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

Oleg Nozhovnik

Potential competing interests: No potential competing interests to declare.

The article discusses the use of Automated Writing Evaluation (AWE) systems for training language skills, particularly writing skills. The study utilizes the PERSUADE corpus, a dataset of 25,000+ student essays annotated by writing professionals, to train and test the models. The dataset undergoes data preprocessing steps such as data cleaning, including removing duplicates, correcting misspellings, and normalizing data elements. Additionally, the article explains the use of stopword removal and stemming techniques to simplify word representation.

Cons of the research:

Limited scope of discourse elements: The research focused on specific discourse elements such as lead, position, claim, counterclaim, rebuttal, evidence, and concluding statement. While these elements are important, the study acknowledges the need for data on additional categories such as rebuttal or counterclaim to further enhance the overall results. The limited scope of discourse elements may restrict the generalizability of the findings to a broader range of writing tasks.

Lack of external validation: The research does not mention external validation of the models' performance on independent datasets or comparison with other existing automated writing evaluation systems. External validation is important to ensure the generalizability and reliability of the models' performance. Comparative analysis with other established systems would provide a more comprehensive understanding of the strengths and weaknesses of the proposed approach.

Limited hyperparameter exploration: The study used a fixed set of hyperparameters for the models, including the number of epochs, batch size, and learning rate scheduler. The choice of hyperparameters can significantly impact the model's performance. It would be beneficial to explore a wider range of hyperparameter values to determine the optimal configuration for achieving the best results.

Lack of qualitative analysis: The research primarily focuses on quantitative evaluation metrics and does not provide a qualitative analysis of the model's outputs. Qualitative analysis, such as examining specific examples of essay classifications and evaluating the model's interpretability, could provide deeper insights into the strengths and limitations of the approach.

Limited generalizability: The research evaluates the models' performance on a specific dataset of student essays. The generalizability of the findings to other domains or writing tasks is not explicitly discussed. The effectiveness of the models may vary when applied to different genres, languages, or writing proficiency levels, which limits the broader applicability of the research findings.

Drawbacks of this research:

Lack of real-world application: While the research focuses on discourse classification in student essays, it does not address the practical application of the models in real-world scenarios. The findings may have limited utility outside the specific context of automated writing evaluation for educational purposes. Future research should consider the potential application and impact of these models beyond the academic setting.

Lack of interpretability: The research does not delve into the interpretability of the models' decisions. Understanding why a model classifies a particular essay in a certain way is crucial for building trust and identifying potential biases or errors. By providing interpretability techniques or analyzing the model's attention mechanisms, researchers can gain insights into how the models make their predictions.

Limited exploration of other models: The study primarily focuses on comparing the performance of two specific models (Longformer and BigBird) for discourse classification. However, there are numerous other advanced models and architectures available in the field of natural language processing. Exploring a wider range of models could provide a more comprehensive understanding of the strengths and weaknesses of different approaches.

Lack of discussion on ethical considerations: The research does not explicitly address the ethical implications of using automated systems for evaluating student essays. Automated evaluation systems have the potential to impact student learning, teaching practices, and educational outcomes. Discussing the ethical considerations, such as potential biases, privacy concerns, and the role of human feedback and intervention, is essential for responsible implementation and use of such systems.

Limited sample size: Although the research utilizes a dataset with over 25,000 student essays, the sample size may still be considered limited when dealing with natural language processing tasks. A larger and more diverse dataset could provide a broader representation of different writing styles, genres, and educational levels, leading to more robust and generalizable results.

Lack of long-term evaluation: The research focuses on evaluating the models' performance at a single point in time, without considering their long-term effectiveness or potential decay in performance. It would be valuable to investigate how the models' performance evolves over time as new data becomes available, and whether regular retraining or updating is necessary to maintain optimal performance.

Limited external validation: The research does not extensively validate the models' performance on external datasets or compare them with other existing state-of-the-art models or systems. External validation is crucial for assessing the generalizability and competitiveness of the proposed models. Conducting comparative evaluations with other established approaches would provide a more comprehensive understanding of their strengths and weaknesses.

Lack of replication information: The research does not provide detailed information on the replication of the experiments. Sharing information about the specific configurations, code, and resources used would enable other researchers to replicate and validate the findings, promoting transparency and fostering scientific progress.

Addressing these drawbacks would help strengthen the research and its practical implications, providing a more comprehensive and insightful analysis of discourse classification in student essays.