English Language Learning

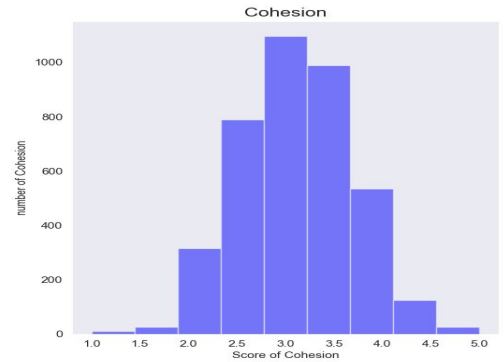
Team 8
Kyuri Kim
Sanchit Vijay

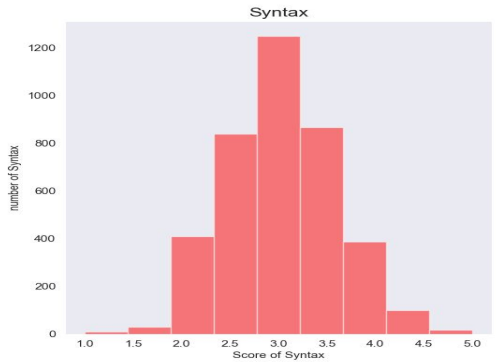
Contents

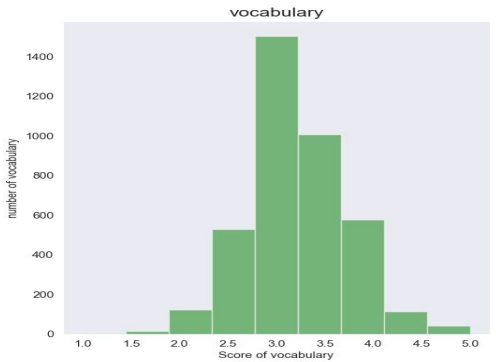
Data Information

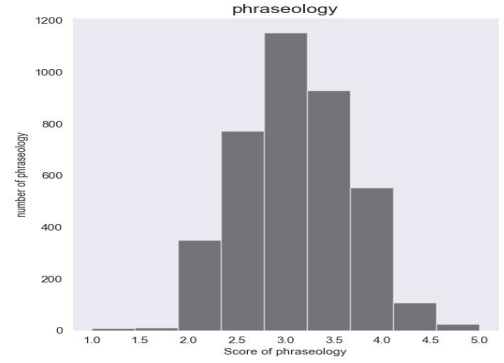
- Collection of argumentative essays crafted by English Language Learners (ELLs) from 8th to 12th grade
- Each essay is scored across six analytical dimensions:
  - Cohesion
  - Syntax
  - Vocabulary
  - Phraseology
  - Grammar
  - Conventions
- Each of these analytical dimensions is assigned a score ranging from 1.0 to 5.0, in 0.5 increments

