Improve Your English Essay with Artificial Intelligence

Iris Lew
iris.lew@ischool.berkeley.edu

Srila Maiti
srila.maiti@ischool.berkeley.edu

Heesuk Jang
jheesuk@ischool.berkeley.edu

UC Berkeley School of Information
W266 - Natural Language Processing with Deep Learning

April 18, 2023

Abstract

Automated essay scoring (AES), the task of employing natural language processing (NLP) technology to automatically assign scores to essays at scale, can be an important tool to help improve people's writing skills as well as reducing the heavy burden from grading activities. Recently with advances in AI, there have been attempts made by experiments with varying degrees of success at replicating human scores, often tackling AES only holistically rather than at the rubric level. Building on previous studies where models were trained to predict both the holistic and rubric scores of essays, this study focuses on predicting the quality of argumentative essays written by 8th - 12th grade English Language Learners (ELLs) at the rubric level by exploring how $\text{BERT}_{\text{base-cased}}$ and $\text{BERTweet}_{\text{base}}$ can be fine-tuned on the AES task. In particular, models that do not contain pre-training performed on par or better than pre-trained transformer models, leading us to investigate whether pre–trained models that rely less on pre-training and models that have been pre-trained on a more informal data source (Twitter) would serve as better predictors of ELLs' essay scores. To develop the AES task and examine the hypothesis, we measures MCRMSE (Mean Column-wise Root Mean Squared Error) as regression loss as well as accuracy with fine-tuning the weights by freezing all or unfreezing the last 6 or all 12 layers of those two base versions of BERT in addition to tuning key hyper parameters such as learning rate and hidden size. In this experiment, we find that unfreezing all the layers and allowing the $\text{BERT}_{\text{base-cased}}$ and $\text{BERTweet}_{\text{base}}$ models to rely more on the pretraining enabled for better performance contrary to our hypothesis, as it allowed for more extreme scores (i.e. 1 or 5 from scores ranged from 1 to 5 with 0.5 increment) to be predicted. To improve the generalizability of the extreme scores, we then clustered the data into low, average, and high scorers as well as performed k-fold cross validation to feed it into our models. Yet, the MCRMSE did not show much improvement on either model type. Additionally, $\text{BERTweet}_{\text{base}}$ did not perform better than $\text{BERT}_{\text{base-cased}}$, which implies that ELLs' essays belong to a different population than Tweets.

1 Introduction

Writing is complex, and grading all students fairly is essential. While it's controversial to use an AES system systematically, there has been encouragement to use it as a tool to help improve writing through automated feedback.[1] Especially with the rise of massive open online courses, valid and reliable AES tools are vital to examine the quality of learning from a large pool of students while helping alleviate educators' workload. This can be particularly useful to English Language Learners (ELLs) in predominantly English speaking schools where they are judged with native English speaking peers or if they are preparing for a test like the TOEFL.[2] Clear evidence in improved educational results with English-language

[1] https://ebookcentral-proquest-com.libproxy.berkeley.edu/lib/berkeley-ebooks/detail.action?docID=1172902

[2] https://ebookcentral-proquest-com.libproxy.berkeley.edu/lib/berkeley-ebooks/detail.action?docID=1172902

proficiency has been witnessed [3] thus, we are particularly interested in developing AES in the use of scoring ELLs to help them improve their skills in English literacy.

AES tasks have been applied to multiple datasets and with different architectures. Using a supervised learning paradigm, the two more predominant architectures are Long Short Term Memory models (LSTMs) and Bidirectional Encoder from Transformer (BERT) [4] , with LSTMs generally performing on par or outperforming BERT (scores for LSTM models ranged from 72.65 percent to 83 percent whereas scores for BERT models ranged from 64.6 percent to 78.2 percent on the same dataset). [5] Although pre-training is often examined to surpass model robustness and uncertainty estimates over training a model without pre-trained weights [6] , the studies lead us to suspect that the pre-training does not provide performance benefit on the AES tasks and architectures over training from scratch like LSTMs. Furthermore, if the data that the BERT model is pre-trained on is more informal, like Tweets from Twitter, and thus could potentially be more similar to student writing, or the model is less reliant on its pre-trained data, then the pre-trained BERT models could display an increase in accuracy when scoring ELLs' essays.

2 Background

Prior AES supervised learning studies have used both LSTMs and BERT-base models on the ASAP (Automated Student Assessment Prize) dataset, which is a set of essays that are written by students from Grade 7-10 in English.[7] The LSTMs model and BERT models both generally perform between 70-85%, with LSTMs performing on par or outperforming BERT models which forgets significant contextual information that impact the scoring[8] and implies that their BERT model was unable to learn the context. Because we expect BERT to perform better than bidirectional LSTMs, we hypothesize the pre-training from the BERT models did not suit the dataset which is composed of ELLs' writing styles. Thus, it is possible that by overriding the learned weights on pre-trained models, using a model trained on a dataset with more varied writing styles, or even doing more fine-tuning to rely less on the training would improve performance.

Additionally, even though many features within a text tend to correlate with one another,[9] it is possible for someone to produce an essay with a low grammar score but a high cohesion score, and therefore it makes sense to use separate models to evaluate different features within an essay and we find few studies that analyze an essays' individual components rather than assigning one holistic score. AES has already been used to assess coherence and writing skills while accounting for spelling mistakes.[10]

Furthermore, ELLs will produce different essays compared to native speakers of the language. ELLs have an additional hurdle to surmount in that their English writing ability can vary greatly and they are sometimes paying more attention to language rather than content.[11] We propose that BERTweet[12] , which is pre-trained on Tweets where people do not have to follow any grammatical rules as long as they stick within the 280 character limit when posting[13], would perhaps perform better since there is a greater sentence variety than in the English found in books and Wikipedia.

We decided to use the pre-trained transformers to improve the representations and then score these improved representations using linear regression as we wanted to preserve the ordinal structure of the score (a 5.0 is a better written essay than a 4.5). If our models were more successful at predicting the scores, then we find ELLs' writing belong to a different category and thus should be assessed with different models.

[3] https://learningenglish.voanews.com/a/number-of-english-learners-in-us-schools-keeps-rising/4635659.html

[4] https://arxiv.org/pdf/1810.04805.pdf

[5] https://thesai.org/Downloads/Volume12No10/Paper_28-Automatic_Essay_Scoring.pdf

[6] https://arxiv.org/pdf/1901.09960.pdf

[7] https://www.kaggle.com/c/asap-aes

[8] https://thesai.org/Downloads/Volume12No10/Paper_28-Automatic_Essay_Scoring.pdf

[9] https://ebookcentral-proquest-com.libproxy.berkeley.edu/lib/berkeley-ebooks/detail.action?docID=1172902

[10] https://thesai.org/Downloads/Volume12No10/Paper_28-Automatic_Essay_Scoring.pdf

[11] https://ebookcentral-proquest-com.libproxy.berkeley.edu/lib/berkeley-ebooks/detail.action?docID=1172902

[12] https://aclanthology.org/2020.emnlp-demos.2.pdf

[13] https://developer.twitter.com/en/docs/counting-characters

3 Methods

The dataset of ELLs' essays comes from a Kaggle competition.[14] Each essay has been assigned six different scores (cohesion, syntax, vocabulary, phraseology, grammar, and convention) because there are generally many components to an essay's grade and separating out the scores will capture the complexity better than just assigning one overall score. We will be assessing these six analytic metrics graded from one to five in half-integer increments on each of the categories, with 1.0 being poor and 5.0 being excellent, against the scores that were provided by human graders. In order to calculate our losses and assess our accuracy, we will be using MCRMSE as shown in Formula 1, where N_t is the number of scored ground truth target columns, and y and $\hat{y}$ are the actual and predicted values respectively.

$$MCRMSE = \frac{1}{N_t}\sum_{j=1}^{N_t}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij}-\hat{y}_{ij})^2} \quad \textbf{(Formular 1)}$$

We used the predicted score to calculate MCRMSE and then to train the model. After predicting the score on our test dataset, we transformed the predicted score by rounding it to the nearest possible score, that is from one to five by increments of one-half as depicted in Formula 2 (e.g., 3.82 scales up to 4.0 and 3.57 scales down to 3.5) and use it to calculate the MCRMSE again to produce an adjusted MCRMSE.

$$Score_{adjusted} = round(\frac{\hat{y}}{0.5}) \times 0.5 \quad \textbf{(Formular 2)}$$

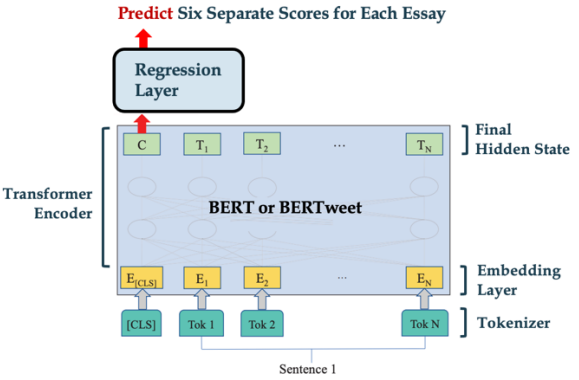
We are unable to use the adjusted score in training our model because the prediction tensor could not be transformed in TensorFlow for a custom loss calculation and it would be considered a discrete value rather than a continuous value.

