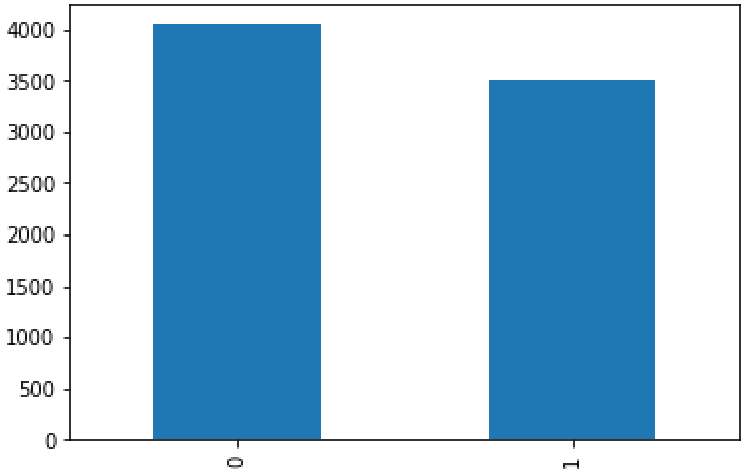
Eeshaan Ravi Tivari (2019465)
Manvi Goel (2019472)

REPORT
Assignment 3

CSE641- Deep Learning

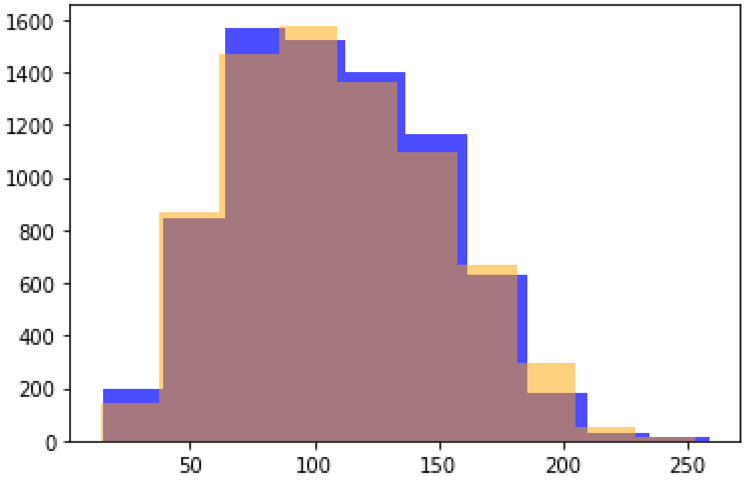
1. Visualization of Data

Distribution of classes in the Training dataset

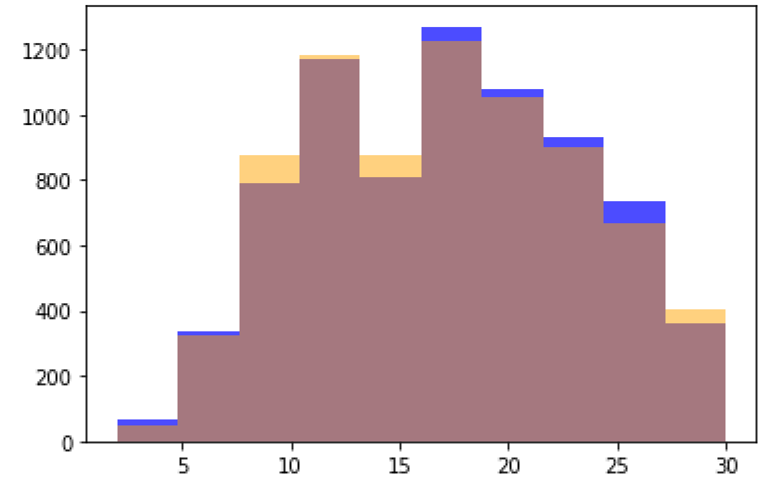


Note - From the above plot we can see that the data is almost evenly distributed between both classes with the non-paraphrased class having a slightly higher no. of samples in the distribution. There are no null values in the dataset.

Distribution of character-level length of Definition 1 (blue) and Definition 2 (orange)