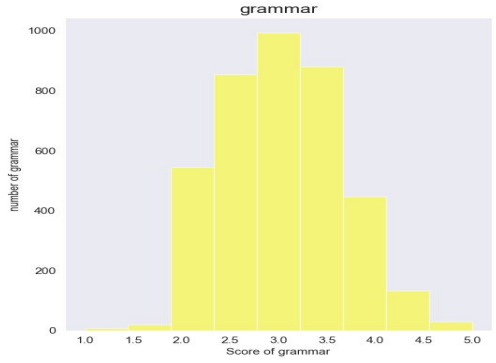
Histogram of Scoring Dimensions

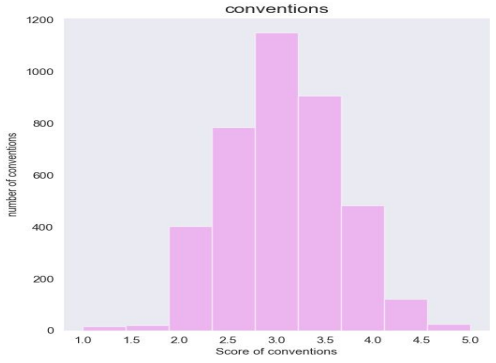








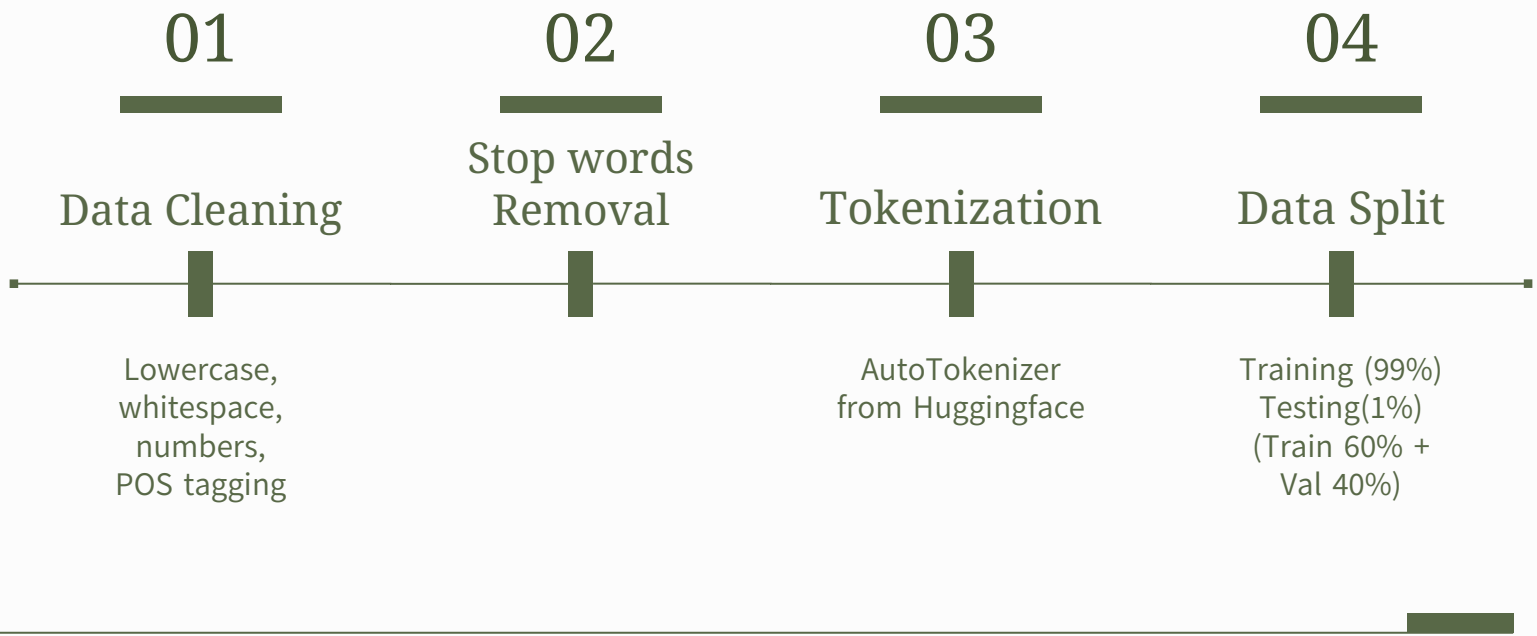




Heatmap


cohesion
syntax
vocabulary
phraseology
grammar
conventions
1
0.7
0.67
0.69
0.64
0.67
0.7
1
0.68
0.73
0.71
0.7
0.67
0.68
1
0.74
0.65
0.66
0.69
0.73
0.74
1
0.72
0.67
0.64
0.71
0.65
0.72
1
0.67
0.67
0.7
0.66
0.67
0.67
1
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65

Data Preprocessing



Modeling - Logistic Regression

Target	Accuracy
cohesion	0.63
syntax	0.62
vocabulary	0.67
phraseology	0.62
grammer	0.60
conventions	0.62

