DeBERTa with hats makes Automated Essay Scoring system better

Shixiao Wang

School of Computing, Newcastle University, Newcastle upon Tyne, Tyne and Wear, NE4 5TG, United Kingdom

s.wang@ldy.edu.rs

Abstract. Automated Essay Scoring (AES) is a rapidly growing field that applies natural language processing (NLP) and machine learning techniques to the analysis and evaluation of academic essays. By automating the process of evaluating essay quality, AES not only greatly reduces the workload of human graders but also ensures consistency and objectivity in the evaluation process. AES systems can evaluate essays based on multiple criteria, including organization, coherence, and content. With the advent of deep learning, AES has shown significant improvements in accuracy and reliability. AES systems have numerous applications in education, particularly in large-scale assessment and feedback loops. In this article, we delve into the use of an improved Bidirectional Encoder Representations from Transformers (BERT) architecture with disentangled attention mechanism known as DeBERTa for student question-based summarization. This is one of the downstream tasks within AES, which is of great significance for student learning assessment. The organic combination of DeBERTa-v3 and diverse hats like Light Gradient Boosting Machine (LGBM) algorithm and Extreme Gradient Boosting algorithm (XGBoost) has proven to be highly effective in achieving excellent results in this task, indicating their significant potential in real-world AES systems.

Keywords: DeBERTa, Automated Essay Scoring, Light Gradient Boosting Machine.

1. Introduction

Due to factors such as population growth and the spread of pandemics, the distribution of educational resources has gradually become highly unequal. The student-to-teacher ratio has significantly increased as a result. The education sector is eagerly anticipating a new educational paradigm that can standardize basic education standards and elevate the quality of education in resource-scarce areas. Digital education has emerged as a compelling solution, as emphasized by Ramesh [1]. Many regions with uneven distribution of educational resources are now utilizing Automated Essay Scoring (AES) systems to assess students' essays. This assessment system employs advanced algorithms to evaluate and score students' essays from multiple perspectives, including text, content, grammar, and logic. It not only dramatically saves teachers' time and enhances teaching efficiency but also addresses the issue of unequal distribution of educational resources, ensuring fairness when dealing with different student groups. Additionally, it provides a convenient learning pathway for students who face difficulties in participating in offline education, including those with learning disabilities.

Reading and summarizing articles are essential tasks for students, which not only help develop students' logical thinking and language expression skills but also foster positive values and worldviews. Therefore, the assessment and scoring of students' summary writing by AES systems can provide teachers with valuable insights into the abilities and perspectives of different students. Indeed, this capability enables a more personalized and efficient approach to educational planning, accommodating individual needs and specific circumstances.

The success of large language models like ChatGPT, Llama, and others based on the Transformer paradigm has pushed traditional Natural Language Processing (NLP) tasks to new heights but has also provided a fresh perspective and approach to tasks such as AES for evaluating student summaries. The Bidirectional Encoder Representations from Transformers (BERT) model achieved state-of-the-art results in 11 downstream tasks as soon as it was introduced, garnering worldwide attention and sparking research and experimentation among scientists [2]. In recent years, the BERT family has welcomed several outstanding members, and among them, the Decoding-enhanced BERT with Disentangled Attention (DeBERTa) series of models has stood out as a top performer in downstream tasks, whether primarily in English text or in multilingual contexts [3].

2. Related Work

The development and advancements in deep learning and neural networks have offered AES a perspective distinct from traditional machine learning and feature engineering. This is particularly evident with the emergence of NLP models based on the transformer paradigm. For instance, Toledo et. al. using BERT as the foundational architecture for training on new datasets has yielded promising results [4, 5].

As the BERT family continues to grow and evolve, the replacement of base models has also contributed to the advancement of AES. Hicke et. al. experimented with various strategies based on Deberta-V3 for their application in the AES task. The combination of fine-tuning and ensemble learning yielded the best results [6].

Similarly, Younes at. el. employed deep learning models based on soft ensemble learning, combining Deberta-v3 with hateBert in various ways [7]. This soft ensemble approach yielded excellent results in downstream NLP tasks related to identifying discriminatory text.

3. Method

3.1. Decoding-enhanced BERT with Disentangled Attention (DeBERTa)

DeBERTa represents a significant advancement in the field of natural language processing [3]. It introduces a series of unique innovations to the BERT family, with the aim of enhancing the performance and efficiency of natural language processing tasks.

