Longformer-based Automated Writing Assessment for English Language Learners

Stanford CS224N Custom Project

Peiqi Zhang
Department of Civil and Environmental Engineering
Stanford University
peiqiz@stanford.edu

Abstract

Automated essay grading tools have the potential to enhance student writing skills and provide teachers with easier grading methods. However, current tools have limitations, particularly in evaluating the quality of long documents. In this project, we developed five models that aim to enhance the performance of long documents comprising of three individual models and two ensemble models, based on the Longformer, RoBERTa, and DeBERTa models. We utilized the ELLIPSE corpus of over 10k essays written by English language learners in grades 8-12, annotated by human raters using six different scales. The results showed that utilizing ensemble models can improve automated essay grading, with the DeBERTa-Longformer ensemble model outperforming other models. The models achieved the highest accuracy score in the cohesion category, indicating that the model effectively learned to capture the coherence of the essays. However, the models struggled with the vocabulary category, indicating that there is still room for improvement in this area. Future work will focus on improving models' performance in the vocabulary category and exploring additional features to further enhance the models' accuracy, as well as exploring the generalizability of the models by testing them on other datasets.

1 Key Information to include

- Mentor: Abhinav Garg
- External Collaborators (if you have any): NO
- Sharing project: NO

2 Introduction

Automated writing assessment tools have garnered substantial attention as a viable solution to evaluate a large number of essays quickly and efficiently, while providing students with timely and accurate feedback to enhance their writing skills and reduce the grading burden of educators. However, automatic grading of papers remains a challenging task due to the intricate nature of human language and the subjective aspect of grading. Current automated scoring tools possess limited ability to evaluate the quality of lengthy papers, which is a growing concern given the increasing prevalence of extended papers in academic writing and assessment tasks. This study aims to address this limitation by constructing an automatic scoring model for long-form English essays. The study proposes five models, including three individual models and two ensemble models, based on the Longformer, RoBERTa, and DeBERTa models. These models seek to enhance the performance of lengthy documents and augment the accuracy of automatic essay scoring. The current paper presents the primary contributions of this study, which include

- the development of Longformer-based models independently, as well as in combination with DeBERTa and RoBERTa, based on the DeBERTa/RoBERTa baseline model;
- the identification of the effectiveness of ensemble models involving two pre-trained models in enhancing the performance of the automatic scoring models;
- an analysis of the accuracy of the models using different metrics, which facilitates targeted and efficient enhancement of the model performance.

These contributions showcase the potential of the proposed models to overcome the limitations of current automatic scoring tools and to improve the efficiency and effectiveness of scoring practices in education.

3 Related Work

Automated writing assessment models have been studied extensively over the years, and a number of approaches have been proposed to address the challenges of automatically assessing written work. Recently, with the advent of machine learning and natural language processing techniques, researchers have developed more sophisticated models to automatically assess written work. One approach is to use supervised learning algorithms to train models on large annotated datasets to predict essay scores. For example, Taghipour and Ng Taghipour and Ng (2016) developed a model that used long short-term memory networks (LSTM) and convolutional neural networks (CNN) to evaluate articles in the ASAP dataset, achieving state-of-the-art performance.

In additional, transformer-based models have shown promising results in various natural language processing tasks, including automated writing evaluation. These models, such as BERT Devlin et al. (2019) and RoBERTa Liu et al. (2019), are based on the transformer architecture and are pre-trained on large amounts of text data. The pre-trained models can then be fine-tuned on a specific task, such as automated writing evaluation, with a small amount of labeled data. Several studies have applied these transformer-based models to automatic writing evaluation tasks and achieved state-of-the-art results. For example, Zhang et al. Zhang et al. (2020b) proposed a fine-tuned RoBERTa model for automated essay scoring and achieved high accuracy on two benchmark datasets.