Modeling and Pooling

01

BERT

base - uncased

02

RoBERTa

large

03

ELECTRA

base - discriminator

04

DeBERTa

V3 - base / V3 - large

Modeling and Pooling

01

Mean Pooling

02

LSTM Pooling

03

Concat Pooling

04

Conv1D Pooling

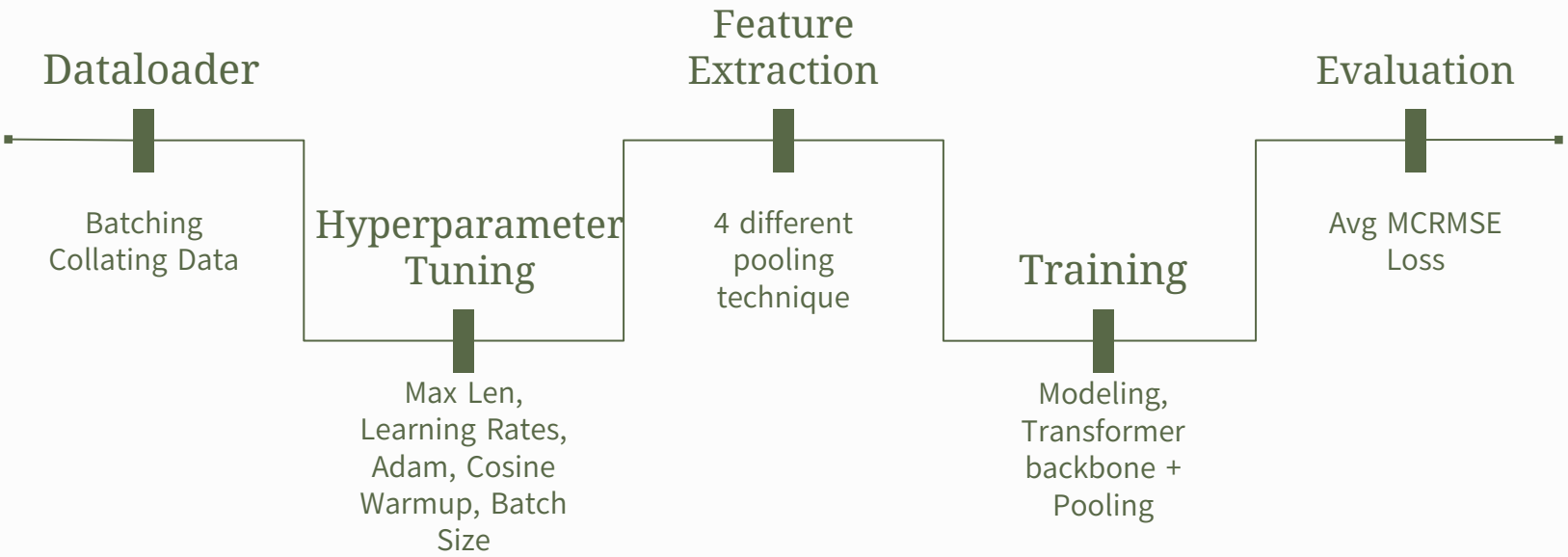
Hyperparameters

	Training batch size	Validation batch size	Max_Length
Bert-base-uncased	32	64	512
Electra-base-discriminator	32	64	512
Roberta-large	12	32	512
Deberta-v3-base	12	32	768
Deberta-v3-large	2	8	1024

