Do English Language Learner Students Write Like Published Authors?

Iris Lew, Srila Maiti
iris.lew@berkeley.edu , srilamaiti@berkeley.edu

Abstract

Previous Automated essay scoring (AES) experiments have shown that models that are trained from scratch performed on par or better than pre-trained transformer models, leading us to investigate whether models that rely less on pre-training and whether models that have been pre-trained on a more informal data source (Twitter) would serve as better predictors of English Language Learners' (ELLs') essay scores when they are graded on six different essay components. We find that when we allowed BERT and a BERT-derived model (BERT-base-cased and BERTweet-base) to learn the training data by unfreezing layers allowed for them to predict a greater range of scores and thus performed better as shown by a lower MCRMSE (mean column-wise root mean squared error) score. The models were unable to learn the more extreme scores despite using K-means to cluster the data into low scores, average scores, and high scores and performing k-fold cross validation with our models the experimental models. Additionally, BERTweet-base did not perform better than BERT-base-cased regardless if we trained it on the entire range or clustered data, which implies that ELLs' essays contain a different vocabulary than Tweet vocabulary.

1 Introduction

Writing is complex and it is essential to grade all students fairly. Automated essay scoring (AES) is the task of employing natural language processing (NLP) technology to automatically assign scores to essays at scale. While it's controversial to use an AES system to systematically grade students, there has been encouragement to use it as a tool to help improve writing through automated feedback. [1] 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] 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 machine learning paradigm, the two more predominant architectures are Long Short Term Memory models (LSTMs) and Bidirectional Encoder from Transformer (BERT)[3] , with LSTMs generally performing on par or outperforming BERT (scores for LSTM models ranged from 72.7 to 83.0 percent whereas scores for BERT models ranged from 64.6 to 78.2 percent on the same dataset).[4] Although we expect pre-training to perform better than models that were only able to learn from the training dataset, these studies lead us to suspect that the pre-training is not able to boost model performance because the pre-training is not suited for the dataset. If the data that the BERT

[1] Mark D. Shermis and Jill Burstein. 2013.

[2] Mark D. Shermis and Jill Burstein. 2013.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019.

[4] Ridha Hussein Chassab. 2021.

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[5], which is a set of essays that are written by students from Grade 7-10 in English. The LSTMs model and BERT models both generally perform between 70 to 85 percent, with LSTMs performing on par or outperforming BERT models which forgets significant contextual information that impact the scoring.[6] In one of the BERT studies [7] , they converted their documents to GLoVe word embeddings to use in their model and found that the GLoVe word embeddings were not performing so well because they lacked access to the specific word-based features, implying that their BERT model was unable to learn the context, which is further supported by another study[8] where a Bag-of-Super-Word-Embeddings model achieved 78.8 percent accuracy on the same task with the ASAP dataset. Using a pre-trained BERT impacted accuracy but with fewer parameters, accuracy increased slightly.[9] Because we expect BERT to perform better than bidirectional LSTMs and a Bag-of-Super-Word-Embeddings as they can solve "out of vocabulary" problems [10] , we hypothesize the pre-training from the BERT models did not suit the dataset. LSTMs and bag-of-words models are trained from scratch as opposed to BERT which is trained from Wikipedia and Google's BooksCorpus[11] , thus we believe that the pre-training for BERT is lowering the accuracy and the style of writing found in Wikipedia and the books in the BooksCorpus is different from student writing styles. Students generally do not have the most refined writing. Thus, it is possible that by overwriting the training 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.

Humans are able to evaluate a piece of writing sentence-by-sentence as well as holistically, but many AES systems are unable to replicate what humans can do and the most common type of automatic feedback is at the sentence-level rather than at the entire essay-level.[12] For example, while it is possible to parse, assess, and correct grammar [13] sentence-by-sentence, it is not the same as assigning a score holistically. Essays contain many sentences which should be evaluated together, thus demonstrating a need for context.

Additionally, even though many features within a text tend to correlate with one another,[14] 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.[15]

Furthermore, the quality of the essays written by ELLs will be different than that of the native English speakers. 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

[5] https://www.kaggle.com/c/asap-aes
[6] Ridha Hussein Chassab. 2021.
[7] Elijah Mayfield and Alan W Black. 2020.
[8] Madalina Cozma, Andrei M Butnaru, and Radu Tudor Ionescu. 2018.
[9] Ridha Hussein Chassab. 2021.
[10] Ridha Hussein Chassab. 2021.
[11] https://huggingface.co/blog/bert-101
[12] Mark D. Shermis and Jill Burstein. 2013.
[13] Hui- Hsien Feng, Aysel Saricaoglu, and Evgeny Chukharev-Hudilainen. 2016.
[14] Mark D. Shermis and Jill Burstein. 2013.
[15] Ridha Hussein Chassab. 2021.

content.[16] We propose that BERTweet[17], which is pre-trained on Tweets where people do not have to follow any grammatical rules as long as they stick within the given character limit when posting, would perhaps perform better since there are greater variety of sentences as opposed to the English found in books and Wikipedia.