One notable feature of DeBERTa is the adoption of a dual-stream architecture, where one stream focuses on self-attention, while the other stream specializes in decoupled attention mechanisms. This novel approach allows the model to effectively capture various relationships and semantics within the text. By decoupling attention, DeBERTa gains a better understanding of the structural complexities and contextual information embedded in the text, making it particularly proficient at handling complex language patterns.

On the other hand, DeBERTa's enhanced decoding process improves its performance in generative tasks such as text summarization, translation, and generative question-answering. This decoding capability positions DeBERTa as a valuable resource in various natural language processing applications. Additionally, during fine-tuning, DeBERTa utilizes a novel adversarial training algorithm that normalizes word embeddings into probability vectors and then introduces perturbations to train them.

He, P at. el. used a new pre-training method from ELECTRA called Replace Token Detection (RTD) to replace MLM training in DeBERTa [7]. RTD includes both a generator and a discriminator, two transformer encoders trained for binary classification tasks on both MLM and RTD. This creative experiment improved the performance of DeBERTa-v3 on same datasets, adding another strong member to the BERT family.

The DeBERTa series is a powerful language model at the forefront of natural language processing research. Its disentangled attention mechanisms, dual-stream architecture, and decoding enhancements contribute significantly to its outstanding performance in a wide range of natural language understanding and generation tasks, solidifying its status as a significant milestone in the continually evolving field of natural language processing. In the subsequent model construction and experiments, DeBERTa-V3 have been selected as the primary model structure for application in one of the downstream tasks of Reading and Summarizing Articles.

3.2. Extreme Gradient Boosting (XGBoost)

XGBoost is a cutting-edge machine learning algorithm renowned for its exceptional performance and versatility across various research and practical domains [8]. This algorithm falls under the category of gradient boosting, which involves ensemble learning by combining multiple weak learners to create a robust predictive model. XGBoost, however, stands out due to its extreme efficiency and predictive power.

XGBoost is a gradient boosting approach, wherein it iteratively builds an ensemble of decision trees. Each tree corrects the errors of its predecessors, gradually improving the model’s accuracy. This technique, coupled with advanced regularization methods such as L1 (Lasso) and L2 (Ridge) regularization, helps prevent overfitting and enhances model generalization.

Such a huge amount of calculation does not slow down the computing speed of XGBoost, it has been meticulously engineered to provide a highly optimized and scalable solution for predictive modeling. Through parallel and distributed computing, it leverages the capabilities of multi-core processors and distributed computing platforms, resulting in lightning-fast training times even on large datasets.

Another noteworthy aspect of XGBoost is its innate ability to handle missing data gracefully. It seamlessly incorporates missing value handling during tree construction, eliminating the need for extensive preprocessing steps.

Furthermore, XGBoost offers an insightful feature importance evaluation mechanism, allowing users to assess the relevance of each feature in their dataset. This feature is invaluable for understanding the factors driving predictive outcomes.

XGBoost’s wide-ranging applications encompass classification, regression, ranking, and more. It has proven its mettle in various data science competitions and has found widespread adoption in real-world applications, thanks to its combination of predictive accuracy and efficiency.

3.3. Light Gradient Boosting Machine (LGBM)

LGBM stands as a formidable machine learning algorithm that has made substantial contributions to both research and practical applications [9]. LGBM uses a gradient boosting framework, and is well known for its effectiveness in tackling classification and regression tasks [10].

A notable hallmark of LGBM is its lightweight design and exceptional efficiency. It achieves this by employing histogram-based algorithms and memory-efficient techniques that facilitate rapid training and reduced memory usage. These attributes make LGBM particularly well-suited for handling large datasets and complex problems.

Parallelization is another strength of LGBM. It takes full advantage of multi-threading and parallel processing, harnessing the power of modern multi-core processors to accelerate training times significantly. This parallelized approach enhances its scalability and enables it to handle massive datasets efficiently.

LGBM has consistently demonstrated high predictive accuracy, making it a popular choice in various machine learning competitions and real-world applications. Its ability to produce accurate results across a wide range of problem domains, coupled with its speed, positions LGBM as a versatile tool for data scientists and researchers alike.

Furthermore, LGBM incorporates built-in mechanisms for handling missing data, eliminating the need for extensive preprocessing. It also adapts to the data distribution and minimizes overfitting through the use of regularization techniques like L1 (Lasso) and L2 (Ridge) regularization.

In summary, LGBM is a powerful and efficient machine learning algorithm that excels in both research and practical applications. Its lightweight yet high-performance nature, coupled with parallel processing capabilities and built-in features for data handling and regularization, makes LGBM a valuable asset in the toolkit of data scientists and researchers working on a diverse array of predictive modeling tasks.

