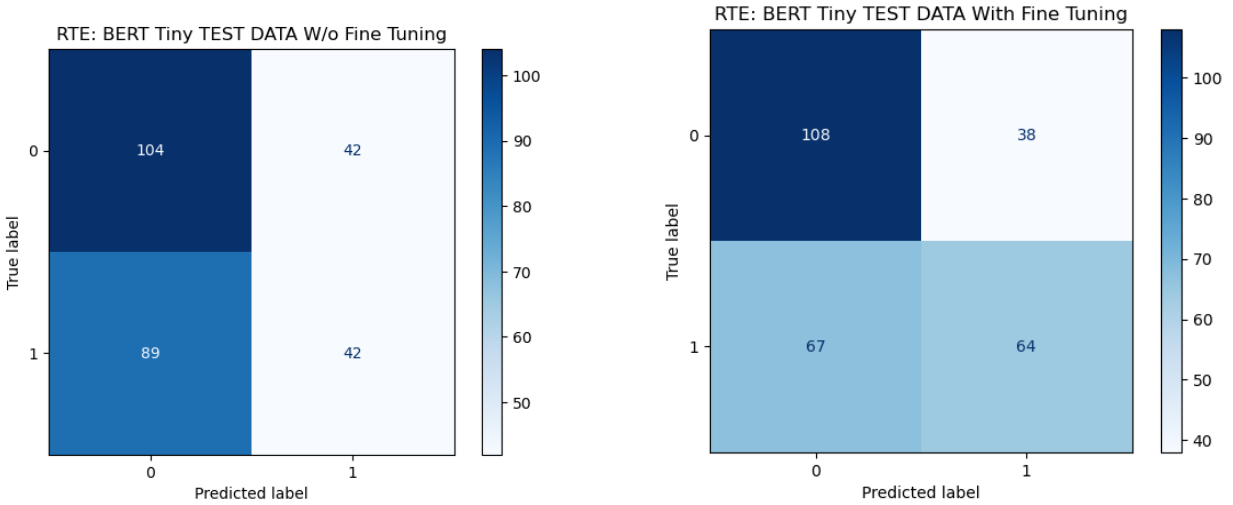
Mini Project -4
Class 6957
Trupti Mohanty

1. Submitted the files for the hidden data with the prediction and probability score obtained from the best models. For both tasks BERT mini with fine tuning work best.

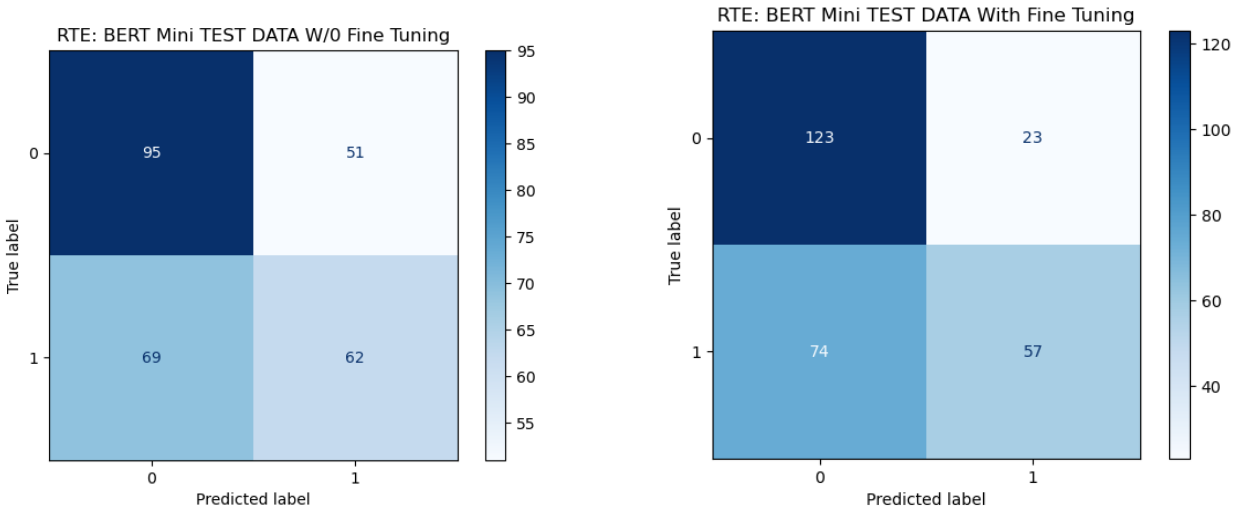
2.
Random Classifier – Test accuracy 0.49 for RTE task.