A study which uses linear regression to predict scores, but even after correcting for imbalanced data, the improvement in assigning scores wasn't much.[18] Thus, 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 (e.g., a score of 5.0 is better than a score of 4.5). If these 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 Methods

Vanderbilt University and the Learning Agency Lab made a dataset of 8^{th}-12^{th} grade ELLs' argumentative essays available through a Kaggle competition.[19] 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. These scores ranged from 1.0 to 5.0 in 0.5 increments, with a higher number reflecting a higher proficiency in that area. The scores fall into an approximately normal distribution for each component (Appendix A).

We will use our models to predict a score for each essay component and then 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.

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

In this case, the ground truth target columns are the six essay components. 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) to produce an adjusted MCRMSE to align the range of possible scores in 0.5 increments.

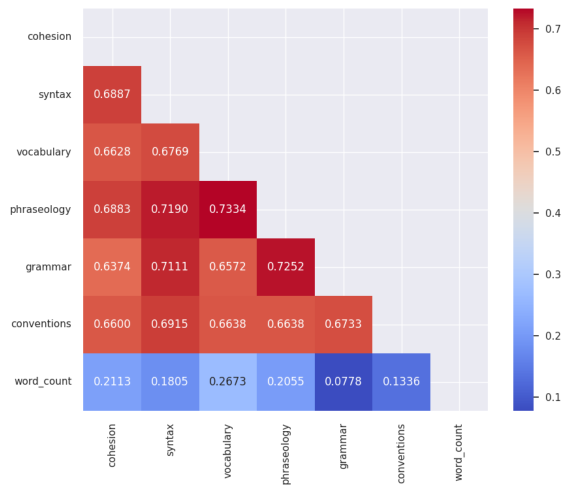
$$Score_{adjusted} = round\left(\frac{\hat{y}}{0.5}\right) \times 0.5 \quad (2)$$

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. The model will then randomly 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 so that we would use only the first 512 tokens produced by the BERT tokenizer and first 128 token produced by the BERTweet tokenizer. When we compared the scores that the human graders provided to the word count, we observed low correlations ($r = 0.0778 - 0.2673$ as indicated by Figure 1); therefore, we believe that it would be reasonable to use only the max number of tokens that each model can consume.

[16] Mark D. Shermis and Jill Burstein. 2013.
[17] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020
[18] Ridha Hussein Chassab. 2021.
[19] https://www.kaggle.com/competitions/feedback-prize-english-language-learning/data

Figure 1 Correlation Matrix



3.1 Models

Each model uses their respective tokenizers to represent the input text. We passed the CLS token generated from the tokenizers through the pre-trained models, followed by a fully connected neural network which finally feeds to a linear regression output layer with a custom MCRMSE loss function for weight optimization.

3.2 Experiment Settings

We ran a BERT-base-cased model with various sets of hyperparameters to choose a set. We ultimately settled on training the model with ten epochs, a batch size of eight, a learning rate of 0.00001, a validation split of 0.2, dropout of 0.1, and two hidden layers with 64 nodes.

3.3 Baseline

For our baseline, we used the BERT-base-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 experiment with unfreezing the layers and using BERTweet-base so that the model can learn from the training set.

4 Results and Discussion

We calculate the following adjusted MCRMSE scores for the different models from our test dataset of 783 records:

Table 1 Adjusted MCRMSE Scores

	BERT-base-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-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 and supported the idea that relying less on pre-training would improve model performance on AES tasks. Contrary to our expectations, BERTweet-base generally did not perform better than BERT-base-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 the 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 score. We see that the percentage that is predicted correctly for all the components was increasing as more layers were unfrozen.

Table 2 Percentage of Test Dataset Records That Were Correctly Predicted Per Essay Component by BERT-base-cased (Score Within a Range of ±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%)

Overall, while the BERT-base-cased models were predicting the correct score between 27.7

percent to 44.2 percent of the dataset, when we expanded to see if the score was 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 for all the key performance indices within a range of 1.5 and 4.5 (crosstabs between predicted and actual scores found in Appendix C). 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 (see Appendix A), 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 a Range of ±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%)

Overall, the BERTweet-base models predicted scores correctly between 25.0 to 39.2 percent. 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, but the differences 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. Similar to BERT-base-cased, when examined whether the predicted score was within half of its actual value, the BERTweet-base models were predicting the correct score the majority of the time (70.5 percent to 86.7 percent) with the most accurate scores for the completely unfrozen model and that it was only able to predict the more extreme scores (from 2.5 and 4.0 to 1.0 and 4.5) when all its layers were unfrozen (crosstabs between predicted and actual scores found in Appendix B).

