Enhancing Student Writing Skills: Leveraging Transfer Learning and Fine-tuned Language Models for Automated Essay Structure Recognition

Sani Abdullahi Sani

Preprint v1

June 7, 2023

https://doi.org/10.32388/W283Y7

Enhancing Student Writing Skills: Leveraging Transfer Learning and Fine-tuned Language Models for Automated Essay Structure Recognition

Sani Abdullahi Sani
Faculty of Information Technology
Almaty, Kazakhstan 30332–0250
Email: s_sani@kbtu.kz

June 7, 2023

***Abstract*—Writing skills are essential for academic and professional success. However, many students struggle to become proficient writers, highlighting the need for effective writing instruction and feedback methods. Automated Writing Evaluation (AWS) systems have emerged as a promising solution to address these challenges. This study proposes a model that utilizes fine-tuned language models to evaluate essay structure, specifically identifying key argumentative and rhetorical elements. The Longformer and Bigbird models were fine-tuned and evaluated for discourse classification. The results demonstrate that the Longformer model outperformed the Bigbird model, achieving an F1 score of 0.634 compared to 0.615. The Longformer model's ability to handle large data inputs without losing vital information contributed to its superior performance. Integrating machine learning models with AWE systems can enhance automated essay evaluation, providing valuable feedback to students. While positional encoding improves discourse classification, future research should focus on expanding data coverage across additional essay categories. This study highlights the significance of leveraging advanced NLP techniques to improve writing skills and lays the foundation for further advancements in automated essay evaluation systems.**

I. Introduction

Writing is a fundamental skill that is crucial for academic and professional success [1] and yet, few students graduate high school as proficient writers with less than a third of high school seniors being proficient writers, according to the National Assessment of Educational Progress (NAEP) [2]. While traditional writing instruction and feedback methods are widely used, they have limitations, such as being time-consuming, Heterogeneity of students, and subjective [1]. In academic contexts, essays serve as a common assessment tool for evaluating students' writing abilities. However, writing a good essay requires more than just a good idea; it also requires a clear and coherent structure that guides the reader through the argument. Teaching students how to structure their essays can be challenging, especially since different types of essays may require different structures.

Before 2004, Automated Writing Evaluation (AWE) systems were only used to help score essays holistically. However, researchers realized that AWE systems could have a greater impact if they could also help students improve their writing skills. A low score on an essay does not tell a student why they received the low score or what they can do to improve. To address this issue, researchers have begun to develop AWE systems that are capable of scoring specific aspects of an essay, for example coherence as in [3], [4], technical mistakes, as well as relevance to the prompt as in [2], [5]. Moreover, as stated in Lagakis [6], even the most sophisticated Automated Writing Evaluation (AWE) systems face challenges when it comes to precisely assessing elements such as coherence, persuasiveness, and argument clarity in an essay. This difficulty arises from two primary factors: the inherent complexity

involved in modeling such aspects and the limited availability of dimension-specific datasets that offer human-graded evaluations, particularly in comparison to datasets that provide holistic scoring. Recently, advancements in natural language processing (NLP) have presented new opportunities in essay evaluation. One promising strategy involves the fine-tuning of pre-trained language models (LMs), enabling the transfer of knowledge acquired from extensive pre-training to novel tasks.

In this study, we propose a model that utilizes fine-tuned language models to evaluate essay structure, specifically it will automatically segment the text of essays written by students and classify the argumentative and rhetorical elements. Our goal is to train the models to identify key structural elements of a quality essay. We used a dataset of student essays annotated by human raters to guide the model's learning process. The Longformer and Bigbird models were fine-tuned and evaluated for discourse classification. The finetuned language models were tested on a separate set of essays to assess their performance in identifying missing structural elements and generating prompts for improvement.

II. Related Works

The concept of Automated Writing Evaluation (AWE) systems was initially introduced by E.B. Page in 1966, who pioneered the development of computer-aided grading systems. Nowadays, AWE is recognized as a prominent application of Natural Language Processing (NLP) in academia, leveraging artificial intelligence to score written documents [6].