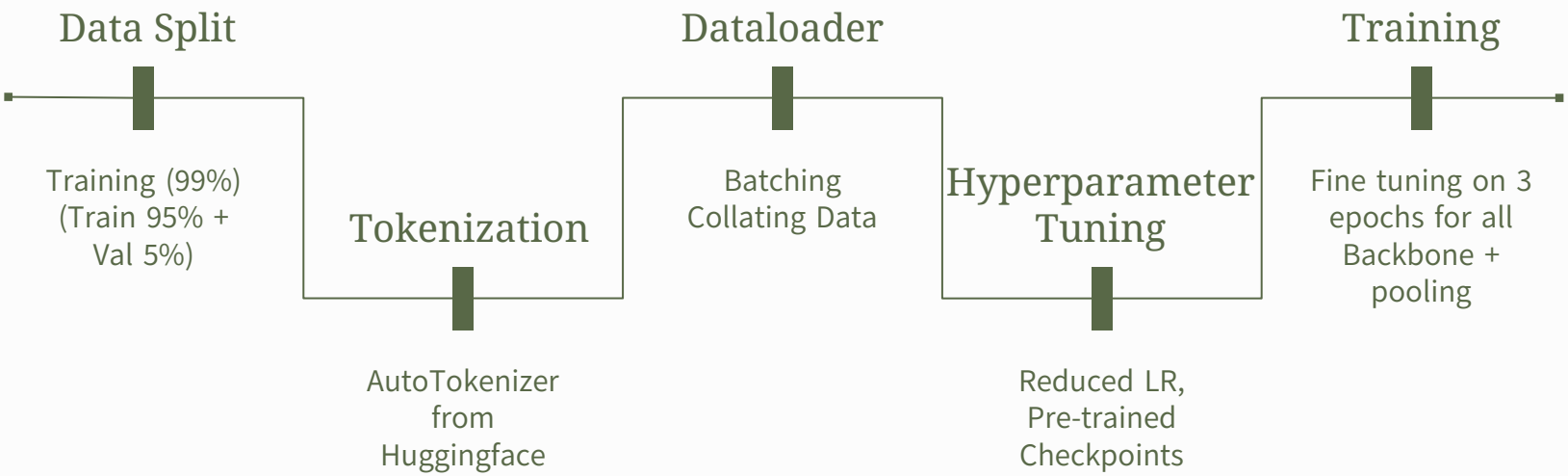
Learning Rates

Layer	Pre-Training	Fine Tuning
Encoder LR	2e-5	2e-6
Embedding LR	7e-6	1.5e-6
Decoder/Pooler LR	3e-5	9e-6
Weight Decay	1e-8	1e-2

