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

Tingting Fan[1]

1 Nanjing University of Aeronautics and Astronautics

Potential competing interests: No potential competing interests to declare.

This study compared the performance of two natural language processing (NLP) models (Longformer*vs.* Bigbird models) in the evaluation of essay structure, specifically the identification of key argumentative and rhetorical elements. The manuscript has its strengths, but also some issues that need to be addressed.

Major issues:

1. The author mentioned repeatedly in the manuscript that this study proposed a model utilizing fine-tuned language models to evaluate essay structure. However, it is important to note that the study did not actually propose a novel model. Instead, its primary focus was on comparing the performance of two NLP models.

2. The author indicated that “Our goal is to train the models to identify key structural elements of a quality essay.” (p.2). However, the training details were not presented in the manuscript.

3. F1 score was used to evaluate and compare the performance of the two NLP models. However, the manuscript does not mention the specific criteria for interpreting the F1 score. For example, what is the range of the F1 score? What does a higher score indicate?

4. The formula for calculating F1 score is displayed in the manuscript. However, what does each element of the formula (i.e., TP, FP, FN) represent?

5. According to the manuscript, “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 …” (p.5). Is the difference significant between the two models? As far as I'm concerned, applying appropriate statistical tests would allow for a rigorous comparison of the F1 scores between the two models.

6. Fig.3 presents the F1 score for each discourse element. In addition to the figure, it would be beneficial to describe the findings in the text. Moreover, Fig. 1 and Fig. 2 were also displayed in the RESULTS section but were not described or discussed in the text.

Minor issues:

1. The format of in-text citations is not appropriate, for example, “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].” (p.1). The author needs to clearly indicate the author names and years of publication of the articles that were cited.

2. Other formatting and writing errors:

- such as being time-consuming, Heterogeneity of students, and subjective (p.1)
- and Hence our approach (p.3)
- ”cohenrence” score (p.3)
- ”traditional” AWE model (p.3)
- in the ”Trained Models and their F1 scores” table (p.5)