Improve Your English Essay with AI

W266 NLP with DL

University of California, Berkeley

April 18, 2023

Iris Lew

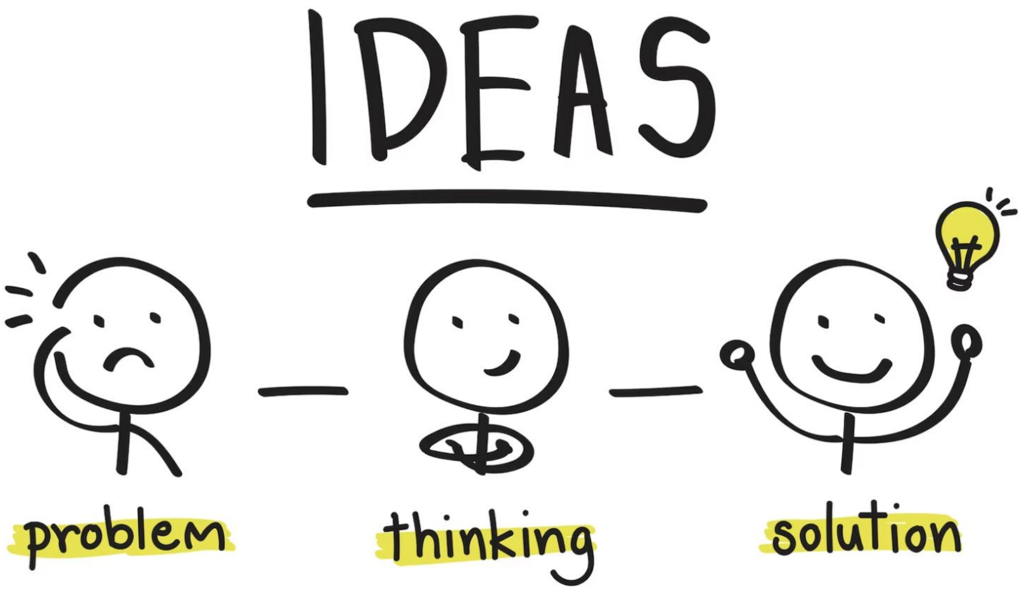
Heesuk Jang

Srila Maiti

Agenda

- Motivation
- Dataset
- Approaches
- Results
- Conclusion & Future Work

Motivation

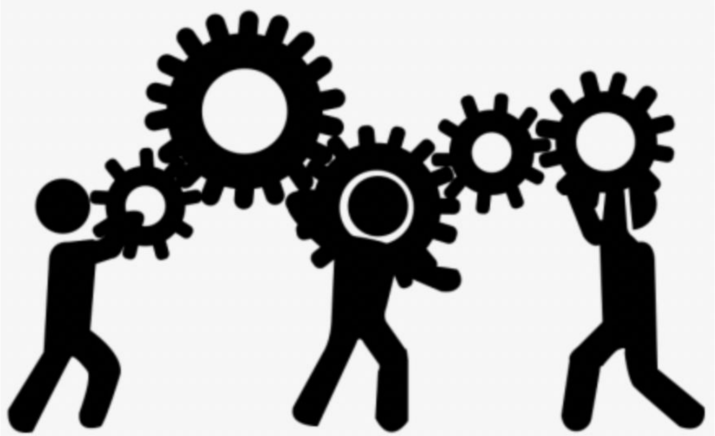


Dataset

The dataset presented here (the **ELLIPSE corpus**) comprises argumentative **essays written by 8th-12th grade English Language Learners (ELLs)**. We have 3911 essays which have been scored according to six analytic measures: **cohesion, syntax, vocabulary, phraseology, grammar, and conventions**.

Each measure represents a component of proficiency in essay writing, with greater scores corresponding to greater proficiency in that measure. The scores **range from 1.0 to 5.0 in increments of 0.5**.

