CSE 556: Natural Language Processing Assignment 3

Sanskar Ranjan 2021096 | Saumil Lakra 2021097 | Jeremiah Malsawmkima Rokhum 2021533 | Vishal Singh 2021575

1. Evaluation for Task 1

1.1. 1A:

- Pearson Score on Validation Data: 0.8641

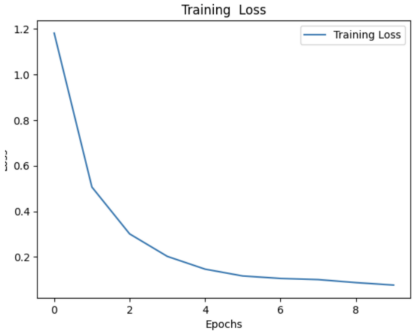

Figure 1. Training Loss of Task 1A

- The training loss decreases with each epoch, which means that the model is learning, i.e., it is able to capture the patterns in the data.

Figure 2. Validation Loss of Task 1A

- The graph of validation loss gives a decreasing curve, which means that the model is increasing its performance with each epoch. It is not overfititing since it performs better on unseen data.

1.2. 1B:

- Pearson Score on Validation Data is 0.792024

1.3. 1C:

- Pearson Score on Validation Data is 0.8126659

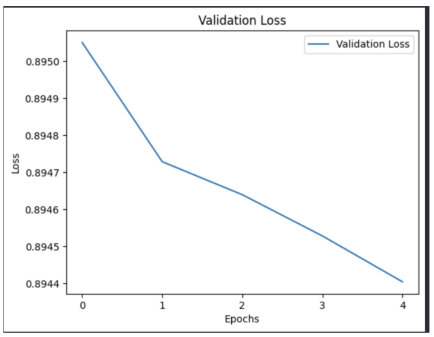

Figure 3. Validation Loss of Task 1C

- The graph of validation loss gives a decreasing curve, which means that the model is increasing its performance with each epoch. It is not overfitting since it performs better on unseen data.

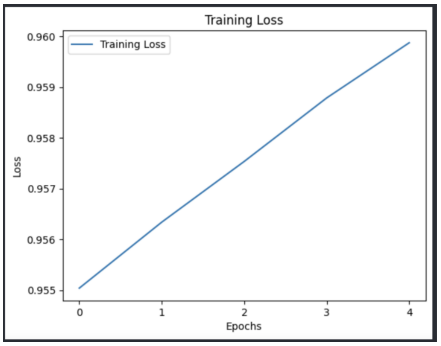

Figure 4. Training Loss of Task 1C

- The training rate increases with each epoch which means that the model is underfitting, i.e., it is unable to capture the patterns in the data. We also tried decreasing the learning rate, but the outcome was not changed. Increasing the number of epochs also didn't help.

1.4. Comparisons between the three subtask

For Task1B, since we used pre-trained 'all-MiniLM-L6-v2', it gave us the Pearson correlation value of 0.792. In the fine-tuned model, the Pearson Correlation on the same test dataset(validation set) was 0.812. This may have happened because the pre-trained model is trained on a more extensive set of data, but fine-tuning made the model more accurate. Fine-tuning the model is also one of the reasons why it gave a better Pearson correlation value.

2. Evaluation for Task 2

2.1. 2A

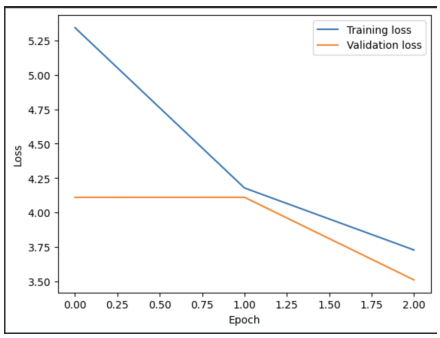

Figure 5. Training and Validation Loss of Task 2A

- The training loss decreases with each epoch, which means that the model is learning, i.e., it is able to capture the patterns in the data.
- The graph of validation loss gives a decreasing curve, which means that the model is increasing its performance with each epoch. It is not overfititing since it performs better on unseen data.
- The BERT score is 0.7631, 0.818, 0.7894 respectively for Precision, Recall and F1 Score.