We have 3,911 records in total, so we decided to randomly split it into two sets: 80 percent in the training and 20 percent in the test set. As we feed the code into the model, we then pull 20 percent of the training set to be used as the validation set with the remaining becoming our final training set. With the average length of 430 words, we truncated all the essays with max length of the first 512 words for $BERT_{base\text{-}cased}$ and the first 128 words for $BERTweet_{base}$, respectively. Training more tokens without additional preprocessing and experiment settings remaining equal would measure the quality of writing in a more meaningful fashion. To further validate the effect of token length on the human rated rubrics, we also examined the correlation between the number of words and the essay component scores in our training dataset (See Appendix A). We observed negligible correlation coefficients (r=.0778-0.2673), which indicate longer essays do not necessarily pilot a likelihood of obtaining higher scores or vice versa. Therefore, we believe that it would be reasonable to use only the max number of tokens that each model can consume.

3.1 Models

As shown in Figure 1, we employed the $BERT_{base\text{-}cased}$ and $BERTweet_{base}$ for our regression task.

Figure 1 Model Architecture



With preprocessing (adding a special token, [CLS], at the beginning of each essay), each token is transformed into its embedding and sent into the models. The representations of all essays are the output vectors mapping to [CLS], which allow the models to learn text representations so ultimately to capture deep semantics. Then a fully connected neural network is used to map the representations to scores while constraining the scores with MCRMSE loss for the weight optimization. The MCRMSE is computed as a single value evaluation metric from averaging across all RMSE values for each of the six analytic measures like cohesion and grammar.

[14] https://www.kaggle.com/competitions/feedback-prize-english-language-learning/data

3.2 Experiment Settings

After testing various sets of hyperparameters using $BERT_{base\text{-}cased}$, we discovered that the best parameters were 10 epochs, a batch size of 8, a learning rate of 0.00001, a validation split of 0.2, dropout of 0.1, and two hidden layers with 64 nodes respectively. The summary of all experiments to derive the optimal set of hyperparameters is outlined in Appendix C. We decided to use these same parameters for all experiments including our baseline model in order to make them as comparable as possible.

3.3 Baseline

For our baseline, we used the $BERT_{base\text{-}cased}$ without altering the weights by freezing all the 12 layers. This would mean that the same model will be used to produce scores for each of the essay components. From there, we fine-tuned with gradually unfreezing the layers so that the model can learn from the training set in an incremental way.

4 Results and Discussion

We calculate the following MCRMSE scores after we rounded our predicted results to the nearest half-integer for the models from our test dataset of 783 records and the results of adjusted MCRMSE scores for the respective 0, 6, or 12 trainable layers are outlined in Table 1.

Table 1 Adjusted MCRMSE Scores

	$BERT_{base\text{-}cased}$	$BERTweet_{base}$
0 trainable layers	0.6350	0.6549
6 trainable layers	0.6271	0.6224
12 trainable layers	0.5254	0.5536

We see that for $BERT_{base\text{-}cased}$ and $BERTweet_{base}$, as we progressively unfroze more layers and allowed their weights to update, the lower the MCRMSE score became, indicating a better model performance. Contrary to our expectations, $BERTweet_{base}$ generally did not perform better than $BERT_{base\text{-}cased}$ and instead performed around the same and even slightly worse. This could be because most of the essays in our test dataset were longer than 280 characters with only two that were shorter, and thus overall belonged to a separate population than Tweets, which are short sentences that $BERTweet_{base}$ was pre-trained on.

In addition to MCRMSE, we wanted to examine the proportion of responses that the models correctly performed after transforming their predicted scores. We see that the percentage that is predicted correctly for all the components was increasing as more layers were unfrozen as shown in Table 2.

Table 2 Percentage of Test Dataset Records That Were Correctly Predicted Per Essay Component by $BERT_{base\text{-}cased}$ (Score Within 0.5)

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

While the $BERT_{base\text{-}cased}$ models were predicting the correct scores between 29.8 percent to 44.1 percent of the dataset, when we expanded to see if the scores were within a given range of +- 0.5 (i.e., if the correct score is 2.0 and the model predicts between 1.5 and 2.5), it was accurate between 72.0 percent to 89.8 percent, with the accuracy increasing from all layers frozen to all the layers unfrozen. When all the layers were frozen and when the last six layers were unfrozen, the model was only predicting scores between 2.5 and 3.5 across all essay components, but when the model was entirely unfrozen, it was able to predict scores between 1.5 and 4.5 (crosstabs between predicted and actual scores found in Appendix D, Table 7-Table 24). This depicts that the models are unable to predict accurately in the extreme edge cases, which is reasonable considering the distribution of the essay components' scores in the training dataset followed a normal distribution (as seen in Appendix B), with few extreme values, so the model could not learn to predict the extreme values. As the completely unfrozen model could learn from the training data, thus it could generate some predictions of the extreme data, even though it struggled.

Table 3 Percentage of Test Dataset Records That Were Correctly Predicted Per Essay Component by $BERTweet_{base}$ (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%)

In Table 3 training $BERTweet_{base}$ models, we see that the proportion predicted correctly was mixed when we go from the model where all the weights were frozen and to the model where the last six layers were unfrozen: syntax and conventions decreased in accuracy; but increased for the rest. The differences in how correctly predicted between these two models were small as it ranged from 0.4 percent to 2.70 percent. However, when we examine accuracy when all the layers were unfrozen, accuracy improves to being within the 35.4 percent to 39.2 percent range. Overall, the $BERTweet_{base}$ models predicted scores correctly between 25.0 percent to 39.2 percent.

When we look at whether the models predicted a score that is within half of the actual value, we see that the models were accurate between 70.5 percent to 86.7 percent, with the most accurate scores for the completely unfrozen model. $BERTweet_{base}$ was able to predict between 2.5 and 4.0 across all essay components regardless of how many layers were unfrozen, but was able to predict the more extreme scores of 1.0 and 4.5 more frequently when all its layers were unfrozen (crosstabs between predicted and actual scores found in Appendix D, Table 25-Table 42). Similar to $BERT_{base\text{-}cased}$, this signals that the models were less capable at predicting extreme scores as there were very few observations of extreme scores in the training data.

When we compare between the two models to see if there were essay components which performed better with $BERTweet_{base}$, we see that it performed worse than $BERT_{base\text{-}cased}$ when it was completely frozen on all essay components except for phraseology, and all essay components when the last six layers were unfrozen. Overall, the $BERT_{base\text{-}cased}$ model with six trainable layers predicted all essay components better than the six-unfrozen $BERTweet_{base}$ model by 0.1 to 2.5 percentage points. On the other hand, we achieved mixed results when all layers were unfrozen: $BERT_{base\text{-}cased}$ was more accurate when predicting cohesion and grammar by 1.8 percentage points and 3.0 percentage points respectively; while $BERTweet_{base}$ was more accurate 0.8 to 8.7 percentage points for the rest.

Due to the differences in magnitude by the percentage points and the number of components which achieved more correct predictions, we determined that $BERT_{base\text{-}cased}$ is a better model than $BERTweet_{base}$ in almost all the components. Thus, while the pretraining is more important when grading ELLs essays as shown by the increase in model accuracy and lower MCRMSE scores, the writing of the ELLs' essays do not belong to the same population as Tweets from Twitter.

4.1 Clustering and K-Fold Cross Validation

Overall, the $BERT_{base\text{-}cased}$ was between 39.8 to 44.1 percent accurate regardless of the number of layers frozen and $BERTweet_{base}$ was between 25.5 to 43.2 percent accurate, we decided to examine whether the models would be more accurate when we grouped scores together into low, average, and high scorers. We see that there is a high correlation between all the components with the lowest correlation coefficient being a r=0.6374 (See Appendix A). This indicates that generally, if a student scored low in one component, they would score low in the others; conversely, if a student scored high in one component, they would also receive a high score in the others.

We wanted to make sure we do our experiments holistically and to try to predict the more extreme scores, so we used k-fold cross validation along three clusters to run the experiment on a representative data set. To create the "clustered" version of the models, we summed up all the scores within each category (lowest possible total score a student could achieve would be 6 and highest total score would be 30), and by using the elbow method, we decided to divide the data into three clusters of scores (6.0-17.0, 17.5-21.5, and 21.5-30.0) through K-means and performed the same

experiments. We decided to use stratified two-folds (split the sample into two as we were testing if this would cause the model to perform better), and ran the same experiments to achieve the following adjusted MCRMSE scores:

Table 4 Adjusted MCRMSE Scores for the Clustered Models

	$BERT_{base\text{-}cased}$	$BERTweet_{base}$
0 trainable layers	0.6763	0.6907
6 trainable layers	0.6681	0.6688
12 trainable layers	0.6798	0.6652

We see that these MCRMSE scores are worse than the corresponding versions without the clusters as shown in Table 4. Thus, grouping the scores into three clusters, and seeing if the model predicts better within the clusters actually decreased its performance. Even though more layers were unfrozen and the model was allowed to learn from the training, we see that both $BERT_{base\text{-}cased}$ and $BERTweet_{base}$ with clustering did not show much improvement.

