1. Done.
2. Test accuracies for each of the four settings:

	w/o BERT Fine-tuning	with BERT Fine-tuning
BERT-tiny	0.5137	0.5076
BERT-mini	0.5582	0.6119

Table 1: Experiment Results for RTE. Random baseline accuracy: 0.5046

	w/o BERT Fine-tuning	with BERT Fine-tuning
BERT-tiny	0.5586	0.7015
BERT-mini	0.5204	0.7206

Table 2: Experiment Results for SST-2. Random baseline accuracy: 0.5272

3. The models without fine-tuning performs bad, almost as bad as a random classifier. This is expected as the Bert model was not fine tuned on the task at all, only the linear classifier was trained.

   On the other hand, the fine-tuned models performed better on both the tasks. Mini models are bigger than tiny and performed better as expected. Test accuracy on SST-2 task is better than on RTE task because sentiment classification is an easier task than language inference.
4. Model predictions:

   RTE:

   a. entailment
   b. entailment
   c. entailment
   d. entailment

   SST-2:

   e. positive
   f. positive
   g. positive
   h. negetive
5. RTE: All the examples correctly identified as ‘entitlement’. From this results it seems like the model doesn’t have any gender bias. But it might be needed to be tested on more data to say for sure.

   SST-2: The first 3 is predicted ‘positive’ correctly but the last one was predicted ‘negetive’. It seems like the model is not changing it output based on he/she. But when the pronoun is ‘they’ it somehow gets confused and predict the wrong output.

6. Theory: Exploration of Layer Norm

   The answer is in the hand written part below:

1. Say, the input x is a d-dimensional vector

Given,

$$\text{LayerNorm}[x] = \frac{x-\bar{x}}{\sqrt{\text{Var}[x]+\varepsilon}} \quad \cdots\cdots\cdots (i)$$

Now, we know,

$$\bar{x} = \frac{1}{d}\sum_{i=1}^{d} x_i$$

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

Now, substituting in equation (i) we get,

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

[Ignoring ε as it's a tiny number]

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

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

$$\therefore \text{LayerNorm}[x] = \sqrt{d}$$

(showed).

2. Say, the 2-d input vector is

$x = [a\ ,\ b]$ where a, b can be any scaler.

So, we can write,

$$\text{LayerNorm}[x] = \left[\frac{a - \frac{a+b}{2}}{\sqrt{\frac{1}{2}\left\{\left(a-\frac{a+b}{2}\right)^2 + \left(b-\frac{a+b}{2}\right)^2\right\}}} , \frac{b - \frac{a+b}{2}}{\sqrt{\frac{1}{2}\left\{\left(a-\frac{a+b}{2}\right)^2 + \left(b-\frac{a+b}{2}\right)^2\right\}}} \right]$$

Now, from the first element we get,

$$\frac{\frac{a-b}{2}}{\sqrt{\frac{1}{2}\left(\frac{a-b}{2}\right)^2 + \left(\frac{b-a}{2}\right)^2}}$$

$$= \frac{\frac{a-b}{2}}{\sqrt{\frac{1}{4}(a-b)^2}}$$

$$= \frac{\frac{a-b}{2}}{\frac{a-b}{2}}$$

$$= 1$$

Similarly, from the second element gives us, -1

So, LayerNorm$[x] = [1, -1]$ or $[-1, 1]$ for any 2-d input vector depending on the sign of the elements. (Showed).

3. If γ & β be any real number then,

the answer of question 1 will be $= \gamma\sqrt{d} + \beta$

the answer of question 2 will be $= [\gamma+\beta, -\gamma+\beta]$ or $[-\gamma+\beta, \gamma+\beta]$