All for One and One for Each: Comparing BERT and DeBERTa Models for Multi-Dimensional Automated Essay Scoring

Alexander Carite, **Kurt Eulau**, and **Thomas Welsh**
UC Berkeley School of Information
alex.carite@berkeley.edu, keulau@berkeley.edu and twelsh@berkeley.edu

Abstract

Although research has shown that pre-trained BERT models improve automated essay scoring (AES), most earlier studies assessed quality with a single "holistic" score. However, multi-dimensional scoring is more valuable to teachers and students because it better pinpoints opportunities for writing improvement. This paper examines the relative effectiveness of BERT-based models across six different rubric dimensions using a new dataset from Vanderbilt University and the Learning Agency Lab which contains essays written by 8th-12th Grade English Language Learners (ELLs). While we expected fine-tuned BERT models to outperform more traditional AES methods and surpass our baselines, we also found that models which predicted individual rubric traits collectively (all-in-ones) outperformed collections of siloed models that focused on one rubric trait each (one-for-each). Additionally, we found that DeBERTa models, with the help of the enhanced masked decoder, further strengthen all-in-one and one-for-each models.

1 Introduction and Background

English language learners require timely and accurate feedback on essays in order to measure progress towards fluency. Ideally, students receive constructive feedback indicating which particular areas require greater focus. Manually grading essays with comprehensive rubrics with multiple traits, however, is a laborious process that prevents timely evaluation and is subject to bias and grader fatigue (Taghipour 2017).

Academic researchers have seen the potential value of AES since the 1960s (Page, 1966; Ajay et al., 1973), starting with assigning a single score to evaluate essay quality. However, the desire for educators to give more insightful feedback to students has driven natural language processing (NLP) researchers to publish papers that go beyond holistic scoring and focus on more granular dimensions of total essay quality, such as coherence (Higgins et al., 2004; Somasundaran et al., 2014), relevance to prompt (Louis and Higgins, 2010; Persing and Ng, 2014), argument strength (Persing and Ng, 2015) and grammar (Heilman et al., ACL 2014) just to name a few. However, across the AES field of study, researchers have not yet agreed on a standard set of dimensions or a standard scale for judging quality.

The development of self-attention-based transformer architectures, and in particular Bidirectional Transformers for Language Understanding (BERT) (Vaswani et al., 2017), have brought greater performance to many areas in computational linguistics. As a result, AES models leveraging BERT have recently overtaken models using Long Short-Term Memory (LSTM) (Hochreiter et al., 1997) as the current state of the art (Devlin et al., 2019; Uto et al., 2020; Ridley et al., 2021). Other current model architectures incorporate a combination of BERT and LSTM architectures and use the BERT CLS token to characterize the entire essay (Wang et al., 2022). However, while the performance of AES systems are still not sufficient for educators to favor them over traditional methods, at the time of writing, no publicly accessible study has made use of improved BERT frameworks like DeBERTa (He et al., 2020) for multi-dimensional AES while contrasting these tools in all-for-one or one-for-each configurations.

The objective of this paper is to address this shortfall, by performing a series of experiments that compare the relative abilities of different BERT-based models to predict the essay scores across six different rubric domains using mean column wise root mean squared error (MCRMSE), which averages the root mean squared errors across each rubric domain.

2 Data

Several foundations, in conjunction with Vanderbilt University and the educational non-profit The Learning Agency Lab, published the ELLIPSE dataset via Kaggle in 2022. Students wrote argumentative essays in response to a variety of different prompts, setting up a cross-prompt multi-trait AES scoring task (see Ridley 2021 for taxonomy of AES tasks). To generate the labels, two teachers evaluated each essay on all six rubrics, scores were averaged, and supervisors addressed essays with highly disparate ratings. Contrary to other AES datasets, notably the canonical ASAP set (Hewlett 2012), essay graders did not assess the ELL essays on the students' ability to make a persuasive argument, but rather to demonstrate command of the English language.

According to the Learning Agency Lab, the ELLIPSE corpus "comprises 3911 argumentative essays written by 8th-12th grade ELLs. The essays were scored according to six analytic measures: cohesion, syntax, vocabulary, phraseology, grammar, and conventions…each measure represents a component of proficiency in essay writing, with greater scores corresponding to greater proficiency in that measure. The scores range from 1.0 to 5.0 in increments of 0.5" (Feedback 2022). Each essay was scored across all domains by two raters and final scores for each domain were computed as an average of the grade from each scorer. Post-scoring multi-facet Rasch analysis ensured reliability in raters, texts, and scales (Boone 2016). Essay scores were distributed along a normal curve, with few essays scoring in the extremes. Two thirds of essays, for example, had an average score of at least 2.5 and at most 3.5. Prior to modeling, we divided the available data into train, validation, and test sets using a stratified random sampling method using frequency counts of individual traits across each rubric to execute a 60/20/20 split of the data. This technique ensured that extreme high and low scores, which appeared less frequently in our dataset, were distributed as equally as mathematically possible among the training, validation, and test sets.