Table 5 Percentage of Test Dataset Records That Were Correctly Predicted Per Essay Component by $BERTweet_{base}$ After Clustering (Score Within 0.5)

	0 trainable layers	6 trainable layers	12 trainable layers
Cohesion	28.0% (70.0%)	26.8% (72.7%)	25.4% (72.0%)
Syntax	30.7% (72.4%)	33.1% (72.9%)	32.2% (73.9%)
Vocabulary	31.2% (77.7%)	29.6% (79.6%)	29.6% (81.2%)
Phraseology	27.2% (71.1%)	30.5% (73.3%)	29.6% (73.2%)
Grammar	24.4% (70.8%)	25.8% (70.5%)	24.8% (71.1%)
Conventions	29.1% (71.6%)	30.9% (73.9%)	30.7% (72.4%)

Table 5 above shows that as the layers unfroze, $BERT_{base\text{-}cased}$ was only predicting the correct score between 24.5 percent and 37.2 percent of the essays in the test set and accuracy increased in all components going from all the layers freezing to all the layers unfrozen with the greatest being vocabulary with a difference of 8.5 percentage points. This is in stark contrast to the unclustered versions where unfreezing all the layers allowed for a better prediction. While the model performed generally well when predicting scores that were within 0.5 of what the actual score was, we do not see that there was much improvement when we go from the model with all the layers frozen (70.0 percent to 78.5 percent), to the model with its last six layers unfrozen (71.3 percent to 79.2 percent), to the model with all its layers unfrozen (70.2 percent to 80.3 percent). The $BERT_{base\text{-}cased}$ clustered models were performing on par with the unclustered version with all its layers frozen, but performed worse than the unclustered model with all its layers unfrozen. When all the layers were frozen and when the last six layers were unfrozen, the model was only predicting scores between 2.5 and 3.5 across all essay components, but when the model was entirely unfrozen, it was able to predict scores between 1.5 and 3.5 (crosstabs between predicted and actual scores found in Appendix D, Table 43-Table 60). Again, this signals that the models were less capable at predicting extreme scores.

Table 6 Percentage of Test Dataset Records That Were Correctly Predicted Per Essay Component by BERT-base After Clustering (Score Within 0.5)

	0 trainable layers	6 trainable layers	12 trainable layers
Cohesion	28.6% (71.9%)	28.2% (73.9%)	28.4% (73.8%)
Syntax	32.7% (73.9%)	30.0% (73.4%)	30.5% (72.7%)
Vocabulary	36.9% (78.5%)	37.2% (79.2%)	28.4% (80.3%)
Phraseology	30.5% (73.4%)	30.5% (73.2%)	29.0% (73.1%)
Grammar	27.8% (70.0%)	27.3% (71.3%)	24.5% (70.2%)
Conventions	30.8% (71.8%)	30.7% (76.2%)	28.6% (70.4%)

In general, the $BERTweet_{base}$ cluster models predicted a correct score more often when more layers unfroze as shown in Table 6; when we compare the all frozen model to the six-unfrozen and completely unfrozen, cohesion and vocabulary decreased in accuracy while the rest increased.

The $BERT_{base\text{-}cased}$ clustered performed better than the $BERTweet_{base}$ clustered model when they were completely frozen on all essay components and on all but conventions when it was six-unfrozen. $BERTweet_{base}$ performed better by 0.6 to 2.1 percentage points on all essay components

except cohesion where it was less accurate by 3.0 percentage points.

The $\text{BERTweet}_{\text{base}}$ clustered models were performing slightly worse than the unclustered versions. The differences between the models were less than 3.0 percentage points; except when it was six unfrozen and predicting vocabulary (the clustered version was less accurate by 6.9 percentage points), and when all the layers were unfrozen, the clustered versions were worse in predicting the correct scores by 3.4 to 11.7 percentage points.

Regardless of how many layers were frozen, the $\text{BERTweet}_{\text{base}}$ clustered model was only predicting scores between 2.5 and 4.0 across all essay components and was overall predicting a greater variety of scores than its $\text{BERT}_{\text{base-cased}}$ clustered counterparts (crosstabs between predicted and actual scores found in Appendix D, Table 61-Table 78). This shows that even when we try a technique to have the models to improve the underrepresented scores, it still struggles to predict them.

5 Conclusion and Future Work

We see that the $\text{BERT}_{\text{base-cased}}$ model and $\text{BERTweet}_{\text{base}}$ models performed the best when all their layers were unfrozen. This signals that relying more on the pre-trained data would behoove automated essay evaluation systems. $\text{BERTweet}_{\text{base}}$ did not perform better than $\text{BERT}_{\text{base-cased}}$ holistically (as shown by the higher MCRMSE score) or correctly score the essay's components by more than 3.0 percentage points no matter how many layers were unfrozen. Through this, we can surmise that ELLs' essays belong to a different population than Tweets from Twitter, and their informal nature using less grammar rules is not enough to predict ELLs' performance on essays. From a technical point of view, supplying less information with just 128 tokens and the limit of 280 characters compared to 512 tokens for $\text{BERT}_{\text{base-cased}}$ would also drive the worse performance of the $\text{BERTweet}_{\text{base}}$. Most of the scores predicted were centered between 2.5 and 3.5, signaling it was harder for the models to predict the more extreme scores. It was only able to predict more extreme scores, such as 1.5 and 4.5 when more layers were unfroze, but was unable to predict the greatest extremes (1.0 and 5.0) in any of the versions. Even when we clustered the scores by low scorers, medium scorers, and high scorers, both the clustered and unclustered models still struggled to predict the low and high scores regardless. This clearly implies that mitigating the effects of class imbalance would be essential to produce a more satisfactory model performance. Future work also might look at increasing the size of our input dataset while employing the larger versions of the base versions we introduced here or jointly optimized pre-trained BERT and multiple BERT-derived models as they allow significant horse-power with much more parameters and attention heads. Finally, we are envisioning fine-tuning even further by developing a human rater-like AES system. We anticipate training student essays together with key topical sentences, which are the essential parts of a high-quality writing accompanied with supportive reasons and evidence to a given topic would significantly enhance the validity and reliability of the AES system.

References

Mark D. Shermis and Jill Burstein. 2013. Handbook of automated essay evaluation: current application and new directions. https://ebookcentral-proquest-com.libproxy.berkeley.edu/lib/berkeley-ebooks/detail.action?docID=1172902

Ridha Hussein Chassab. 2021. Automatic essay scoring: A review on the feature analysis techniques. International Journal of Advanced Computer Science and Applications, 12(10): 252-264. https://thesai.org/Downloads/Volume12No10/Paper_28-Automatic_Essay_Scoring.pdf

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 9–14, Online. Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-demos.2.pdf

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/pdf/1810.04805.pdf

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, 151-162. https://aclanthology.org/2020.bea-1.15.pdf

Madalina Cozma, Andrei M Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. Proceedings of the

56th Annual Meeting of the Association for Computational Linguistics (Short Papers), 503-509. https://aclanthology.org/P18-2080.pdf

https://www.kaggle.com/competitions/feedback-prize-english-language-learning/data

https://www.kaggle.com/c/asap-aes

Number of English Learners in US Schools Keeps Rising. 2018. https://learningenglish.voanews.com/a/number-of-english-learners-in-us-schools-keeps-rising/4635659.html

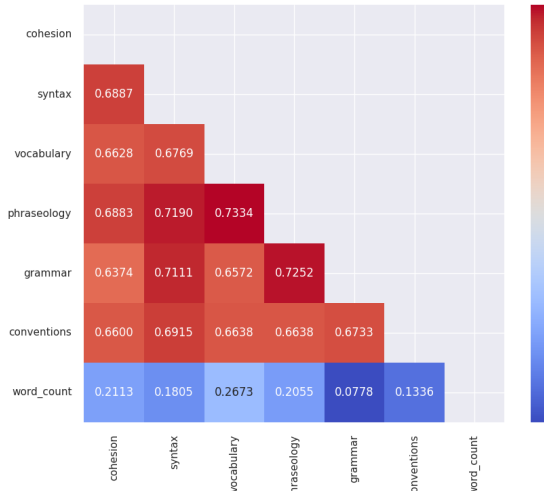
https://developer.twitter.com/en/docs/counting-characters

Appendices

A Correlation Matrix

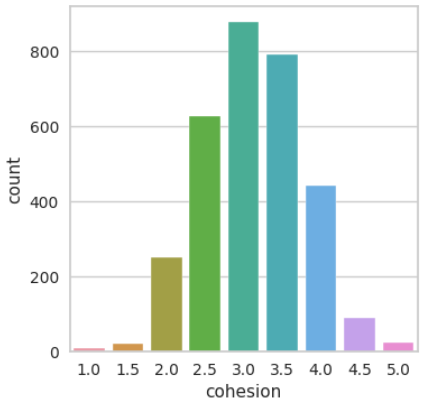
Correlation matrix between the number of words the essay component scores in our training dataset is presented below.