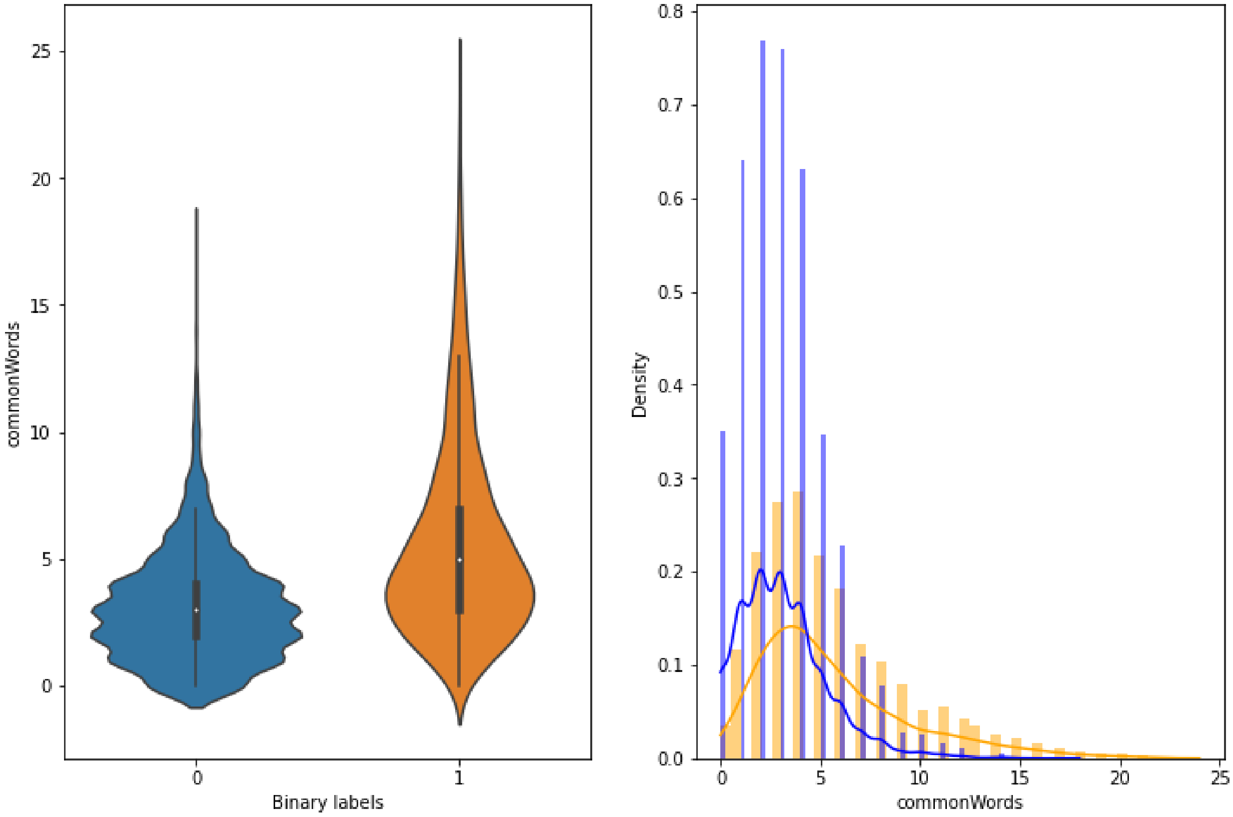


Distribution of word-level length of Definition 1 (blue) and Definition 2 (orange)



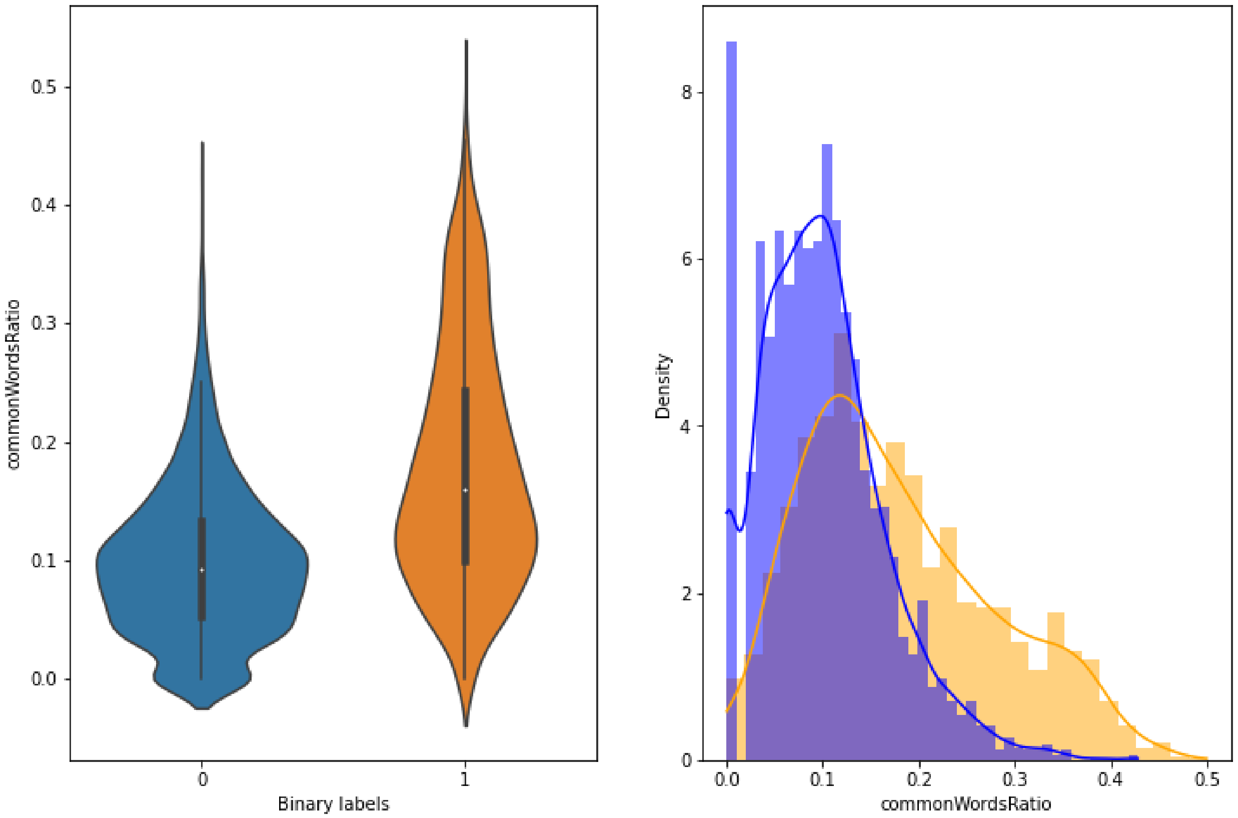
Note - We can see that the maximum number of in a definition is 30. Thus, the maximum length for the tokenizer embedding can be 35*2 + 3.

The number of common unique words in the two definitions for the two classes.



