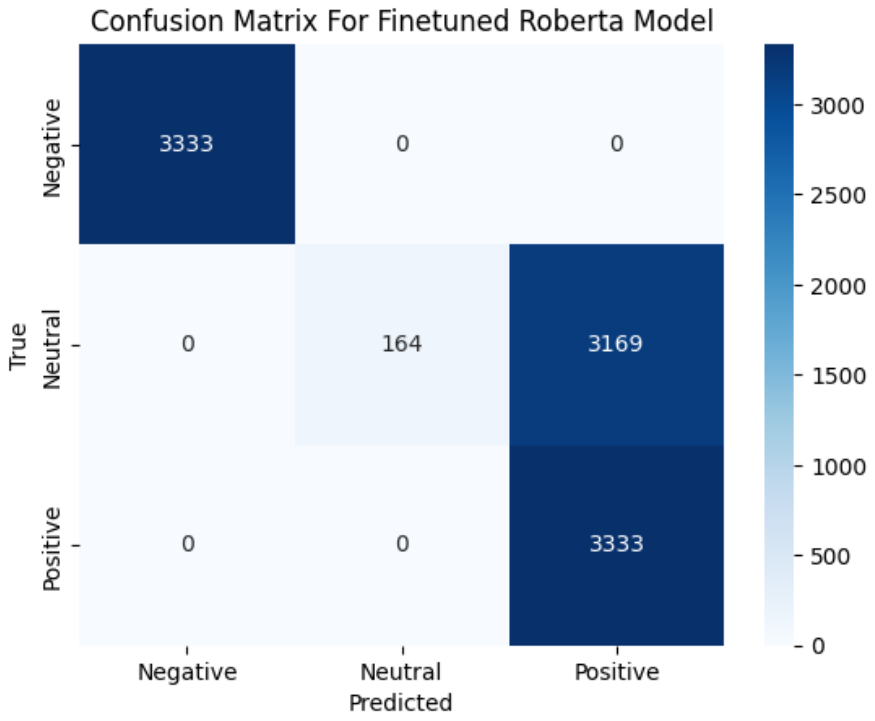
Pretrained Model (Roberta-base)

Metrics:

Class	Precision	Recall	F1-Score	Support
Negative	1.00	1.00	1.00	3333
Neutral	1.00	0.05	0.09	3333
Positive	0.51	1.00	0.68	3333
Accuracy			**0.68**	**9999**
Macro Avg	0.84	0.68	0.59	9999
Weighted Avg	0.84	0.68	0.59	9999



- No false positives for negative and positive reviews. However, out of 3,333 neutral reviews, only 164 were correctly predicted as neutral.
- Neutral sentiments had the least recall, 5%
- Positive sentiments had the least precision, 51%

Error analysis:

Examples:

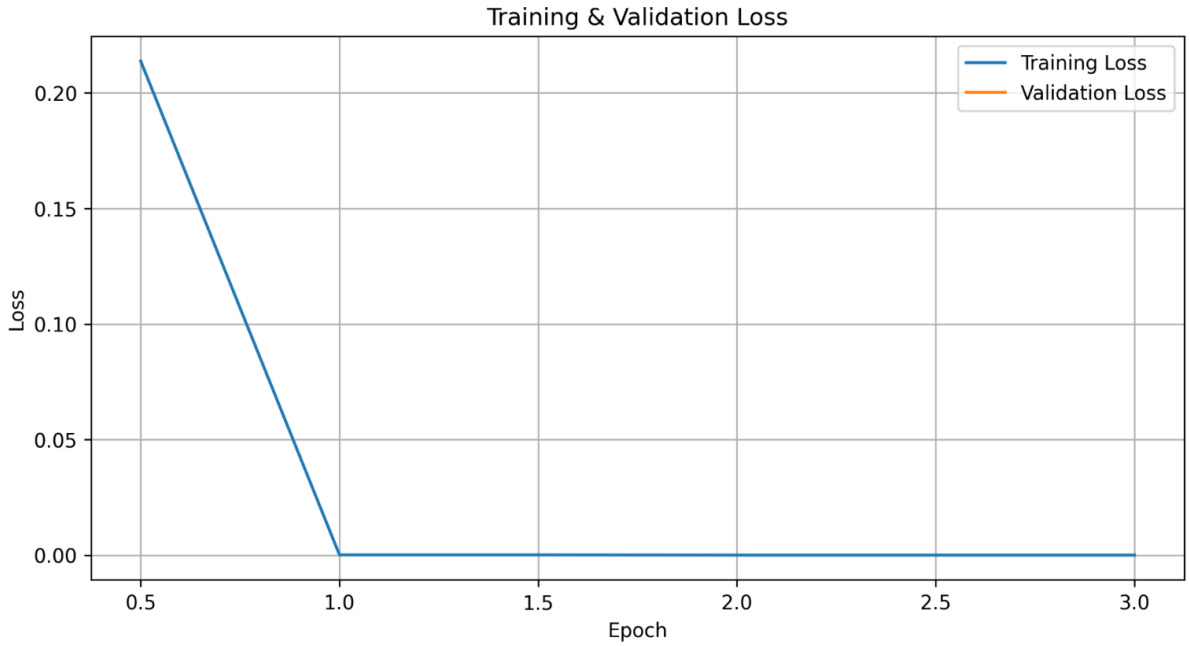
- ***This headphones is alright. If you don't expect much from the sound quality, it's fine.***
  The terms "is alright" and "it's fine" seem to carry more positive weight.
- ***The tablet works fine. The response time is okay but nothing exceptional.***
  The term "woks fine" and "is okay" seems to carry more positive weight.
- ***It's a usable smartphone. The screen meets minimum expectations.***
  The term "usable" likely triggers a positive sentiment.
- ***This smartphone is alright. If you don't expect much from the camera, it's fine.***
  "Is alright" and "it's fine" seem to carry more positive weight.
- ***This headphones is alright. If you don't expect much from the connectivity, it's fine.***
  "Is alright" and "it's fine" seem to trigger positive sentiment prediction.
- ***Decent tablet. It gets the job done though the battery could be better.***
  "Decent", "gets the job done," and "be better" likely trigger positive sentiment prediction.