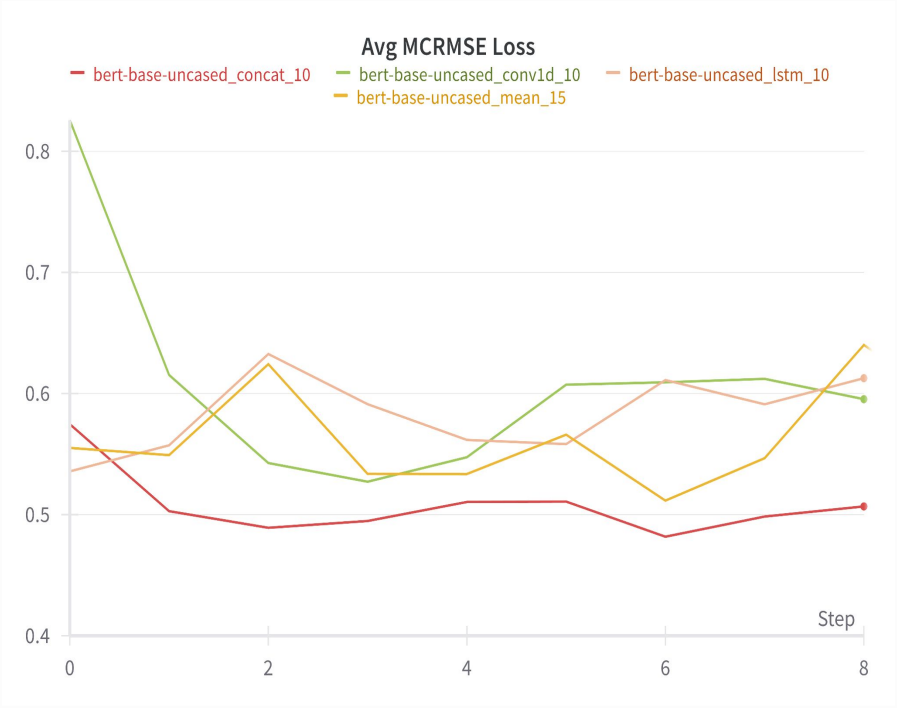
Pre-training

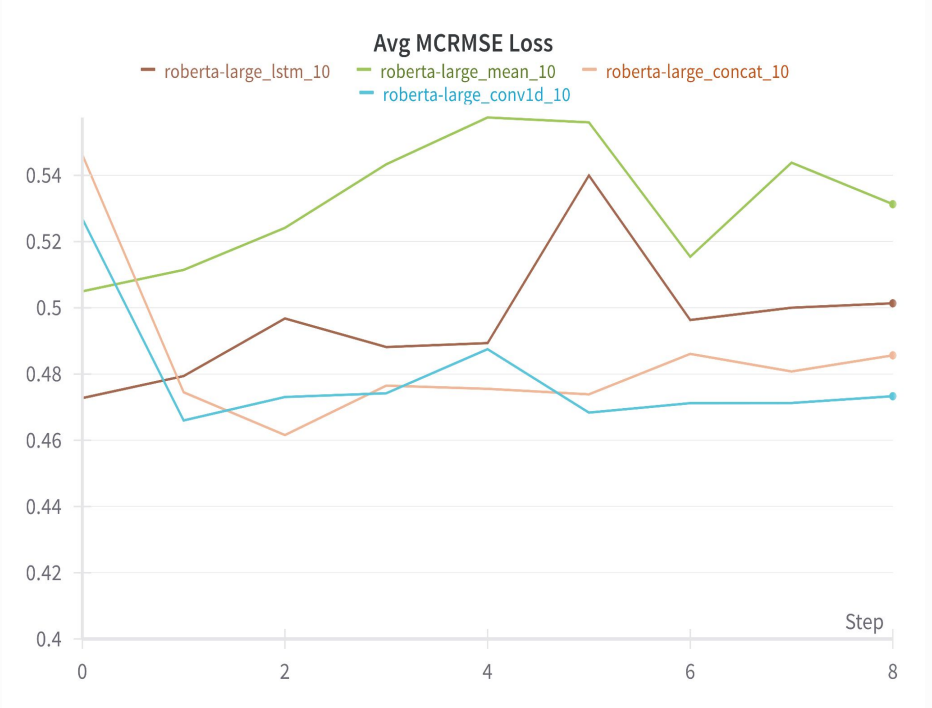


Fine Tuning


Data Split
Training (99%) (Train 95% + Val 5%)
Tokenization
AutoTokenizer from Huggingface
Dataloader
Batching Collating Data
Hyperparameter Tuning
Reduced LR, Pre-trained Checkpoints
Training
Fine tuning on 3 epochs for all Backbone + pooling

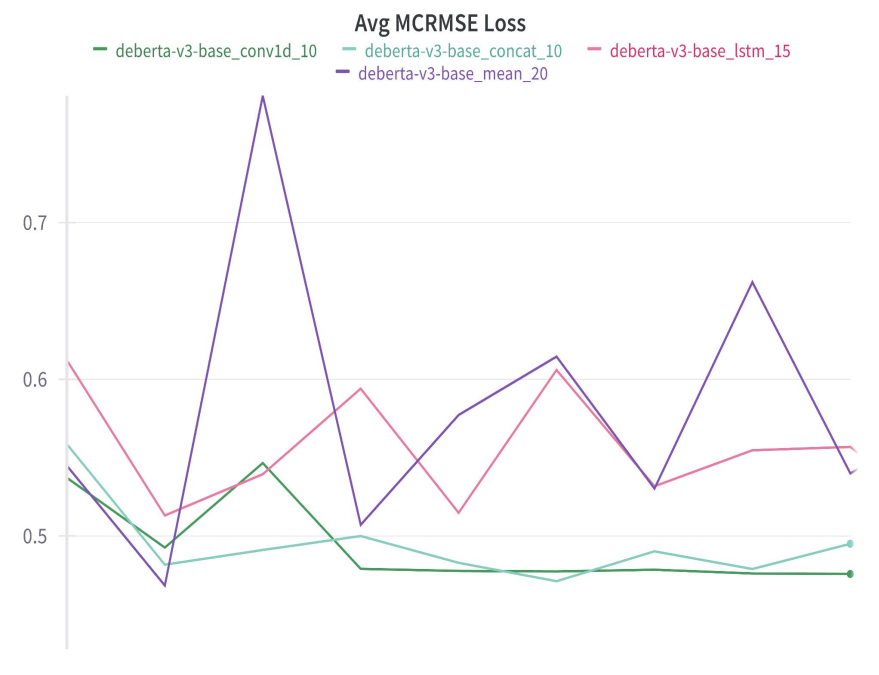
Comparing a Model with different Poolings