Early AWE systems relied on handcrafted features [6]. For instance, Page's Project Essay Grader TM (PEG) is regarded as the first AWE system, employing supervised learning with linear regression. It incorporated length-based features such as average word length and text length [7], as well as lexical features such as punctuation frequency. Other notable early systems, including IEA, E-rater, IntelliMetric, and BETSY, also employed handcrafted features [6]. In addition, some systems incorporated syntactic features like parse trees [8], with one system utilizing LambdaMART for ranking and using parse tree depth as a measure of syntactic complexity within sentences [6]. Moreover, there exists another classification of AWE systems that utilizes neural approaches for automated feature extraction instead of relying on traditional feature engineering methods. For example, In [9] The model generates the necessary features in an automated manner by utilizing an input consisting of one-hot vectors representing the words that are present in the essay that requires evaluation, which is subsequently forward to a convolution layer that extracts n-gram level features. Afterward, these features are forwarded through a recurrent layer within a Long-Short Term Memory (LSTM) network, which captures a separate set of features. The n-gram level features focus on the immediate word relations, while the second vector captures the long-distance relationships between words in the essay. Both sets of features are then combined and serve as the input to a dense layer that produces the ultimate holistic score for the essay. Additionally, [10] equally used LSTM-based approach, accompanied by an additional layer aiming to extract coherence features from the essay. Similar strategies, like the one described in [11], utilize word embeddings instead of one-hot word vectors.

In recent years, the NLP community has extensively embraced the application of deep neural networks. The transformer architecture, introduced in 2017 [12] and popularized by BERT (Bidirectional Encoder Representations from Transformers) [13], has emerged as the prevailing trend. It has replaced older recurrent neural network (RNN) models like long short-term memory (LSTM) due to its widespread adoption and effectiveness.

It is important to note that Transformer models make use of extensive datasets comprising general text data, such as the Wikipedia Corpus and Common Crawl. Models like BERT and GPT (Generative Pre-trained Transformer) undergo a pretraining phase on these datasets to acquire an understanding of contextual meanings and the interplay between words. Following pretraining, the models are fine-tuned using specific labeled datasets and employed in diverse tasks, with classification or structured

prediction tasks being among the most prevalent applications.

Among the notable AWE research using the state of the art transformers is The Two-Stage Learning Framework (TSLF) [14]. It is a system that employs a two-component model. In the first component, a pre-trained BERT model is utilized to generate sentence embeddings. The specific BERT model used has 12 layers, 768 hidden units, 12 attention heads, and 110M parameters. These sentence embeddings are then passed as input to a recurrent neural network (RNN). The second component of the model incorporates hand-crafted features. This hybrid approach has shown promise in the Two-Stage Learning Framework (TSLF) in which the use of BERT sentence representations enables the learning of an essay score, prompt-relevance score, and "coherence" score. These scores are trained on both original and permuted essays. In conjunction with document representations from the neural network and hand-crafted features, a gradient-boosting decision tree is employed to predict the final essay score.

In a different study [15], the relevance of discourse-aware structures and discourse-related pretraining in neural network AWE systems is demonstrated. The researchers explore two neural models, HAN [16] and Bidirectional context with attention (BCA) [17], both based on LSTM, to map essays into vectors for holistic evaluation using ordinal regression. The BCA model integrates discourse awareness by taking into account the interdependencies among sentences, calculating contextual similarities, and adapting the ultimate representation of each word accordingly. It leverages pretraining with BERT embeddings and posits that the task of predicting the next sentence can capture the discourse coherence elements of the essay.

Another AWE system utilizing the BERT model is R2BERT [18]. It introduces the multi-loss approach for fine-tuning BERT models in AWE systems, combining regression and ranking models. Experimental results indicate performance improvements using the multi-loss approach.

A recently published AWE system [7] adopts a deep neural network (DNN) framework in conjunction with item response theory (IRT) [19]. This approach aims to mitigate human rater bias and is particularly useful for low or medium stakes tests with potentially lower quality training data. The framework evaluates the effectiveness of combining IRT with both a "traditional" AWE model consisting of convolutional neural networks with LSTM and the more recent approach of using BERT.

In another hybrid AWE system [20], DNNs are combined with hand-crafted essay-level features. Multiple methods, including LSTM and BERT, are tested, with BERT demonstrating the most favorable experimental results.