XGBoost and LGBM, as outstanding and advanced gradient boosting algorithms, have been widely applied and have demonstrated their inherent value in both the industry and academia. In our experiments, we will employ them as the structure for the HAT model to explore the application and performance of the organic integration of deep learning and machine learning in the context of paper reading and abstract generation tasks.

4. Experiment and Result

4.1. Dataset

The Kaggle commit-lit dataset, created by commit-lit, is a student summary dataset compiled from multiple renowned articles and books. The distributions are demonstrated in Figure 1 and Figure 2 respectively. It evaluates student-generated summaries from two perspectives: text content and text vocabulary.

The strengths of this dataset lie in its provision of specific questions paired with corresponding summary responses. It also incorporates negative samples to enhance the training's generalization capability. However, due to its complexity and focus on accuracy, the dataset's size is relatively small. This inevitably leads to overfitting issues during training, resulting in reduced model generalization and robustness when directly trained on it.

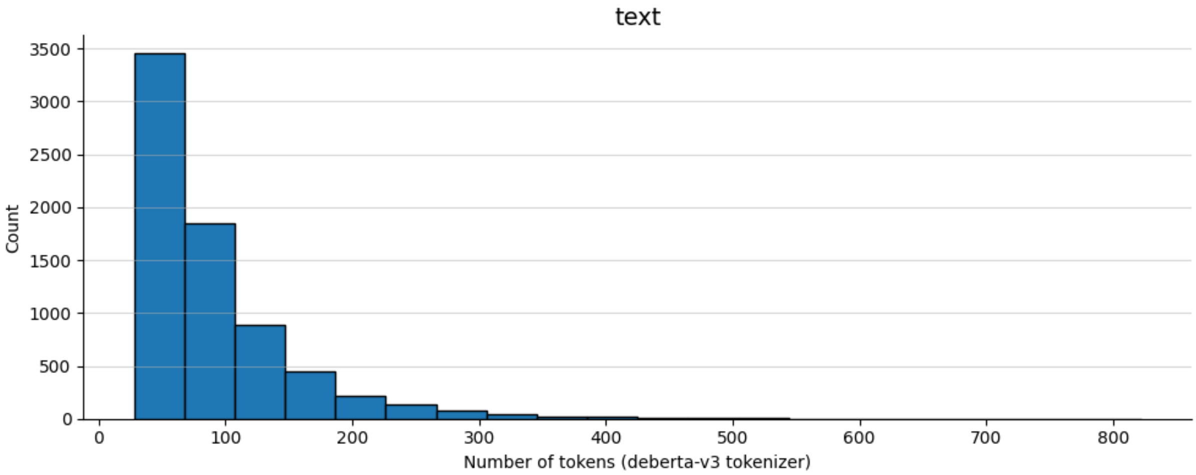

Figure 1. Length distribution of abstract text in the dataset (Figure Credits: Original).

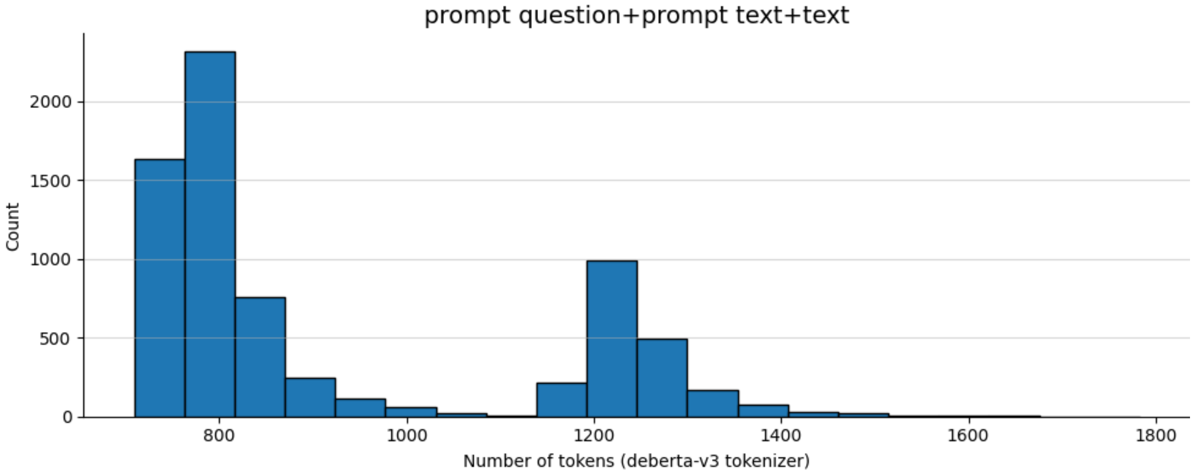

Figure 2. The total length of target questions and target text in the dataset, along with the student abstract text (Figure Credits: Original).