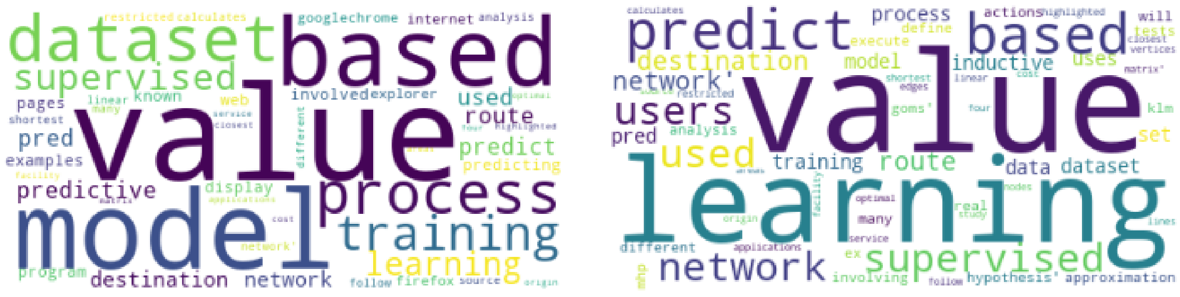
Note - We can see that the number of common words is higher for the definitions that are paraphrases of each other.

The number of common unique words in the two definitions for the two classes is divided by the sum of the length of the definitions.



Note - The sharing of words is more apparent from the ratio which shows the ratio of common words is significantly higher in the paraphrased sentences.

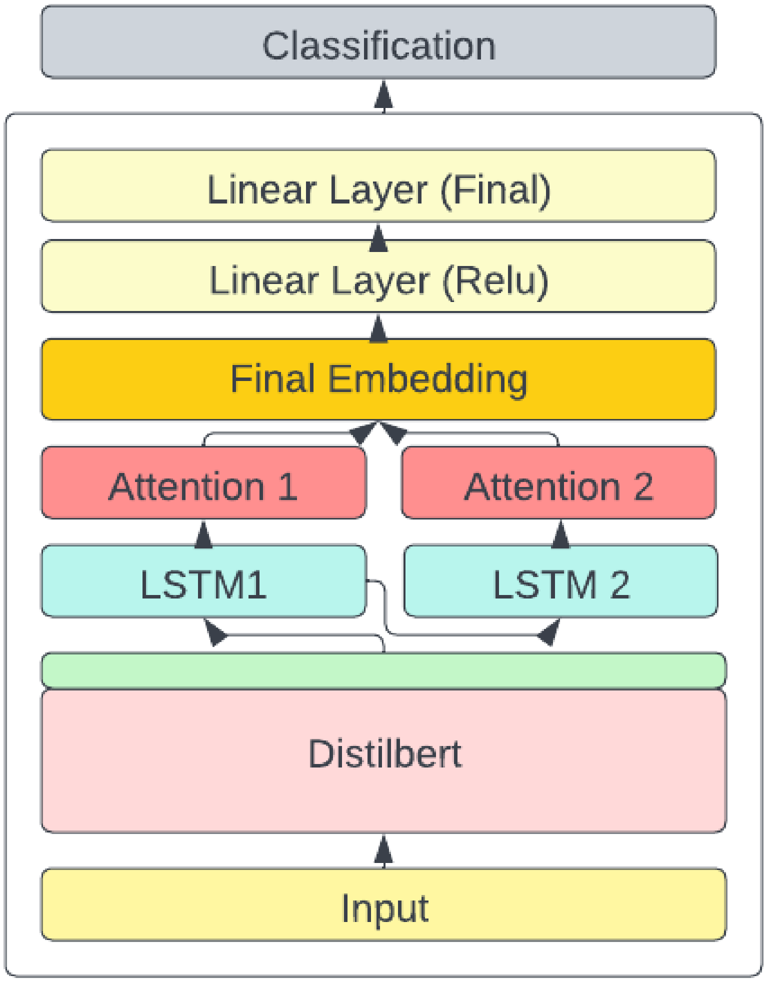
Word Clouds for the definitions that are paraphrased and for the definitions that are not paraphrased.



Note - We can see that similar words are used in the two classes, thus an approximation cannot be made using this and a semantic understanding of the definitions is required.

2. Your Model

Final Model Architecture



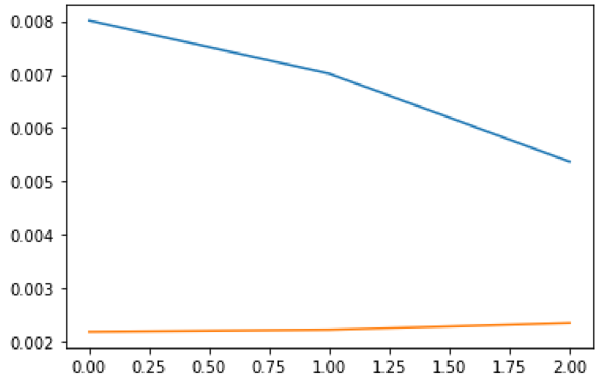
Evaluation metrics on Test Data
F1 Score - 0.7163232963549919
Accuracy - 0.7361827560795873
Recall - 0.6953846153846154
Precision - 0.738562091503268
Loss - 0.0088959587744781

Experiments on the test dataset for the fully trainable model.

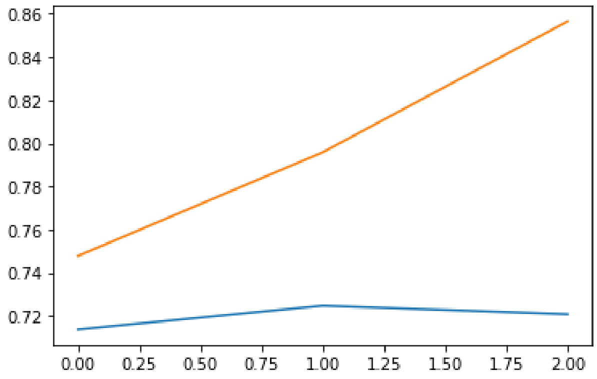
Model	F1	Accuracy	Recall	Precision
DistilBert - Attention-MLP(2 linear layer(400,200)), [ReLu] with dropout	0.66153	0.70817	0.59538	0.74423
DistilBert - Attention-MLP(2 linear layer(500,500)), [tanh] with dropout	0.62175	0.71039	0.49692	0.83033
DistilBert - Attention-MLP(2 linear layer(400,200)), [ReLu] without dropout	0.57545	0.68901	0.44000	0.83139
Final Model (without dropout in fully connected layers)	0.68959	0.72733	0.63230	0.75830
Final Model (with unidirectional LSTMs)	0.70349	0.73102	0.66615	0.74526
Final Model	0.71632	0.73618	0.69538	0.73856

Plots for the best model.
When all layers in the distilbert models are trainable.

