Assignment 5 Report
Sentence Pair Classification using Transformer-Based Models

Name: Gopi Trinadh Maddikunta
PSID: 2409404

The main objective is to fine-tune the transformer-based models like roberta-base for sentence pair classification tasks. Here we explored two tasks. Paraphrase Detection which helps in prediction when two sentences are paraphrases on MRPC dataset and Entailment Recognition which helps in the prediction when one sentence logically entails another on RTE dataset. Where both tasks are good at real world applications like semantic search, QA systems and detection of duplicate.

By starting with step 1, Initiated with Dataset preprocessing where dataset is Microsoft Research Paraphrase Corpus (MRPC) which is loaded using Huggingface Datasets. Preprocessing were done by using RobertaTokenizer with no additional cleaning.

Fine-tuning in two way strategy for initial step, Starting with freezing all layers except classification head then fine-tuned for 10-20 epochs and I have applied early stopping. I have some hyperparameter, learning rate: 5e-5, Batch size : 16 and AdamW as optimizer for the sake monitoring training loss and validation metrics recorded per epoch. Second strategy is unfreeze all parameters of the model and finr-tuned with 3-5 epochs and lowered learning rate to 2e-5 for preventing catastrophic forgetting. Then focus on the evaluation metrics like Accuracy, Precision, Recall, F1-score.

Model Type	Accuracy	Precision	Recall	F1 Score
Frozen Model	0.6838	0.6838	1.00	0.8122
Unfrozen Model	0.8775	0.8908	0.9355	0.9126

Table: Metrics to evaluate performance

Best Model: Unfrozen RoBERTa-base which is fine-tuned for 4 epochs and I have trained full model that is no freezing which resulted better F1 Score.

For task 2, the main objective is to fine-tune the transformer-based mode like RoBERTa-base for the sentence pair classification task of Entailment Recognition. Here I have used the RTE dataset, where the goal is predict whether the second sentence logically follows from the first. This task is useful in real world applications like textual inference, automated reasoning and document understanding.

Starting with, dataset preprocessing done by loading the Recognizing Textual Entailment (RTE) dataset using Huggingface Datasets library. The dataset were preprocessed using RoBERTaTokenizer with truncation and padding without any additional cleaning.

Fine-tuning was performed with two-way strategy. Initially, freezing all layers except the classification head and fine-tuned for 10-20 epochs. Early stopping applied for preventing overfitting. The hyperparameters chosen included learning rate 2e-5, batch size 32 and AdamW optimizer, and training loss and validation metrics monitored per epoch. In the second strategy, all parameters of the model were unfrozen and the model fine-tuned for 3-5 epochs. Early stopping was again applied carefully to avoid overfitting on small dataset. Evaluation based on accuracy, precision, Recall and F1 Score, and final testing metrics reported on fully 277 examples from RTE validation set.

Model Type	Accuracy	Precision	Recall	F1 Score
Frozen Model	0.5060	0.5052	0.9602	0.6621
Unfrozen Model	0.7148	0.7955	0.5344	0.6393

Table: Metrics to evaluate performance

Best Model: Unfrozen RoBERTa model which is fine-tuned with all parameters unfrozen because frozen model is over-predicting the positive class with high recall and low precision as well as Unfrozen model gives better balance between precision & recall and it is correctly generalizes better on unseen data samples. The overall balance and improved accuracy made the unfrozen model for the better choice.

For each task, I have finetuned two models and below I have mentioned number of trained parameters for each.

Model Type	Trainable Parameters	Total Parameters
Frozen model	592,130	124,647,170
Unfrozen model	124,647170	124, 647, 170

In Frozen model, only classification head was trained which results in 592,130 parameters in trainable, while the total parameters of model stayed same as 124,647, 170. In unfrozen model, all parameters were made trainable including RoBERTa backbone, so the number of trainable parameters becomes 124,647,170. This clearly shows that frozen fine-tuning is much lightweight compare to full fine-tuning.

Performance metrics on unseen data should be clearly mentioned and discussed (accuracy, precision, recall, F1 score). Below table shows that

Task	Model	Accuracy	Precision	Recall	F1 Score
MRPC	Frozen	0.6838	0.6838	1.000	0.8122
MRPC	Unfrozen	0.8775	0.8908	0.9355	0.9126
RTE	Frozen	0.5060	0.5052	0.9602	0.6621
RTE	Unfrozen	0.7148	0.7955	0.5344	0.6393

The Per class precision, recall and F-1 Score

Classification Report:

	precision	recall	f1-score	support
0	0.6615	0.8836	0.7566	146
1	0.7927	0.4962	0.6103	131
accuracy			0.7004	277
macro avg	0.7271	0.6899	0.6835	277
weighted avg	0.7236	0.7004	0.6874	277

277 samples achieved an accuracy of 70.04%. The model showed a precision of 79.27% and recall of 49.62% for the entailment class, while the non-entailment class had a recall of 88.36%. The overall weighted F1 score was 68.74%, showing a good balance between both classes.

Hyperparameter used for all 4 models. Are shown in below table.

Hyperparameter	Value	Justification
Learning Rate	2e-5	Small lr is better for stable finetuning.
Batch size	32	Gives balance between fast train and fit.
Optimizer	AdamW	Handles weight decay, improves generalization
Epochs (frozen)	10-20	Classification head is trainable
Epochs(unfrozen)	3-5	Less epochs avoid overfitting.
Early Stopping	Frozen	If validation loss not improving.

Describe fine-tuning process and challenges.

Task	Challenges	Solution	LoRA
MRPC	Overfitting risks, Small dataset	Early stopping, fewer unfrozen epochs	No
RTE	Tiny validation set, fluctuation.	Careful epoch limits, early stopping	No

If I used LoRA had been used, trainable parameters would drop from ~125M → 1-2M.

Ways to Improve Further:

- Longer training with stronger early stopping.
- Data Agumentation
- Using larger models like RoBERTa-large
- Ensemble of multiple models.
- Use LoRA for larger future models.

Additional Insights:

- ➢ Full Fine-tuning is critical for small complex tasks.
- ➢ Precision-Recall Tradeoff Needs careful monitoring.
- ➢ Early Stopping is Mandatory for small Dataset.

Fine-tuning roberta-base with careful strategies helped to improve sentence pair classification performance even on small datasets.