LSTM network, which offers several advantages over traditional unidirectional LSTMs:

1. **Bidirectional Context**: By processing sequences in both forward and backward directions, the Bi-LSTM captures contextual information from both past and future tokens, providing a more comprehensive understanding of text semantics.

2. **Enhanced Feature Extraction**: The bidirectional approach allows the model to detect more subtle patterns that might be directionally dependent, such as linguistic structures that are common in human or AI writing.

The Bi-LSTM implementation uses:

- Word embeddings with dimension 300
- Hidden state dimension of 128
- 2 stacked Bi-LSTM layers for deeper feature extraction
- Dropout rate of 0.5 for regularization

2.2 Attention Mechanism

A critical enhancement to the basic Bi-LSTM model is the addition of an attention mechanism. This component allows the model to:

1. **Focus on Key Information**: By learning to weigh the importance of different parts of the text, the model can identify the most discriminative features for classification.
2. **Handle Long Dependencies**: Attention helps mitigate the challenge of capturing long-range dependencies in text, which is particularly important when distinguishing between human and AI writing styles.
3. **Improve Interpretability**: The attention weights can provide insights into which parts of the text most strongly influence the classification decision.

The attention mechanism is implemented as a learnable parameter matrix that computes a weighted sum over the Bi-LSTM outputs, producing a context vector that emphasizes the most relevant parts of the text for classification.

2.3 Model Training

The model was trained with the following configuration:

- Binary Cross-Entropy loss function
- Adam optimizer with learning rate 0.001
- Batch size of 256
- Maximum of 10 epochs with early stopping
- Patience of 3 epochs for early stopping

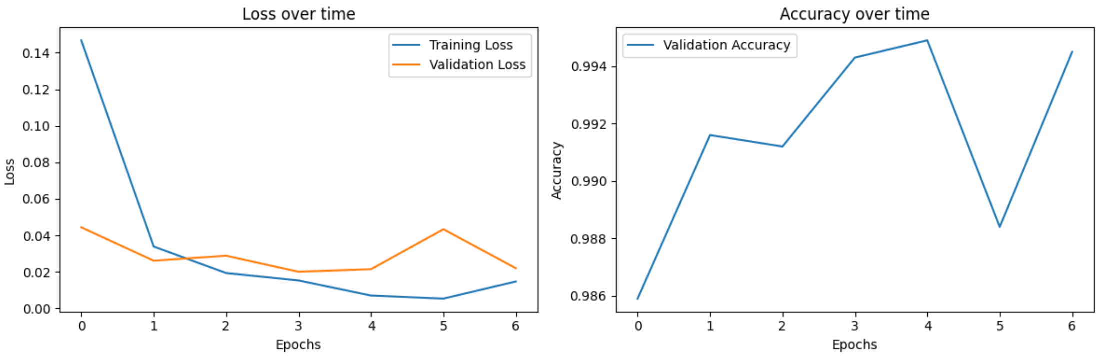
Early stopping was implemented to prevent overfitting by monitoring validation loss. Training was terminated when validation loss failed to improve for 3 consecutive epochs, and the best model state was restored.

3. Results and Analysis

3.1 Performance Metrics

The final model achieved the following performance on the test set:

- **Accuracy**: 99.6%
- **Loss**: 0.016



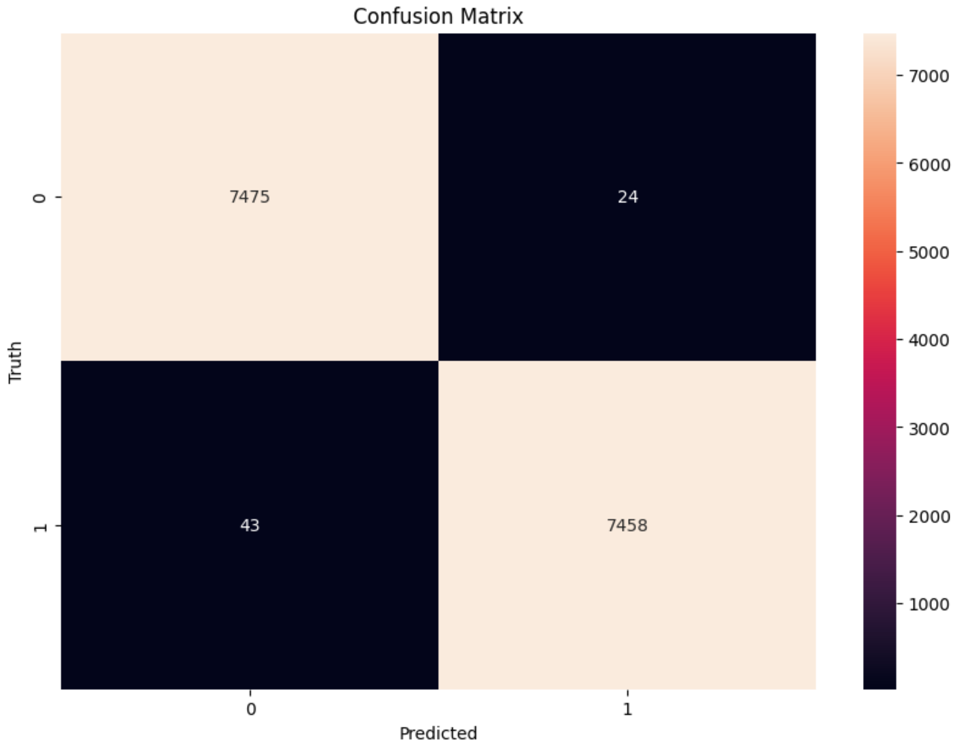

```
Total Params: 9992697
Epoch 1/10: 100%|██████████| 293/293 [14:51<00:00,  3.04s/batch, training_loss=0.070]
Validation Loss: 0.044, Validation Accuracy: 0.986
Epoch 2/10: 100%|██████████| 293/293 [14:51<00:00,  3.04s/batch, training_loss=0.048]
Validation Loss: 0.026, Validation Accuracy: 0.992
Epoch 3/10: 100%|██████████| 293/293 [14:52<00:00,  3.05s/batch, training_loss=0.003]
Validation Loss: 0.029, Validation Accuracy: 0.991
Epoch 4/10: 100%|██████████| 293/293 [14:42<00:00,  3.01s/batch, training_loss=0.003]
Validation Loss: 0.020, Validation Accuracy: 0.994
Epoch 5/10: 100%|██████████| 293/293 [14:41<00:00,  3.01s/batch, training_loss=0.006]
Validation Loss: 0.021, Validation Accuracy: 0.995
Epoch 6/10: 100%|██████████| 293/293 [14:35<00:00,  2.99s/batch, training_loss=0.006]
Validation Loss: 0.043, Validation Accuracy: 0.988
Epoch 7/10: 100%|██████████| 293/293 [14:42<00:00,  3.01s/batch, training_loss=0.002]
Validation Loss: 0.022, Validation Accuracy: 0.995
Early stopping triggered. No improvement for 3 epochs.
Early stopping triggered after 7 epochs
Loaded best model with validation loss: 0.020
```

3.2 Confusion Matrix Analysis

The confusion matrix shows:

- Very low false positive rate (AI texts classified as human)
- Very low false negative rate (Human texts classified as AI)

This balanced performance suggests the model is not biased toward either class and has learned robust discriminative features.



3.3 Comparative Advantage of Bi-LSTM with Attention

The Bi-LSTM with attention architecture provides several advantages over simpler models for this classification task:

1. **Contextual Understanding**: The bidirectional processing captures nuanced patterns in text flow that may differ between human and AI writing.
2. **Focus on Discriminative Features**: The attention mechanism helps identify the most telling indicators of AI generation, such as repetitive patterns, unusual word combinations, or specific rhetorical structures.
3. **Robustness to Text Length**: The combination of LSTM's sequential processing and attention's global view makes the model effective across texts of varying lengths.