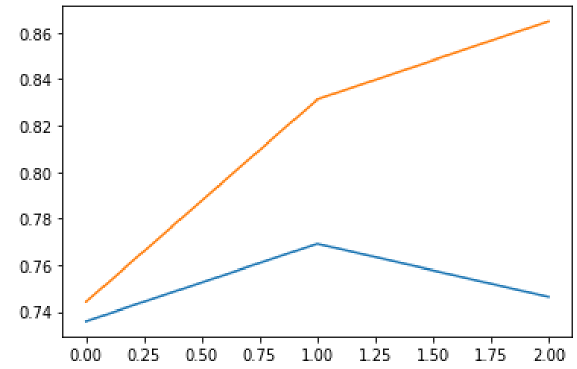
Loss plots



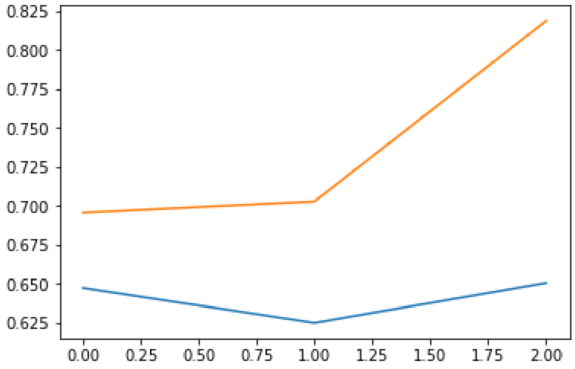
Accuracy plots.



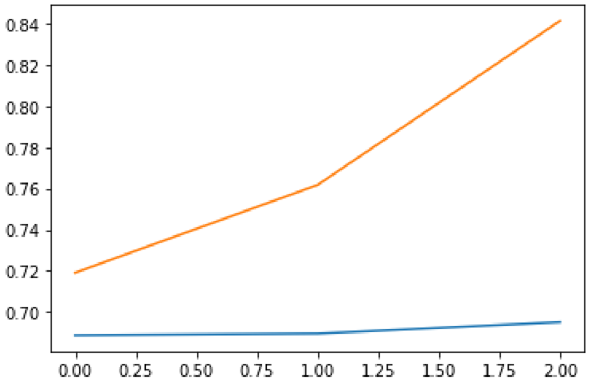
Precision plots



Recall plots



F1 Score plots for the training and the validation dataset.



3. All layers of Distilbert are frozen

Evaluation metrics for test data

Note - The number of epochs taken is 6 to allow the model to completely converge.

F1 score - 0.6816816816816818

Accuracy - 0.6875460574797347

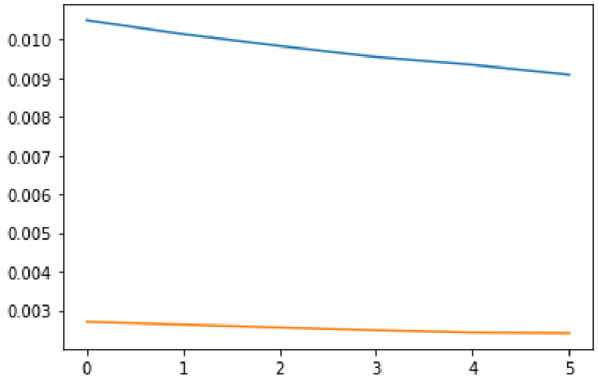
Recall - 0.6984615384615385

Precision - 0.6656891495601173

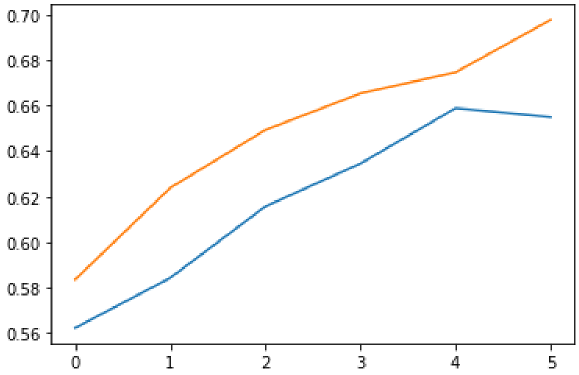
Loss - 0.009413266981359606

Plots.