3 Preprocessing

In the dataset, human readers manually anonymized identifiable information in the essays, including names of students, schools, and cities. We standardized this anonymization process so that the replacement words were consistent across essays. We also filtered out select non-alphanumeric characters that were artifacts of the data curation process. Further preprocessing was deemed unnecessary as the dataset was otherwise relatively well-formatted.

4 Performance Metric

MCRMSE coalesces the overall performance of a model by giving equal weight to each of the rubric traits, which was a natural choice in the absence of evidence for unequal weighting. Here we present the formula where N_t is the number of scored ground truth target columns, y and $\hat{y}$ are the actual and predicted values, respectively.

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

5 Experiments & Results

After preprocessing our text inputs and selecting a performance metric, we developed three families of models:

1. Baseline experiments that relied on handcrafted features similar to the approach taken by Uto 2020.
2. The most performant BERT and DeBERTa models that were fine-tuned for our particular task.
3. Additional BERT and DeBERTa models that were used to illustrate alternative approaches.

A table of the baseline models, subsequent transformer-based models, and accompanying MCRMSE for the final, held-out test sets run with the same number of epochs and batch size follow below. All

DeBERTa models employed DeBERTaV3 (He et al., 2021).

Model	MCRMSE
Baseline Experiments	
Guess the Mode	0.6618
Linear Regression	0.5809
XGBoost Regression	0.5535
Main Experiments	
BERT Nominal Classification	0.6289
BERT One-For-Each Regression	0.5919
DeBERTa One-For-Each Regression	0.5012
BERT All-In-One Regression	0.3937
DeBERTa All-In-One Regression	**0.3767**
Additional Experiments	
Frozen BERT All-In-One Regression	0.5080
DeBERTa XS All-In-One Regression	0.3982
DeBERTa L All-In-One Regression	0.3847

Table 1: Experimental Results

5.1 Baseline Experiments

In Table 1, Guess the Mode simply predicts the most common class for each rubric trait, which was 3.0 for all rubric dimensions. For less naive baselines models, we engineered 36 text features using spaCy for our Linear Regression and XGBoost model (see Appendix A.2 Handcrafted Feature List for details) to further reduce MCRMSE.

5.2 Main Experiments

We expected that models relying solely on statistical features would underperform BERT-based models and would struggle to adequately address word meaning ambiguities and word order. However, Mayfield and Black (2020) found "that fine-tuning BERT produces similar performance to classical models at significant additional cost…while state-of-the-art strategies do match existing best results, they come with opportunity costs in computational resources." Results from initial explorations tended to support the view of Mayfield and Black, where a variety of BERT models performed no better than eXtreme Gradient Boosting (XGBoost). Framing our problem as a classification task, which we recorded as BERT Nominal Classification, did not improve performance either, even after extensive tuning.

Reframing our AES task as a regression problem, we hypothesized that training six independent (one-for-each) models would allow the model to hone in on the specific signals associated with each individual trait. For example, since cohesion and vocabulary represent distinct attributes of an essay, training models focused on each attribute could provide the latitude for the model to optimize for a specific trait. We also hypothesized that changing weights to optimize for cohesion could potentially alter weights utilized heavily for vocabulary predictions so models trained in parallel offered some potential advantages. However, developing six independent BERT models, one-for-each rubric trait, and then combining the results into an ensemble failed to make adequate progress, as evidenced by the MCRMSE listed for BERT One-For-Each Regression. An analogous approach similarly failed to make adequate progress towards lowering the topline metric.

The source of greatest progress came from "going all-in" with DeBERTa All-In-One Regression, as evidenced by its 0.3767 MCRMSE. Contrary to the one-for-each models where models were trained in parallel for each rubric trait, these all-in-one models predicted all six traits at once from the same network with a simple six node regression dense layer. Since essays that scored well in cohesion also tended to score highly in vocabulary, we hypothesize that the all-in-one models tended to perform better because one trait is highly predictive of another such that mutually reinforcing or shared signals propagate through a shared architecture, boosting performance. For example, an essay with excellent cohesion is also likely to have advanced vocabulary, and thus feedback to improve vocabulary performance would also improve that of cohesion and performance and vice versa.