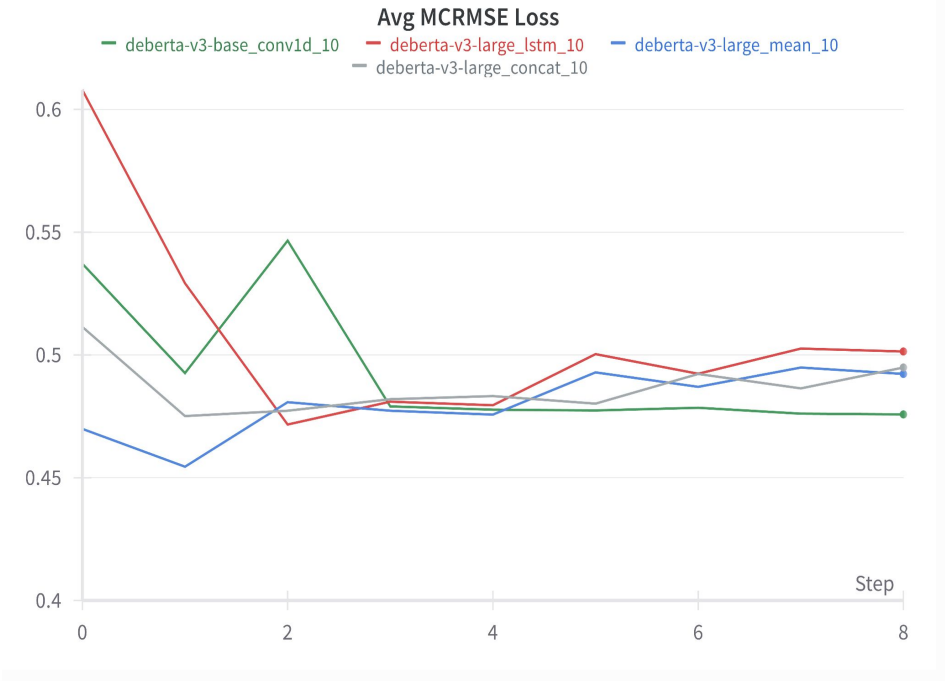




- Bert-base-uncased
- RoBERTa - large

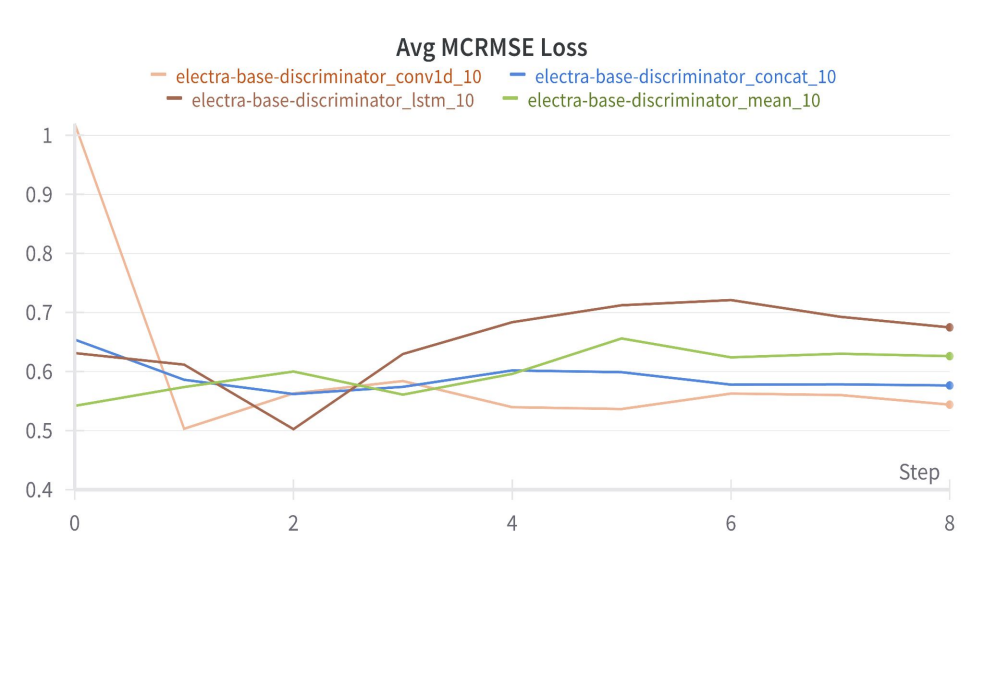
Comparing a Model with different Poolings





