# Lab 2 Report

*Aravind Nair, aanair@kth.se*

*Mikolaj Konrad Bochenski, mkbo@kth.se*

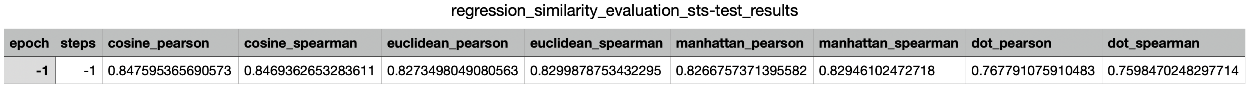
# Regression Task

Pretrained Model: ***bert-base-uncased***

Dataset used - STS Benchmark dataset

Run command - python3 regression.py

Result -



regression_similarity_evaluation_sts-test_results

| epoch | steps | cosine_pearson | cosine_spearman | euclidean_pearson | euclidean_spearman | manhattan_pearson | manhattan_spearman | dot_pearson | dot_spearman |
|---|---|---|---|---|---|---|---|---|---|
| **-1** | -1 | 0.847595365690573 | 0.8469362653283611 | 0.8273498049080563 | 0.8299878753432295 | 0.8266757371395582 | 0.82946102472718 | 0.767791075910483 | 0.7598470248297714 |

Fig 1. Regression task results on STS benchmark test dataset with

# Classification Task

Pretrained Model: **roberta-base**

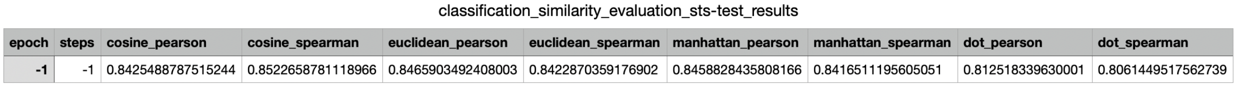
Dataset used for training - All NLI Dataset

Dataset used for fine-tuning - STS Benchmark dev dataset

Dataset used for evaluation - STS Benchmark test dataset

Run command - python3 classification.py

Result -



classification_similarity_evaluation_sts-test_results

| epoch | steps | cosine_pearson | cosine_spearman | euclidean_pearson | euclidean_spearman | manhattan_pearson | manhattan_spearman | dot_pearson | dot_spearman |
|---|---|---|---|---|---|---|---|---|---|
| **-1** | -1 | 0.8425488787515244 | 0.8522658781118966 | 0.8465903492408003 | 0.8422870359176902 | 0.8458828435808166 | 0.8416511195605051 | 0.812518339630001 | 0.8061449517562739 |

Fig 2. Evaluation on STS benchmark test dataset with model trained on NLI dataset

From Fig 1 and Fig 2 we can infer that the accuracy of the model does gets better when it is trained on NLI dataset and then fine-tuned using  STS benchmark dev dataset. However the classification task was run for just 4 epochs with the fine-tuning set every 2 epochs due to computational constraints. The actual impact of such an experiment could be understood only by training for a longer period (say 100 epochs).