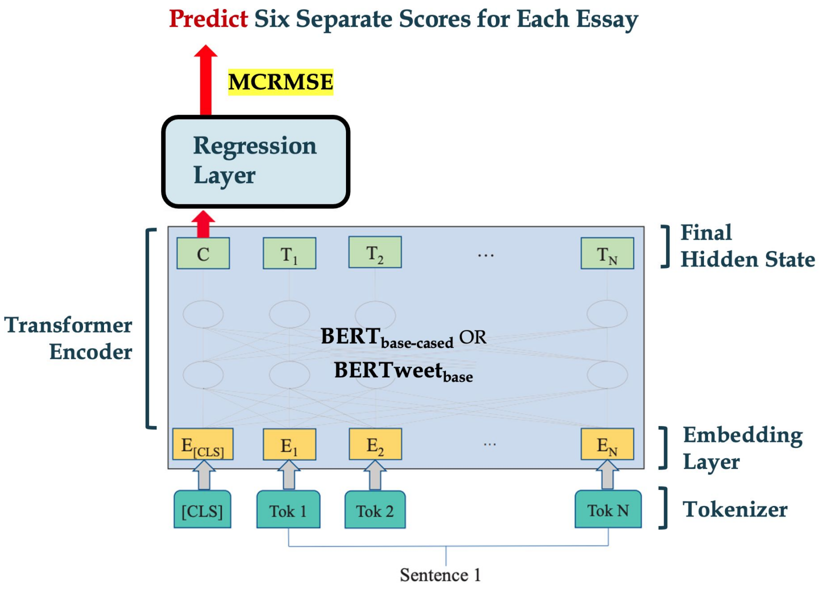
Approaches



- Transformer based language models (BERT base-cased and BERTweet base)
- Number of frozen and unfrozen layers
- Clustering
- Stratified Sampling

Model - Architecture & Token Length

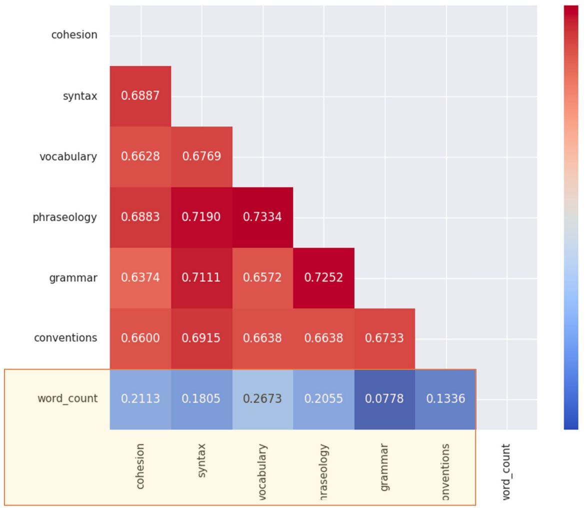
Model Architecture



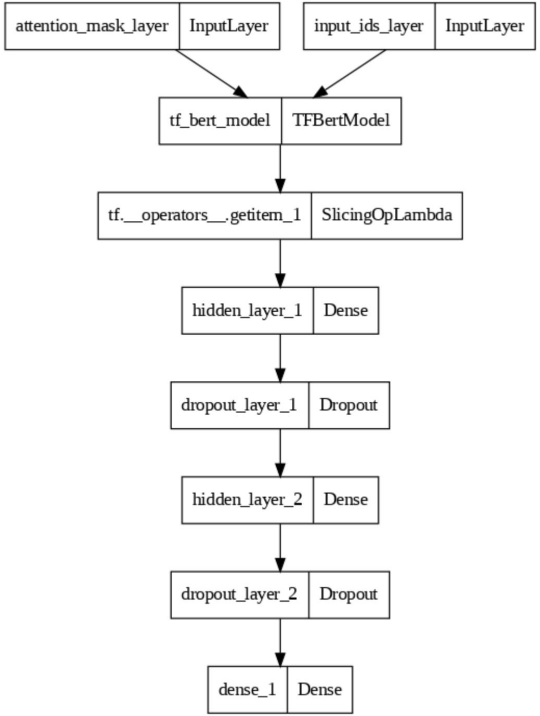
Input Token Length at Max

$\text{BERT}_{\text{base-cased}}$ 512

$\text{BERTweet}_{\text{base}}$ 128



Model - Best Model



Best Set of Parameters from Training BERT$_{\text{base-cased}}$

Test MCRMSE	Trainable Layers	Learning Rate	Hidden Layers	Hidden Units	Batch Size	Dropout	Epochs
0.4590	12	0.00001	2	64	8	0.1	10

Hyper Parameter Tuning

01 Learning Rate
[5e-4, 1e-4 , 5e-5, 1e-5]

02 Hidden Layer
[1, 2]

03 Hidden Units
[64, 128, 256]

04 Dropout Rate
[0.1, 0.2, 0.3]

05 Batch Size
[8, 16]

06 Trainable Layers
[min=0, max=12, step=2]

Results - BERT$_{\text{base-cased}}$ VS. BERTweet$_{\text{base}}$

Adjusted MCRMSE Scores

	BERT$_{\text{base-cased}}$	BERTweet$_{\text{base}}$
0 trainable layers	0.6350 Baseline	0.6549
6 trainable layers	0.6271	0.6224
12 trainable layers	**0.5254**	**0.5536**

BERT$_{\text{base-cased}}$

% of Test Dataset Records that were ***Correctly*** Predicted Per Analytic Measure (Score within 0.5)