Overall, In [6], it has been demonstrated that the integration of BERT and manually engineered features in AWE systems employing transformer models yields the most optimal outcomes, establishing it as the current pinnacle in the field, and Hence our approach.

III. Methodology

A. Dataset

In this study, we used the PERSUADE corpus [21], a dataset created by the Learning Agency Lab, to train and test my models. This corpus was designed specifically for the discourse classification problem and contains 25,000+ student essays that have been annotated by writing professionals. To ensure the accuracy of the dataset, each essay was annotated using a double-blind rating process and adjudicated by a third writing professional. Additionally, The PERSUADE corpus is an outstanding resource for training and testing models for discourse classification. However, I believe that some changes can be made to the formatting of the dataset through data preprocessing. As explained in [21], The list of discourse elements was put together by a team of teachers and professional writers at The Learning Agency Lab.

- Lead - an introduction that begins with a statistic, a quotation, a description, or some other device to grab the reader's attention and point toward the thesis
- Position - an opinion or conclusion on the main question
- Claim - a claim that supports the position

- Counterclaim - a claim that refutes another claim or gives an opposing reason to the position
- Rebuttal - a claim that refutes a counterclaim
- Evidence - ideas or examples that support claims, counterclaims, or rebuttals.
- Concluding Statement - a concluding statement that restates the claims

B. Data Cleaning and Preprocessing

C. Data Preprocessing

In the preprocessing stage, we performed the following steps to prepare the data for further analysis. First, we read the contents of the Test folder and converted it into a Pandas dataframe. This allowed us to efficiently manipulate and organize the data for subsequent processing. Similarly, we read the contents of the Train folder and transformed it into a Pandas dataframe as well. This preprocessing step ensured that the data from both folders were easily accessible and ready for subsequent analysis.

D. Data Cleaning

Data cleaning involves identifying and addressing duplicates, misspellings, superfluous symbols, and inconsistent notations within a dataset. These issues are resolved through the removal, correction, or normalization of problematic data elements.

- Stopwords (stop): Stopword removal is another data cleaning technique in NLP that involves eliminating common words that do not carry significant meaning or contribute much to the overall understanding of a text. Therefore, we remove all the stopwords using the NLTK library.
- Stemming (stem): Stemming is a data cleaning technique used in natural language processing (NLP) to reduce words to their base or root form, called stems. It involves removing prefixes and suffixes from words to simplify their representation and consolidate variations of the same word. For example, stemming would convert ”running,” ”runs,” and ”ran” to the common stem ”run.” We use NLTK Snowball stemmer for stemming our training corpus.

  Next, we converted all text words into Named Entity recongition (NER) labels and save in a dataframe. Subsequently, we developed a PyTorch dataset function that consistently generates two outputs, namely tokens and attention. In addition to these outputs, during the training phase, it also includes labels to facilitate the training process. Similarly, during the inference phase, the function further provides word IDs, which help in converting token predictions into accurate word predictions.

Moreover Babanejad et. al, in [22] noted that a more careful consideration of the sequence in which preprocessing techniques are applied showed to obtain a more stable result. For instance, pos-tagging should be applied before stemming in order for the tagger to work well, or negation should be performed prior to removing stopwords. For that, we consider the following ordering when combining the aforementioned Data preprocessing: removing stopwords, and stemming.

E. Two Models Used - PyTorch BigBird and Tensoflow Longformer

In this study, we used two models namely Google BigBird-v26 specifically, bigbird-roberta-base and Longformer. The motive behind choosing the models is explained as follows: Bigbird is a Transformer-based neural network that can process sequences up to eight times longer than previous models [23]. Our motivation stemmed from the performance of the model. As explained by choko in [23], Bigbird achieved better performance by using a new self-attention mechanism that reduces the computational complexity of the Transformer architecture. BigBird has been shown to achieve state-of-the-art results in natural language processing (NLP) and genomics tasks. Further, Bigbird demonstrated the ability to have outperformed previous models on several question-answering and document classification datasets. For example, on the Arxiv dataset, BigBird achieved an F1 score of 92.31%, which is a new high. In genomics, BigBird has outperformed previous models on two genome classification tasks: promoter region prediction and chromatin-profile prediction. In the former task,