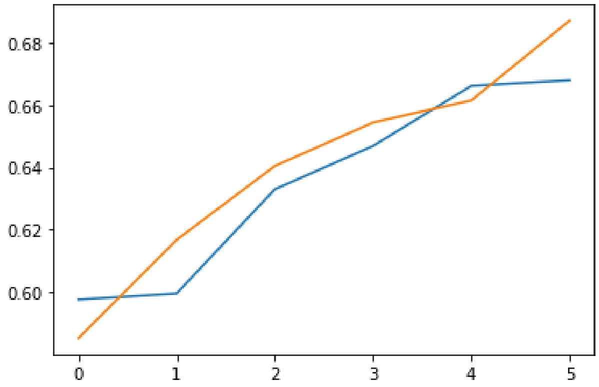
Loss



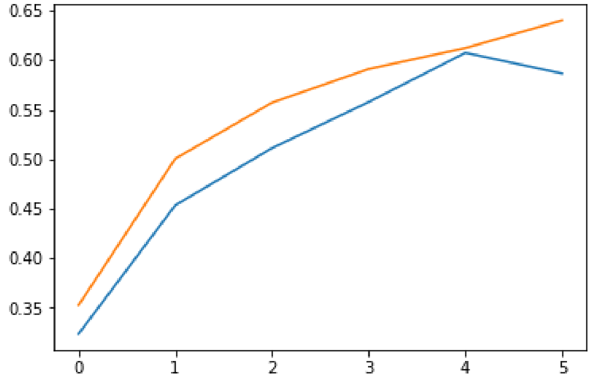
Accuracy



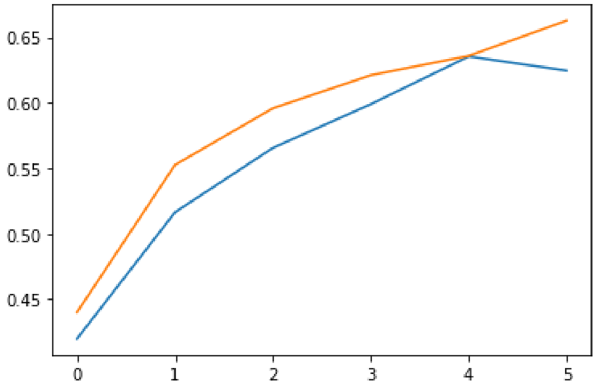
Precision



Recall



F1 Score



Example of two misclassified samples from the test dataset.

1. Def1: the use of it tools and methods to collect, process, consolidate, store, and secure data from sources that are often fragmented and inconsistent

Def2: facilities, processes, and procedures used to collect, store, and distribute information between producers and consumers of information in physical or electronic format.
Correct Class: 0 (Non paraphrased)

Explanation: a possible reason can be that there are many significant words that are common in both the definitions (processes, use, collect, store, and etc) which might be giving the model the indication that both are similar sentence and hence it classified them as ‘paraphrased’.

2. Def1: source code is compiled into this, and it is then processed by a computer
Def2: a code that is processed by a computer and is copied from a source code.
Correct Class: 1 (Paraphrased)

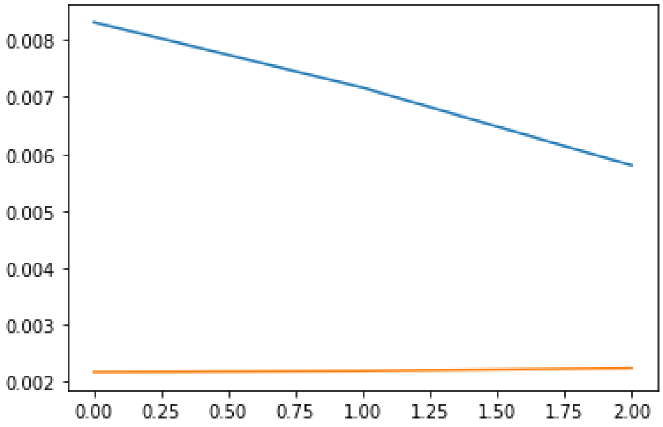
Explanation: the reason can be since the transformer is not getting trained, the resulting embedding is not as good as it can be and may miss out on the features representing similarity in both these definitions and which are necessary to further direct the remaining network that both these definitions are similar and are in reality a paraphrased version of each other. Along with this since both the sentence structures are almost opposite to each other (position of “source code”), LSTM is unable to interpret the meaning of this and handle it due to additional long-term dependency problems and hence may be giving the wrong output.

4. Some layers of Distilbert are frozen

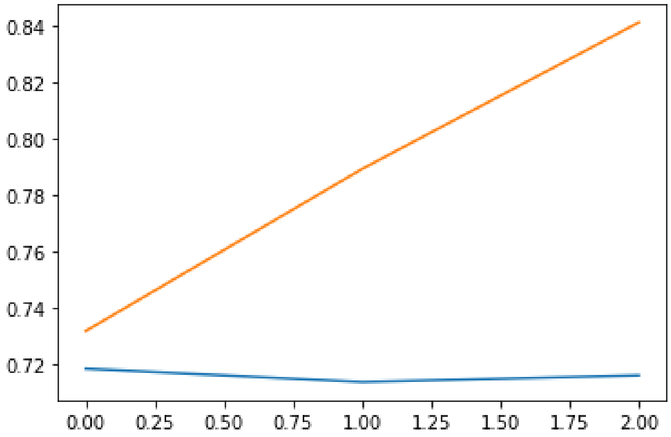
We have tried 3 different combinations for this part-
Model1 - Model where distilBert’s transformer layers 3, 4, and 5 are frozen.
Model2 - Model where distilBert’s transformer layers 3, and 5 are frozen.
Model3 - Model where distilBert’s transformer layers 2, 3, and 4 are frozen.