Within the realm of all-in-one models, DeBERTa performed better than BERT. Since the release of the DeBERTa paper in 2020, the DeBERTa architecture has reached new state-of-the-art levels in many NLP tasks.

The experimental results shown in Table 1 uphold this hypothesis regarding the advantage of DeBERTa vis-à-vis BERT.

We also note two hyperparameter optimizations that demonstrated significant diminutions in MCRMSE and were applied across all BERT and DeBERTa experiments: fine-tuning via unfreezing all transformer layers and setting batch size to one.

Merchant et al. (2020) illustrated the effect of progressively unfreezing BERT layers on several canonical NLP tasks, finding that "fine-tuning—as currently practiced—is a conservative process: largely preserving linguistic features, affecting only a few layers, and specific to domain examples" and that "there appears to be room for improvement" by unfreezing more than just the top few layers. Our results align with their analysis, where unfreezing all transformer layers translated to smaller residuals. One potential explanation for this phenomenon relates to the differences between the pre-training text and the student essays. Pre-training on sources like Wikipedia will produce embeddings typical of that type of writing. High school ELL students writing essays in the very recent past, however, are likely to write more similarly to each other than to the pre-training text. The word use, syntactic structures, and other text features from a relatively specific sub-group differ meaningfully from Wikipedia or the Toronto Book Corpus, which the original BERT model used during training (Devlin et al., 2019). Allowing the model to adjust to the idiosyncrasies of the sub-group in question could account for the benefits in unfreezing all transformer layers.

We hypothesize that a batch size of one improves the performance of the model for similar reasons. Small batch sizes allow for more frequent backpropagation and weight adjustments. In cases where the held out sets are dissimilar from the training set, such an approach can lead to over-fitting. Here, however, argumentative essays written by a particular student population are similar to each other. Essays tend to be written about the same set of topics (should students wear uniforms, should schools mandate that all students take art, etc.) and are composed in a single relatively structured fashion, as taught by teachers who adhere to a common set of learning standards. While the fact that essays are more similar to each other than they are to text on Wikipedia has implications for broader generalizability, we optimized models for the task at hand.

5.3 Additional Experiments

Frozen BERT All-In-One Regression highlights the impact of fine-tuning, where the non-fine tuned model failed to perform as well as the fine-tuned model for the same number of epochs. Tests comparing DeBERTa XS (extra-small) and DeBERTa L (large) to the default DeBERTa embedding size and number of transformer layers demonstrated that a smaller network does lead to worse results but that a large network does not necessarily lead to better results.

Other experiments not included in the table either failed to demonstrate advancement or even worsened performance. The most performant all-in-one models utilized the last hidden layer of embeddings rendered by the transformer architecture along with a global average pooling layer, normalization layer, and finally a simple, six node dense prediction layer. Adding on any combination of few or many and/or narrow or wide dense layers after transformer outputs or convolution neural networks only hampered the model's ability to make accurate predictions. Max pooling, dropout, or the use of CLS token similarly failed to progress the model. Splitting essays longer than 512 tokens provided limited benefit. Models listed above used a truncated set of the first 512 tokens, though future research could explore alternatives by bootstrapping sections of the given essays from the current corpus to augment the dataset.

6 Discussion

Automated essay scoring, especially multi-prompt cross-trait scoring, remains an unsolved problem. Our findings both echo previous research and provide new insights.

Handcrafted features vs. BERT vs. DeBERTa
Non-transformer machine learning methods to automate essay scoring proved less capable than approaches employing BERT and/or DeBERTa, but only after numerous trials demonstrated the effectiveness of fine-tuning and low batch sizes for those transformer models. Ultimately DeBERTa outperformed BERT, with

a top MCRMSE scores of 0.3767 and 0.3937, respectively. The residuals of particular rubric traits illustrate why DeBERTa exhibits better overall performance. Figure 1, for example, shows the average absolute value of the residuals for phraseology grouped by the true label essay score. Both models struggle with very high and low scores, but DeBERTa struggles less. See Appendix A.4 and A.5 for similar plots of all rubric traits.