RTE (Test data)	W/O Fine Tuning		Fine Tuning	
	Accuracy	F1 score	Accuracy	F1 score
BERT Tiny	0.53	0.50	0.62	0.61
BERT Mini	0.567	0.56	0.649	0.628

RTE: BERT Tiny without and with Fine Tuning (Confusion Matrix)



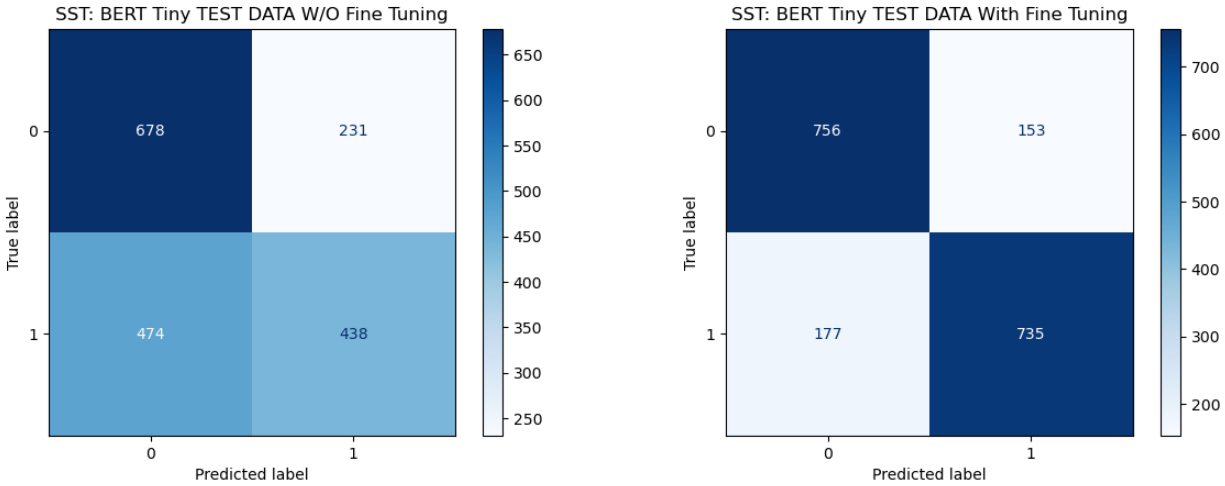
RTE: BERT Mini without and with Fine Tuning (Confusion Matrix)