BigBird achieved an accuracy of 99.9%, which is 5% better than the best model ever. Similarly, By utilizing Longformer's self-attention mechanism, the computational complexity associated with the query-key matrix multiplication, which typically presents a significant memory and time bottleneck, can be mitigated. Specifically, the complexity can be reduced from (O(n$_s \cdot n_s$)$toO(n_s \cdot w)$ where n_s represents the sequence length and w denotes the average window size. This transformation assumes that the number of tokens attending at a "global" level is relatively small compared to the number of tokens attending at a "local" level.

F. Hyperparameters

For the model hyperparameters, we used

- EPOCHS = 5
- BATCH SIZE = 4
- Learning Rates Scheduler (LRS) = [0.25e-4, 0.25e-4, 0.25e-4, 0.25e-4, 0.25e-5]

G. Model Evaluation - F1-score

To evaluate the models, i used the F1 score as an evaluation metric.

$$\frac{TP}{TP + 0.5 \cdot (FP + FN)}$$

The motive behind that is to assess the overall performance of our model in capturing both positive and negative instances accurately. It is also worthy to note that it provides a single value that summarizes the trade-off between precision and recall, allowing us to compare different models or variations of our model based on their ability to balance both aspects of classification.

IV. Results

In our experiment, we performed fine-tuning on two transformer models sourced from the Huggingface library: Longformer [21], and Bigbird [22]. Based on the results presented in the "Trained Models and their F1 scores" table, it is evident that the Longformer model achieved the highest performance, attaining an F1 score of 0.634. The Bigbird model ranked second with an F1 score of 0.615. These findings indicate that the Longformer model outperformed the Bigbird models for discourse classification towards improving student writing skills. Furthermore, It is likely that the Longformer's exceptional capacity to handle substantial data inputs without losing crucial information contributed to its success in our experiments.

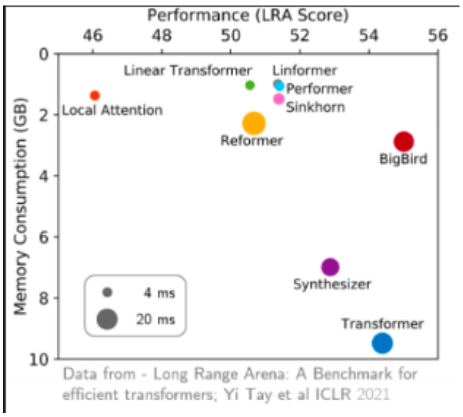

Fig. 1. Bigbird

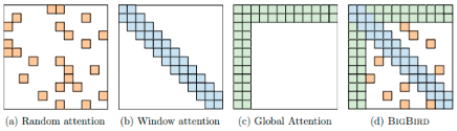

Fig. 2. Bigbird

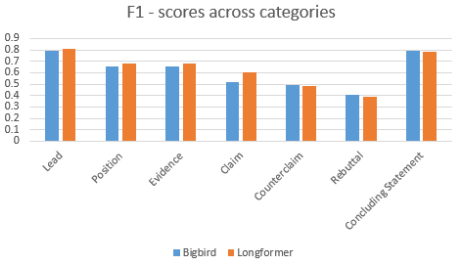

Fig. 3. Average F-1 scores across all models (except baseline) for each discourse element

V. Discussion

In our study, we have observed similarities between our work and the research conducted by

Alkabool et. al [24] in terms of the models employed for fine-tuning. Alkabool also utilized three models, namely BERT, Longformer, and GPT-2, similar to our approach. Notably, the Longformer model exhibited the highest performance, yielding an impressive F1 score of 0.535, establishing it as the top-performing model in Alkabool's work. The BERT model followed in second place with an F1 score of 0.395, while the GPT-2 model demonstrated the lowest performance, achieving an F1 score of 0.362.

Our work aligns with Alkabool's study in terms of the superior performance of the Longformer model. However, we have achieved even more promising results in our research, as our Longformer model attained a higher F1 score of 0.634, surpassing the performance reported by Alkabool. This signifies an enhanced performance and highlights the effectiveness of our proposed approach.