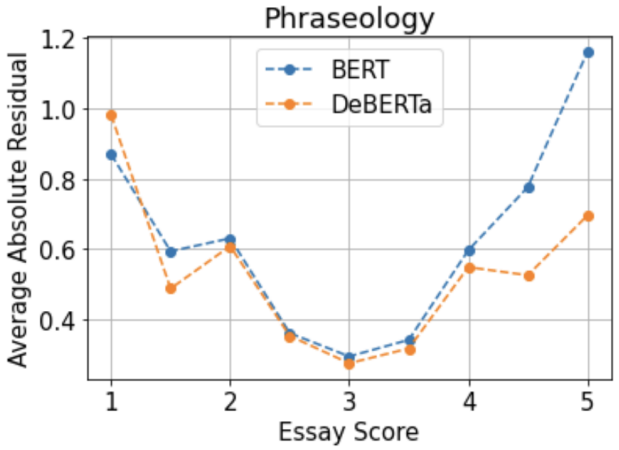


Figure 1: Mean of the absolute values of the residuals for phraseology grouped by label for best BERT and DeBERTa models

One hypothesis for why DeBERTa is better able to score phraseology in general is due to its enhanced mask decoder in which both absolute and relative positions of the words are taken into account instead of just the absolute position embeddings that BERT monitors. He et al. (2020) argued in their original DeBERTa paper that BERT's strategy of using the absolute position vectors in the input layer hinders the ability for the model to appropriately learn information regarding relative positions. DeBERTa incorporates absolute position vectors after the transformer blocks and before the softmax layer for the predicting masked tokens (Figure 2). In doing this, DeBERTa gathers the relative position information from all transformer blocks and only uses absolute position as complimentary information when predicting masked words. This understanding of relational position is critical in assessing the phraseology rubric. Lexicon bundles, a component of phraseology, is a prime example of why DeBERTa can assess how these words are used in relation to one another.

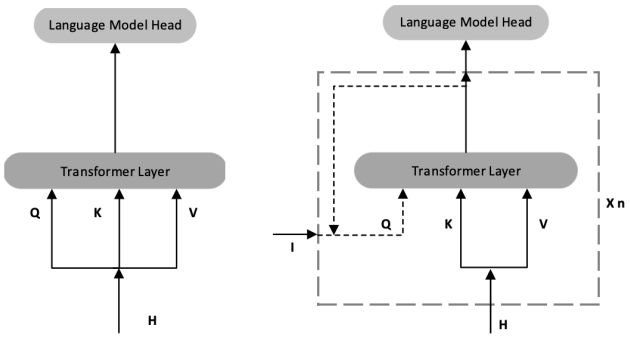


Figure 2: BERT vs DeBERTa decoding comparison (He et al., 2020)

The same hypothesis holds true for the grammar rubric trait. Consider the sentence, "*Frustrated, the chairs took the boy forever to set up"*. This grammatically incorrect sentence can be better understood if the relative position of the modifier *frustrated* was closer to its subject *boy.* The more grammatically correct way to use this sentence would be, "The boy was frustrated by how long the chairs took to set up". DeBERTa is much more capable of understanding the local relationship between the tokens and thus can score the grammar rubric more accurately.

Regression vs. classification Since essay rubric scores are discrete, ordinal response variables, treating AES as a classification task. Similar to Berggren et al. (2019), however, we find that regression approaches yield better results than classification frameworks.

All-in-one vs. one-for-each Models that predict individual rubric traits collectively (all-in-one) outperform a collection of siloed models that have the latitude to focus on one rubric trait each (one-for-each). We hypothesize that "going all-in" on a single large model produces better results because rubric traits for a particular essay tend to be highly correlated with one another.

Bias-variance trade-off Models that were given the latitude to adjust away from pre-training weights in BERT and DeBERTa managed to make gains in performance on both the training, validation, and test sets. Given the differences between the text that ELL students produce in American classrooms and the pre-training text that transformer models learn on,

extensive fine-tuning to a particular essay corpus yielded meaningful progress. Dependent on generalizability concerns, optimizing for lower bias with less emphasis on overfitting presents at least one perspective to inform subsequent research.

Given the challenges in the field, significant future research is needed to advance AES. In particular, we advocate for partnerships between ELL teachers and school districts and technical researchers seeking to make progress in this area. Research partnerships with primary stakeholders such as teachers could yield benefits for both the model development cycle and the generation of labeled datasets for two reasons.

First, researchers with the ability to construct and test deep learning networks might not be able to distinguish the text features that explain differences in specific grades. For example, what are the exact, ground-truth reasons why one essay earns a 3.0 for a trait like grammar while another earns a 3.5? See Appendix A.3 for example essay excerpts. Since we found it difficult to differentiate between factors that lead to particular essay scores, it became difficult to instruct the model on how to proceed after viewing model results. The inability to grade essays accurately ourselves hampered the iterative process.

