Natural Language Processing

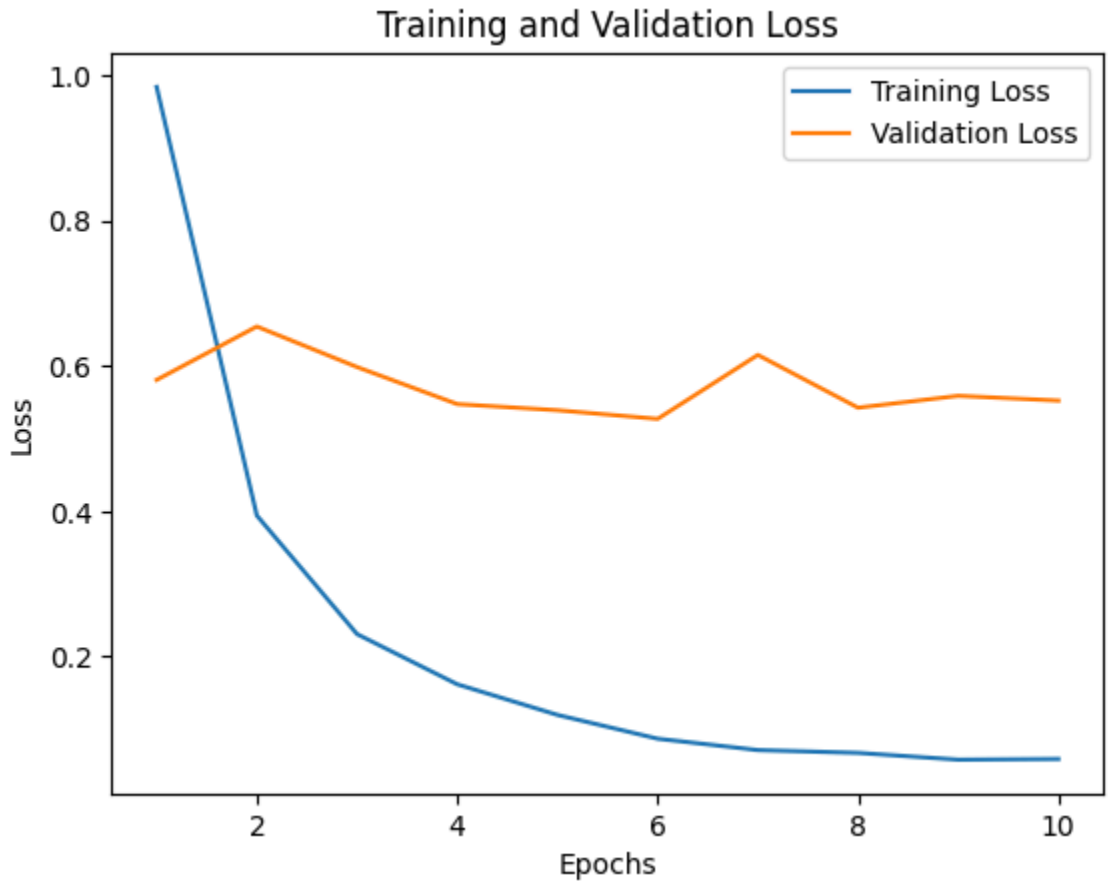
Assignment-1

Manan Chugh 2021335
Abhijeet Anand 2021509
Ankit Gautam 2021518
Samridh Girdhar 2021282

Task 1

Setup 1A: -

Plot:



Analysis:

In the graph we can see that as the number of epochs are increasing the training loss is decreasing and it went down to 0.05 without any unusual behavior and spikes, this signifies that the model has been trained efficiently. However the Validation loss has minimal changes during the training, it ranges from 0.52 - 0.61.

Pearson Correlation:

```
Validation: 100%|██████████| 92/92 [01:26<00:00,  1.06it/s]
Pearson correlation coefficient: 0.8700755408169226
```