BERTweet-base generally performed the same or worse than BERT-base-cased. The BERTweet-base models generally produced similar MCRMSE scores and failed to produce a better score on most of the analytic measures when the model was completely frozen and on all of them when the model was half-unfrozen (difference of 0.2 to 2.5 percentage points). The completely unfrozen BERTweet-base model was able to correctly predict its score more accurately than its BERT-base-cased counterpart by 1.8 and 3 percentage points on cohesion and phraseology respectively, but was unable to outperform it on syntax (-2.2 percentage points), vocabulary (-5 percentage points), phraseology (-8.7 percentage points), and conventions (-2.5 percentage points). 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-cased is a better model than BERTweet-base and do not recommend using one model over another to predict scores on certain essay components.

4.1 Clustering and K-Fold Cross Validation

Since both BERT-base-cased and BERTweet-base struggled to generate more extreme predictions and we wanted to make sure we do our experiments holistically, we investigated whether we could use stratified k-folds cross-validation along clusters to run the same experiments in order for the models to learn the lowest and highest scores.

There are high correlations among all the components with the lowest correlation being an $r = 0.6374$ (see Figure 1). 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 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), divided the data into three clusters of scores (6.0-17.0, 17.5-21.5, and 21.5-30.0) using K-means and the elbow method, and performed the same experiments using k-folds cross-validation. We decided to use two k-folds (we decided to split the sample into two as we were testing to see if this would cause the model to perform better). We achieved the following adjusted MCRMSE scores on the same test dataset:

Table 4 Adjusted MCRMSE Scores for the Clustered Models

	BERT-base-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 indicating that we were unable to teach the models to recognize different clusters of scores.

Table 5 Percentage of Test Dataset Records That Were Correctly Predicted Per Essay Component by BERT-base After Clustering (Score Within a Range of ±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%)

The BERT-base-cased model was only able to predict the score between 24.5 to 37.2 percent of the time when the data was clustered and failed to show any improvement in producing the correct score as the layers unfroze and were allowed to learn the clustered training data except vocabulary when it went from 0 to 6 trainable layers with a difference of 0.3 percentage points as shown in Table 5. These three models were only able to predict a score between 2.0 and 3.5 regardless of how many layers were unfrozen (see Appendix C).

Table 6 Percentage of Test Dataset Records That Were Correctly Predicted Per Essay Component by BERTweet-base After Clustering (Score Within a Range of ± 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%)

When all the layers were frozen and when the last six layers were unfrozen, the BERTweet-base model was only predicting scores between 2.5 and 4.0 (crosstabs between predicted and actual scores found in Appendix C). The model was unable to predict the correct score well (between 24.4 percent and 33.1 percent) and generally well when predicting scores within 0.5 of what the actual score was (between 70.0 and 81.2 percent), we did not see vast improvement in scores as the layers unfroze (see Table 6).

In comparing the "clustered" BERT-base-cased model with its BERTweet-base counterpart, we noticed that when the layers were completely frozen, the BERT-base-cased model performed better on all analytic measures by 0.6 to 7.6 percentage points, but BERTweet-base performed better on syntax, vocabulary, phraseology, grammar, and convention (from 0.3 to 2.1 percentage points) when the model was completely unfrozen. Due to the magnitude of the points difference, we could not conclude whether BERT-base-cased or BERTweet-base was better after clustering the data and performing k-folds cross-validation. As the "clustered" models did not predict any score of 1.0 or 5.0, we would need to examine if there was something wrong in its implementation or if another method could be used to fix the class imbalance in its predictions.

5 Conclusion and Future Work

We see that the BERT-base-cased model and BERTweet-base models performed the best when all their layers were unfrozen. This signals that relying less on the pre-trained data would behoove automated essay evaluation systems. BERTweet-base did not perform better than BERT-base-cased holistically (as shown by the higher MCRMSE score) or correctly score the essays 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. Even when we clustered the scores into low, average, and high scores, both models still struggled to predict the low and high scores regardless. To improve the models, we would like to use a larger dataset such as TOEFL essays and examine if we could use weights to fix the class imbalance so that the model would predict the extreme edge cases.

