Fine Tuning BERT for Text Classification

Chantel and Vaishak

Problem

- Use BERT for Text Classification
- Paper tries Question Classification, Sentiment Analysis, and Topic Classification
- Attempts methods of Fine Tuning
- Attempts other methods of Pretraining

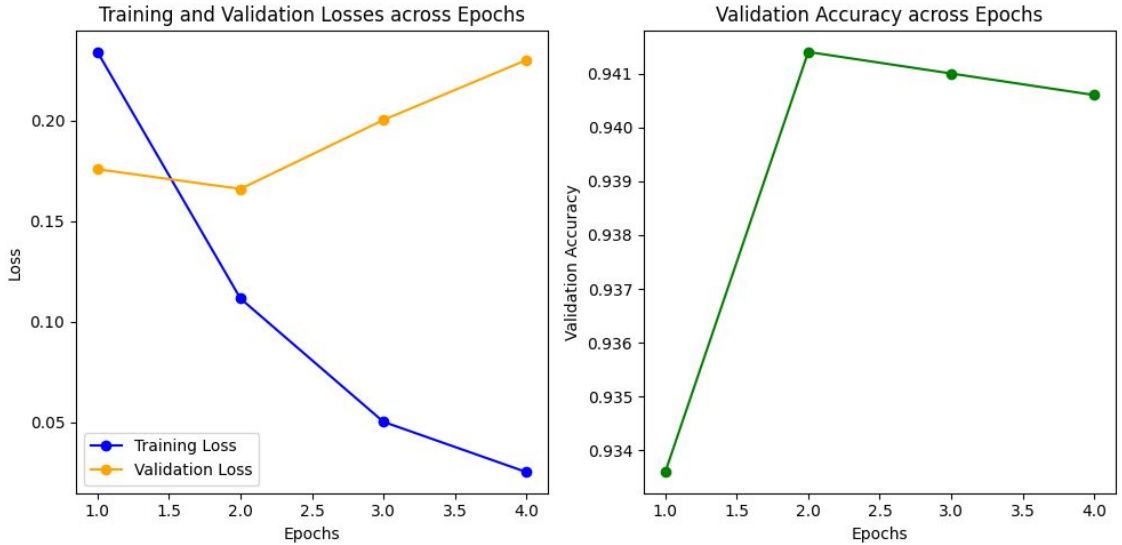
Sentiment Analysis

Implementer: Vaishak Menon

Tried Methods:

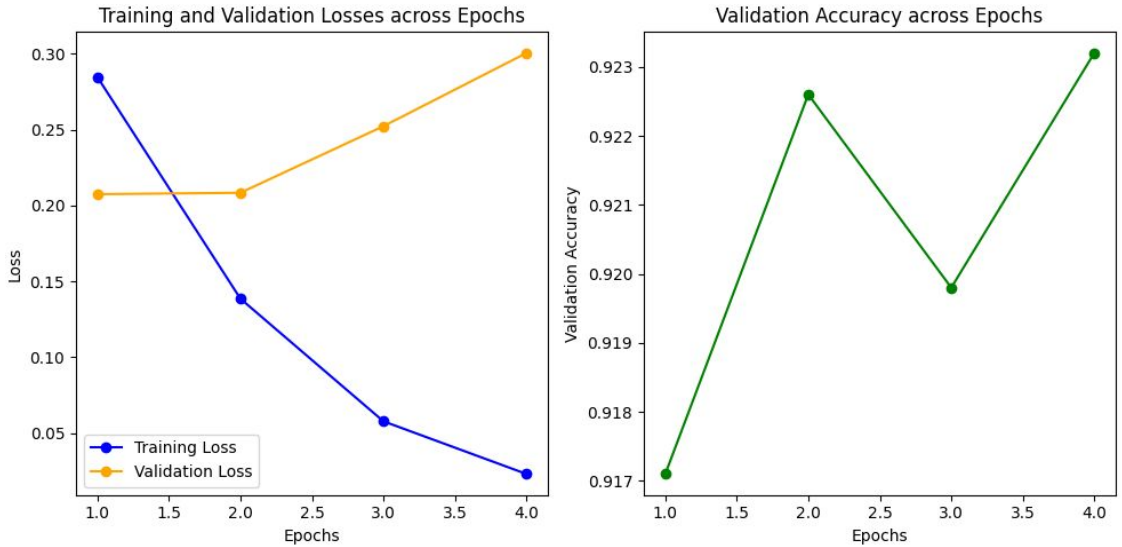
- Implemented Head-Tail Truncation
- Implemented Catastrophic Forgetting by ensuring that learning rate was always 2e-5
- Implemented Extracting Last Layer for Fine tuning
- Implemented Layer-wise Decreasing Layer Rate

