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. At the same time, we also tried another model, RoBERTa, which used more data and longer training time during training, as well as some other improvement strategies, making it better than BERT on a variety of natural language processing tasks.

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. In this example, we set the learning rate to be 1e-4 and the batch_size to be 8.

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	1.097500	1.050579	0.583333	0.372222	0.583333	0.454167
2	1.049100	0.965508	0.583333	0.375975	0.583333	0.456798
3	0.972700	0.841939	0.722222	0.802778	0.722222	0.699228
4	0.851000	0.667357	0.833333	0.880800	0.833333	0.836786
5	0.720500	0.587198	0.805556	0.850529	0.805556	0.808615
6	0.614300	0.527127	0.777778	0.827381	0.777778	0.777778
7	0.522000	0.487976	0.805556	0.839744	0.805556	0.806397
8	0.400200	0.477824	0.805556	0.839744	0.805556	0.806397
9	0.446800	0.459182	0.833333	0.860119	0.833333	0.832598
10	0.374800	0.454741	0.805556	0.841204	0.805556	0.802364
11	0.402400	0.453201	0.805556	0.841204	0.805556	0.802364
12	0.502200	0.449475	0.805556	0.841204	0.805556	0.802364
13	0.389700	0.460272	0.805556	0.841204	0.805556	0.802364
14	0.386700	0.460059	0.805556	0.841204	0.805556	0.802364
15	0.351500	0.445623	0.805556	0.841204	0.805556	0.802364
16	0.291500	0.445408	0.805556	0.841204	0.805556	0.802364
17	0.266900	0.465500	0.805556	0.841204	0.805556	0.802364
18	0.251400	0.453360	0.805556	0.841204	0.805556	0.802364
19	0.259800	0.447598	0.805556	0.841204	0.805556	0.802364
20	0.281600	0.447639	0.805556	0.841204	0.805556	0.802364

Based on the table provided, we can observe the following changes in the performance metrics of the model over the course of 20 epochs:

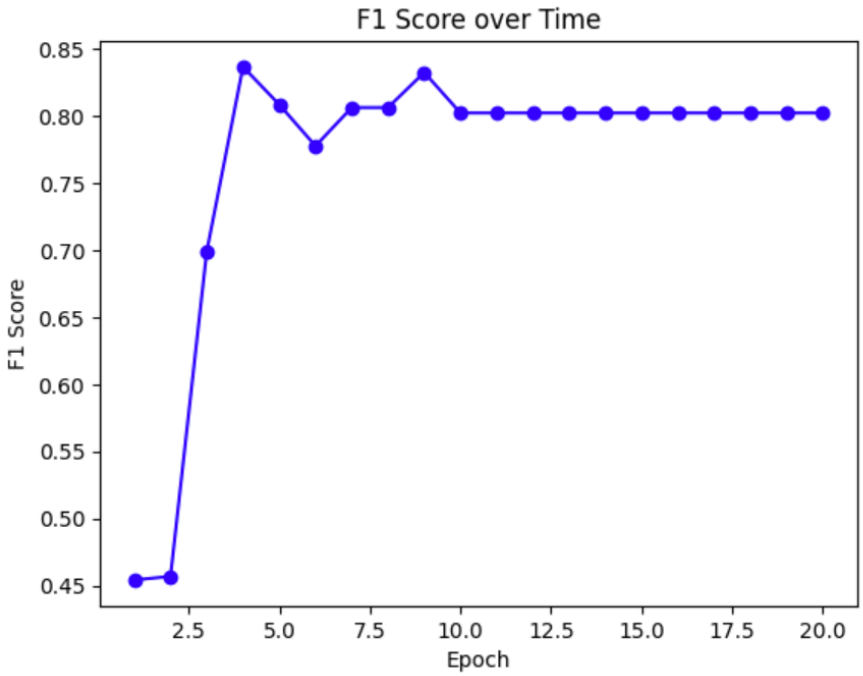
The values of Training Loss and Validation Loss are gradually decreasing, indicating that the model is converging during the training process. Specifically, the Training Loss is decreasing from 1.0975 to 0.2816, while the Validation Loss is decreasing from 1.0506 to 0.4476.

The value of Accuracy initially starts at 0.5833, but begins to increase noticeably after the third epoch, reaching a maximum of 0.8333 at the ninth epoch. This indicates that the model's overall performance is improving over time.