Second, there exists an opportunity to produce a large, expert-labeled dataset because hundreds of thousands of ELL and English teachers are already paid to grade essays. In some fields, obtaining enough quality labeled data is the single largest impediment to progress. In this instance, manually labeled data is being produced everyday, that data is just not standardized or collected. An enterprising research lab or educational tech startup could partner with school districts or some other educational organization to standardize rubrics, normalize grading practices, and ensure expert scorers save all grades to a repository for future research.

7 Conclusion

Experiments using all-in-one and one-for-each architectures with both BERT and DeBERTa demonstrate that the most performant approach is to use an all-in-one DeBERTa regression model to predict scores for multi-trait essays written by high school ELLs. Meaningful hyperparameter considerations include fine-tuning DeBERTa on the training set and using small batch sizes.

Acknowledgments

We thank Mark Butler, Natalie Ahn, and Peter Grabowski for their patient guidance and thoughtful suggestions which greatly helped to improve this paper.

References

ELLIPSE Dataset: https://www.kaggle.com/competitions/feedback-prize-english-language-learning/data (2022)

Ajay, H.B., 1973. Strategies for content analysis of essays by computer. University of Connecticut.

Berggren, S.J., Rama, T. and Øvrelid, L., 2019, August. Regression or classification? automated essay scoring for Norwegian. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 92-102).

Boone, W.J., 2016. Rasch analysis for instrument development: Why, when, and how?. CBE—Life Sciences Education, 15(4), p.rm4.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Feedback Prize English Language Learning. https://www.kaggle.com/competitions/feedback-prize-english-language-learning (2022)/

He, P., Liu, X., Gao, J. and Chen, W., 2020. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

He, P., Gao, J. and Chen, W., 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543.

Hewlett ASAP Dataset at https://www.kaggle.com/c/asap-aes (2012)

Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M. and Tetreault, J.R., 2014, January. Predicting grammaticality on an ordinal scale. In ACL (2).

Higgins, D., Burstein, J., Marcu, D. and Gentile, C., 2004. Evaluating multiple aspects of coherence in student essays. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004 (pp. 185-192).

Mayfield, E. and Black, A.W., 2020, July. Should you fine-tune BERT for automated essay scoring?. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 151-162).

Merchant, A., Rahimtoroghi, E., Pavlick, E. and Tenney, I., 2020. What happens to bert embeddings during fine-tuning?. arXiv preprint arXiv:2004.14448.

Page, E.B., 1966. The imminence of... grading essays by computer. The Phi Delta Kappan, 47(5), pp.238-243.

Persing, I. and Ng, V., 2014, June. Modeling prompt adherence in student essays. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1534-1543).

Persing, I. and Ng, V., 2015, July. Modeling argument strength in student essays. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 543-552).

Ridley, R., He, L., Dai, X.Y., Huang, S. and Chen, J., 2021, May. Automated cross-prompt scoring of essay traits. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 15, pp. 13745-13753).

Somasundaran, S., Burstein, J. and Chodorow, M., 2014, August. Lexical chaining for measuring discourse coherence quality in test-taker essays. In Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical papers (pp. 950-961).

Taghipour, K., 2017. Robust trait-specific essay scoring using neural networks and density estimators (Doctoral dissertation, National University of Singapore (Singapore)).

Uto, M., Xie, Y. and Ueno, M., 2020, December. Neural automated essay scoring incorporating handcrafted features. In Proceedings of the 28th International

Conference on Computational Linguistics (pp. 6077-6088).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Wang, Y., Wang, C., Li, R. and Lin, H., 2022. On the Use of BERT for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation. arXiv preprint arXiv:2205.03835.

Abbreviations

AES: automated essay scoring
BERT: Bidirectional Transformers for Language Understanding
ELL: English language learner
DeBERTa: Decoding-enhanced BERT with Disentangled Attention
LSTM: Long-Short Term Memory
MCRMSE: mean column-wise root mean squared error
NLP: natural language processing
RoBERTa: Robustly Optimized BERT Pretraining Approach