Sentiment Analysis Results Thus Far



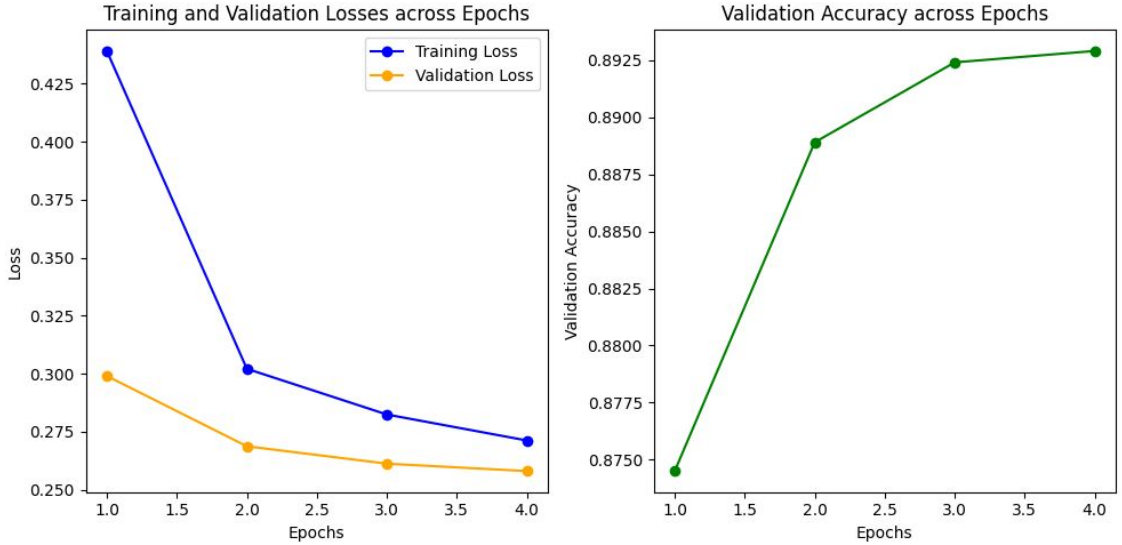
Base Bert Model

Sentiment Analysis Results Thus Far



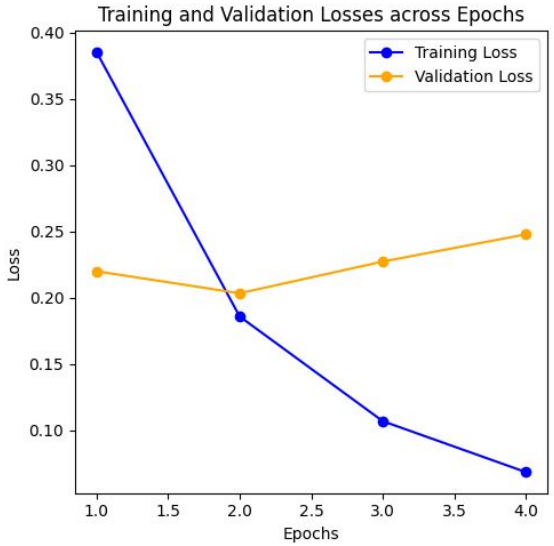
Base Bert Model with Head-Tail Truncation