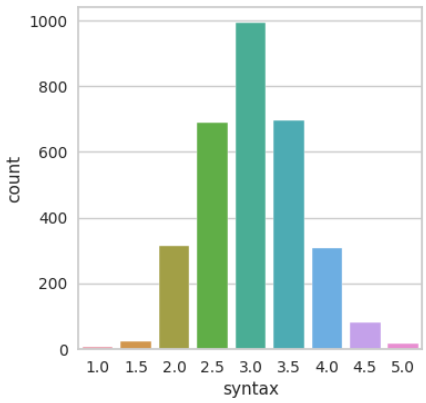
Figure 1 Correlation Matrix

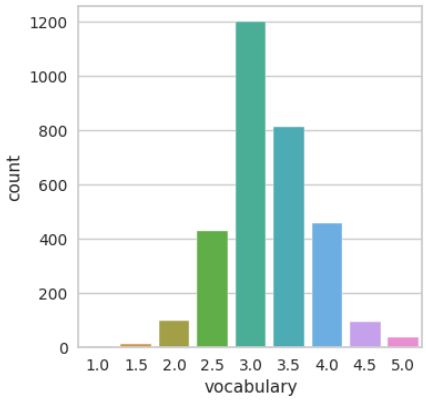


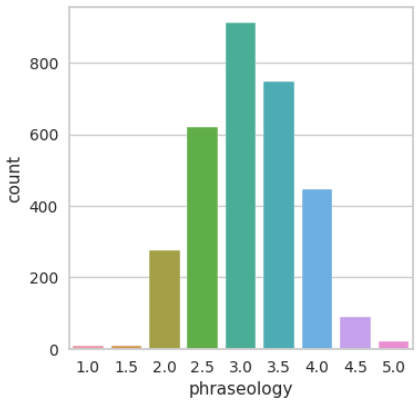
B Score Distribution

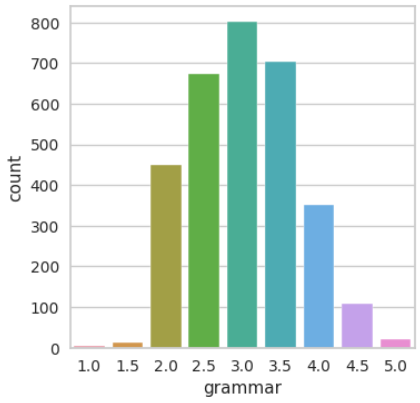
The Gaussian distribution of the essay components' scores in the training dataset is presented below.

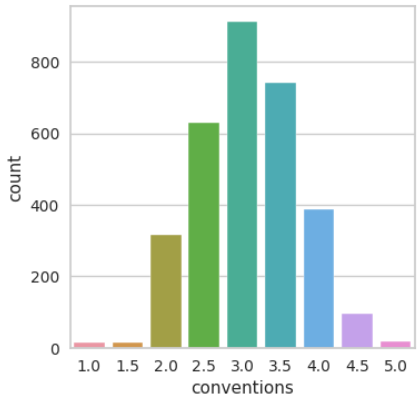












C Hyperparameters

The following table contains the experiments from adjusting hyperparameters using $BERT_{base\text{-}cased}$. Due to the limit of time and GPU memory, we fine-tuned the learning rate from 5e-4 to 1e-5, one or two hidden layers with hidden nodes ranged from 64 to 256, and dropout rate from 0.1 to 0.3 with batch size of 8 or 16 for 10 epochs while adjusting the number of unfreezing layers ranged from 0 to 12. Finally, the best model is defined based on the lowest performance of MCRMSE (0.4590) on the test set that is fine-tuned for 8 batches at a time with learning rate of 0.00001, two hidden layers with 64 nodes respectively, and 0.1 dropout rate (See the yellow-highlighted last row in the table below). Although the size of our training set was relatively small, a very low learning rate with small batch size and low hidden nodes and dropout to retrain the fully trainable $BERT_{base\text{-}cased}$ model was sensible to yield the best performance.

Num Trainable layers	Epochs	Test MCRMSE	Test Loss	Learning Rate	Num Hidden Layers	Num hidden Units	Dropout	Batch Size
0	10	0.4722	0.4779	0.0001	1	256	0.2	16
2	10	0.4736	0.4797	0.0001	1	256	0.2	16
4	10	0.4715	0.4778	0.0001	1	256	0.2	16
6	10	0.4707	0.4762	0.0001	1	256	0.2	16
0	10	0.4982	0.5006	0.0005	1	64	0.2	16
4	10	0.4754	0.4806	0.0005	1	64	0.2	16
8	10	0.4813	0.4858	0.0005	1	64	0.2	16
12	10	0.5352	0.5325	0.0005	1	64	0.2	16
0	10	0.4722	0.4779	0.0001	1	256	0.2	16
2	10	0.4736	0.4797	0.0001	1	256	0.2	16
4	10	0.4715	0.4778	0.0001	1	256	0.2	16
6	10	0.4707	0.4762	0.0001	1	256	0.2	16
0	10	0.4804	0.4856	0.00005	1	256	0.3	16
2	10	0.479	0.484	0.00005	1	256	0.3	16
4	10	0.4761	0.4813	0.00005	1	256	0.3	16
6	10	0.4789	0.4835	0.00005	1	256	0.3	16
2	10	0.5088	0.5092	0.00005	1	256	0.1	16
4	10	0.5667	0.5635	0.00005	1	256	0.1	16
0	10	0.4759	0.4809	0.00005	1	256	0.3	8
2	10	0.484	0.4887	0.00005	1	256	0.3	8
4	10	0.4845	0.4892	0.00005	1	256	0.3	8
6	10	0.4826	0.4872	0.00005	1	256	0.3	8
8	10	0.4766	0.482	0.00005	1	256	0.3	8
10	10	0.4798	0.4856	0.00005	1	256	0.3	8
12	10	0.5192	0.52	0.00005	1	256	0.3	8
0	10	0.4827	0.486	0.00005	1	256	0.3	16
6	10	0.4752	0.4789	0.00005	1	256	0.3	16
12	10	0.4839	0.4854	0.00005	1	256	0.3	16
0	10	0.6	0.6026	0.00001	2	64	0.1	16
2	10	0.5958	0.5977	0.00001	2	64	0.1	16
4	10	0.6012	0.6057	0.00001	2	64	0.1	16
6	10	0.57	0.5733	0.00001	2	64	0.1	16
8	10	0.6248	0.6307	0.00001	2	64	0.1	16
10	10	0.5859	0.5885	0.00001	2	64	0.1	16
12	10	0.4646	0.4717	0.00001	2	64	0.1	16
0	10	0.4764	0.4826	0.00005	1	256	0.3	8
6	10	0.4787	0.4845	0.00005	1	256	0.3	8
12	10	0.4688	0.4701	0.00005	1	256	0.3	8
2	10	0.5088	0.5092	0.00005	1	256	0.1	16
4	10	0.5667	0.5635	0.00005	1	256	0.1	16
0	10	0.4982	0.5006	0.0005	1	64	0.2	16
4	10	0.4754	0.4806	0.0005	1	64	0.2	16
8	10	0.4813	0.4858	0.0005	1	64	0.2	16
12	10	0.5352	0.5325	0.0005	1	64	0.2	16
0	10	0.4804	0.4846	0.0005	1	256	0.3	16
6	10	0.4792	0.4855	0.0005	1	256	0.3	16
12	10	0.4816	0.4855	0.0005	1	256	0.3	16
0	10	0.5065	0.5095	0.0005	1	128	0.2	16
4	10	0.505	0.5084	0.0005	1	128	0.2	16
8	10	0.4919	0.4954	0.0005	1	128	0.2	16
12	**10**	**0.459**	**0.464**	**0.00001**	**2**	**64**	**0.1**	**8**

D Crosstabs

The top row represents the predicted values while the left column represents the actual values.

Table 7 $BERT_{base\text{-}cased}$, 0 Layers Trainable: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	2	0	0	0	0	0
1.5	0	0	0	5	1	0	0	0	0
2.0	0	0	0	13	51	0	0	0	0
2.5	0	0	0	24	138	0	0	0	0
3.0	0	0	0	13	200	6	0	0	0
3.5	0	0	0	7	182	9	0	0	0
4.0	0	0	0	5	79	9	0	0	0
4.5	0	0	0	0	28	7	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 8 $BERT_{base\text{-}cased}$, 0 Layers Trainable: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	4	0	0	0	0	0
1.5	0	0	0	3	2	0	0	0	0
2.0	0	0	0	27	69	0	0	0	0
2.5	0	0	0	23	126	0	0	0	0
3.0	0	0	0	13	244	0	0	0	0
3.5	0	0	0	5	165	0	0	0	0
4.0	0	0	0	1	77	3	0	0	0
4.5	0	0	0	0	19	0	0	0	0
5.0	0	0	0	0	2	0	0	0	0

Table 9 $BERT_{base\text{-}cased}$, 0 Layers Trainable: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	4	0	0	0	0
1.5	0	0	0	0	24	2	0	0	0
2.0	0	0	0	0	86	15	0	0	0
2.5	0	0	0	0	213	92	0	0	0
3.0	0	0	0	0	113	81	0	0	0
3.5	0	0	0	0	64	58	0	0	0
4.0	0	0	0	0	11	13	0	0	0
4.5	0	0	0	0	3	4	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 10 B $BERT_{base\text{-}cased}$ 0 Layers Trainable: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	2	1	0	0	0	0
1.5	0	0	0	2	1	0	0	0	0
2.0	0	0	0	9	62	3	0	0	0
2.5	0	0	0	4	124	24	0	0	0
3.0	0	0	0	2	180	60	0	0	0
3.5	0	0	0	0	119	61	0	0	0
4.0	0	0	0	0	52	55	0	0	0
4.5	0	0	0	0	5	13	0	0	0
5.0	0	0	0	0	2	2	0	0	0