Appendix

A.1 Rubric

The original essay scoring rubric shared by the dataset curator is shown below

English Proficiency Scoring Rubric for English Language Learners

	HOLISTIC	ANALYTIC					
	Overall	**Cohesion**	**Syntax**	**Vocabulary**	**Phraseology**	**Grammar**	**Conventions**
5	Native-like facility in the use of language with syntactic variety, Appropriate word choice and phrases; well-controlled text organization; precise use of grammar and conventions; rare language inaccuracies that do not impede communication.	Text organization consistently well controlled using a variety of effective linguistic features such as reference and transitional words and phrases to connect ideas across sentences and paragraphs; appropriate overlap of ideas.	Flexible and effective use of a full range of syntactic structures including simple, compound, and complex sentences; There may be rare minor and negligible errors in sentence formation.	Wide range of vocabulary flexibly and effectively used to convey precise meanings; skillful use of topic-related terms and less common words; rare negligible inaccuracies in word use.	Flexible and effective use of a variety of phrases, such as idioms, collocations, and lexical bundles, to convey precise and subtle meanings; rare minor inaccuracies that are negligible.	Command of grammar and usage with few or no errors.	Consistent use of appropriate conventions to convey meaning; spelling, capitalization, and punctuation errors nonexistent or negligible.
4	Facility in the use of language with syntactic variety and range of words and phrases; controlled organization; accuracy in grammar and conventions; occasional language inaccuracies that rarely impede communication.	Organization generally well controlled; a range of cohesive devices used appropriately such as reference and transitional words and phrases to connect ideas; generally appropriate overlap of ideas.	Appropriate use of a variety of syntactic structures, such as simple, compound, and complex sentences; occasional errors or inappropriateness in sentence formation.	Sufficient range of vocabulary to allow flexibility and precision; appropriate use of topic-related terms and less common lexical items.	Appropriate use of a variety of phrases, such as idioms, collocations, and lexical bundles; occasional inaccuracies and colloquialisms.	Minimal errors in grammar and usage.	Generally consistent use of appropriate conventions to convey meaning; spelling, capitalization, and punctuation errors few and not distracting.
3	Facility limited to the use of common structures and generic vocabulary; organization generally controlled although connection sometimes absent or unsuccessful; errors in grammar and syntax and usage. Communication is impeded by language inaccuracies in some cases.	Organization generally controlled; cohesive devices used but limited in type; Some repetitive, mechanical, or faulty use of cohesion within and/or between sentences and paragraphs.	Simple, compound, and complex syntactic structures present although the range may be limited; some apparent errors in sentence formation, especially in more complex sentences.	Minimally adequate range of vocabulary for the topic; no precise use of subtle word meanings; topic related terms only used occasionally; attempts to use less common vocabulary but with some inaccuracy.	Evident use of phrases such as idioms, collocations, and lexical bundles but without much variety; some noticeable repetitions and misuses.	Some errors in grammar and usage.	Developing use of conventions to convey meaning; errors in spelling, capitalization, and punctuation that are sometimes distracting.
2	Inconsistent facility in sentence formation, word choice, and mechanics; organization partially developed but may be missing or unsuccessful. Communication impeded in many instances by language inaccuracies.	Organization only partially developed with a lack of logical sequencing of ideas; some basic cohesive devices used but with inaccuracy or repetition.	Some sentence variation used; many sentence structure problems.	Narrow range of vocabulary to convey basic and elementary meaning; topic related terms used inappropriately; errors in word formation and word choice that may distort meanings.	Narrow range of phrases, such as collocations and lexical bundles, used to convey basic and elementary meaning; many repetitions and /or misuses of phrases.	Many errors in grammar and usage.	Variable use of conventions; spelling, capitalization, and punctuation errors frequent and distracting.
1	A limited range of familiar words or phrases loosely strung together; frequent errors in grammar (including syntax) and usage. Communication impeded in most cases by language inaccuracies.	No clear control of organization; cohesive devices not present or unsuccessfully used; presentation of ideas unclear.	Pervasive and basic errors in sentence structure and word order that cause confusion; basic sentence errors are common.	Limited vocabulary often inappropriately used; limited control of word choice and word forms; little attempt to use topic-related terms.	Memorized chunks of language, or simple phrasal patterns predominate; many repetitions and misuses of phrases.	Errors in grammar and usage throughout.	Minimal use of conventions; spelling, capitalization, and punctuation errors throughout.

Rubric Key Terms and Definitions

Phrase: Multiple word units

Grammar: The rules by which words change their forms, including the use of word classes and grammatical morphology in English. Word classes include prepositions, pronouns, nouns, verbs, etc… Grammatical morphology includes third person, plural, possessive, etc…

Syntax: Structuring sentences according to syntactic rules related to coordinating clauses, developing syntactic phrases (noun, verb, preposition phrases), phrasal and clausal dependency, and transformations such as passives, relative clauses, and negations.