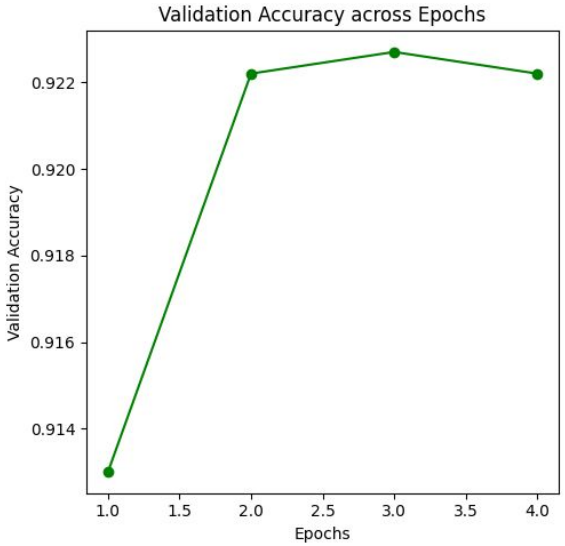
Sentiment Analysis Results Thus Far



Base Bert Model with Head-Tail Truncation and Last Layer Tuning

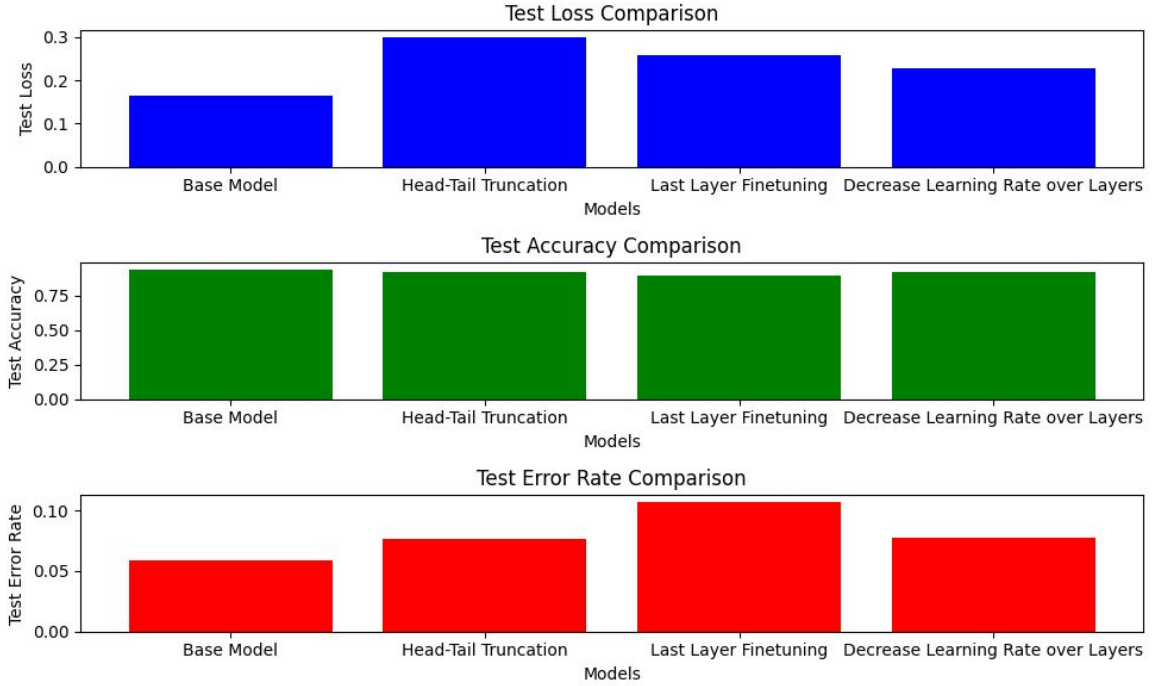
Sentiment Analysis Results Thus Far





Base Bert Model with Head-Tail Truncation, Last Layer Tuning, and Layer Wise Decreasing Learning Rate

Sentiment Analysis Test Data Comparison



Question Classification

Implementers: Chantel and Vaishak

Tried Methods:

- Implemented Head-Tail Truncation
- Implemented Catastrophic Forgetting by ensuring that learning rate was always 2e-5

Output - TREC (above) and Yahoo (below)

Epoch 1/4:	Epoch 2/4:	Epoch 3/4:	Epoch 4/4:	Test Epoch
Average Training Loss: 0.7755	Average Training Loss: 0.1718	Average Training Loss: 0.0828	Average Training Loss: 0.0492	Test Loss: 0.1243
Validation Loss: 0.2081, Validation Accuracy: 0.9341	Validation Loss: 0.1458, Validation Accuracy: 0.9635	Validation Loss: 0.1243, Validation Accuracy: 0.9698	Validation Loss: 0.1216, Validation Accuracy: 0.9698	Test Accuracy: 0.9700
				Test Error Rate: 0.0300