Table 11 $BERT_{base\text{-}cased}$, 0 Layers Trainable: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	3	0	0	0	0	0
1.5	0	0	0	5	1	0	0	0	0
2.0	0	0	0	38	57	0	0	0	0
2.5	0	0	0	47	134	0	0	0	0
3.0	0	0	0	24	169	0	0	0	0
3.5	0	0	0	8	167	1	0	0	0
4.0	0	0	0	3	92	1	0	0	0
4.5	0	0	0	0	25	0	0	0	0
5.0	0	0	0	0	8	0	0	0	0

Table 12 $BERT_{base\text{-}cased}$, 0 Layers Trainable: Conventions

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	0	0	0	0
1.5	0	0	0	1	4	0	0	0	0
2.0	0	0	0	7	76	2	0	0	0
2.5	0	0	0	5	148	2	0	0	0
3.0	0	0	0	4	235	1	0	0	0
3.5	0	0	0	2	158	8	0	0	0
4.0	0	0	0	0	92	4	0	0	0
4.5	0	0	0	0	23	3	0	0	0
5.0	0	0	0	0	7	0	0	0	0

Table 13 $BERT_{base\text{-}cased}$, 6 Layers Trainable: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	2	0	0	0	0	0
1.5	0	0	0	5	1	0	0	0	0
2.0	0	0	0	18	43	3	0	0	0
2.5	0	0	0	24	119	19	0	0	0
3.0	0	0	0	16	180	23	0	0	0
3.5	0	0	0	9	158	31	0	0	0
4.0	0	0	0	4	65	24	0	0	0
4.5	0	0	0	0	20	15	0	0	0
5.0	0	0	0	0	2	2	0	0	0

Table 14 $BERT_{base\text{-}cased}$, 6 Layers Trainable: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	4	0	0	0	0	0
1.5	0	0	0	3	2	0	0	0	0
2.0	0	0	0	24	70	2	0	0	0
2.5	0	0	0	27	119	3	0	0	0
3.0	0	0	0	21	200	36	0	0	0
3.5	0	0	0	9	128	33	0	0	0
4.0	0	0	0	1	62	18	0	0	0
4.5	0	0	0	0	13	6	0	0	0
5.0	0	0	0	0	2	0	0	0	0

Table 15 $BERT_{base\text{-}cased}$, 6 Layers Trainable: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	2	2	0	0	0	0
1.5	0	0	0	8	17	1	0	0	0
2.0	0	0	0	15	77	9	0	0	0
2.5	0	0	0	11	215	79	0	0	0
3.0	0	0	0	1	118	75	0	0	0
3.5	0	0	0	0	68	54	0	0	0
4.0	0	0	0	0	10	14	0	0	0
4.5	0	0	0	0	2	5	0	0	0
5.0	0	0	0	2	2	0	0	0	0

Table 16 $BERT_{base\text{-}cased}$, 6 Layers Trainable: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	3	0	0	0	0	0
1.5	0	0	0	1	2	0	0	0	0
2.0	0	0	0	24	48	2	0	0	0
2.5	0	0	0	11	115	26	0	0	0
3.0	0	0	0	5	180	57	0	0	0
3.5	0	0	0	1	114	65	0	0	0
4.0	0	0	0	0	55	52	0	0	0
4.5	0	0	0	0	3	15	0	0	0
5.0	0	0	0	0	2	2	0	0	0

Table 17 $BERT_{base\text{-}cased}$, 6 Layers Trainable: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	3	0	0	0	0	0
1.5	0	0	0	5	1	0	0	0	0
2.0	0	0	0	30	65	0	0	0	0
2.5	0	0	0	51	129	1	0	0	0
3.0	0	0	0	28	164	1	0	0	0
3.5	0	0	0	17	156	3	0	0	0
4.0	0	0	0	8	86	2	0	0	0
4.5	0	0	0	0	25	0	0	0	0
5.0	0	0	0	1	7	0	0	0	0

Table 18 $BERT_{base\text{-}cased}$, 6 Layers Trainable: Conventions

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	0	0	0	0
1.5	0	0	0	3	2	0	0	0	0
2.0	0	0	0	11	71	3	0	0	0
2.5	0	0	0	4	144	7	0	0	0
3.0	0	0	0	3	209	28	0	0	0
3.5	0	0	0	4	143	21	0	0	0
4.0	0	0	0	0	76	20	0	0	0
4.5	0	0	0	0	20	6	0	0	0
5.0	0	0	0	0	4	3	0	0	0

Table 19 $BERT_{base\text{-}cased}$, 12 Layers Trainable: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	1	0	1	0	0	0	0	0
1.5	0	1	5	0	0	0	0	0	0
2.0	0	4	16	27	14	3	0	0	0
2.5	0	1	23	70	54	14	0	0	0
3.0	0	0	5	71	101	40	2	0	0
3.5	0	0	1	25	101	71	0	0	0
4.0	0	0	0	4	32	51	6	0	0
4.5	0	0	0	0	1	21	13	0	0
5.0	0	0	0	0	1	3	0	0	0

Table 20 $BERT_{base\text{-}cased}$, 12 Layers Trainable: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	2	1	1	0	0	0	0	0
1.5	0	2	3	0	0	0	0	0	0
2.0	0	3	33	46	14	0	0	0	0
2.5	0	0	23	87	38	1	0	0	0
3.0	0	0	14	88	131	24	0	0	0
3.5	0	0	1	34	93	39	3	0	0
4.0	0	0	0	5	31	41	4	0	0
4.5	0	0	0	0	2	14	3	0	0
5.0	0	0	0	0	0	1	1	0	0

Table 21 $BERT_{base\text{-}cased}$, 12 Layers Trainable: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	3	1	0	0	0	0	0
1.5	0	0	11	12	3	0	0	0	0
2.0	0	0	2	48	44	7	0	0	0
2.5	0	0	5	56	178	64	2	0	0
3.0	0	0	0	6	88	96	4	0	0
3.5	0	0	0	2	31	76	13	0	0
4.0	0	0	0	0	2	14	8	0	0
4.5	0	0	0	0	0	2	5	0	0
5.0	0	0	3	1	0	0	0	0	0

Table 22 $BERT_{base\text{-}cased}$, 12 Layers Trainable: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	2	1	0	0	0	0	0
1.5	0	1	2	0	0	0	0	0	0
2.0	0	1	20	34	16	3	0	0	0
2.5	0	0	4	65	58	21	4	0	0
3.0	0	0	2	40	125	68	7	0	0
3.5	0	0	0	11	60	92	16	1	0
4.0	0	0	0	0	11	52	41	3	0
4.5	0	0	0	0	0	5	12	1	0
5.0	0	0	0	0	0	0	4	0	0

Table 23 $BERT_{base\text{-}cased}$, 12 Layers Trainable: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	2	0	1	0	0	0	0	0
1.5	0	5	1	0	0	0	0	0	0
2.0	0	7	31	37	18	2	0	0	0
2.5	0	4	34	83	48	12	0	0	0
3.0	0	0	14	68	73	37	1	0	0
3.5	0	0	3	26	79	62	6	0	0
4.0	0	0	0	11	31	46	8	0	0
4.5	0	0	0	0	2	14	9	0	0
5.0	0	0	0	0	1	4	3	0	0

Table 24 $BERT_{base\text{-}cased}$, 12 Layers Trainable: Convention

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	1	0	0	0	0	0	0	0
1.5	0	3	1	1	0	0	0	0	0
2.0	0	16	32	29	6	2	0	0	0
2.5	0	1	20	68	54	12	0	0	0
3.0	0	0	10	55	104	65	6	0	0
3.5	0	0	0	20	70	65	13	0	0
4.0	0	0	0	2	16	52	26	0	0
4.5	0	0	0	0	0	16	10	0	0
5.0	0	0	0	0	0	1	5	1	0

Table 25 BERTweet$_{base}$, 0 Layers Trainable: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	2	0	0	0	0
1.5	0	0	0	1	4	1	0	0	0
2.0	0	0	0	0	61	3	0	0	0
2.5	0	0	0	1	132	29	0	0	0
3.0	0	0	0	0	183	36	0	0	0
3.5	0	0	0	0	168	30	0	0	0
4.0	0	0	0	0	74	19	0	0	0
4.5	0	0	0	0	27	8	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 26 BERTweet$_{base}$, 0 Layers Trainable: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	1	3	0	0	0	0
1.5	0	0	0	1	4	0	0	0	0
2.0	0	0	0	3	92	1	0	0	0
2.5	0	0	0	2	143	4	0	0	0
3.0	0	0	0	0	250	7	0	0	0
3.5	0	0	0	1	159	10	0	0	0
4.0	0	0	0	0	80	1	0	0	0
4.5	0	0	0	0	18	1	0	0	0
5.0	0	0	0	0	2	0	0	0	0