To address the shortcomings of this dataset and enable the model to handle complex scenarios simultaneously, two parallel strategies will be applied during the training process:

(1) Data Augmentation: A series of data augmentation techniques will be employed. This involves generating additional training data through various transformations of the existing data, such as paraphrasing, synonym replacement, or introducing small variations to the text. Data augmentation helps increase the diversity of the training set and can mitigate overfitting.

(2) Model Architecture Exploration: Multiple combinations of different head structures and base models will be experimented with to construct a new model architecture. By exploring various combinations, it’s possible to discover a model configuration that is better suited to the characteristics of the commit-lit dataset. This approach aims to improve the model’s generalization and robustness.

These parallel strategies are intended to enhance the model’s performance on the dataset and make it more adaptable to complex scenarios.

4.2. Experimental Settings

Large language models have achieved several state-of-the-art results on downstream NLP tasks. However, the native performance of these models on some complex downstream tasks may not be as impressive as expected, necessitating fine-tuning of the base model. For large models with billions of neurons, obtaining a sufficiently large pre-training dataset can be challenging. As a result, fine-tuning often needs to be performed on smaller, more complex datasets, which can lead to less noticeable or suboptimal improvements in pre-training performance.

On the other hand, traditional machine learning models and feature engineering approaches tend to perform well and adapt quickly on small datasets. They can effectively capture the unique features of the dataset. However, these methods often have poor robustness and generalization and may struggle to adapt to real-world and complex usage scenarios.

Ensemble learning improves the overall performance and robustness of a model by combining predictions from multiple base learners, often referred to as weak learners. It can significantly mitigate the weaknesses of both types of models. Based on the concept of ensemble learning, the original Deberta-v3-base and Deberta-v3-large are used as base models, and applying different “hats” to them, representing various machine learning algorithm strategies. Fine-tuning and ensemble learning were conducted on the dataset. The architecture of hats is displayed in Figure 3.

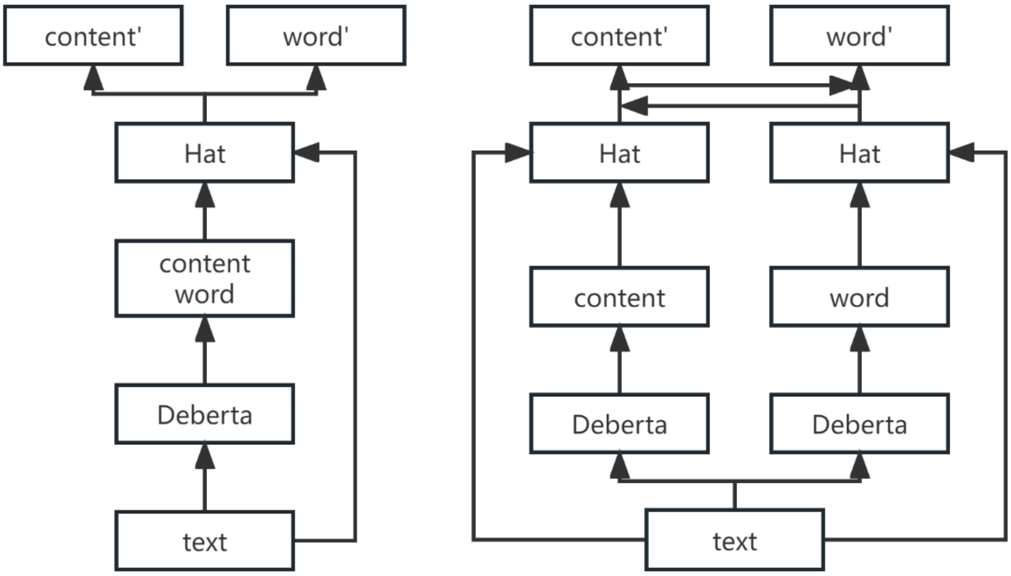

Figure 3. Architecture of the ‘hat’ model (Figure Credits: Original).