Cohesive device: Cohesive devices are used as links between two or more items (e.g., words, phrases, clauses) in a text to enhance text cohesion. These include the use of conjunctions (and, but, if, on the other hand), transitions (first, next, finally, for example), repetition of words, phrases, and ideas across sentences and paragraphs, and the use of anaphor (pronouns replacing nouns).

Simple, complex, and compound sentences :

- Simple: Independent clauses
- Complex: Independent and dependent clauses
- Compound: Two of more independent clauses

Chunks: Multiple words that combine to have a single meaning. Often memorized without knowing what the individual words mean (e.g., “How are you” for “Hello”)

Lexical bundles: Multiple word units that are common in English but are not idiomatic (“There is”). More common than collocations

Collocations: Two or more words that are often used together (e.g., save time, go to bed, fast food)

Idioms: multi-word unit where meaning not deducible from those of the individual words (kick the bucket, rain cats and dogs)

A.2 Handcrafted Feature List

Using spaCy, we created a variety of text features to use in models leveraging linear regression and XGBoost. The parts of speech features are the counts of each part of speech, where the parts of speech are implemented by spaCy and defined by the Universal Dependencies framework. Specifically, spaCy uses the default Universal POS tags V2. More information can be found at: https://universaldependencies.org/u/pos/

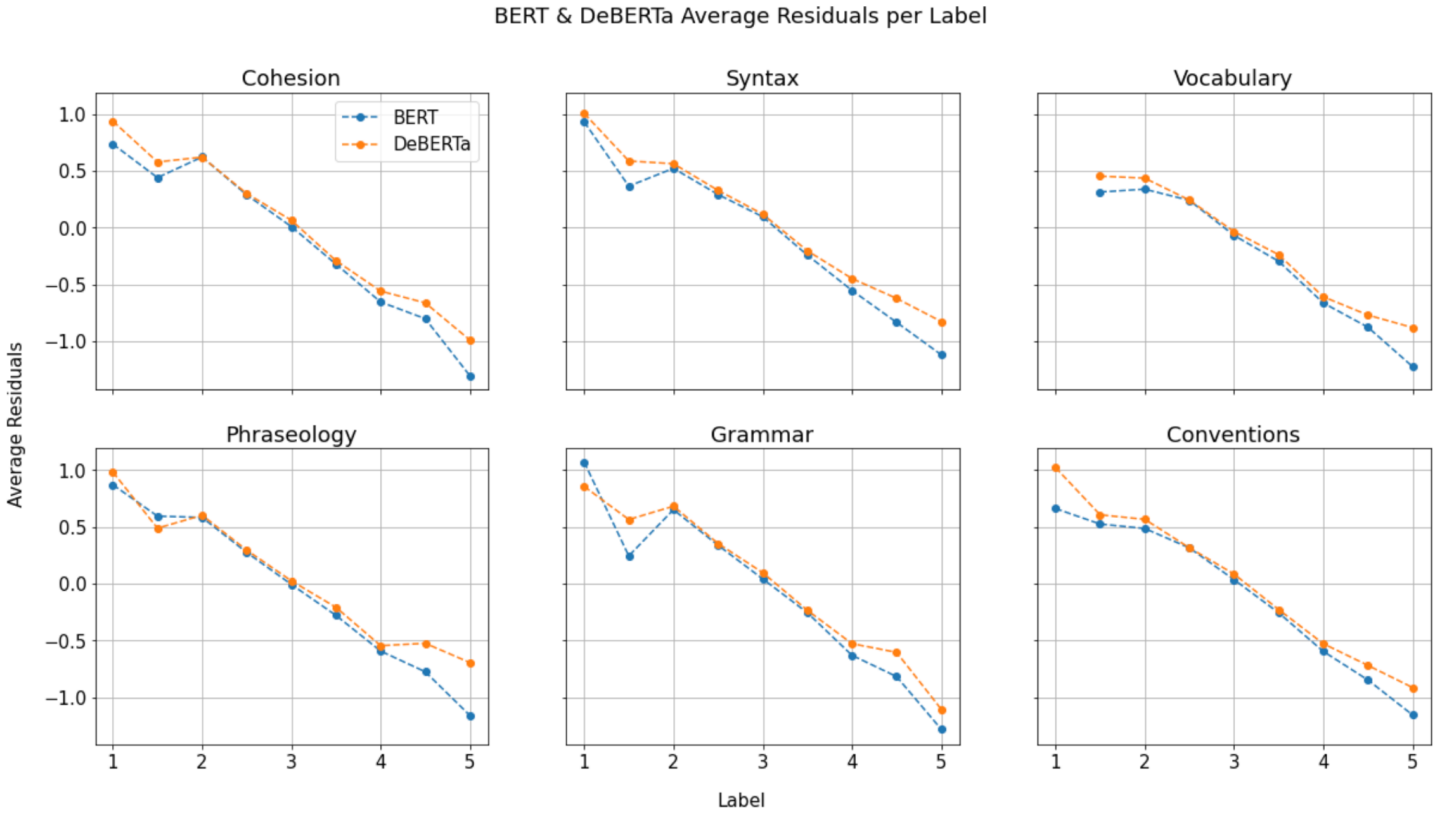
Parts of Speech Features	**Other Features**
ADJ	character_count
ADP	contraction_count
ADV	flesch_kincaid_score
AUX	mean_sentence_length
CCONJ	mean_word_length
DET	paragraph_count
INTJ	polarity
NOUN	punctuation_count
NUM	sentence_count
PART	stopwords
PRON	subjectivity
PROPN	syllable_count
PUNCT	title_count
SCONJ	variance_sentence_length
SPACE	variance_word_length
SYM	vocabulary_size
VERB	word_count
X	