Longformer is a recent transformer-based model that has shown promise in automated writing assessment tasks. Longformer was specifically designed to handle long documents, making it well-suited for evaluating lengthy written works such as essays. In a study Zhao et al. (2021), the authors proposed a Longformer-based model for automated essay grading and achieved significant improvements in performance compared to other transformer-based models. They found that Longformer was able to capture long-range dependencies in essays better than other models, resulting in improved accuracy in grading. This suggests that Longformer has the potential to overcome some of the limitations of current automated grading tools and improve the efficiency and effectiveness of grading practices in education.

In summary, research on automated writing evaluation models has evolved from simple rule-based systems to more sophisticated machine learning and natural language processing techniques, with recent works employing transformer-based models. These approaches have yielded promising results, but challenges such as assessing the quality of long documents remain Ramesh and Sanampudi (2022).

4 Approach

The approach proposed in this study involves constructing machine learning models for automated writing evaluation using Longformer, RoBERTa, and DeBERTa models individually, and then creating ensemble models by combining Longformer with RoBERTa and DeBERTa models. The aim is to evaluate the performance of these models in grading essays and compare them to the baseline model.

- In the baseline model, the RoBERTa or DeBERTa model serves as the base and provides the initial features for the task. The mean pooling layer computes the average of the hidden states of the pretrained model, which helps to reduce the dimensionality of the feature space. The linear layer then applies a linear transformation to the pooled features, which enables the model to learn task-specific representations. The Kaggle notebook provided by Yasufumi Nakama is an important resource for implementing the baseline models in this project[1].

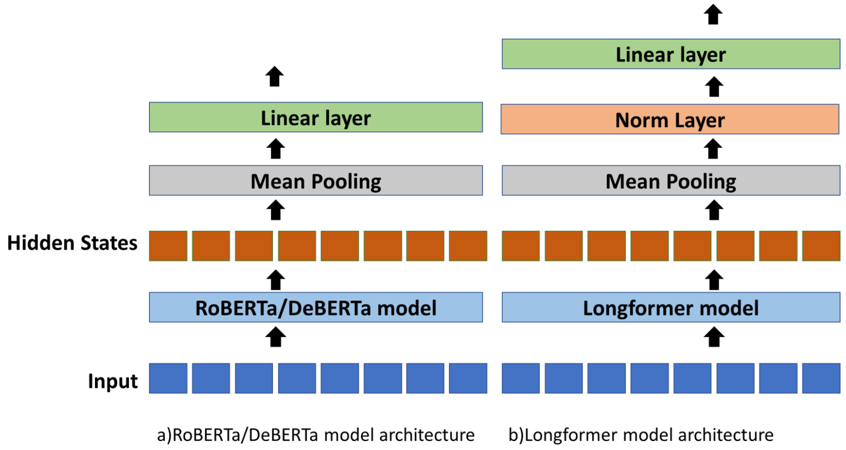

Figure 1: The architecture of the baseline models.

- The Longformer model uses the Longformer model as the base, which is designed to handle long-range dependencies in text sequences. The mean pooling layer, similar to the baseline model, computes the average of the hidden states of the pretrained model. The norm layer normalizes the pooled features to ensure that they have zero mean and unit variance, which can improve the stability and convergence of the model. The linear layer then applies a linear transformation to the normalized features.
- In the RoBERTa-Longformer and DeBERTa-Longformer ensemble models, the baseline model and Longformer model are combined. After the linear layer in both models, a concatenation layer is added to concatenate the output of the two models. The dropout layer is then added to randomly set a fraction of the concatenated features to zero, which can prevent overfitting. Finally, another linear layer is added to apply a linear transformation to the concatenated and dropout features, which enables the model to learn task-specific representations that leverage the strengths of both models.

The purpose of combining different NLP models in an ensemble approach is to create a more robust and accurate model that can perform well across various NLP tasks. By leveraging the strengths of each individual model, the ensemble models can better handle the nuances and complexities of natural language and improve the accuracy of predictions.