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

Hui- Hsien Feng, Aysel Saricaoglu, and Evgeny Chukharev-Hudilainen. 2016. Calico Journal (Online), 33(1): 49-70. https://www.jstor.org/stable/calicojournal.33.1.49

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

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

https://huggingface.co/blog/bert-101

A Score Distribution

From the training dataset, we observe that each of the essay components follow a normal distribution.

Figure 2 Distribution of Cohesion Scores in the Training Dataset

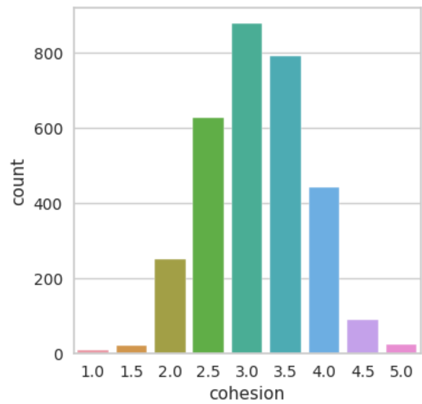


Figure 3 Distribution of Cohesion Scores in the Training Dataset

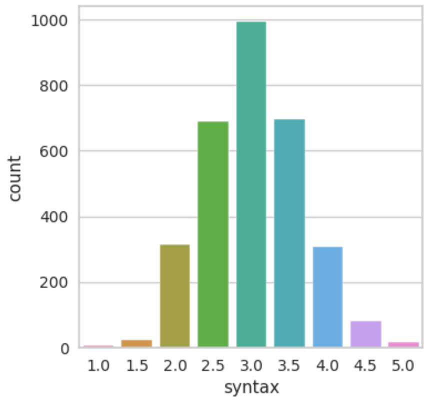


Figure 4 Distribution of Vocabulary Scores in the Training Dataset

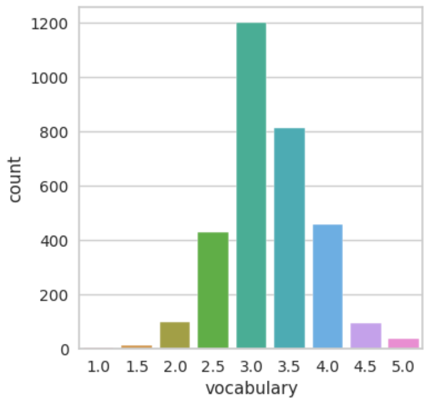


Figure 5 Distribution of Phraseology Scores in the Training Dataset

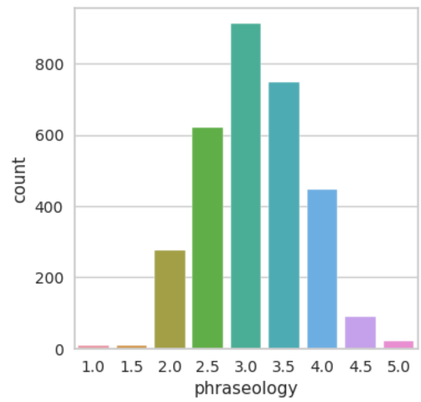


Figure 6 Distribution of Grammar Scores in the Training Dataset

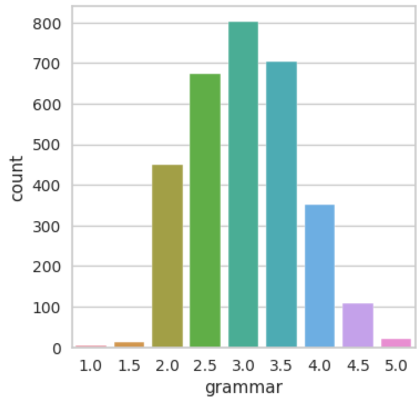
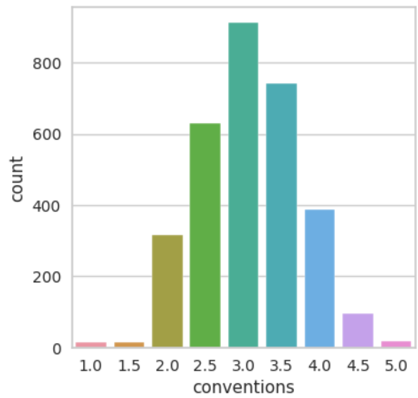


Figure 7 Distribution of Conventions Scores in the Training Dataset



B Crosstabs

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

Table 7 BERT-base-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-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-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 BERT-base-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-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-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-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-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-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-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-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-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-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-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-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-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-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-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, 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-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-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-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-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-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-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-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-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-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-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-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-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-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-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-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-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-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-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