VI. Conclusion

Developing strong writing skills is crucial for young students, and automated essay evaluation systems can play a significant role in nurturing their talent by providing detailed analysis of their writing. To enhance current automated essay evaluation, one approach is to integrate them with machine learning models that can effectively differentiate between various writing elements in a student's essay. In this study, the effectiveness of Longformer models was evaluated in comparison to Bigbird models for discourse classification. The results indicated that Longformer models outperformed BigBird models in the given context. Furthermore, it was demonstrated that fine-tuning the Longformer models with the entire essay as input allowed the model to capture positional relationships between discourse elements, particularly in the Lead and Concluding Statement classes. However, it is important to note that while positional encoding contributes to discourse classification, additional efforts are needed to gather data on more categories, such as rebuttal or counterclaim, to further enhance overall results.

References

[1] M. D. Shermis and J. Burstein, "Automated writing evaluation for improving writing skills: An instructional framework and review of automated writing systems," *Educational Psychology Review*, vol. 15, no. 3, pp. 377–396, 2003.

[2] I. Persing and V. Ng, "Modeling prompt adherence in student essays," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 1534–1543.

[3] D. Higgins, J. Burstein, D. Marcu, and C. Gentile, "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*. Boston, Massachusetts, USA: Association for Computational Linguistics, 2004, pp. 185–192.

[4] S. Somasundaran, J. Burstein, and M. Chodorow, "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*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014, pp. 950–961.

[5] A. Louis and D. Higgins, "Off-topic essay detection using short prompt texts," in *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*. Los Angeles, California: Association for Computational Linguistics, 2010, pp. 92–95.

[6] P. Lagakis and S. Demetriadis, "Automated essay scoring: A review of the field." Institute of Electrical and Electronics Engineers Inc., 2021.

[7] M. Uto and M. Okano, "Robust neural automated essay scoring using item response theory," in *Proceedings of the Conference Name*. Springer International Publishing, 2020.

[8] H. Chen and B. He, "Automated essay scoring by maximizing human-machine agreement," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, 2013, pp. 1741–1752.

[9] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1882–1891.

[10] Y. Tay, M. Phan, L. A. Tuan, and S. C. Hui, "Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 756–765.

[11] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016, pp. 715–725.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, December 2017, pp. 5999–6009.

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

[14] J. Liu, Y. Xu, and Y. Zhu, "Automated essay scoring based on two-stage learning," in *Proceedings of the Conference Name*, 2019, pp. 1–7.

[15] F. Nadeem, H. Nguyen, Y. Liu, and M. Ostendorf, "Automated essay scoring with discourse-aware neural models," in *Proceedings of the Conference Name*, 2019, pp. 484–493.

[16] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016, pp. 1480–1489.

[17] F. Nadeem and M. Ostendorf, "Estimating linguistic complexity for science texts," in *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 45–55.

[18] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He, "Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 1560–1569.

[19] F. M. Lord, *Applications of Item Response Theory To Practical Testing Problems*, 1st ed. Hillsdale, N.J: Lawrence Erlbaum Associates, 1980.

[20] M. Uto, Y. Xie, and M. Ueno, "Neural automated essay scoring incorporating handcrafted features," in *Proceedings of the Conference Name*, 2021, pp. 6077–6088.

[21] "The feedback prize," https://www.the-learning-agency-lab.com/the-feedback-prize/, accessed on May 2021.

[22] N. Babanejad, A. Agrawal, A. An, and M. Papagelis, "A comprehensive analysis of preprocessing for word representation learning in affective tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 5799–5810. [Online]. Available: https://aclanthology.org/2020.acl-main.514

[23] TensorChoko, "Feedback Prize - EDA Train JP/EN," Kaggle [Notebook], Year, retrieved from https://www.kaggle.com/code/tensorchoko/feedback-prize-eda-train-jp-en.

[24] A. Alkabool, S. Abdullah, S. Zadeh, and H. Mahfooz, "Identifying discourse elements in writing by longformer for ner token classification," *Iraqi Journal for Electrical and Electronic Engineering*, vol. 19, pp. 87–92, 6 2023.