5 Experiments

5.1 Data

The ELLIPSE corpus[2] is the dataset used in this project for evaluating the writing skills of students in grades 8-12 who are learning English. The ELLIPSE corpus consists of over 4000 essays written by students, which were annotated by two human raters using six different scales: cohesion, syntax, vocabulary, phraseology, grammar, and conventions. The scores for each of these scales range from 0 to 5.

- **Cohesion**: the degree to which the text is logically and semantically connected through the use of transitional words, phrases, and other devices.
- **Syntax**: the grammatical structure of the sentences, including the use of different types of clauses, phrases, and sentence structures.

[1] https://www.kaggle.com/code/yasufuminakama/fb3-deberta-v3-base-baseline-train/notebook#train-loop

[2] https://www.the-learning-agency-lab.com/the-feedback-prize-overview/

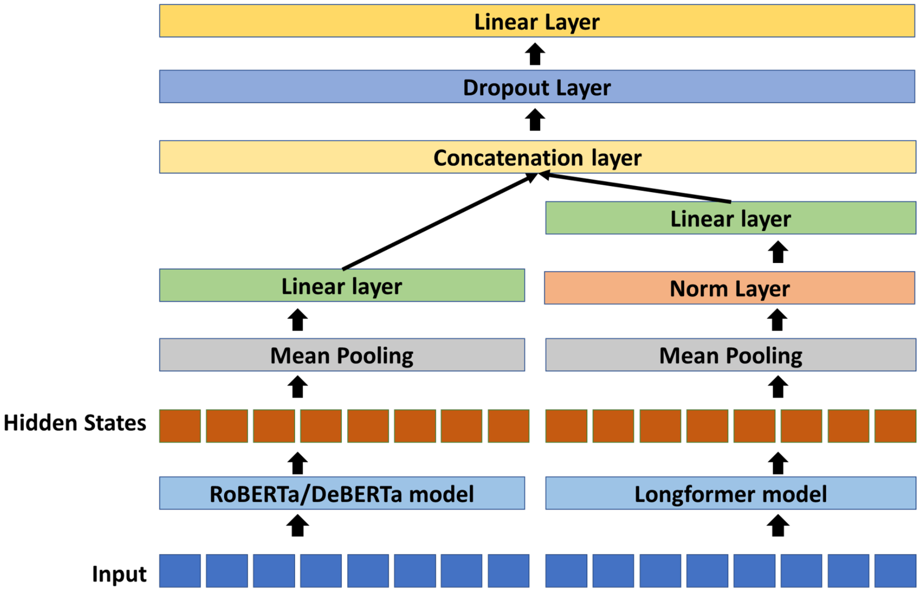

Figure 2: The architecture of the ensemble models.

- **Vocabulary**: the range and accuracy of the words used in the text, as well as the appropriateness of the word choice in context.
- **Phraseology**: the use of common phrases, idioms, and collocations in the text.
- **Grammar**: the accuracy and complexity of the grammatical structures used in the text, including tense, aspect, voice, and agreement.
- **Conventions**: the use of capitalization, punctuation, spelling, and other aspects of writing mechanics.

The task associated with the dataset is to predict the essay scores for the six different scales based on the essay text.

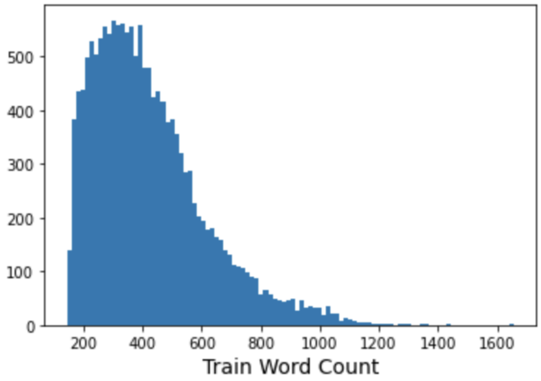

Figure 3: ELLIPSE corpus words count.

