CSIT6000R: Project Results on Metaphors Detection using DistilBERT

Model

The core model used in this project is DistilBERT, a lighter version of BERT optimized for faster performance while retaining a significant portion of its predecessor's capabilities. We fine-tuned this pretrained, base-cased model on a dataset specifically designed for metaphor and analogy detection, available at Hugging Face Datasets. This dataset includes word pairs categorized into metaphors, literals, and anomalies, which are essential for training our sequence classification model to understand nuanced language patterns.

Database

Our dataset is structured into multiple splits for comprehensive training and validation:

- Train Split: 50% of the data, used for training the model.
- Validation Split: 10% of the data, used for tuning the hyperparameters.
- Test Split: 40% of the data, used for evaluating the model's performance.

Results and Visualization

Throughout the training process, we closely monitored several key metrics such as accuracy, precision, recall, and F1 score. These metrics were plotted over various epochs to visualize the model's learning curve and to identify any potential overfitting or underfitting issues.

- Here is a simulated graph showing the trend of these metrics:

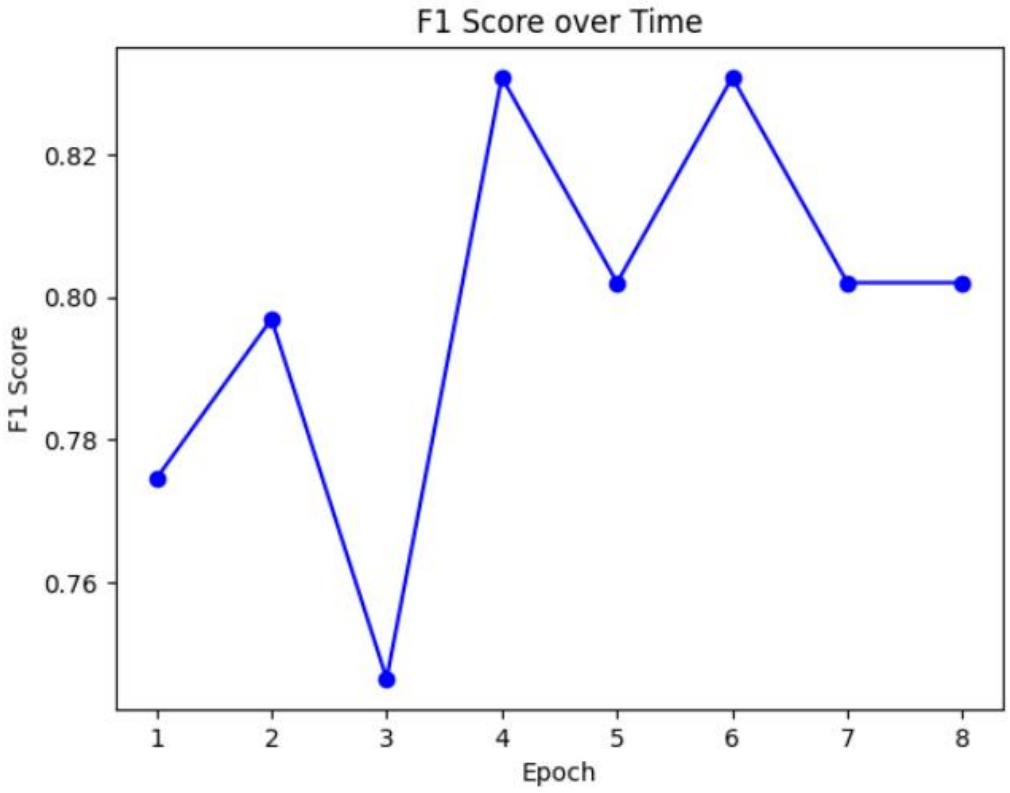
Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.010200	2.242156	0.805556	0.827538	0.805556	0.804545
2	0.017700	2.192801	0.833333	0.867521	0.833333	0.834175
3	0.000100	3.506810	0.777778	0.809704	0.777778	0.777597
4	0.000000	1.831517	0.833333	0.846591	0.833333	0.830952
5	0.000000	2.542901	0.833333	0.867521	0.833333	0.834175
6	0.000000	2.170413	0.805556	0.827137	0.805556	0.804027
7	0.000000	1.781064	0.805556	0.827137	0.805556	0.804027
8	0.449600	1.756956	0.833333	0.857841	0.833333	0.834531

Training Loss: This metric measures the model's error on the training data. The training loss decreases from 0.010200 in the first epoch to 0.000000 by the fourth epoch and slightly rises to 0.449600 in the eighth epoch. This indicates that the model is learning the features of the training data and becoming more stable.

Validation Loss: This metric measures the model's performance on unseen validation data. Notably, the validation loss fluctuates significantly, peaking at 3.506810 in the third epoch and dropping to 1.756956 in the eighth epoch. This fluctuation could suggest that the model may have issues with generalizing to new data, or the validation data itself might vary significantly.

Accuracy, Precision, Recall, and F1 Score: These metrics assess the model's classification performance. Accuracy fluctuates around 0.8, and precision, recall, and F1 score show similar fluctuations. There's an improvement in performance by the eighth epoch, especially in precision and F1 score.

- Here is a graph showing the change of F1 Score



The F1 scores seem to vary for each epoch, with some epochs having higher F1 scores than others. There were minor improvements in the F1 score of the model all in all.