Table 27 BERTweet$_{base}$, 0 Layers Trainable: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	4	0	0	0	0
1.5	0	0	0	1	18	7	0	0	0
2.0	0	0	0	0	61	40	0	0	0
2.5	0	0	0	1	158	146	0	0	0
3.0	0	0	0	0	87	107	0	0	0
3.5	0	0	0	0	55	67	0	0	0
4.0	0	0	0	0	10	14	0	0	0
4.5	0	0	0	0	2	5	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 28 BERTweet$_{base}$, 0 Layers Trainable: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	3	0	0	0	0
1.5	0	0	0	2	1	0	0	0	0
2.0	0	0	0	1	59	14	0	0	0
2.5	0	0	0	0	105	47	0	0	0
3.0	0	0	0	0	134	108	0	0	0
3.5	0	0	0	0	81	99	0	0	0
4.0	0	0	0	0	36	71	0	0	0
4.5	0	0	0	0	6	12	0	0	0
5.0	0	0	0	0	1	3	0	0	0

Table 29 BERTweet$_{base}$, 0 Layers Trainable: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	3	0	0	0	0
1.5	0	0	0	2	4	0	0	0	0
2.0	0	0	0	2	92	1	0	0	0
2.5	0	0	0	0	180	1	0	0	0
3.0	0	0	0	0	192	1	0	0	0
3.5	0	0	0	0	172	4	0	0	0
4.0	0	0	0	0	95	1	0	0	0
4.5	0	0	0	0	24	1	0	0	0
5.0	0	0	0	0	8	0	0	0	0

Table 30 BERTweet$_{base}$, 0 Layers Trainable: Conventions

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	0	0	0	0
1.5	0	0	0	1	4	0	0	0	0
2.0	0	0	0	5	77	3	0	0	0
2.5	0	0	0	0	138	17	0	0	0
3.0	0	0	0	2	199	39	0	0	0
3.5	0	0	0	0	125	43	0	0	0
4.0	0	0	0	0	63	33	0	0	0
4.5	0	0	0	0	17	9	0	0	0
5.0	0	0	0	0	3	4	0	0	0

Table 31 BERTweet$_{base}$, 6 Layers Trainable: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	2	0	0	0	0
1.5	0	0	0	2	4	0	0	0	0
2.0	0	0	0	1	58	5	0	0	0
2.5	0	0	0	3	115	44	0	0	0
3.0	0	0	0	0	131	88	0	0	0
3.5	0	0	0	0	102	96	0	0	0
4.0	0	0	0	0	37	56	0	0	0
4.5	0	0	0	0	6	29	0	0	0
5.0	0	0	0	0	2	2	0	0	0

Table 32 BERTweet$_{base}$, 6 Layers Trainable: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	1	3	0	0	0	0
1.5	0	0	0	3	2	0	0	0	0
2.0	0	0	0	3	89	4	0	0	0
2.5	0	0	0	2	130	17	0	0	0
3.0	0	0	0	2	199	56	0	0	0
3.5	0	0	0	0	112	58	0	0	0
4.0	0	0	0	0	55	26	0	0	0
4.5	0	0	0	0	9	10	0	0	0
5.0	0	0	0	0	1	1	0	0	0

Table 33 BERTweet$_{base}$, 6 Layers Trainable: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	4	0	0	0	0
1.5	0	0	0	2	21	3	0	0	0
2.0	0	0	0	0	64	37	0	0	0
2.5	0	0	0	2	135	168	0	0	0
3.0	0	0	0	0	44	150	0	0	0
3.5	0	0	0	0	18	103	1	0	0
4.0	0	0	0	0	0	24	0	0	0
4.5	0	0	0	0	0	7	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 34 BERTweet$_{base}$, 6 Layers Trainable: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	1	2	0	0	0	0
1.5	0	0	0	2	1	0	0	0	0
2.0	0	0	0	1	61	12	0	0	0
2.5	0	0	0	1	112	39	0	0	0
3.0	0	0	0	0	136	106	0	0	0
3.5	0	0	0	0	79	101	0	0	0
4.0	0	0	0	0	23	84	0	0	0
4.5	0	0	0	0	5	13	0	0	0
5.0	0	0	0	0	2	2	0	0	0

Table 35 BERTweet$_{base}$, 6 Layers Trainable: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	2	1	0	0	0	0
1.5	0	0	0	5	1	0	0	0	0
2.0	0	0	0	15	74	6	0	0	0
2.5	0	0	0	21	147	13	0	0	0
3.0	0	0	0	3	142	48	0	0	0
3.5	0	0	0	5	119	52	0	0	0
4.0	0	0	0	0	61	35	0	0	0
4.5	0	0	0	0	13	12	0	0	0
5.0	0	0	0	0	6	2	0	0	0

Table 36 BERTweet$_{base}$, 6 Layers Trainable: Convention

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	0	0	0	0
1.5	0	0	0	2	2	1	0	0	0
2.0	0	0	0	6	74	5	0	0	0
2.5	0	0	0	1	139	15	0	0	0
3.0	0	0	0	3	180	57	0	0	0
3.5	0	0	0	4	115	49	0	0	0
4.0	0	0	0	0	58	38	0	0	0
4.5	0	0	0	0	17	9	0	0	0
5.0	0	0	0	0	3	4	0	0	0

Table 37 BERTweet$_{base}$, 12 Layers Trainable: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	1	0	0	1	0	0	0	0
1.5	0	2	4	0	0	0	0	0	0
2.0	0	2	9	26	21	5	1	0	0
2.5	0	0	14	44	67	36	1	0	0
3.0	0	0	3	34	99	80	3	0	0
3.5	0	0	1	14	64	114	5	0	0
4.0	0	0	0	1	18	63	11	0	0
4.5	0	0	0	0	0	23	12	0	0
5.0	0	0	0	0	0	4	0	0	0

Table 38 BERTweet$_{base}$, 12 Layers Trainable: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	1	2	0	1	0	0	0	0
1.5	0	2	2	1	0	0	0	0	0
2.0	0	1	10	44	33	7	1	0	0
2.5	0	0	5	41	70	25	7	1	0
3.0	0	0	1	37	89	103	27	0	0
3.5	0	0	0	4	34	93	39	0	0
4.0	0	0	0	0	7	30	43	1	0
4.5	0	0	0	0	0	6	12	1	0
5.0	0	0	0	0	0	1	1	0	0

Table 39 BERTweet$_{base}$, 12 Layers Trainable: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	1	1	1	1	0	0	0	0
1.5	0	2	5	11	6	2	0	0	0
2.0	0	0	0	21	48	29	3	0	0
2.5	0	0	0	18	104	140	43	0	0
3.0	0	0	0	1	28	110	52	3	0
3.5	0	0	0	0	6	52	59	5	0
4.0	0	0	0	0	1	2	14	7	0
4.5	0	0	0	0	0	0	6	1	0
5.0	0	1	1	1	1	0	0	0	0

Table 40 BERTweet$_{base}$, 12 Layers Trainable: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	1	1	0	1	0	0	0	0
1.5	0	1	1	1	0	0	0	0	0
2.0	0	2	11	36	15	9	1	0	0
2.5	0	0	2	42	63	36	9	0	0
3.0	0	0	1	20	71	112	38	0	0
3.5	0	0	0	2	36	83	55	4	0
4.0	0	0	0	1	4	34	65	3	0
4.5	0	0	0	0	0	1	13	4	0
5.0	0	0	0	0	0	1	2	1	0

Table 41 BERTweet$_{base}$, 12 Layers Trainable: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	2	0	1	0	0	0	0
1.5	0	4	1	1	0	0	0	0	0
2.0	0	1	18	34	31	10	1	0	0
2.5	0	0	13	60	59	40	9	0	0
3.0	0	0	2	19	65	81	26	0	0
3.5	0	0	1	3	29	96	47	0	0
4.0	0	0	0	3	11	39	43	0	0
4.5	0	0	0	0	0	5	20	0	0
5.0	0	0	0	0	0	1	7	0	0

Table 42 BERTweet$_{base}$, 12 Layers Trainable: Conventions

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	1	0	0	0	0	0	0	0	0
1.5	1	1	2	1	0	0	0	0	0
2.0	1	6	26	31	18	3	0	0	0
2.5	0	0	12	50	69	19	5	0	0
3.0	0	0	2	45	104	82	7	0	0
3.5	0	0	0	11	56	90	11	0	0
4.0	0	0	0	1	13	62	20	0	0
4.5	0	0	0	0	2	12	12	0	0
5.0	0	0	0	0	0	2	5	0	0

Table 43 $BERT_{base\text{-}cased}$, 0 Layers Trainable, Clustered: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	2	0	0	0	0
1.5	0	0	0	1	5	0	0	0	0
2.0	0	0	0	2	51	11	0	0	0
2.5	0	0	0	4	128	30	0	0	0
3.0	0	0	0	1	180	38	0	0	0
3.5	0	0	0	3	155	40	0	0	0
4.0	0	0	0	0	78	15	0	0	0
4.5	0	0	0	1	27	7	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 44 $BERT_{base\text{-}cased}$, 0 Layers Trainable, Clustered: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	1	3	0	0	0	0
1.5	0	0	0	2	3	0	0	0	0
2.0	0	0	0	12	80	4	0	0	0
2.5	0	0	0	19	125	5	0	0	0
3.0	0	0	0	16	229	12	0	0	0
3.5	0	0	0	9	153	8	0	0	0
4.0	0	0	0	4	72	5	0	0	0
4.5	0	0	0	0	19	0	0	0	0
5.0	0	0	0	0	2	0	0	0	0