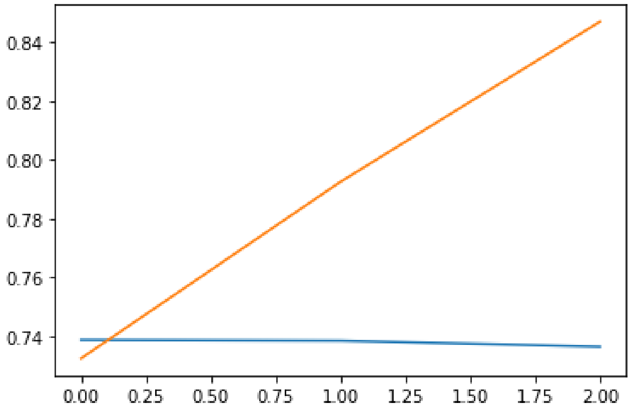
Model1 -

Loss plot



Accuracy plot

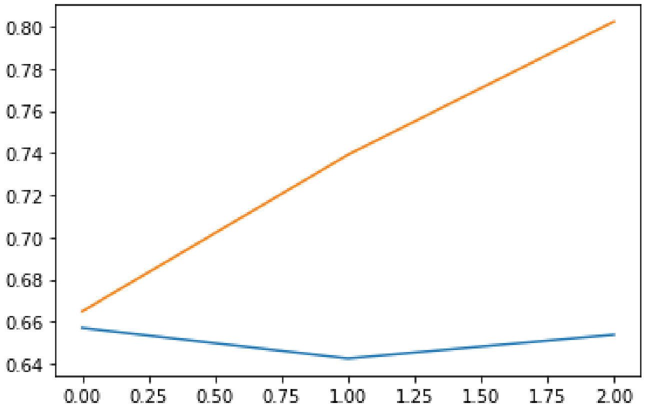


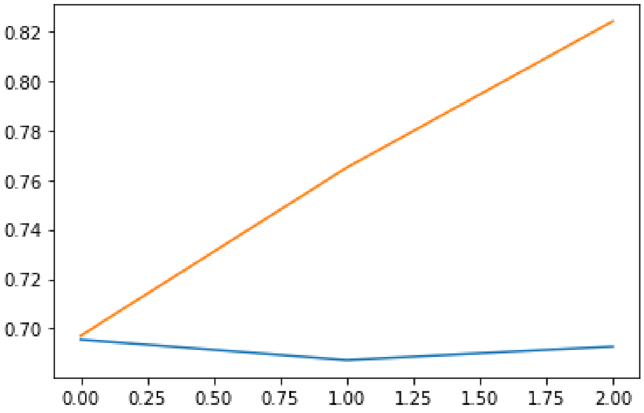


Recall plot

F1 plot

Precision plot

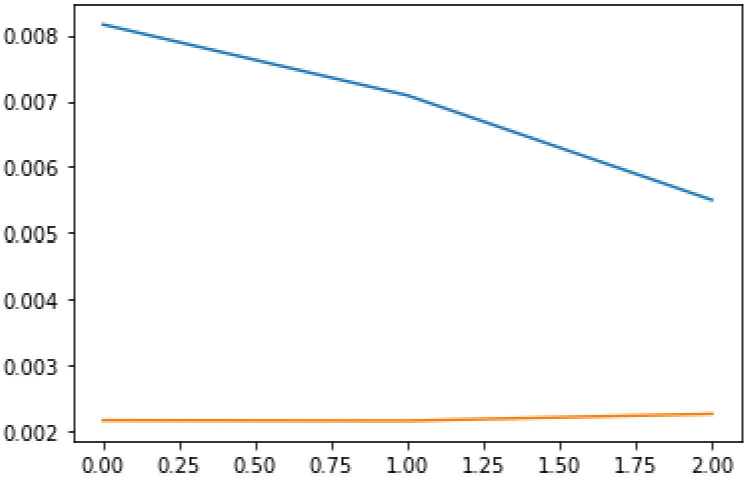


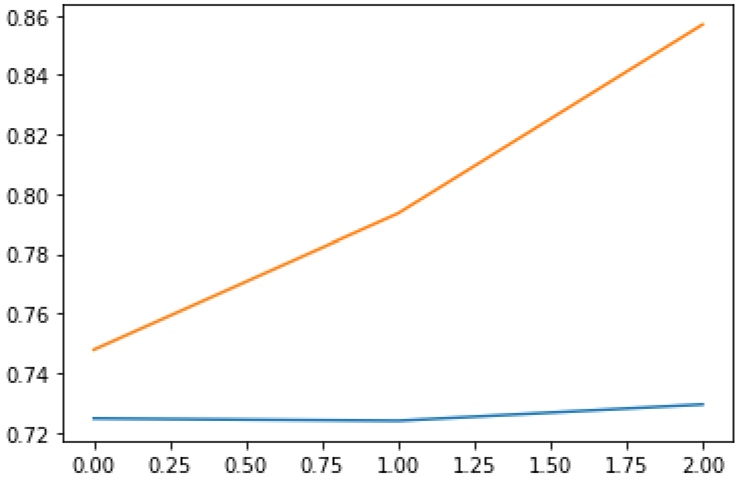


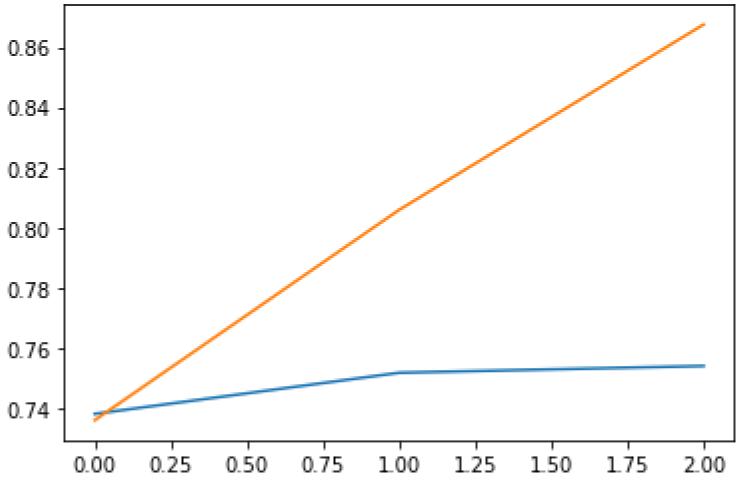
Model2 -

Loss plot

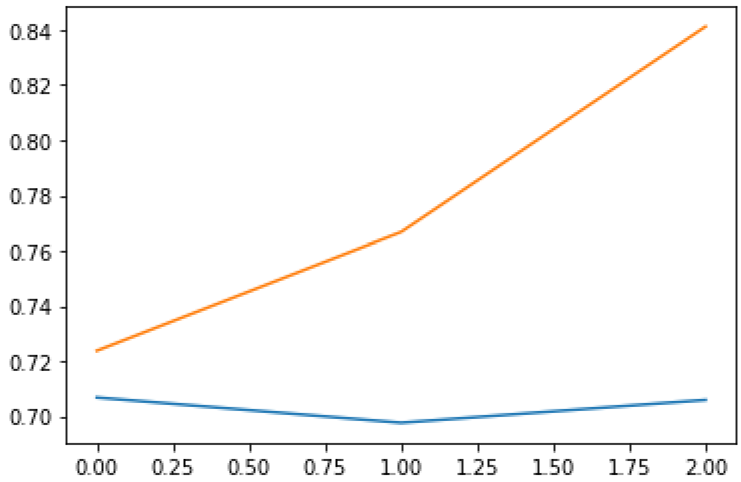
Accuracy plot

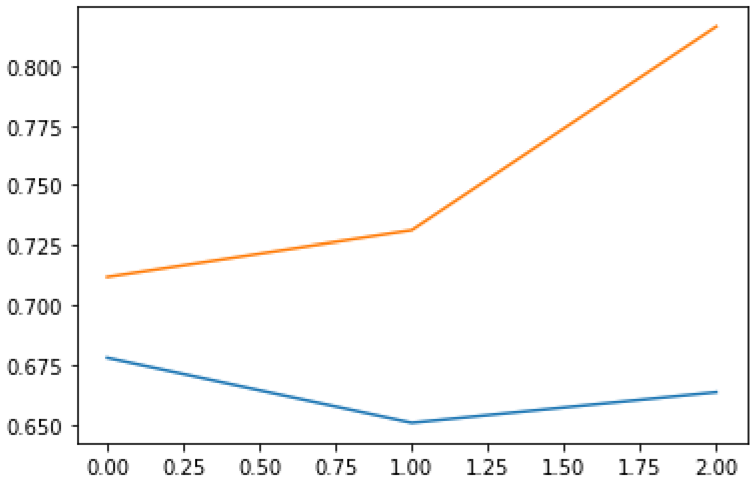






Recall plot Precision plot F1 plot

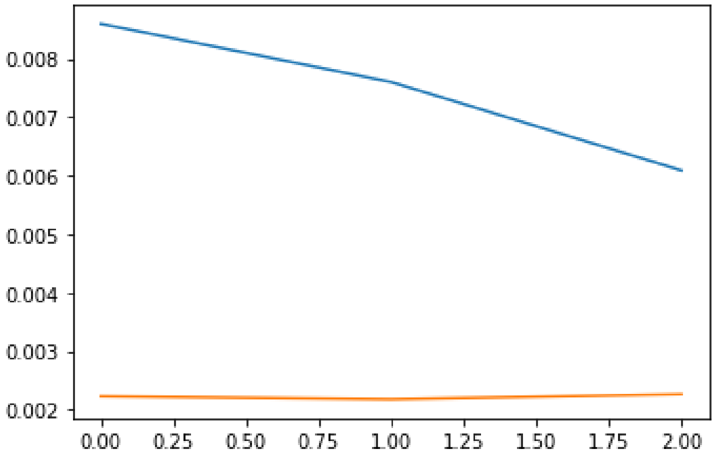


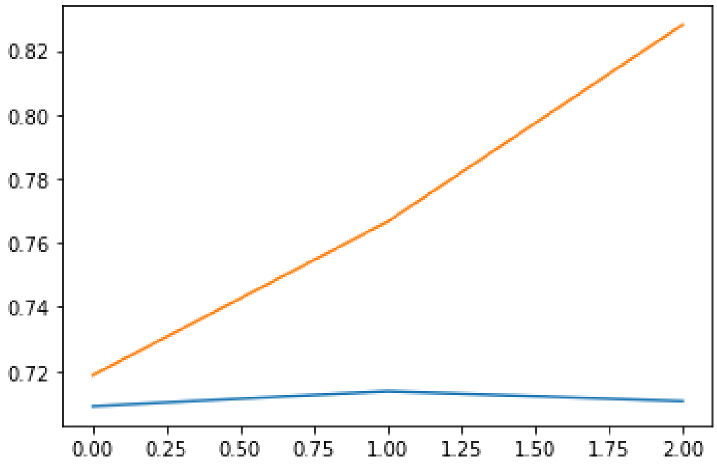


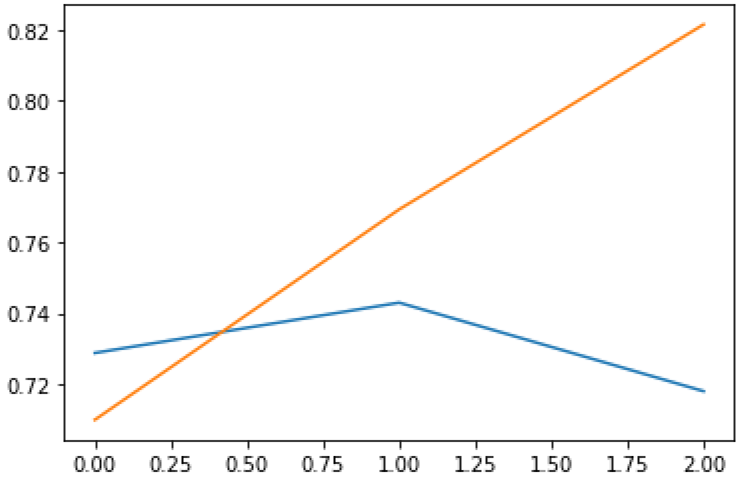
Model3 -

Loss plot

Accuracy plot



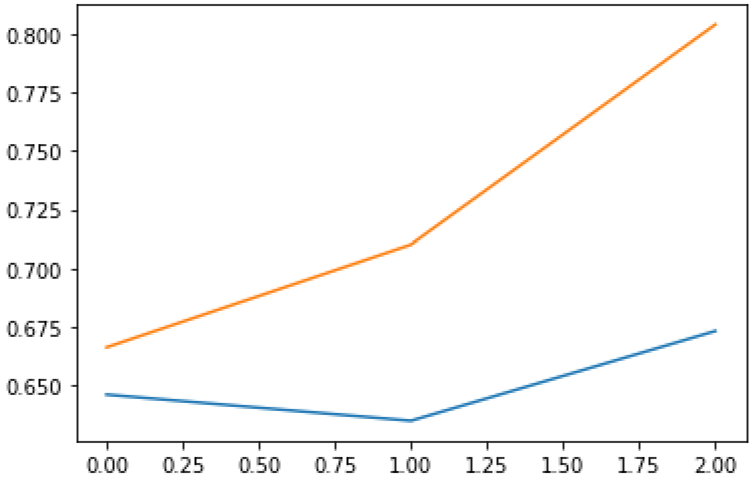


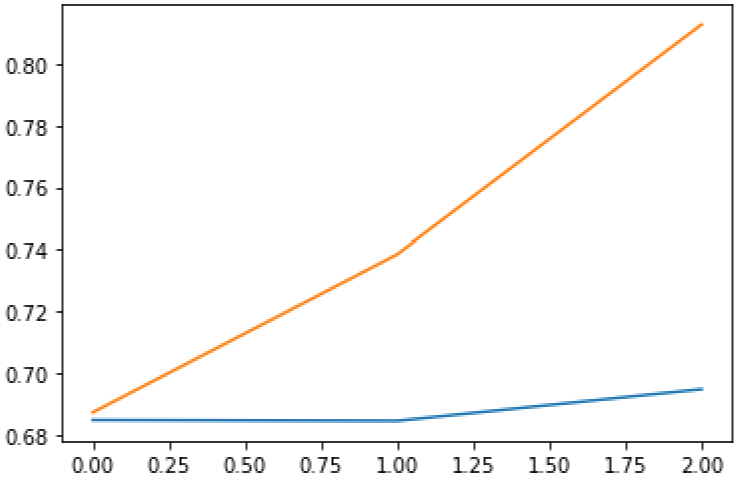


Recall plot

Precision plot

F1 plot





Results on test set-

Model	**Loss**	**Accuracy**	**F1**	**Precision**	**Recall**
1	0.00894	0.73618	0.71129	0.74745	0.67846
2	0.00927	0.73176	0.70926	0.73754	0.68307
3	0.00904	0.72586	0.69951	0.73639	0.66615

Misclassified samples with model (best)-

1. Def1: a computer generated simulation of an environment that can be interacted with using specialized equipment ; often immersive

   Def2: simulate a real situation on screen and allow the user to interact it in a near-natural way.

   Correct class: 1

   Explanation: The sentences have very less similar words and an understanding of the topic (Virtual Reality) in the embedding space would be required to make a correct prediction like "specialized equipment to the screen" or "environment to natural way".

2. Def1: network can be constructed and enforced through several logical controls to provide segregation and protection of an environment. technical control.

