**Observations Criteria:**

Note:

1. For the Text Classification task, the IMDB dataset is used and loaded from the built-in hugging face dataset i.e., “**noob123/imdb\_test**” from [noob123/imdb\_test · Datasets at Hugging Face](https://huggingface.co/datasets/noob123/imdb_test). For all three models’ comparisons, the same IMDB dataset has been used.
2. Tokenizer for the associated models has been used to convert the loaded dataset feature into the format to match with the format of the dataset of the pre-trained models on which it is trained
3. 100 samples of test dataset have been used to avoid the time consumption problem

* **Text Classification Task:**
  + **Model 1 – Roberta base**
    1. The pre-trained model is trained on the IMBD dataset ([imdb · Datasets at Hugging Face](https://huggingface.co/datasets/imdb))

All metrics:

{'eval\_loss': 0.8647624254226685, 'eval\_accuracy': 0.82, 'eval\_recall': 0.8518518518518519, 'eval\_precision': 0.8214285714285714, 'eval\_F1': 0.8363636363636364, 'eval\_runtime': 12.3165, 'eval\_samples\_per\_second': 8.119, 'eval\_steps\_per\_second': 1.055}

* + **Model 2 – Distilbert base uncased fine-tuned**
    1. The pre-trained model is trained on the IMBD dataset ([imdb · Datasets at Hugging Face](https://huggingface.co/datasets/imdb))

All metrics:

{'eval\_loss': 0.5687190890312195, 'eval\_accuracy': 0.78, 'eval\_recall': 0.8518518518518519, 'eval\_precision': 0.7666666666666667, 'eval\_F1': 0.8070175438596491, 'eval\_runtime': 5.5461, 'eval\_samples\_per\_second': 18.031, 'eval\_steps\_per\_second': 2.344}

* + **Model 3 – Albert Base V2**

All metrics:

{'eval\_loss': 0.6518518924713135, 'eval\_accuracy': 0.85, 'eval\_recall': 0.9074074074074074, 'eval\_precision': 0.8305084745762712, 'eval\_F1': 0.8672566371681415, 'eval\_runtime': 11.0557, 'eval\_samples\_per\_second': 9.045, 'eval\_steps\_per\_second': 1.176}

**Models Results:**

* + Below table shows the final comparison between all three models with the same test dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| **Pre-trained Transformer Model** | **Accuracy** | **Loss** | **F1** |
| Simple Roberta Base | 82% | 86% | 0.83 |
| Distilbert-base-uncased (fine-tuned version) | 78% | 56% | 0.80 |
| Albert-base-v2 (fine-tuned version) | **85%** | 65% | **0.86** |

**SUMMARY OF ANALYSIS OF RESULTS:**

1. Overall accuracy of Albert-base-v2 85% which is better than the other two models, logically expected from the ALBERT model
2. ALBERT is a kind of lite BERT, which has fewer parameters as compared to the other two models Roberta and DistilBert, to make the model parameter efficient by achieving better results
3. ALBERT uses factorized embedding parameterization, where the embedding matrix is broken down into smaller matrices and cross-layer parameter sharing, reusing the same parameters through the layers
4. F1 score of 0.86 also concluded that the ALBERT Base V2 model performed well, one strange thing to notice is the evaluation loss of the ALBERT model i.e., 65% much higher than expected.
5. The increase in loss could be because of overfitting, the test dataset that has been used for evaluation testing might be closely related to the trained model
6. This is surprising to see that the Simple Roberta base model loss is increased with accuracy. This is a clear sign of overfitting, as it is said earlier that it could be because bias is introduced to the model as it is closely related to the dataset
7. Another reason, might be because of cross-entropy loss for classification, which leads to bad predictions sometimes. A single misclassification target label will have a high loss and which results in a higher mean loss
8. Although DistilBert model accuracy is not better than the other two models, its loss is comparatively low, which could be possible as the model is a fine-tuned version and retains 90% of the functionality
9. DistilBert model is used because it’s 40% smaller than the original BERT-base mode, 60% faster than it
10. When the dataset samples increased from 100 to 1k for Roberta the accuracy increased by 1% and there was a considerable drop in the loss from 86% to 74%, indicating the testing data size does matter for the loss, with further increase in the number of samples to 2k the accuracy becomes better and a further drop in the loss has been seen
11. Same happened with the DistilBert model, an increase in the number of samples improved the accuracy and a noticeable drop has been observed in the loss

**Limitations of the comparisons:**

* + Large dataset usage constraint, because of the training time consumption
  + Hugging face dataset has around 2k samples for most of the types of the dataset, which is not more than enough and restricts understanding the model in depth
  + DistilBert Model has been chosen as one of the models because it’s a lightweight model as compared to the other two models, and its performance is better than the other two models
  + Choosing the correct model was a bit of challenge that does the same classification task and is trained on a similar kind of dataset,
  + However, the decision of choosing the dataset from the same field i.e., IMDB forced the model to overfit with a higher loss value

**Future work using pre-trained models:**

* The transformer-based pre-trained models could be useful for detecting or monitoring road accidents using previous road accidents data
* What could be the likelihood of a person getting into a road accident through observing different crucial factors like traffic density, speed, weather, roadway conditions, and could be many more deciding factors