Table 45 $BERT_{base\text{-}cased}$, 0 Layers Trainable, Clustered: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	1	3	0	0	0	0
1.5	0	0	0	0	23	3	0	0	0
2.0	0	0	0	1	92	8	0	0	0
2.5	0	0	0	6	256	43	0	0	0
3.0	0	0	0	1	161	32	0	0	0
3.5	0	0	0	2	96	24	0	0	0
4.0	0	0	0	0	17	7	0	0	0
4.5	0	0	0	0	6	1	0	0	0
5.0	0	0	0	1	3	0	0	0	0

Table 46 $BERT_{base\text{-}cased}$, 0 Layers Trainable, Clustered: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	1	2	0	0	0	0
1.5	0	0	0	0	3	0	0	0	0
2.0	0	0	0	5	67	2	0	0	0
2.5	0	0	0	6	142	4	0	0	0
3.0	0	0	0	6	224	12	0	0	0
3.5	0	0	0	4	167	9	0	0	0
4.0	0	0	0	3	100	4	0	0	0
4.5	0	0	0	0	17	1	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 47 $BERT_{base\text{-}cased}$, 0 Layers Trainable, Clustered: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	3	0	0	0	0	0
1.5	0	0	0	4	2	0	0	0	0
2.0	0	0	0	38	57	0	0	0	0
2.5	0	0	0	57	123	1	0	0	0
3.0	0	0	0	34	159	0	0	0	0
3.5	0	0	0	39	135	2	0	0	0
4.0	0	0	0	22	74	0	0	0	0
4.5	0	0	0	2	22	1	0	0	0
5.0	0	0	0	3	5	0	0	0	0

Table 48 $BERT_{base\text{-}cased}$, 0 Layers Trainable, Clustered: Conventions

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	0	0	0	0
1.5	0	0	0	1	4	0	0	0	0
2.0	0	0	0	16	68	1	0	0	0
2.5	0	0	0	24	128	3	0	0	0
3.0	0	0	0	19	214	7	0	0	0
3.5	0	0	0	18	147	3	0	0	0
4.0	0	0	0	10	82	4	0	0	0
4.5	0	0	0	0	26	0	0	0	0
5.0	0	0	0	1	6	0	0	0	0

Table 49 $BERT_{base\text{-}cased}$, 6 Layers Trainable, Clustered: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	1	1	0	0	0	0
1.5	0	0	0	1	5	0	0	0	0
2.0	0	0	0	5	57	2	0	0	0
2.5	0	0	0	7	151	4	0	0	0
3.0	0	0	0	10	198	11	0	0	0
3.5	0	0	0	4	178	16	0	0	0
4.0	0	0	0	2	88	3	0	0	0
4.5	0	0	0	0	30	5	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 50 $BERT_{base\text{-}cased}$, 6 Layers Trainable, Clustered: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	3	1	0	0	0	0
1.5	0	0	0	2	3	0	0	0	0
2.0	0	0	0	38	58	0	0	0	0
2.5	0	0	0	49	100	0	0	0	0
3.0	0	0	0	71	186	0	0	0	0
3.5	0	0	0	39	131	0	0	0	0
4.0	0	0	0	22	59	0	0	0	0
4.5	0	0	0	4	15	0	0	0	0
5.0	0	0	0	1	1	0	0	0	0

Table 51 $BERT_{base\text{-}cased}$, 6 Layers Trainable, Clustered: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	0	0	0	0	0
1.5	0	0	0	0	4	0	0	0	0
2.0	0	0	0	0	20	6	0	0	0
2.5	0	0	0	0	85	16	0	0	0
3.0	0	0	0	0	232	73	0	0	0
3.5	0	0	0	0	135	59	0	0	0
4.0	0	0	0	0	86	36	0	0	0
4.5	0	0	0	0	17	7	0	0	0
5.0	0	0	0	0	4	3	0	0	0

Table 52 $BERT_{base\text{-}cased}$, 6 Layers Trainable, Clustered: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	1	2	0	0	0	0
1.5	0	0	0	0	3	0	0	0	0
2.0	0	0	0	0	67	7	0	0	0
2.5	0	0	0	3	134	15	0	0	0
3.0	0	0	0	1	207	34	0	0	0
3.5	0	0	0	2	149	29	0	0	0
4.0	0	0	0	1	90	16	0	0	0
4.5	0	0	0	0	17	1	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 53 $BERT_{base\text{-}cased}$, 6 Layers Trainable, Clustered: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	2	1	0	0	0	0
1.5	0	0	0	2	4	0	0	0	0
2.0	0	0	0	29	66	0	0	0	0
2.5	0	0	0	46	133	2	0	0	0
3.0	0	0	0	27	161	5	0	0	0
3.5	0	0	0	24	145	7	0	0	0
4.0	0	0	0	10	81	5	0	0	0
4.5	0	0	0	2	21	2	0	0	0
5.0	0	0	0	1	7	0	0	0	0

Table 54 $BERT_{base\text{-}cased}$, 6 Layers Trainable, Clustered: Convention

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	0	0	0	0
1.5	0	0	0	1	4	0	0	0	0
2.0	0	0	0	30	53	2	0	0	0
2.5	0	0	0	16	133	6	0	0	0
3.0	0	0	0	15	202	23	0	0	0
3.5	0	0	0	5	141	22	0	0	0
4.0	0	0	0	2	79	15	0	0	0
4.5	0	0	0	0	22	4	0	0	0
5.0	0	0	0	1	5	1	0	0	0

Table 55 $BERT_{base\text{-}cased}$, 12 Layers Trainable, Clustered: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	2	0	0	0	0
1.5	0	0	0	0	6	0	0	0	0
2.0	0	0	0	1	59	4	0	0	0
2.5	0	0	0	4	145	13	0	0	0
3.0	0	0	0	5	196	18	0	0	0
3.5	0	0	0	0	176	22	0	0	0
4.0	0	0	0	1	81	11	0	0	0
4.5	0	0	0	0	26	9	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 56 $BERT_{base\text{-}cased}$, 12 Layers Trainable, Clustered: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	3	1	0	0	0	0
1.5	0	0	0	2	3	0	0	0	0
2.0	0	0	1	22	73	0	0	0	0
2.5	0	0	2	21	126	0	0	0	0
3.0	0	0	2	38	217	0	0	0	0
3.5	0	0	1	27	142	0	0	0	0
4.0	0	0	0	17	64	0	0	0	0
4.5	0	0	0	2	17	0	0	0	0
5.0	0	0	0	0	2	0	0	0	0

Table 57 $BERT_{base\text{-}cased}$, 12 Layers Trainable, Clustered: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	3	1	0	0	0
1.5	0	0	0	0	9	17	0	0	0
2.0	0	0	0	0	36	65	0	0	0
2.5	0	0	0	3	87	215	0	0	0
3.0	0	0	0	0	59	135	0	0	0
3.5	0	0	0	0	28	94	0	0	0
4.0	0	0	0	0	5	19	0	0	0
4.5	0	0	0	0	2	5	0	0	0
5.0	0	0	0	0	3	1	0	0	0

Table 58 $BERT_{base\text{-}cased}$, 12 Layers Trainable, Clustered: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	1	2	0	0	0	0
1.5	0	0	0	0	3	0	0	0	0
2.0	0	0	0	4	59	11	0	0	0
2.5	0	0	0	4	131	17	0	0	0
3.0	0	0	0	2	204	36	0	0	0
3.5	0	0	0	4	157	19	0	0	0
4.0	0	0	0	1	91	15	0	0	0
4.5	0	0	0	0	17	1	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 59 $BERT_{base\text{-}cased}$, 12 Layers Trainable, Clustered: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	2	1	0	0	0	0
1.5	0	0	0	0	6	0	0	0	0
2.0	0	0	0	10	81	4	0	0	0
2.5	0	0	0	9	169	3	0	0	0
3.0	0	0	0	8	180	5	0	0	0
3.5	0	0	0	9	164	3	0	0	0
4.0	0	0	0	7	87	2	0	0	0
4.5	0	0	0	1	24	0	0	0	0
5.0	0	0	0	0	8	0	0	0	0

Table 60 $BERT_{base\text{-}cased}$, 12 Layers Trainable, Clustered: Conventions

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	0	0	0	0
1.5	0	0	0	0	5	0	0	0	0
2.0	0	0	0	3	65	17	0	0	0
2.5	0	0	0	6	121	28	0	0	0
3.0	0	0	0	3	179	58	0	0	0
3.5	0	0	0	4	125	39	0	0	0
4.0	0	0	0	2	77	17	0	0	0
4.5	0	0	0	0	19	7	0	0	0
5.0	0	0	0	0	6	1	0	0	0

Table 61 $BERTweet_{base}$, 0 Layers Trainable, Clustered: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	1	0	0	0
1.5	0	0	0	0	3	3	0	0	0
2.0	0	0	0	0	37	27	0	0	0
2.5	0	0	0	1	106	55	0	0	0
3.0	0	0	0	0	153	66	0	0	0
3.5	0	0	0	0	133	65	0	0	0
4.0	0	0	0	0	69	24	0	0	0
4.5	0	0	0	0	19	16	0	0	0
5.0	0	0	0	0	3	1	0	0	0