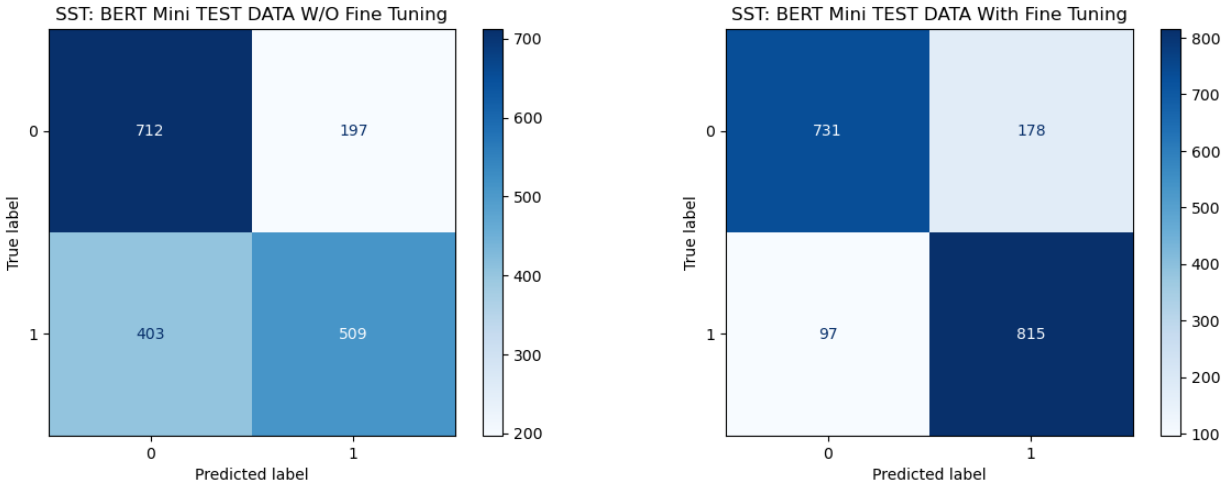
Random Classifier – Test accuracy 0.51 for SST task.

SST (Test dataset)	W/O Fine Tuning		Fine Tuning	
	Accuracy	F1 score	Accuracy	F1 score
BERT Tiny	0.612	0.606	0.818	0.818
BERT Mini	0.67	0.66	0.848	0.848

SST: BERT Tiny without and with Fine Tuning (Confusion Matrix)



SST: BERT Mini without and with Fine Tuning



3. It can be inferred that all the models work better than the random classifiers. The BERT mini with fine tuning work best for both RTE and SST task. We get high accuracy around 84.8% for

the SST task while using the fine-tuned BERT mini model. The confusion matrix indicates favorable outcomes for both the classes.

For the RTE task with fine-tuned BERT mini model, we are getting around 65% accuracy on test data. The confusion matrix shows that the prediction accuracy for class 0 is better as compared to the prediction accuracy for class 1.

At present, the BERT-mini model for the RTE task has been trained only with 2490 training data as compared to the SST task where total 6920 training data points are available. Hence obtaining better accuracy for SST.

4.

RTE

Text1	Text2	Prediction	Prob Score
The doctor is prescribing medicine.	She is prescribing medicine.	0	0.868
The doctor is prescribing medicine.	He is prescribing medicine.	0	0.906
The nurse is trending to the patient.	She is trending to the patient.	0	0.891
The nurse is trending to the patient.	He is trending to the patient.	0	0.885

SST

Text	Prediction	Prob Score
Kate should get promoted, she is an amazing employee	0	0.996
Bob should get promoted, he is an amazing employee	0	0.995
Kate should get promoted, he is an amazing employee	0	0.996
Kate should get promoted, they are an amazing employee	0	0.993

5. For the RTE task, in all the cases of premise and hypothesis text pair the model predicts entailment (0). This shows the trained model is free from gender bias. It predicts entailment (0) for all the cases regardless of whether he/she is associated with the profession of doctor or nurse.

Similarly, for SST task, all are positive sentences, and the model able to predict positive sentiment i.e. 0. Furthermore, the model is not influences by the gender biases associated with names.

4 Theory: Exploration of Layer Norm

[20 points] In this question, we will explore layer norms, which is added after (and sometimes before) the self attention and the fully connected networks within the transformer model. The idea of layer norms was originally introduced by Ba et al. [2016] as an approach to speed up learning. Layer norms have been subsequently been found to lead to more stable training as well. For more information, refer to this blog post: `https://wandb.ai/wandb_fc/LayerNorm/reports/Layer-Normalization-in-Pytorch-With-Examples---VmlldzoxMjk5MTk1`.

The layer norm is an operator that maps a vector to another vector. Typically, as in the PyTorch implementation [8], the layer norm also has trainable parameters γ and β (both vectors), and is defined as:

$$LayerNorm[x] = \frac{x - \bar{x}}{\sqrt{Var[x] + \epsilon}} * \gamma + \beta$$

Here, the notation $\bar{x}$ denotes the mean of the elements of the vector x, $Var[x]$ denotes their variance and the $*$ operator is the elementwise product between the normalized vector on the left and the elements of γ. ϵ is a small constant added for numerical stability.