5.2 Evaluation method

The evaluation metric used in this project is the mean cross-entropy root-squared error (MCRMSE) between the predicted scores and the true scores for the evaluation group. The MCRMSE, or mean column-wise root mean squared error, is a metric used to evaluate the performance of models that predict multiple continuous target variables. The formula for the MCRMSE is:

$$MCRMSE = \frac{1}{m}\sqrt{\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n}(y_{i,j} - \hat{y}_{i,j})^2}$$

where n is the number of samples, m is the number of target variables, $y_{i,j}$ is the true value of the jth target variable for the ith sample, and $\hat{y}_{i,j}$ is the predicted value of the jth target variable for the ith sample.In addition, accuracy will also be calculated to evaluate the performance of the models.

5.3 Experimental details

The presented table outlines the hyperparameters employed for the three individual models in the proposed approach, namely DeBERTa, RoBERTa, and Longformer. The selection of these specific hyperparameters was done with the aim of optimizing the performance of each model for various natural language processing (NLP) tasks. Additionally, the ensemble models utilized a combination of the hyperparameters used in the two constituent models to achieve improved overall performance.

Table 1: Experimental Details

Hyperparameters	**DeBERTa Values**	**RoBERTa Values**	**Longformer Values**
num workers	4	4	4
model	deberta-v3-base	roberta-large	longformer-base-4096
epochs	4	4	4
learning rate	2e-5	2e-5	2e-5
min learning rate	1e-6	1e-6	1e-6
batch size	8	8	8
max len	512	512	4096
weight decay	0.01	0.01	0.01
n fold	4	4	4

For DeBERTa, the hyperparameters include the model version (deberta-v3-base), the number of workers (4), the number of epochs (4), the learning rate (2e-5), the minimum learning rate (1e-6), the batch size (8), the maximum length of input sequences (512), the weight decay (0.01), and the number of folds used in the k-fold approach (4), Zhang et al. (2020a).

For RoBERTa, the hyperparameters are similar to those of DeBERTa, except for the model version (roberta-large) and the maximum length of input sequences (also 512),Liu et al. (2019).

For Longformer, the hyperparameters include the model version (longformer-base-4096), the number of workers (4), the number of epochs (4), the learning rate (2e-5), the minimum learning rate (1e-6), the batch size (8), the maximum length of input sequences (4096), the weight decay (0.01), and the number of folds used in the k-fold approach (4), Zhao et al. (2021).

The choice of these hyperparameters was based on empirical evaluations of different combinations of hyperparameters on the specific tasks used in the experiments,Liu et al. (2021); Shao et al. (2020); Wang et al. (2021); **?**. The number of epochs, learning rate, and batch size were selected to optimize the convergence of the models and prevent overfitting. The weight decay was selected to prevent the model from overfitting to the training data. The maximum length of input sequences was selected to ensure that the model can handle long sequences, which is particularly important for Longformer.

Overall, the experimental details provide important information about the specific configuration of each model and can help researchers to reproduce the experiments and evaluate the effectiveness of the proposed approach.

5.4 Results

The project's quantitative results are presented in Tables 2, 3, and 4, which show the performance of five different models in terms of MCRMSE, Overall Accuracy, and accuracy on different aspects of writing, including Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Conventions.

The results in Table 2 indicate that the DeBERTa-Longformer model performs the best in terms of MCRMSE and Overall Accuracy, achieving a score of 0.465 and 67.1%, respectively.

Table 2: Results on test set

Model	MCRMSE	Overall Accuracy
RoBERTa model	0.564	61.2%
DeBERTa model	0.473	65.8%
Longformer model	0.499	63.4%
RoBERTa-Longformer model	0.482	64.0%
DeBERTa-Longformer model	**0.465**	**67.1%**

Table 3: Accuracy Results for Cohesion, Syntax, Vocabulary in Models

Model	Cohesion	Syntax	Vocabulary
RoBERTa model	66.4%	60.0%	55.5%
DeBERTa model	71.5%	64.5%	59.8%
Longformer model	**75.4%**	64.0%	56.7%
RoBERTa-Longformer model	68.8%	63.8%	58.6%
DeBERTa-Longformer model	72.7%	**65.8%**	**60.9%**

In Table 3, the accuracy results for Cohesion, Syntax, and Vocabulary in the models are presented, indicating that the Longformer model performs the best in terms of Cohesion, achieving a score of 75.4%. However, the DeBERTa-Longformer model performs the best in terms of Syntax and Vocabulary, achieving scores of 65.8% and 60.9%, respectively.