Table 62 $BERTweet_{base}$, 0 Layers Trainable, Clustered: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	1	2	1	0	0	0
1.5	0	0	0	0	5	0	0	0	0
2.0	0	0	0	0	77	19	0	0	0
2.5	0	0	0	0	133	16	0	0	0
3.0	0	0	0	1	225	31	0	0	0
3.5	0	0	0	0	155	15	0	0	0
4.0	0	0	0	0	74	7	0	0	0
4.5	0	0	0	0	17	2	0	0	0
5.0	0	0	0	0	2	0	0	0	0

Table 63 $BERTweet_{base}$, 0 Layers Trainable, Clustered: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	3	0	0	0
1.5	0	0	0	0	9	17	0	0	0
2.0	0	0	0	0	35	66	0	0	0
2.5	0	0	0	1	130	174	0	0	0
3.0	0	0	0	1	79	114	0	0	0
3.5	0	0	0	0	47	75	0	0	0
4.0	0	0	0	0	6	18	0	0	0
4.5	0	0	0	0	5	2	0	0	0
5.0	0	0	0	0	1	3	0	0	0

Table 64 $BERTweet_{base}$, 0 Layers Trainable, Clustered: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	3	0	0	0	0
1.5	0	0	0	1	2	0	0	0	0
2.0	0	0	0	0	53	21	0	0	0
2.5	0	0	0	1	102	49	0	0	0
3.0	0	0	0	0	171	71	0	0	0
3.5	0	0	0	1	138	41	0	0	0
4.0	0	0	0	0	74	33	0	0	0
4.5	0	0	0	0	13	5	0	0	0
5.0	0	0	0	0	2	2	0	0	0

Table 65 $BERTweet_{base}$, 0 Layers Trainable, Clustered: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	3	0	0	0	0
1.5	0	0	0	0	6	0	0	0	0
2.0	0	0	0	1	85	9	0	0	0
2.5	0	0	0	2	167	12	0	0	0
3.0	0	0	0	1	167	25	0	0	0
3.5	0	0	0	0	154	22	0	0	0
4.0	0	0	0	0	81	15	0	0	0
4.5	0	0	0	0	21	4	0	0	0
5.0	0	0	0	0	8	0	0	0	0

Table 66 $BERTweet_{base}$, 0 Layers Trainable, Clustered: Convention

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	0	0	0	0
1.5	0	0	0	2	3	0	0	0	0
2.0	0	0	0	1	67	17	0	0	0
2.5	0	0	0	0	125	30	0	0	0
3.0	0	0	0	0	190	50	0	0	0
3.5	0	0	0	0	130	38	0	0	0
4.0	0	0	0	0	69	27	0	0	0
4.5	0	0	0	0	20	6	0	0	0
5.0	0	0	0	0	5	2	0	0	0

Table 67 $BERTweet_{base}$, 6 Layers Trainable, Clustered: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	1	0	0	0
1.5	0	0	0	0	6	0	0	0	0
2.0	0	0	0	0	52	12	0	0	0
2.5	0	0	0	2	137	23	0	0	0
3.0	0	0	0	0	185	34	0	0	0
3.5	0	0	0	0	175	23	0	0	0
4.0	0	0	0	0	80	13	0	0	0
4.5	0	0	0	0	31	4	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 68 $BERTweet_{base}$, 6 Layers Trainable, Clustered: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	4	0	0	0	0
1.5	0	0	0	0	5	0	0	0	0
2.0	0	0	0	0	89	7	0	0	0
2.5	0	0	0	0	141	8	0	0	0
3.0	0	0	0	0	242	15	0	0	0
3.5	0	0	0	0	153	17	0	0	0
4.0	0	0	0	0	78	3	0	0	0
4.5	0	0	0	0	15	4	0	0	0
5.0	0	0	0	0	1	1	0	0	0

Table 69 $BERTweet_{base}$, 6 Layers Trainable, Clustered: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	2	2	0	0	0
1.5	0	0	0	0	6	20	0	0	0
2.0	0	0	0	0	25	76	0	0	0
2.5	0	0	0	0	90	215	0	0	0
3.0	0	0	0	0	52	142	0	0	0
3.5	0	0	0	0	24	98	0	0	0
4.0	0	0	0	0	6	17	1	0	0
4.5	0	0	0	0	1	6	0	0	0
5.0	0	0	0	0	2	2	0	0	0

Table 70 $BERTweet_{base}$, 6 Layers Trainable, Clustered: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	3	0	0	0	0
1.5	0	0	0	1	2	0	0	0	0
2.0	0	0	0	0	67	7	0	0	0
2.5	0	0	0	0	122	30	0	0	0
3.0	0	0	0	0	188	54	0	0	0
3.5	0	0	0	0	129	51	0	0	0
4.0	0	0	0	0	77	30	0	0	0
4.5	0	0	0	0	7	11	0	0	0
5.0	0	0	0	0	3	1	0	0	0

Table 71 BERTweet$_{base}$, 6 Layers Trainable, Clustered: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	3	0	0	0	0
1.5	0	0	0	2	4	0	0	0	0
2.0	0	0	0	8	86	1	0	0	0
2.5	0	0	0	12	167	2	0	0	0
3.0	0	0	0	4	181	8	0	0	0
3.5	0	0	0	6	161	9	0	0	0
4.0	0	0	0	2	92	2	0	0	0
4.5	0	0	0	1	24	0	0	0	0
5.0	0	0	0	0	8	0	0	0	0

Table 72 BERTweet$_{base}$, 6 Layers Trainable, Clustered: Conventions

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	0	0	0	0
1.5	0	0	0	0	5	0	0	0	0
2.0	0	0	0	4	76	5	0	0	0
2.5	0	0	0	4	137	14	0	0	0
3.0	0	0	0	1	195	44	0	0	0
3.5	0	0	0	3	122	43	0	0	0
4.0	0	0	0	2	65	29	0	0	0
4.5	0	0	0	0	20	6	0	0	0
5.0	0	0	0	0	3	4	0	0	0

Table 73 BERTweet$_{base}$, 12 Layers Trainable, Clustered: Cohesion

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	2	0	0	0	0
1.5	0	0	0	0	5	1	0	0	0
2.0	0	0	0	0	42	22	0	0	0
2.5	0	0	0	0	98	64	0	0	0
3.0	0	0	0	0	107	112	0	0	0
3.5	0	0	0	0	106	92	0	0	0
4.0	0	0	0	0	44	49	0	0	0
4.5	0	0	0	0	11	24	0	0	0
5.0	0	0	0	0	2	2	0	0	0

Table 74 BERTweet$_{base}$, 12 Layers Trainable, Clustered: Syntax

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	2	2	0	0	0	0
1.5	0	0	0	1	4	0	0	0	0
2.0	0	0	0	2	93	1	0	0	0
2.5	0	0	0	0	145	4	0	0	0
3.0	0	0	0	0	245	12	0	0	0
3.5	0	0	0	1	162	7	0	0	0
4.0	0	0	0	0	75	6	0	0	0
4.5	0	0	0	0	16	3	0	0	0
5.0	0	0	0	0	2	0	0	0	0

Table 75 BERTweet$_{base}$, 12 Layers Trainable, Clustered: Vocabulary

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	3	0	0	0
1.5	0	0	0	1	10	15	0	0	0
2.0	0	0	0	0	35	65	1	0	0
2.5	0	0	0	0	90	215	0	0	0
3.0	0	0	0	0	52	142	0	0	0
3.5	0	0	0	0	21	101	0	0	0
4.0	0	0	0	0	5	19	0	0	0
4.5	0	0	0	0	1	6	0	0	0
5.0	0	0	0	0	1	3	0	0	0

Table 76 BERTweet$_{base}$, 12 Layers Trainable, Clustered: Phraseology

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	3	0	0	0	0
1.5	0	0	0	0	3	0	0	0	0
2.0	0	0	0	0	62	12	0	0	0
2.5	0	0	0	0	125	27	0	0	0
3.0	0	0	0	0	191	51	0	0	0
3.5	0	0	0	0	139	41	0	0	0
4.0	0	0	0	0	81	26	0	0	0
4.5	0	0	0	0	13	5	0	0	0
5.0	0	0	0	0	4	0	0	0	0

Table 77 BERTweet$_{base}$, 12 Layers Trainable, Clustered: Grammar

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	3	0	0	0	0
1.5	0	0	0	0	6	0	0	0	0
2.0	0	0	0	0	92	3	0	0	0
2.5	0	0	0	1	168	12	0	0	0
3.0	0	0	0	0	158	35	0	0	0
3.5	0	0	0	0	141	35	0	0	0
4.0	0	0	0	0	77	19	0	0	0
4.5	0	0	0	0	16	9	0	0	0
5.0	0	0	0	0	7	1	0	0	0

Table 78 BERTweet$_{base}$, 12 Layers Trainable, Clustered: Conventions

	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0	0	0	0	1	0	0	0	0
1.5	0	0	0	2	3	0	0	0	0
2.0	0	0	0	1	77	7	0	0	0
2.5	0	0	0	1	129	25	0	0	0
3.0	0	0	0	0	193	47	0	0	0
3.5	0	0	0	1	121	46	0	0	0
4.0	0	0	0	0	67	29	0	0	0
4.5	0	0	0	0	21	5	0	0	0
5.0	0	0	0	0	4	3	0	0	0