This work conducted experiments to evaluate the model’s performance under various configurations. Here are the design details:

(1) This work maintained a configuration that involves using DeBERTaV3 for centralized processing of the raw data. This includes simultaneous training on content and words, and it serves as the baseline.

(2) This work also employed ensemble learning as an approach to adapt the application of DeBERTaV3 to complex scenarios. Whether this adaptation could yield significant improvements when integrating traditional learning “hats” into the process is examined.

Root Mean Square Error (RMSE) and Mean Columnwise Root Mean Square Error (MCRMSE) are used to record and analyze the performance of model training. RMSE is a statistical metric used to assess regression models, measuring the difference between predicted values and actual observations. Due to the small dataset, we employ the K-fold validation method to train the model on the dataset and use various early stopping techniques to save on training costs. Consequently, we will further average the results using MCRMSE as the primary evaluation metric over RMSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

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

4.3. Results

Intermediate results, including the training loss, validation mean square error, and error using exponential moving average techniques are demonstrated in Figure 4, Figure 5, and Figure 6 respectively. Texts are input to DeBERTa to obtain two scores, one for ‘content’ and one for ‘word.’ At the same time, they are input into two separate DeBERTa models to obtain ‘content’ and ‘word’ scores individually. It could be found that using the first strategy results in scores that are closer to the ground truth compared to the latter.

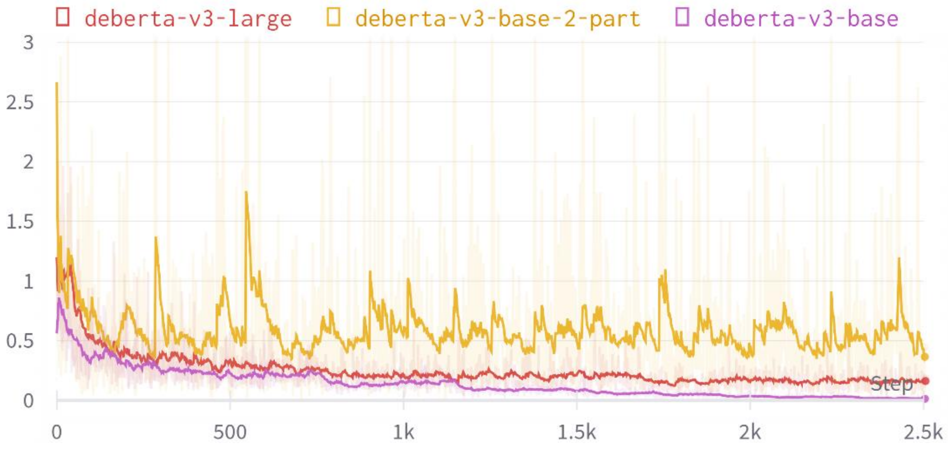

Figure 4. Train Loss of Deberta-V3 family (Figure Credits: Original).

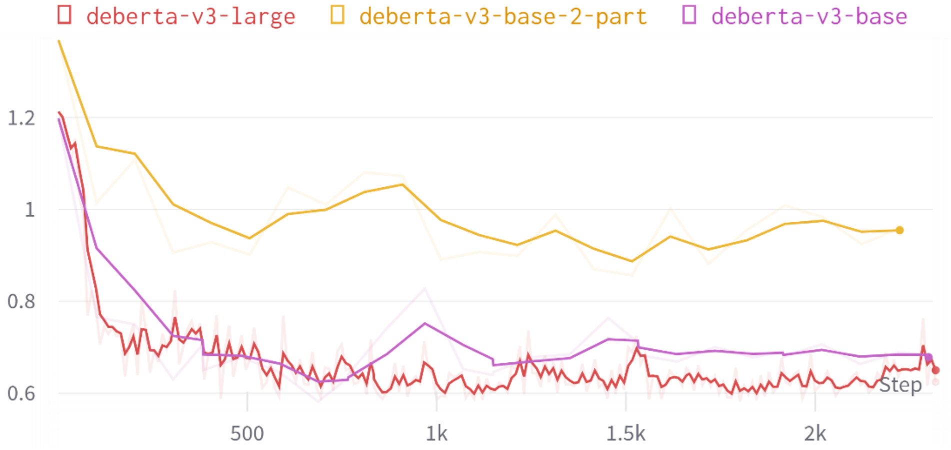

Figure 5. Validation MCRMSE of Deberta-V3 family (Figure Credits: Original).