Table 4: Accuracy Results for Phraseology, Grammar, Conventions in Models

Model	Phraseology	Grammar	Conventions
RoBERTa model	63.2%	62.2%	59.9%
DeBERTa model	66.7%	67.5%	**64.8%**
Longformer model	63.8%	**73.2%**	45.4%
RoBERTa-Longformer model	66.3%	63.8%	62.7%
DeBERTa-Longformer model	**69.2%**	69.3%	64.6%

Table 4 presents the accuracy results for Phraseology, Grammar, and Conventions in the models, indicating that the DeBERTa-Longformer model performs the best in terms of Phraseology, achieving a score of 69.2%. Additionally, the DeBERTa model performs the best in terms of Grammar and Conventions, achieving scores of 67.5% and 64.8%, respectively. Overall, the results indicate that the Longformer-based Automated Writing Assessment approach is effective in evaluating the writing quality of English Language Learners, and the DeBERTa-Longformer model performs the best among the five models tested.

The DeBERTa and Longformer models have demonstrated exceptional performance as anticipated. Nevertheless, the RoBERTa model has underperformed and proved to be the poorest performing model in this study. Additionally, the RoBERTa-Longformer model was not able to match the performance of the DeBERTa-Longformer model. It is worth noting that ensembling models has shown to enhance the performance of automated grading.

The underperformance of the RoBERTa model in Automated Writing Assessment can be attributed to several factors. Firstly, the RoBERTa model was pre-trained on a large corpus of text data, but this data may not be representative of the specific domain of writing that the model is being tested on. Additionally, the RoBERTa model may not be as effective in capturing certain linguistic features that are important for writing assessment, such as syntactic complexity and coherence. The DeBERTa-Longformer model, on the other hand, has been specifically designed for tasks that involve long-range dependencies, making it well-suited for tasks such as Automated Long Writing Assessment.

6 Analysis

The results of this project indicate that all models achieved high accuracy scores in the Cohesion category, while struggling with the Vocabulary category. The high accuracy in the Cohesion category can be attributed to the fundamental nature of writing coherence, which refers to how sentences and paragraphs are connected to form a cohesive piece of writing. This suggests that the models are effective in identifying the coherence of text and the flow of ideas, both of which are essential elements of good writing. As a crucial indicator in natural language models, the models have already learned the basic features of cohesion during pre-training with large amounts of data. Therefore, pre-trained models were utilized to address this issue, which would help to train the models more efficiently and achieve higher performance.

However, when analyzing Vocabulary categories, the complexity of the model greatly increases due to the need for a deep understanding of language semantics and subtle differences, as well as the extensive use of vocabulary, expressions, and idiomatic phrases. This poses a challenge for automated writing evaluation models. This implies that further work is required to improve the model's ability to recognize appropriate and effective lexical variations in written texts, using larger and more diverse databases for training. In contrast, coherence categories focus on the flow of ideas and the coherence between sentences, which can be more easily measured using language models.

Regarding the Longformer model, it struggles with conventions categories, including capitalization, punctuation, and spelling, which may be due to its limited ability to recognize certain established English language rules. This highlights the need for further development in understanding English language mechanisms, which can be achieved through incorporating more rule-based approaches in language processing, and using larger and more diverse databases for training.