	0 trainable layers	6 trainable layers	12 trainable layers
Cohesion	29.8% (75.9%) Baseline	30.0% (75.7%)	**33.8% (83.7%)**
Syntax	34.1% (76.8%) Baseline	33.2% (77.4%)	**37.8% (84.7%)**
Vocabulary	37.5% (82.1%) Baseline	39.0% (83.3%)	**44.2% (89.8%)**
Phraseology	31.3% (78.2%) Baseline	32.7% (79.6%)	**44.1%(88.8%)**
Grammar	27.7% (74.2%) Baseline	27.8% (72.0%)	**33.5% (81.2%)**
Conventions	31.7% (72.8%) Baseline	29.9% (74.5%)	**38.1% (87.5%)**

BERTweet$_{\text{base}}$

% of Test Dataset Records that were ***Correctly*** Predicted Per Analytic Measure (Score within 0.5)

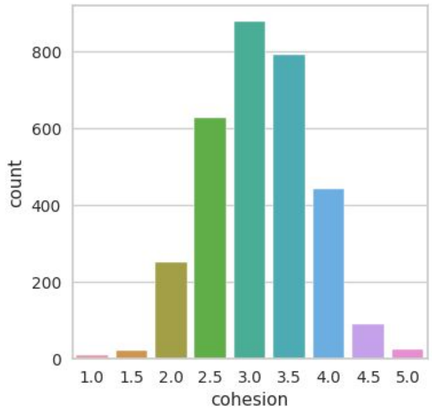
	0 trainable layers	6 trainable layers	12 trainable layers
Cohesion	27.3% (72.7%)	29.4% (75.6%)	**35.6% (83.1%)**
Syntax	33.5% (73.4%)	33.1% (75.1%)	**35.6% (84.0%)**
Vocabulary	33.8% (80.2%)	36.5% (85.4%)	**39.2% (86.7%)**
Phraseology	29.8% (76.5%)	30.4% (79.2%)	**35.4% (83.8%)**
Grammar	25.0% (70.5%)	27.5% (74.3%)	**36.5% (80.3%)**
Conventions	30.9% (74.6%)	29.4% (75.0%)	**35.6% (87.0%)**

Results - BERTbase-cased VS. BERTweetbase

Clustering & Stratified Two-Fold Cross Validation

Adjusted MCRMSE Scores

	$\text{BERT}_{\text{base-cased}}$	$\text{BERTweet}_{\text{base}}$
0 trainable layers	0.6763	0.6907
6 trainable layers	0.6681	0.6688
12 trainable layers	0.6798	0.6652



(6.0-17.0) (17.5-21.5) (21.5-30.0)

	$\text{BERT}_{\text{base-cased}}$ — % of Test Dataset Records that were ***Correctly*** Predicted Per Analytic Measure (Score within 0.5)			$\text{BERTweet}_{\text{base}}$ — % of Test Dataset Records that were ***Correctly*** Predicted Per Analytic Measure (Score within 0.5)		
	0 trainable layers	6 trainable layers	12 trainable layers	0 trainable layers	6 trainable layers	12 trainable layers
Cohesion	28.6% (71.9%)	28.2% (73.9%)	28.4% (73.8%)	28.0% (70.0%)	26.8% (72.7%)	25.4% (72.0%)
Syntax	32.7% (73.9%)	30.0% (73.4%)	30.5% (72.7%)	30.7% (72.4%)	33.1% (72.9%)	32.2% (73.9%)
Vocabulary	36.9% (78.5%)	37.2% (79.2%)	28.4% (80.3%)	31.2% (77.7%)	29.6% (79.6%)	29.6% (81.2%)
Phraseology	30.5% (73.4%)	30.5% (73.2%)	29.0% (73.1%)	27.2% (71.1%)	30.5% (73.3%)	29.6% (73.2%)
Grammar	27.8% (70.0%)	27.3% (71.3%)	24.5% (70.2%)	24.4% (70.8%)	25.8% (70.5%)	24.8% (71.1%)
Conventions	30.8% (71.8%)	30.7% (76.2%)	28.6% (70.4%)	29.1% (71.6%)	30.9% (73.9%)	30.7% (72.4%)

Conclusion

- With more unfrozen layers, the models were able to learn the training data
- Models struggled to predict extreme scores.
- Clustering through K-Means and K-fold cross validation to account for the lower and higher ends of the scores did not improve model performance

Future Work

01 Increase size of the input dataset

02 Introduce larger versions of BERT and BERT-derived models

03 Evaluate essays together with key topical information

04 Explore pre-transformer-based state-of-the-art models as well as models with BERT-based combinations

References

1. https://www.kaggle.com/competitions/feedback-prize-english-language-learning/overview
2. https://www.freepik.com/free-vector/illustration-light-bulb-ideas_3139696.htm#query=motivation%20icon&position=22&from_view=keyword&track=ais
3. https://fontawesome.com/
4. https://thenounproject.com/browse/icons/term/dataset/
5. https://www.kindpng.com/imgv/TbiwiRx_team-work-png-transparent-png/
6.


?


QUESTIONS
ANSWERS