   Def2: network constructed to provide and enforce through logical controls.

   Correct class: 0

   Explanation: Since there are many significant words in definition-2 that repeats in definition-1 the network is probably getting the indication that both of the sentences are similar from their embeddings and hence classified them under the paraphrased class.

5. Comparison between 3 and 4

From the various test metrics calculated from the two models, we can see that the model in which some layers are trained outperformers the model where none of the layers are trained. We know that the *Parade* dataset contains sentences that are specific to certain topics and we can see from the misclassified samples that some semantic understanding in the embedding space that is relevant to the topic at hand is required to make predictions which is to be learned front he embedding itself and depends on its quality.

On a more detailed level, we can see that only the Recall is slightly higher for the completely frozen model and all other metrics are significantly worse. A higher recall can be attributed to a lower number of false negatives in the dataset which can be due to the skew in the dataset towards the class 0 and the model predicts prefers the same. But the flow of gradient in some layers of distilbert allows it to finetune its weights in accordance with the dataset, and better generalize to the unseen data.

On a more theoretical level, the only information provided to the lstm is the embeddings given by the transformer, which, if devoid of a certain understanding of that resembles the topic of sentences then the lstm will be unable to learn the meaning as well. As such, fine-tuning the distilbert model provides better predictions. Also, since LSTM's learning is only based on the embedding passed into it, and in the fully-frozen case, these embeddings are of very poor quality learning from them can't be much better or improved.

6. Variations in hyperparameter

a) Learning Rate

Learning Rate	Test Accuracy	Test F1 Score
1e-3 (Recall = 1)	0.47899778924097275	0.6477329347284505
1e-2 (Recall = 0)	0.5210022107590273	0.0
1e-1 (Recall = 0)	0.5210022107590273	0.0

Remarks.

The original learning for the best model is 3e-5 which is much lower than the three learning rates provided, which implies the need for a small learning rate. We can see that the lowest value of the three actually tries to learn the data but the value is oscillating. But for the others, the model outputs 0 for all the samples and is not able to learn anything due to the high learning rates.

b) Optimizer

Optimizer	Test Accuracy	Test F1 Score
AdamW	0.73618	0.71632
SGD	0.5217391304347826	0.40184331797235023
SGD with momentum = 0.9	0.5187914517317612	0.3786869647954329

Remarks

For this case, we tried both SGD and SGD with momentum but AdamW outperforms both significantly for all metrics. This implies that our model benefits from adaptive learning rates and weight decay. Since the model is deep and complex, the adaptive learning rates also assist the sparse weights. Since the data also shows an inclination to get stuck at a point when the learning rates are not properly managed, AdamW is our best choice.

Thank You