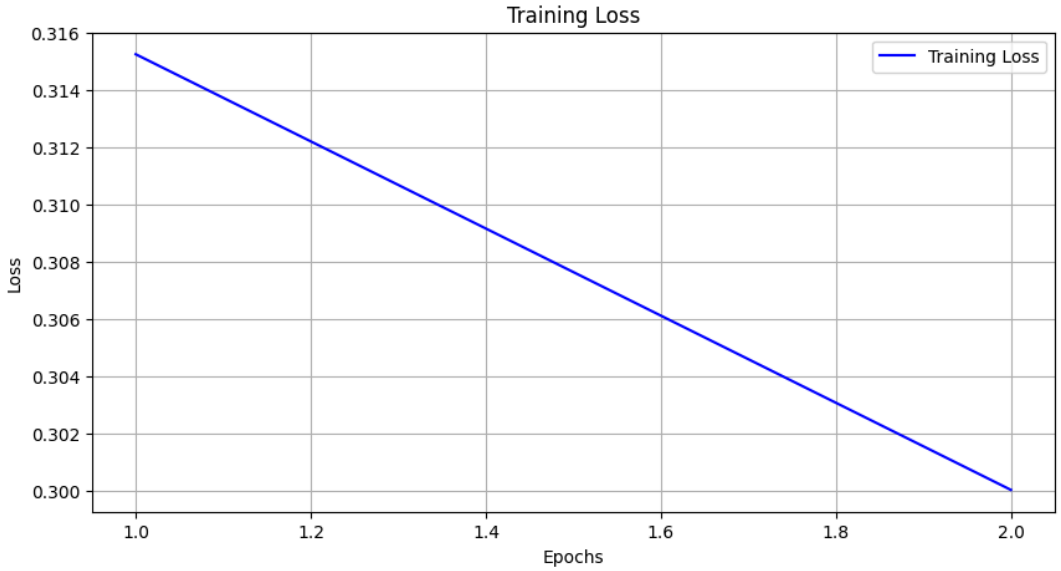
Setup 1B: -

Pearson Correlation:

```
Pearson's correlation coefficient on validation set: 0.8631423871595579
```

Setup 1C: -

Plots:



Analysis:
As we can see in the graphs, both the training and validation loss are linearly decreasing as the epochs increase. The training loss ranges from 0.315 - 0.300 and the valdition loss ranges from 0.268 - 0.236. As we can see that there is very little difference between the training and the validation loss this means there is no overfitting or underfitting and the model has been trained perfectly.

Pearson Correlation:
As we can see from the 1st epoch only the Pearson correlation coefficient is surpassing the same of setup 1B, and in the 2nd epoch it increases even more.

```
Epoch : 0 Train loss : 0.3152616332640056 Val loss : 0.26872962171040315 Pearson Correlation : 0.8705970383070621
```

```
Epoch : 1 Train loss : 0.3000498040994424 Val loss : 0.23635153107614582 Pearson Correlation : 0.8806409824206274
```

Comparison:

Setup 1A (BERT) has a high Pearson Correlation when compared to Setup 1B (pre-trained model) this signifies that the BERT model is performing better than the Sentence-BERT model in finding the text similarities. However, as we did fine-tuning of the Sentence-BERT model in Setup 1C, we can see that not only it is performing better than Setup 1B (pre-trained model) but it is performing better than the Setup 1A (BERT).

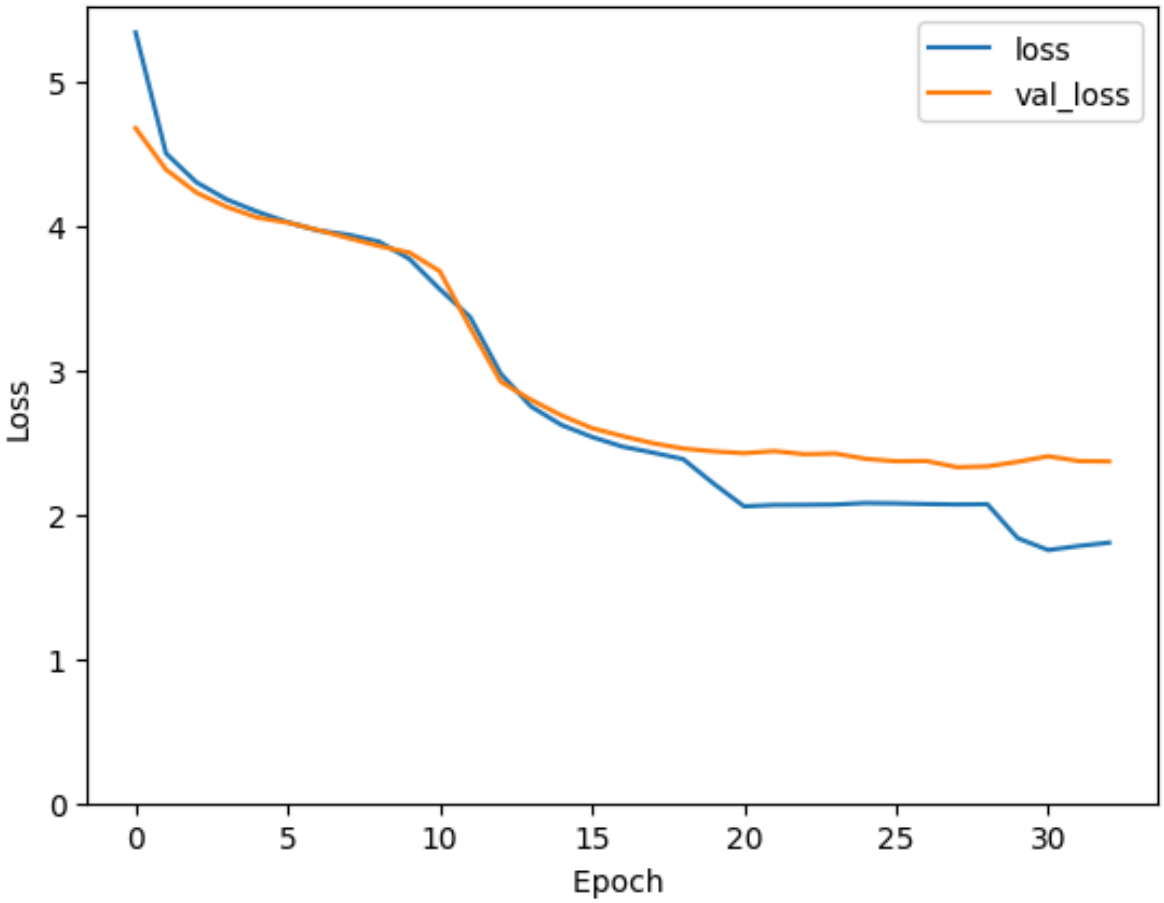
The higher the Pearson Correlation gets the stronger a linear relationship gets between the two variables. All of our setups are showing a positive correlation meaning that when one variable increases the other also increases, which is required in our task as we need to find similarities in the texts. Eventually at the end the Setup 1C is performing the best out of all three models.