These reviews tend to use mildly approving terms such "fine", "not bad", and "decent", which the model, likely influenced by its pretraining data, interprets as positive.

Fine-Tuning

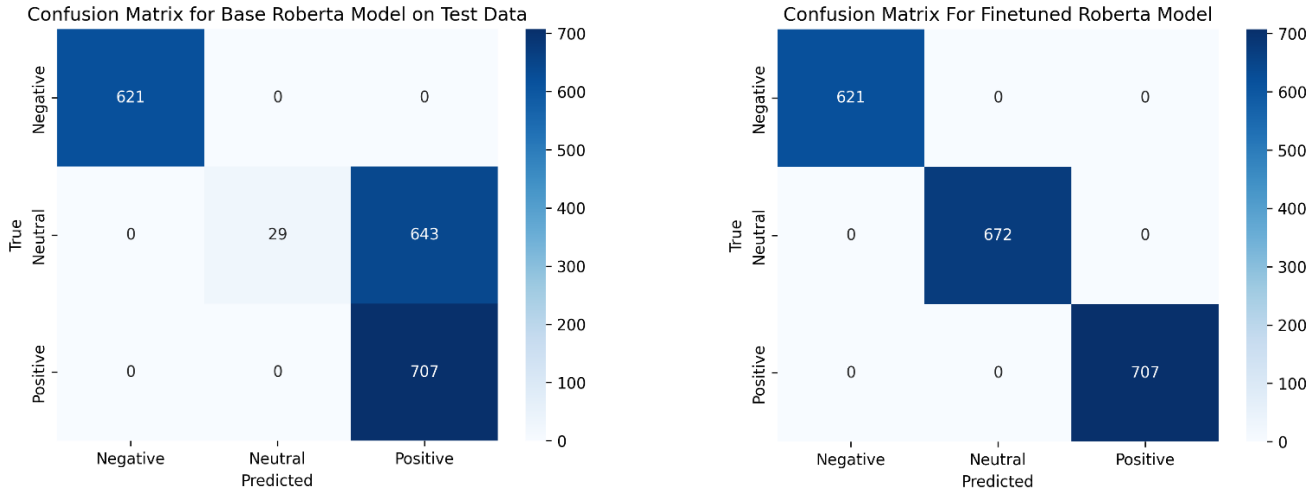
The training of the base model saw an initial loss of over 0.2, which reduced to almost 0.0 in the following epochs.

Seeing this drastic improvement, there was no need to train the model longer, hence, the 3 epochs limit.



Comparison of metrics before and after fine-tuning Roberta-base model:

Roberta-base				**Fine-Tuned**			
Class	**Precision (Before)**	**Recall (Before)**	**F1-Score (Before)**	**Precision (After)**	**Recall (After)**	**F1-Score (After)**	**Support**
Negative	1.00	1.00	1.00	1.00	1.00	1.00	621
Neutral	1.00	0.04	0.08	1.00	1.00	1.00	672
Positive	0.52	1.00	0.69	1.00	1.00	1.00	707
Accuracy	—	—	**0.68**	—	—	**1.00**	2000
Macro Avg	0.84	0.68	0.59	1.00	1.00	1.00	2000
Weighted Avg	0.83	0.68	0.58	1.00	1.00	1.00	2000



A comparison of the confusion matrices after prediction of test data by both models

- From the above, fine-tuning the Roberta-base model by using 80% of the review's dataset sample leads to an increase in sentiment classification accuracy from 68% to 100%.
- The base model could accurately identify and recall negative sentiments, while only accurately identifying neutral sentiments which it also showed the least capacity to recall, recalling only 4% of the neutral sentiments. It failed to accurately identify positive sentiments, only being precise with 52% classification but it did record 100% recall.
- The fine-tuned model marks an improvement, of the recall of neutral sentiments by 96% while also improving the positive sentiments' precision by 48%.

Is the improvement in accuracy statistically significant?

- McNemar's test evaluates whether two classifiers disagree significantly on the same test set.
- Create a 2x2 contingency table from predictions:

	Model B Correct	**Model B Wrong**
Model A Correct	a	b
Model A Wrong	c	d

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

- If χ^2 is large and p-value < 0.05, there is significant difference between the models.
- Thus, with χ^2 and p-value being 0.0000, the improvement of the fine-tuned model's accuracy is statistically significant.

Insights and Observations:

- Fine-tuning allowed the model to learn how to differentiate neutral from positive sentiments which was a major flaw in the base model.
- The neutral sentiments improved the most, seeing that the model could recall neutral sentiments 96% more accurately than before.
- The fine-tuning process was largely straightforward and easy to complete. However, the perfect accuracy, recall, and precision recorded raises some issues:
  I ruled out data leakage since the base model was only trained with 80% of the data and tested on new data, 20%.
- However, I did notice that some of the reviews were quite similar in structure and wording which could have allowed the model to easily generalize or even memorize patterns.
  For instance:
  "This headphones is alright. If you don't expect much from the sound quality, it's fine."
  "This headphones is alright. If you don't expect much from the connectivity, it's fine."
- Therefore, the model could have benefited from redundancy and could perform poorly on a different dataset. It is crucial that it learns from diverse and rich data that can prevent this.