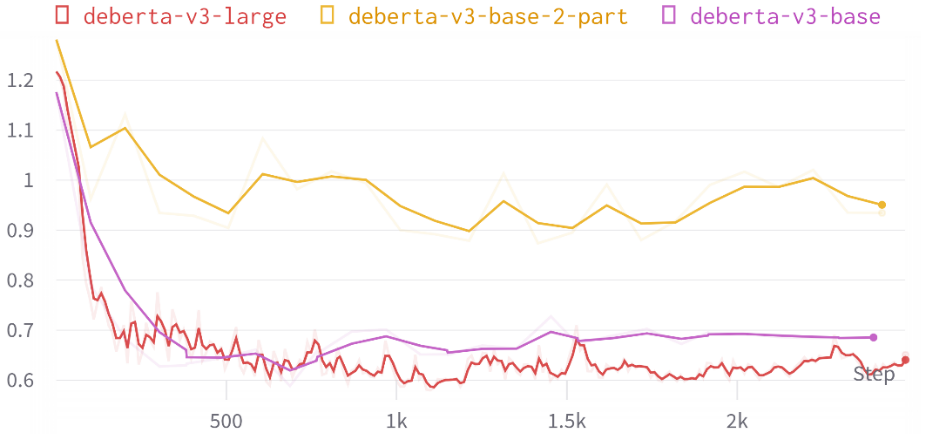

Figure 6. The validation MCRMSE for training DeBERTa using Exponential Moving Average (EMA) technique (Figure Credits: Original).

Additionally, following the previous model architecture, this work incorporated 'hats' onto DeBERTa separately. Light Gradient Boosting Machine (LGBM) is leveraged to combine the aggregated scores from DeBERTa with various statistical features, such as word accuracy and word repetition rate. Since deep learning models are often considered as black-box models lacking interpretability, in order to address this limitation, this work employed machine learning models with better interpretability and feature engineering to complement and adjust the outputs of the deep learning model. This is akin to putting different 'hats' on the deep learning black-box model, allowing them to perform better in complex scenarios. The results are demonstrated in Figure 7.

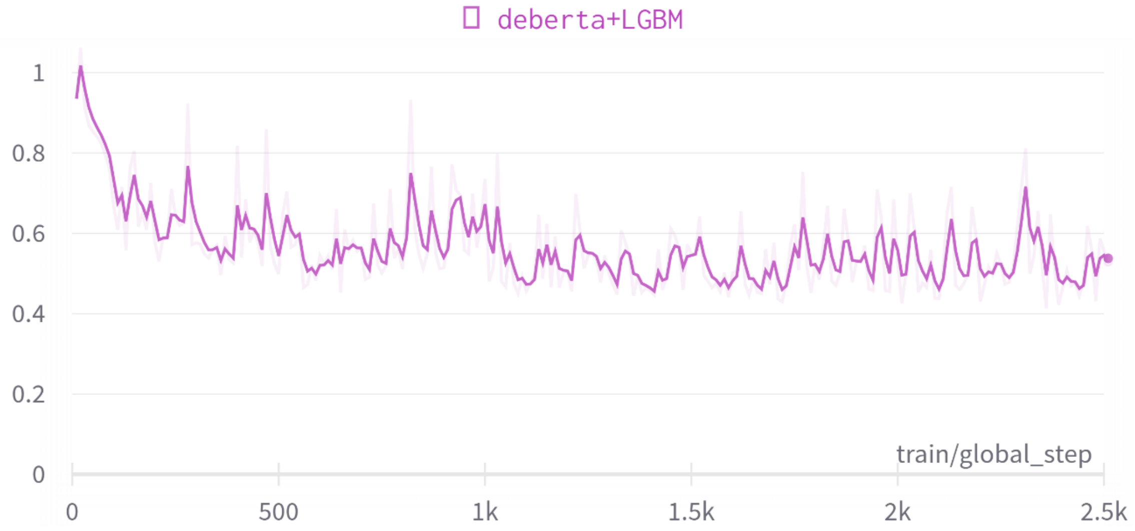

Figure 7. The validation RMSE for training DeBERTa and single LGBM head.

(Figure Credits: Original).