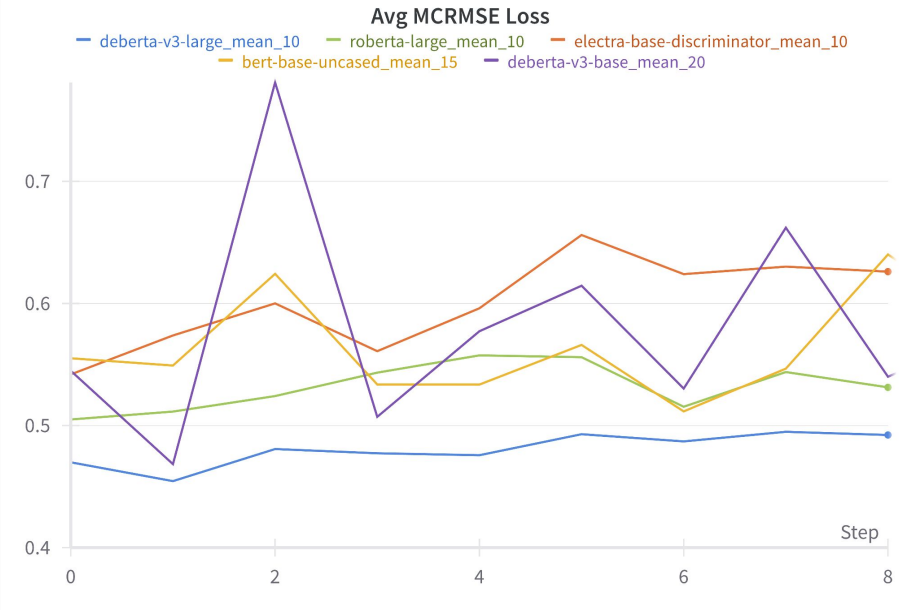
- DeBERTa - v3 - base
- DeBERTa - v3 - large