Epoch 1/1
Train Loss: 0.9030
Validation Loss: 0.8534, Validation Accuracy: 0.7152, Validation Error Rate: 0.2848

Topic Classification

Implementer: Chantel Rose Walia

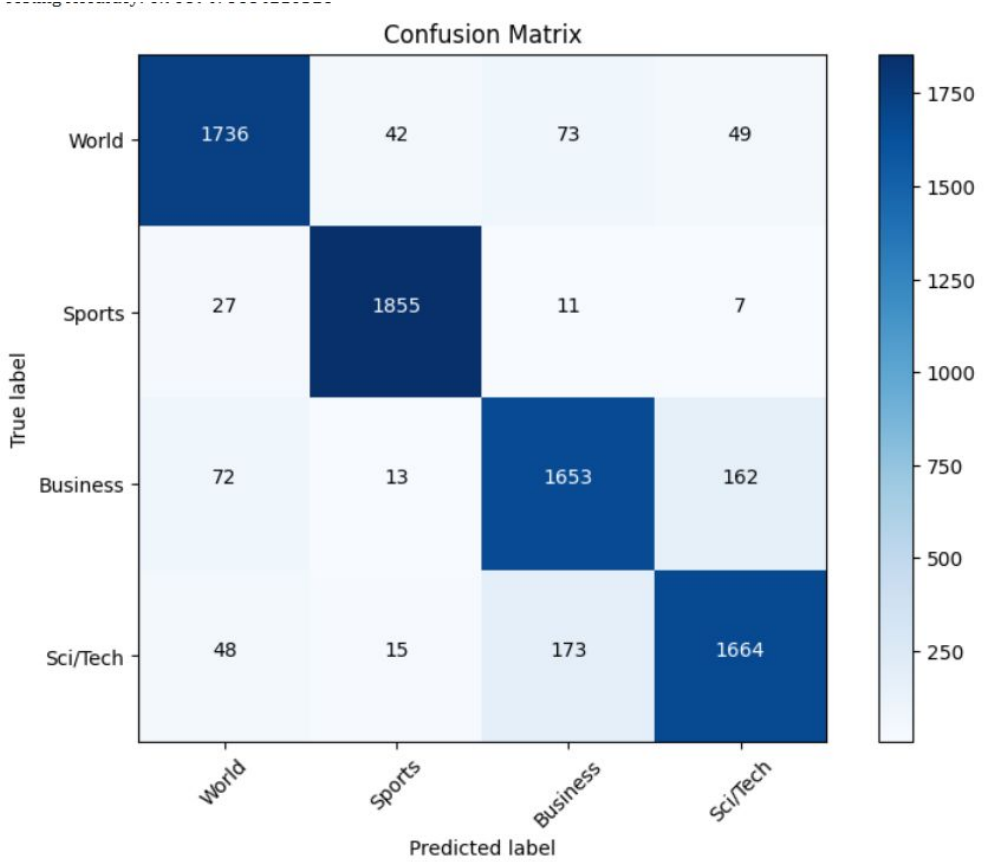
Tried Methods:

- Implemented a base BERT model that does not follow the paper to see the difference.
- Implemented Fine-Tuning according to the paper.
- Implemented Classification layer on top for topic prediction.
- Trying to evaluate model performance for few-shot learning.