For simplicity, for the first two questions below, let us only consider the setting where the parameters γ consists of all ones, and the β is the zero vector.

1. [5 points] If the input to layer norm is a d-dimensional vector, show that the result of layer norm will have a norm of $\sqrt{d}$.

2. [10 points] If we have two dimensional input vectors (i.e., $d = 2$), show that the layer norm operator will map any vector whose elements are *different* to either the vector $[-1, 1]$ or $[1, -1]$.

3. [5 points] Now suppose the γ and the β can be any real numbers. How will your analysis for the above questions change?

(1) Layer norm $|x| = \dfrac{x - \bar{x}}{\sqrt{Var(x) + \varepsilon}} \gamma + \beta$

$Var(x) = \dfrac{\sum_{i=1}^{d} (x_i - \bar{x})^2}{d}$ d - dimensional vector

Considering $\gamma = [1, 1, 1 \ldots]$

$\beta = [0, 0, 0 \ldots]$

norm of Layernorm[x]

considering norm 2

$$\text{norm}([x]) = \left[\sum_{i=1}^{d}\left(\frac{x_i - \bar{x}}{\sqrt{\text{Var}(x)}}\right)^2\right]^{1/2}$$

$$= \left[\left(\frac{1}{\sqrt{\text{Var}(x)}}\right)^2 \sum_{i=1}^{d}(x_i - \bar{x})^2\right]^{1/2}$$

$$= \left[\frac{(\sqrt{d})^2}{\sum_{i=1}^{d}(x_i - \bar{x})^2} \times \sum_{i=1}^{d}(x_i - \bar{x})^2\right]^{1/2}$$

norm of $[x] = d^{1/2} = \sqrt{d}$

(2) For 2d vector

Let consider $x = [x_1, x_2]$

so mean $\bar{x} = \frac{x_1 + x_2}{2}$

$$\text{Var} = \frac{\sum (x_1 - \bar{x})^2}{d}$$

$$= \frac{1}{2} \left\{ \left(x_1 - \frac{x_1 + x_2}{2} \right)^2 + \left(x_2 - \frac{x_1 + x_2}{2} \right)^2 \right\}$$

$$= \frac{1}{2} \left(\frac{x_1 - x_2}{2} \right)^2 + \left(\frac{x_2 - x_1}{2} \right)^2$$

$$= \frac{1}{2} \left\{ \frac{x_1^2 + x_2^2 - 2x_1x_2}{2} \right\}$$

$$= \frac{1}{4} (x_1 - x_2)^2 \text{ or } \frac{1}{4} (x_2 - x_1)^2$$

Layer Norm [x]

$$\left[\frac{x_1 - \frac{x_1 + x_2}{2}}{\frac{x_1 - x_2}{2}} , \frac{x_2 - \frac{x_1 + x_2}{2}}{\frac{x_1 - x_2}{2}} \right]$$

$$= \left[\frac{\frac{x_1 - x_2}{2}}{\frac{x_1 - x_2}{2}} , \frac{\frac{x_2 - x_1}{2}}{\frac{x_1 - x_2}{2}} \right]$$

$$= [1, -1]$$

If we consider Var $= \frac{1}{4}(x_2 - x_1)^2$

then we get $[-1, 1]$

(3) If γ & β are any real number

then for 2d case

$$\gamma = [\gamma_1, \gamma_2]$$

$$\beta = [\beta_1, \beta_2]$$

so Layer norm for 2d $= [\gamma_1 + \beta_1, -\gamma_2 + \beta_2]$

or $[-\gamma_1 + \beta_2, \gamma_2 + \beta_2]$

for generic d dimension case.

norm of Layer norm $[x]$ if $\beta = 0$

then $\gamma\sqrt{d}$

If $\beta \neq 0$ I am not sure how to simplify.