Comparing a Model with different Poolings

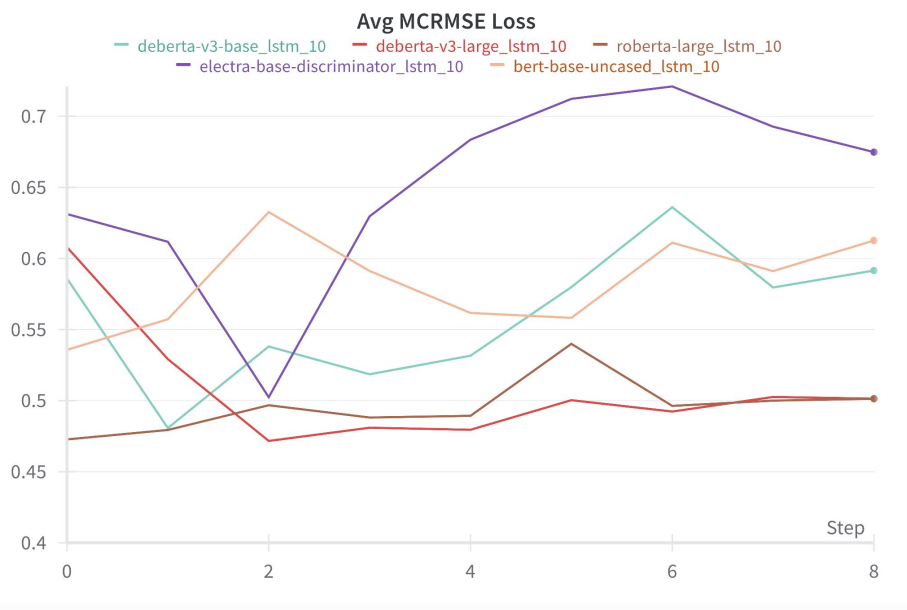


- Electra - base - discriminator

Comparing a Pooling with different Models

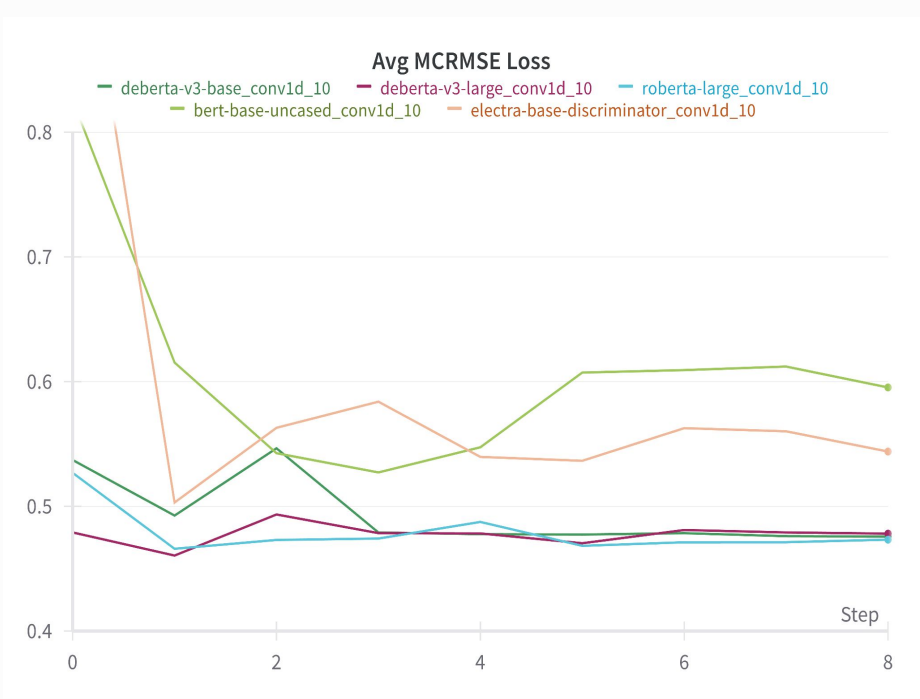


- Mean Pooling

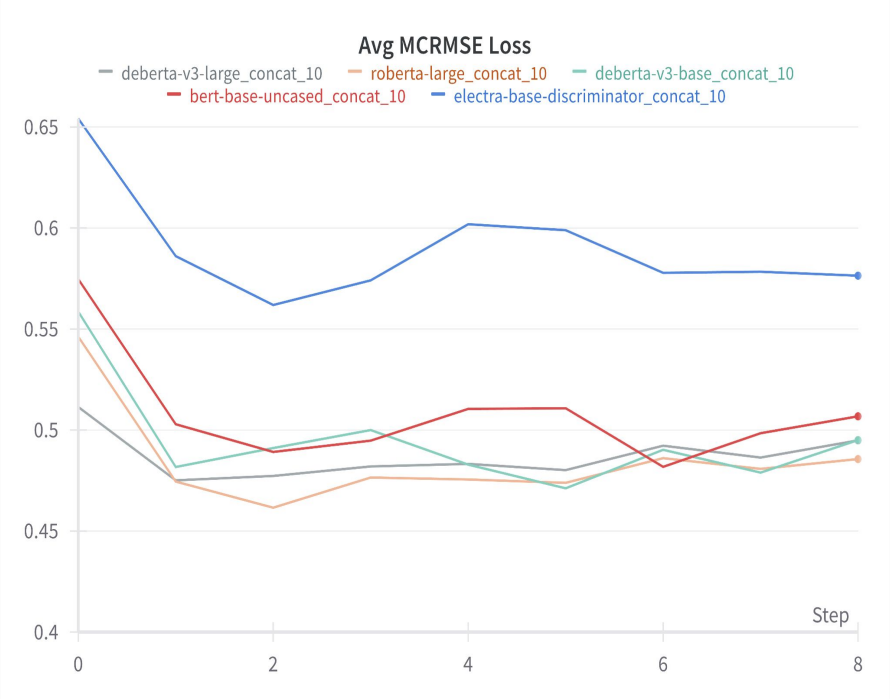


- LSTM Pooling

Comparing a Pooling with different Models



- Conv1D Pooling



- Concat Pooling

Results - Pretraining

	Mean	LSTM	Concat	Conv1D
Bert-base-uncased	0.5116	0.5357	0.4818	0.5272
Electra-base-discriminator	0.5419	0.5024	0.5619	0.5031
Roberta-large	0.5050	0.4727	0.4616	0.4660
Deberta-v3-base	0.4684	0.4807	0.4712	0.4758
Deberta-v3-large	0.4545	0.4717	0.4751	0.4606

Results - Fine tuning

	Mean	LSTM	Concat	Conv1D
Roberta-large	0.4815	0.4928	0.4130	0.494
Deberta-v3-base	0.4492	0.5125	0.4644	0.4340
Deberta-v3-large	0.3986	0.4221	0.4358	0.4131

Let’s explore our App!

Conclusion

- Implemented baseline logistic regression model and developed NLP techniques (BERT, RoBERTa, DeBERTa, ELECTRA) to enhance language proficiency assessment for ELLs
- Despite challenges, utilized diverse backbones, multiple pooling methods, and differential learning rates for accurate language proficiency prediction in ELL essays
- For future improvements, a multi-pronged approach can enhance model performance and robustness like pseudo-labeling during pretraining.
- Additionally, experimenting with a broader range of pooling methods could uncover more effective strategies for data representation, particularly in complex models.

Thank you!