2.2. 2B

- The BERT score for validation is 'P': 0.8086380362510681, 'R': 0.7521953582763672, 'F1': 0.7783204317092896 respectively for Precision, Recall and F1 Score.
- The BERT score for test is 'P': 0.8144293427467346, 'R': 0.754437267780304, 'F1': 0.782230794429779 respectively for Precision, Recall and F1 Score.
- The BLUE1, it is {'bleu': 0.1121480385157142, 'precisions': [0.5469280831730179, 0.2994789327104483, 0.19122257053291536, 0.1298020630052969], 'brevity_penalty': 0.4441323116836975, 'length_ratio': 0.551988251242657, 'translation_length': 24431, 'reference_length': 44260}
- for validation. {'bleu': 0.11971767311094333, 'precisions': [0.5625072784441598, 0.3244019138755981, 0.21118713409042816, 0.1442485902440948], 'brevity_penalty': 0.438440455293214, 'length_ratio': 0.548085975522188, 'translation_length': 34348, 'reference_length': 62669}
- For BLUE2, it is {'bleu': 0.17974657362098445, 'precisions': [0.5469280831730179, 0.2994789327104483], 'brevity_penalty': 0.4441323116836975, 'length_ratio': 0.551988251242657, 'translation_length': 24431, 'reference_length': 44260}
- {'bleu': 0.18729077945943867, 'precisions': [0.5625072784441598, 0.3244019138755981], 'brevity_penalty': 0.438440455293214, 'length_ratio': 0.548085975522188, 'translation_length': 34348, 'reference_length': 62669}
- For BLUE3, it is {'bleu': 0.13999878106724378, 'precisions': [0.5469280831730179, 0.2994789327104483, 0.19122257053291536], 'brevity_penalty': 0.4441323116836975, 'length_ratio': 0.551988251242657, 'translation_length': 24431, 'reference_length': 44260}
- {'bleu': 0.14809382726383616, 'precisions': [0.5625072784441598, 0.3244019138755981, 0.21118713409042816], 'brevity_penalty': 0.438440455293214, 'length_ratio': 0.548085975522188, 'translation_length': 34348, 'reference_length': 62669}
- For BLUE4, it is {'bleu': 0.1121480385157142, 'precisions': [0.5469280831730179, 0.2994789327104483, 0.19122257053291536, 0.1298020630052969], 'brevity_penalty': 0.4441323116836975, 'length_ratio': 0.551988251242657, 'translation_length': 24431, 'reference_length': 44260}
- 'bleu': 0.11971767311094333, 'precisions': [0.5625072784441598, 0.3244019138755981, 0.21118713409042816, 0.1442485902440948], 'brevity_penalty': 0.438440455293214, 'length_ratio': 0.548085975522188, 'translation_length': 34348, 'reference_length': 62669}

2.3. 2C

- The BERT score {'P': 0.7592755556106567, 'R': 0.6897149085998535, 'F1': 0.7214769721031189} {'P': 0.782316267490387, 'R': 0.7103234529495239, 'F1': 0.7432899475097656} done in 8.80 seconds, 340.78 sentences/sec {'P': 0.7592755556106567, 'R': 0.6897149085998535, 'F1': 0.7214769721031189} {'P': 0.782316267490387, 'R': 0.7103234529495239, 'F1': 0.7432899475097656}
- The BLUE1, it is{'bleu': 0.25241930612068536, 'precisions': [0.7265163839181966], 'brevity_penalty': 0.3474378716132394, 'length_ratio': 0.48610483506552193, 'translation_length': 21515, 'reference_length': 44260} {'bleu': 0.25502451941772625, 'precisions': [0.7262989451683485], 'brevity_penalty': 0.3511288583224009, 'length_ratio': 0.48861478561968436, 'translation_length': 30621, 'reference_length': 62669} {'bleu': 0.25241930612068536, 'precisions': [0.7265163839181966], 'brevity_penalty': 0.3474378716132394, 'length_ratio': 0.48610483506552193, 'translation_length': 21515, 'reference_length': 44260} {'bleu': 0.25502451941772625, 'precisions': [0.7262989451683485], 'brevity_penalty': 0.3511288583224009, 'length_ratio': 0.48861478561968436, 'translation_length': 30621, 'reference_length': 62669
- for validation. {'bleu': 0.11971767311094333, 'precisions': [0.5625072784441598, 0.3244019138755981, 0.21118713409042816, 0.1442485902440948], 'brevity_penalty': 0.438440455293214, 'length_ratio': 0.548085975522188, 'translation_length': 34348, 'reference_length': 62669}

- For BLUE2, it is {'bleu': 0.17974657362098445, 'precisions': [0.5469280831730179, 0.2994789327104483], 'brevity_penalty': 0.4441323116836975, 'length_ratio': 0.551988251242657, 'translation_length': 24431, 'reference_length': 44260}

- {'bleu': 0.18729077945943867, 'precisions': [0.5625072784441598, 0.3244019138755981], 'brevity_penalty': 0.438440455293214, 'length_ratio': 0.548085975522188, 'translation_length': 34348, 'reference_length': 62669}

- For BLUE3, it is {'bleu': 0.13999878106724378, 'precisions': [0.5469280831730179, 0.2994789327104483, 0.19122257053291536], 'brevity_penalty': 0.4441323116836975, 'length_ratio': 0.551988251242657, 'translation_length': 24431, 'reference_length': 44260}

- {'bleu': 0.14809382726383616, 'precisions': [0.5625072784441598, 0.3244019138755981, 0.2111871340904 2816], 'brevity_penalty': 0.438440455293214, 'length_ratio': 0.548085975522188, 'translation_length': 34348, 'reference_length': 62669}

- For BLUE4, it is {'bleu': 0.1121480385157142, 'precisions': [0.5469280831730179, 0.2994789327104483, 0.19122257053291536, 0.1298020630052969], 'brevity_penalty': 0.4441323116836975, 'length_ratio': 0.551988251242657, 'translation_length': 24431, 'reference_length': 44260}

- 'bleu': 0.11971767311094333, 'precisions': [0.5625072784441598, 0.3244019138755981, 0.2111871340904 2816, 0.1442485902440948], 'brevity_penalty': 0.438440455293214, 'length_ratio': 0.548085975522188, 'translation_length': 34348, 'reference_length': 62669}

2.4. Credit Statement

Our contribution towards this assignment was fairly equal because, we discussed whenever each one of us faced issues and while proceeding with an idea.

1. **Saumil Lakra:** Task 1 B and Task 1C
2. **Jeremiah Rokhum:** Task 2A and Task2B
3. **Sanskar Ranjan:** Task 2C and Task 2A
4. **Vishal Singh:** Task 1A and Task 1B