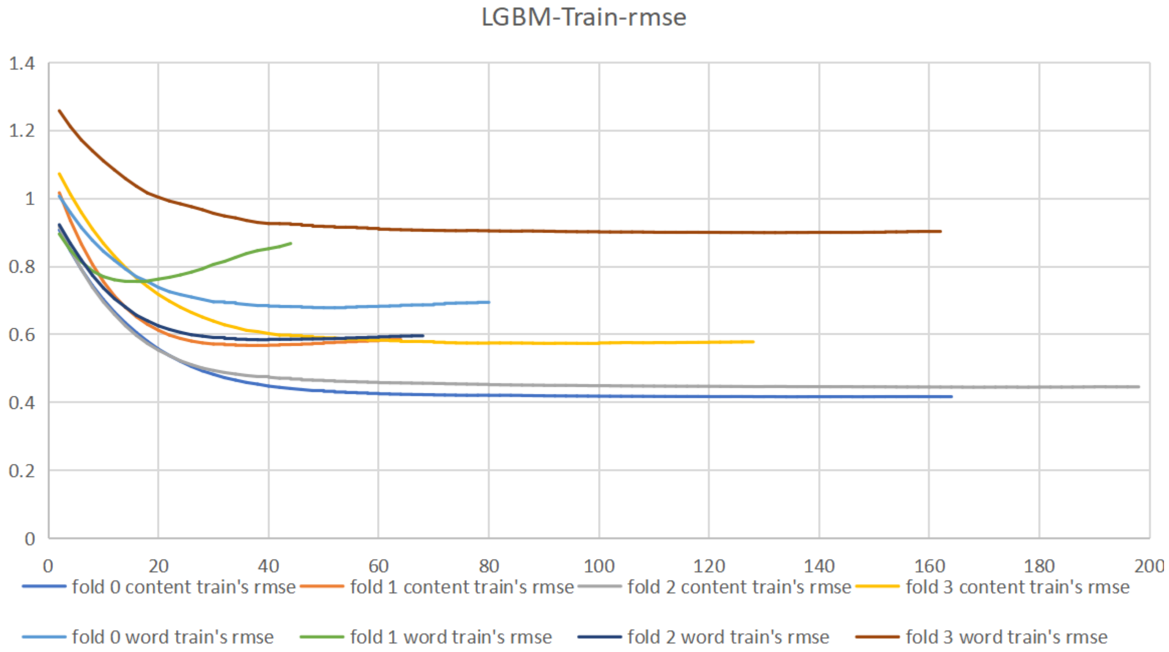

Figure 8. The validation RMSE for Training LGBM and double heads hats (Figure Credits: Original).

Following the design in Figure 3, we conducted experiments with both a single "hat" structure, as shown in Figure 7 and a dual "hat" structure as illustrated in Figure 8. Comparing the results of these experiments, we observed that models trained with a single DeBERTa structure are better at recognizing the connection between the text and word scores. For instance, sometimes students may use highly sophisticated words that do not align with the text, resulting in lower scores.

On the other hand, models trained with a dual DeBERTa structure cannot capture the connection between the abstract text and word scores at the DeBERTa level. However, through the linkage of the dual “hats” and corrections made through feature engineering, the entire model’s scores can be adjusted.