Evaluation Metrics:

Model	Pearson Correlation
Setup 1A	**0.870075540816 (after 10th epoch)**
Setup 1B	**0.863142387159 (from pre-trained model)**
Setup 1C	**0.880640982420 (after 2nd epoch)**

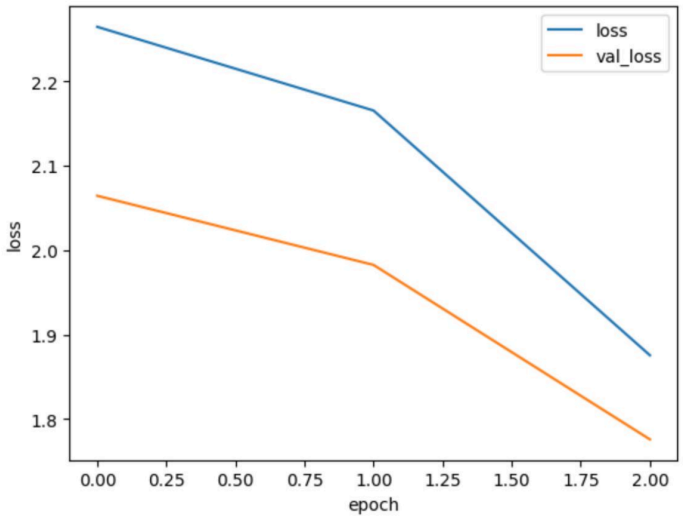
Task 2

1.) 2A - Training Loss and Validation Loss



The above plot represents Training Loss and Validation Loss V/s Epoch. From the above inference, we can see that the loss is constantly decreasing with an increase in the number of epochs. This suggests that the model effectively learns from the training data over successive epochs. As the loss decreases, it indicates that the model's predictions are becoming more accurate compared to the actual target values. The decreasing validation loss also shows that the model is generalizing well to unseen data, as it performs well not only on the training set but also on the validation set.

2.) 2C - Training Loss and Validation Loss



The above plot represents Training Loss and Validation Loss V/s Epoch. From the above inference, we can see that the loss is constantly decreasing with an increase in the number of epochs. A rapid decrease in loss early on suggests the model is grasping the task quickly. Rapidly decreasing validation loss indicates the model isn't just memorizing the training data but can generalize its learned patterns to unseen data, a crucial aspect of robust NLP models.

Comparison and explanation of the performance differences between the three setups

From the provided Meteor scores, we can observe that Model 3(2c) has a significantly lower Meteor score than Models 1(2a) and 2(2b). It also has better BLEU and BERTScore. This suggests that Model 3 performs better according to the Meteor metric, as it produces outputs that are closer to human-generated references compared to Models 1 and 2. This is because model 3 understands English better than models 1 and 2.

The poor performance of model 1 is because it is a relatively small model and needs more optimization. The model fails to distinguish between nouns and produces vague/incorrect translations.

Model 2 is just a pre-trained model that performs English-to-German translation. It does not support German to English translation hence, it cannot be compared with the other two models in the context of German to English translation.

Model 3 is a fine-tuned model that has been fine-tuned on the T5 model. Due to these fine tunings, the model performs better translations and has provided good results.

All evaluation metrics for all the setups

Validation scores:

setups	BLEU-1	BLEU-2	BLEU-3	BLUE-4	METEOR	BERT-f1	BERT precision	BERT recall
2A	0.377	0.214	0.131	0.083	0.3127	0.721	0.716	0.717
2B	0.475	0.306	0.209	0.147	0.530	0.861	0.863	0.866
2C	0.513	0.327	0.230	0.174	0.461	0.912	0.921	0.911

Test Scores:

setups	BLEU-1	BLEU-2	BLEU-3	BLUE-4	METEOR	BERT-f1	BERT precision	BERT recall
2A	0.375	0.217	0.134	0.086	0.318	0.721	0.723	0.718
2B	0.503	0.338	0.236	0.168	0.569	0.872	0.874	0.881
2C	0.531	0.362	0.277	0.201	0.491	0.912	0.911	0.923

Contributions

Task1 - Abhijeet and Manan

Task2 - Ankit and Samridh