A.3 Excerpts from Example Training Set Essays

The table below displays excerpts from example essays taken from the training set. The essays are arranged by average rubric score across the six rubric traits. For example: an essay with scores of 1, 2, 2, 3, 2, and 2 would have an average score of 2.

Avg. Score	Raw Excerpt
1	he is a good because they are the prescient and us. now the are more procession a other prescient and us. because and the more many education for student and school. the more school for one because he have a one women the have many education for a other because the good more the one is information for access because is not like the brazen giant of Greek fame with conquering one nation for is the imprison the eyes command the one book for people the us give more your lied your huddled Farmer with silent lips the more the one is for you because the are a one solution for you and a other people he have a more the one solution for you here ancient lands your storied pound Aries she with silent lips pledge of a in order people he have a education
3	I believe that is not a good idea for a student to finish high school in three year. The reason for this is because they might have a change of heart in what they want to be. Also having four year of high school gives them time to see if they really made a good choices on their career. Also "will they go along with take courses during the summer?" when it time to have fun,relax and enjoy the hot nice weather with family and friends. These are one of the reason that its not good idea. Another way to look at this is to find out if there going to stick with the career they pick. For example, when pick a job everyone look for better pay and benefits to see if the best fit for them . For a student it is totally different because they are look for better pay instead of look for a fitting job they will like an enjoy and stay at . Instead they look at a job that didn't suit them, then just quit looking and jumping from job to job.
3.5	It better working as a groups because the teacher will think that we will benefit more thing as a project or classwork. And also that we will work together and talk as a groups for the teacher because working alone is hard to think for the one question that you don't know what is about. So many of students think working as a groups is better will help each other However, teacher will thinks many from us that will improve from talking each other and together. Because the student will do more stuff and talk a lot also as a groups and alone because many student will do that to get focus and pay attention the class. This why that people prefer working with a group is better. Because the reason that we want working as a group is that we will work together and help out the question to talk and give a convince from the work.
5	I agree with Michelangelo's statement as I have found through experience that it benefits me more to set high expectations and not reach my goal, rather than settling on a low goal and achieving it. When setting high goals, I find that I learn more and progress my abilties further than I do with a lower goal. This is because setting high goals requires confidence, challenges, and pride. Firstly, hard work is required to achieve high goals, and requires confidence. Confidence involves having trust and believing that you are capable of accomplishing something. I have found that when I set high goals, I feel more confident in my abilities. With lower goals however, my confidence is lower as I settle for the easier path and do not trust myself to do better. With a higher aim, my confidence motivates me to not give up. I tend to try harder, and always believe in myself. For example, at school, I had to choose whether I wanted to try out for the varsity tennis team or remain in the club team. The varsity team was a higher reach, and required confidence in my abilities for me to try out. The club team was a lower reach, as I knew that I could simply continue with it. I decided to try out for the varsity team, and was confident in my abilities. I practiced hard and did not give up. Despite not making it onto the team, I found that I actually greatly improved my abilities through the confidence I had gained, and tried harder than I had ever before. I managed to progress my skills, and was able to play more confidently on the club team and excel. Therefore, setting high aims requires confidence, which is a very beneficial characteristic to have in life.

A.4 Residuals Plots



A.5 Absolute Residuals Plots