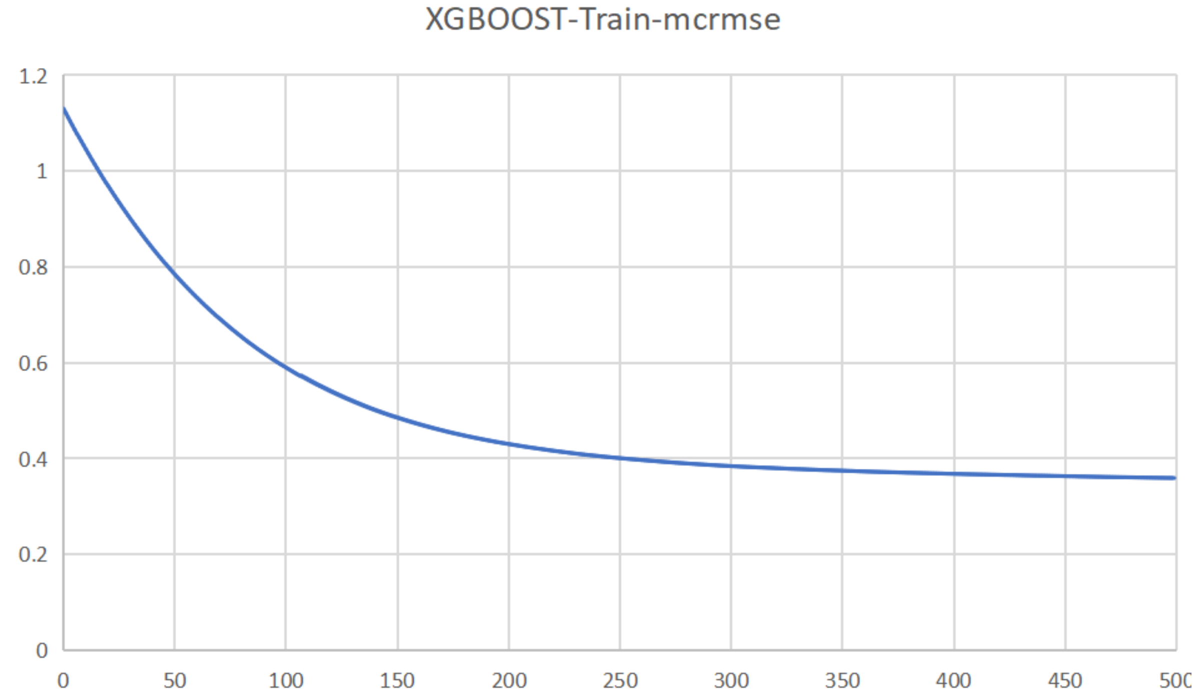

Figure 9. The validation MCRMSE for Training DeBERTa and single XGBoost (Figure Credits: Original).

XGBoost, as another style of “hat”, was also experimented with and placed on top of DeBERTa. Through the organic combination of feature engineering and deep learning, its RMSE results (as shown in Figure 9) even surpassed those of the DeBERTa structure with LGBM. However, in comparison to LGBM, it exhibited slightly slower execution speeds.

Table 1. The validation MCRMSE for Training DeBERTa with different hats.

MODEL	MCRMSE
Single Deberta-V3$_{Base}$	0.686
Single Deberta-V3$_{Large}$	0.651
Double Deberta-V3$_{Large}$	0.935
Single Deberta-V3$_{Large}$ + LGBM	0.588
Single Deberta-V3$_{Large}$ + XGBoost	**0.359**
Double Deberta-V3$_{Large}$ + LGBM	0.636

The experimental results in Table 1 indicate that the single DeBERTa-V3 model with XGBoost outperforms on this task. This may be attributed to two factors. Firstly, the dataset is small and contains missing data, where XGBoost exhibits a clear advantage when dealing with such datasets. Secondly, a single DeBERTa model, when trained for regression tasks, can simultaneously consider both content and words, resulting in improved training outcomes.

5. Conclusion

Wearing the “hat,” BertaV3 achieved outstanding results in complex student abbreviation summarization, allowing it to integrate a broader range of information when dealing with intricate scenarios. While this improved the model interpretability and accuracy, it might reduce its generalizability.

Indeed, the future development of NLP models can involve the use of various auxiliary heads for different tasks, similar to the “backbone+neck+head” structure in the computer vision field. Building upon large, general-purpose models, these diverse auxiliary heads can assist in different tasks, thereby enhancing both the model’s generalizability and accuracy simultaneously.

Through the concepts and experiments described above, we have observed that ensemble learning can effectively combine feature engineering, machine learning, and deep learning. The strong interpretability and operational capabilities of the former two can provide human control and adjustments to the latter’s black-box models. This addresses the limitations of deep learning in cases of small datasets, poor generalization, and complexities in specialized domains, ultimately enhancing the performance of general large language models like GPT in specialized, small dataset scenarios.

In the future, further integration of excellent machine learning and feature engineering methods through ensemble learning with large AI models can expand their applicability across various scenarios and domains. Only by doing so can AI better serve humanity in a multitude of contexts and fields.

References

[1] Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, 55(3), 2495-2527.

[2] Rodriguez, P. U., Jafari, A., & Ormerod, C. M. (2019). Language models and automated essay scoring. arXiv preprint arXiv:1909.09482.

[3] He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

[4] Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., ... & Slonim, N. (2019). Automatic Argument Quality Assessment--New Datasets and Methods. arXiv preprint arXiv:1909.01007.

[5] He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543.

[6] Hicke, Y., Tian, T., Jha, K., & Kim, C. H. (2023). Automated Essay Scoring in Argumentative Writing: DeBERTeachingAssistant. arXiv preprint arXiv:2307.04276.

[7] Younes, M., Kharabsheh, A., & Younes, M. B. (2023). Alexa at SemEval-2023 Task 10: Ensemble Modeling of DeBERTa and BERT Variations for Identifying Sexist Text. In Proceedings of the The 17th International Workshop on Semantic Evaluation, 1644-1649.

[8] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

[9] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, 1-8.

[10] Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., & Zeng, W. (2019). Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agricultural water management, 225, 105758.