Fine-Tuning for Topic Classification

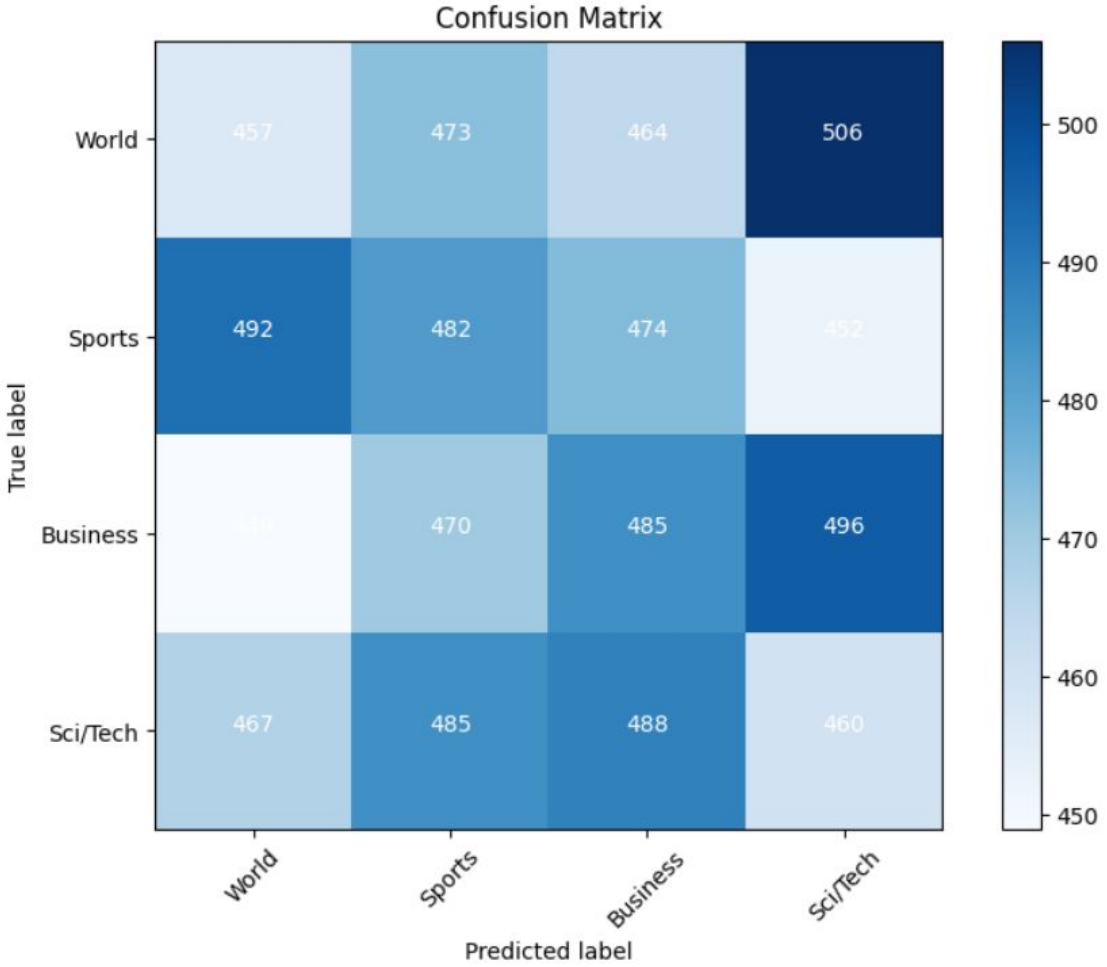
- Fine-tune entire BERT model end-to-end on topic classification data
- Use slanted triangular learning rate,batch size 24, 4 X Titan Xp GPUs (I had access only to 2 A100s and 8 cpus)
- LR 2e-5, 10% warm-up proportion, max seq length 512
- Select features from final hidden state of [CLS] token
- Train for 4 epochs, save best model on validation set

Base model that does not follow the paper



Base model that follows the paper but has different tuning parameters

Epoch 1/3 - Average Training Loss: 0.21024938592215378
Epoch 2/3 - Average Training Loss: 0.1280215780497839
Epoch 3/3 - Average Training Loss: 0.0861933432665343 1
Average Testing Accuracy: 0.9470850840336135


Confusion Matrix
World
Sports
Business
Sci/Tech
True label
457
473
464
506
492
482
474
452
470
485
496
467
485
488
460
500
490
480
470
460
450
World
Sports
Business
Sci/Tech
Predicted label

Problems so far

Sentiment Analysis: The base model seems to be doing the best; Could be due to errors in incorporating fine tuning techniques.

Question Classification: Code for model was not working at first, we updated the code so that at least 1 epoch can be run on both datasets.

Topic Classification: I initially tried using the chinese dataset provided - Sogou dataset, but was unable to get it working. Additionally using a validation set somehow made my code worse-(I am still in the process of debugging it)

Future Improvements

Try to apply other experiments.

Try to ensure that all of the preprocessing and tokenization is equivalent.

Test more on data to find out why overfitting is occurring.

Find out why model accuracy was decreasing

Try to complete the topic classification following their steps.

Github

https://github.com/vaishakkmenon/FineTuneBert

ChatGPT Links are provided in code