The values of F1 starts off relatively low, but begin to increase significantly after the third epoch. After the tenth epoch, the value of F1 remain relatively stable, indicating that the model has converged and reached a relatively stable level of performance.

In summary, the model's performance metrics show a clear improvement over the course of 20 epochs, with Training Loss, Validation Loss, Accuracy, Precision, Recall, and F1 all showing positive changes. The model's F1 Score reach a relatively stable level after the tenth epoch, indicating that the model has converged and can be used for practical applications.

- Here is a graph showing the change of F1 Score with the same parameters.

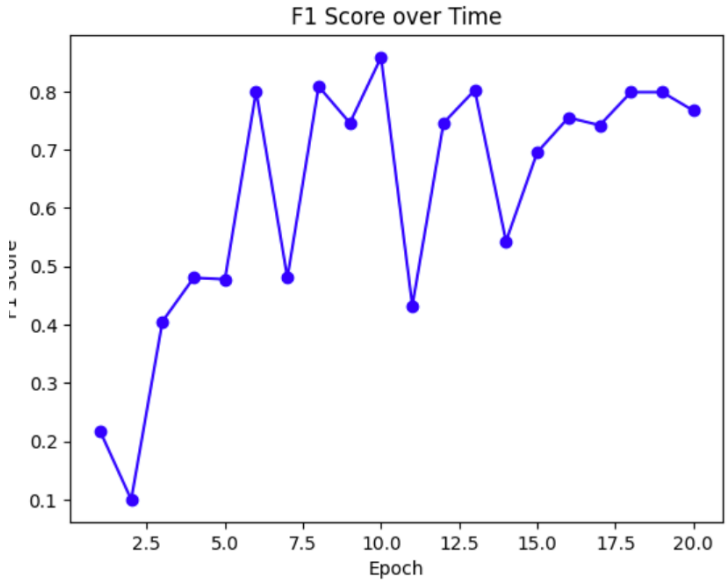


Then, we try to change the learning rate and batch_size hyperparameters to find the conditions that make F1score the highest, and here are our results.

Group number	Model	Learning rare	Batch_size	Accuracy	F1score
1	DistilBERT	1e-4	8	0.805556	0.802364
2	DistilBERT	2e-5	8	0.666667	0.614583
3	DistilBERT	1e-3	8	0.833333	0.832234
4	DistilBERT	5e-4	8	0.833333	0.832598
5	DistilBERT	5e-4	4	0.833333	0.830761
6	DistilBERT	5e-4	16	0.777778	0.773442

Finally, we tried to use and fine-tune another model RoBERTa to complete this sequence classification task. Here is the result of this model, the parameters used are learning rate = 5e-4, batch_size = 8.

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	1.227900	1.079495	0.388889	0.151235	0.388889	0.217778
2	1.158400	1.112495	0.250000	0.062500	0.250000	0.100000
3	1.069500	1.069407	0.500000	0.385417	0.500000	0.405594
4	1.115500	1.028007	0.611111	0.398148	0.611111	0.480368
5	1.003300	1.040963	0.611111	0.392884	0.611111	0.477819
6	1.025100	0.973785	0.805556	0.819360	0.805556	0.800397
7	0.939000	0.956760	0.611111	0.398148	0.611111	0.480368
8	0.900800	0.931565	0.805556	0.850529	0.805556	0.808615
9	0.916800	0.909373	0.750000	0.807540	0.750000	0.745970
10	0.929900	0.860718	0.861111	0.882323	0.861111	0.858598
11	0.792900	0.886110	0.555556	0.357499	0.555556	0.432336
12	0.875500	0.833992	0.750000	0.807540	0.750000	0.745970
13	0.849400	0.809849	0.805556	0.841204	0.805556	0.802364
14	0.840900	0.813040	0.611111	0.736257	0.611111	0.543056
15	0.880900	0.789219	0.722222	0.748513	0.722222	0.695679
16	0.794700	0.819505	0.750000	0.824968	0.750000	0.755181
17	0.762500	0.780734	0.750000	0.782242	0.750000	0.741948
18	0.757600	0.770498	0.805556	0.819599	0.805556	0.798751
19	0.711100	0.767768	0.805556	0.819599	0.805556	0.798751
20	0.760400	0.769692	0.777778	0.800160	0.777778	0.767125



It can be seen that the F1Score fluctuates greatly when using this model, and the results are very unstable. Perhaps this model is not as suitable for the task of this experiment as DistilBERT.