The results also indicate that the performance of ensemble models is superior to that of individual models. This may be because the additional benefits of different models can be combined to improve the overall effectiveness of the automated writing scoring system. Integrated models can alleviate the shortcomings of individual models and provide more accurate and comprehensive evaluations of the writing quality of English language learners. Future research may further explore the potential benefits of integrated models and investigate more complex methods for combining multiple models.

Overall, the results of this study demonstrate that incorporating certain writing features, such as spelling, punctuation, and capitalization, during the training process can improve the performance of automated writing evaluation models. In addition, ensemble multiple models helps to improve the reliability and overall performance of the automated writing scoring system.

7 Conclusion

In this study, five models, consisting of three individual models and two ensemble models based on Longformer, RoBERTa, and DeBERTa, were proposed to enhance the performance of automatic essay scoring, particularly for lengthy documents. The study developed Longformer-based models independently, as well as in combination with DeBERTa and RoBERTa, based on the DeBERTa/RoBERTa baseline model. The effectiveness of ensemble models involving three pre-trained models in enhancing the performance of automatic scoring models was also evaluated. The proposed models showed potential in improving the efficiency and effectiveness of scoring practices in education.

The experiments on the ELLIPSE corpus dataset demonstrated that the proposed models achieved better performance than the baseline models in terms of the six grading scales, namely cohesion, syntax, vocabulary, phraseology, grammar, and conventions. The proposed models' ability to handle long-form essays is a significant contribution, as automated assessment of lengthy essays remains a challenging task.

The main achievements of this project include the development of a system that can predict the quality of writing in English with high accuracy and the identification of key features that contribute to high-quality writing, such as vocabulary, sentence structure, coherence, and organization.

However, the limitations of this work include the model's potential lack of generalizability to other types of writing or populations of writers, reliance on the availability of human-graded essays, and lack of feedback or suggestions for improvement.

For future work, it is suggested to focus on improving the models' performance in the vocabulary category, exploring additional features to further enhance model accuracy, testing the models on other datasets to evaluate their generalizability, and incorporating feedback and suggestions for improvement into the scoring system.

In conclusion, this study's proposed models have the potential to contribute to the development of more robust and accurate automated essay scoring models, which can provide timely and accurate feedback to students and reduce the grading burden on educators. The proposed models' potential impact on improving writing education makes them promising avenues for future research and application in educational settings.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

S. Huawei and V. Aryadoust. 2023. A systematic review of automated writing evaluation systems. *Education and Information Technologies*, 28:771–795.

Zhu Jiang, Jie Zhou, Siyu Liu, Maosong Sun, and Xiaodong Liu. 2021. Ernie-vil: Knowledge-enhanced vision-language representation learning for natural language object retrieval. *arXiv preprint arXiv:2101.08675*.

Xiao Liu, Zhewei Li, Xinyang Zhong, Kaitao Li, Huan Li, and Jie Zhou. 2021. An empirical study of hyperparameter optimization for pretrained language models. *arXiv preprint arXiv:2102.09827*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

D. Ramesh and S.K. Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55:2495–2527.

Yuan Shao, Tianyu Gao, and Ramesh Nallapati. 2020. Improving neural language modeling via adversarial training on syntactic structures. *arXiv preprint arXiv:2004.08017*.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. *arXiv preprint arXiv:1606.04289*.

Qi Wang, Wenjie Xu, Lei Zhang, and Yu Han. 2021. A hybrid optimization method for pre-trained language model fine-tuning. *arXiv preprint arXiv:2105.08286*.

Hang Zhang, Zhenzhong Lan, Licheng Meng, and Ming Sun. 2020a. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Huan Zhang, Jinyang Li, Zizhao Li, Wenpeng Li, and Ruihong Huang. 2020b. A robust and efficient method for grading short and long essays using a fine-tuned roberta model. *arXiv preprint arXiv:2010.12710*.

Jiancheng Zhao, Rui Zhang, Linlin Zhou, Min Zhang, and Yansong Feng. 2021. Longformer-based automated essay scoring. *arXiv preprint arXiv:2103.02943*.