A Robustness Analysis of Automated Essay Scoring Methods

Rafael T. Anchiêta[1], **Rogério F. de Sousa**[1], **Raimundo S. Moura**[2]

[1]Federal Institute of Piauí – (IFPI - Picos)
Laboratory of Artificial Intelligence, Robotics, and Automation (LIARA)

[2]Federal University of Piauí – (UFPI - Teresina)
Laboratory of Natural Language Processing (LPLN)

{rta, rogerio.sousa}@ifpi.edu.br, rsm@ufpi.edu.br

Abstract. *This paper analyzed the robustness of a state-of-the-art Automated Essay Scoring (AES) model by applying various linguistically motivated perturbations to the Essay-BR corpus. Our findings reveal that the AES model failed to detect these adversarial modifications, often assigning higher scores to the disturbed essays than to the original ones.*

1. Introduction

Automated Essay Scoring (AES) aims to provide computational models for automatically grading essays or with minimal involvement of humans [Page 1966]. Although this research area is over fifty years old [Beigman Klebanov and Madnani 2020], it has recently gained the attention of the Brazilian community because of publicly available corpora [Marinho et al. 2021, Marinho et al. 2022a]. Several methods to grade an essay or its characteristics arose based on these resources [de Sousa et al. 2024, Oliveira et al. 2023, Marinho et al. 2022b]. Besides, there is a growing interest in the area. For instance, recently occurred the PROPOR'24 Competition, whose goal was to develop computer systems capable of automatically evaluating essays [Mello et al. 2024].

Despite the advances achieved, the Brazilian community has made little effort to evaluate the robustness of AES methods, including analyzing their sensitivity to adversarial perturbations. [Liu et al. 2024] define robustness as the capacity to remain stable and reliable under different circumstances. Studies demonstrate that AES methods for the English language are easily fooled [Perelman 2014], reducing the trustworthiness of AI-based automated scoring systems [Kabra et al. 2022]. Based on these limitations of AES methods for English, we investigated whether AES methods for Portuguese suffer from robustness problems.

Our objective is to analyze AES methods using adversarial essays. For that, we applied a set of perturbations to an essay corpus, including adding unrelated texts, shuffling, deleting, and repeating paragraphs of an essay. With these linguistically-motivated disturbances, we evaluated a state-of-the-art AES strategy for Portuguese and found that the analyzed model could not deal with adversarial essays, producing, in fact, better results for undisturbed essays.

The remainder of this paper is organized as follows: Section 2 briefly presents related work. In Section 3, we detailed the performed analysis to verify the robustness of an Automated Essay Scoring method for Portuguese. Finally, Section 4 concludes the paper and indicates future directions.

2. Related Work

[Kabra et al. 2022] proposed a model agnostic adversarial evaluation scheme and associated metrics for AES systems to test their natural language understanding capabilities and overall robustness. They evaluated models ranging from feature-engineering-based approaches to the latest deep-learning algorithms. The authors found that AES models are highly overstable such that even heavy modifications (as much as 25%) with content unrelated to the topic of the questions do not decrease the score produced by the models.

[Liu et al. 2024] evaluated Automatic Essay Scoring models' robustness and generalization capabilities through a comprehensive series of experiments to validate various models' efficacy. The authors randomly select a part of the essays and shuffle the order of the sentences or delete a sentence randomly to construct a Chinese adversarial sample set for evaluating the robustness of the models. The results showed that the advanced AES models have poor robustness and generalization ability, and Large Language Models have better performance but still need to be improved.

3. Robustness Analysis

The Essay-BR corpus [Marinho et al. 2021] is organized into training, development, and testing sets, each with 3,198, 686, and 686 essays. We used the test set to generate adversarial essays. First, we extracted the essays with a score greater than or equal to 680 since the average score of ENEM 2023 was 641.6 points[1], resulting in 305 essays. We adopted the strategy of selecting the best essays, avoiding those with several grammatical, structural, and argumentative issues. After that, we applied several perturbations to the essays to produce adversarial essays. From the original and adversarial essays set, we evaluated the robustness of an Automated Essay Scoring (AES) model.

We implemented linguistically motivated perturbations to analyze the robustness of an AES model, i.e., to check whether the model can detect any difference between original and modified responses. The perturbations are detailed below.

Add unrelated text. We added an unrelated paragraph in each essay. We create three sets of essays with unrelated content, each indicating the position where an unrelated text was added. The sets are with unrelated texts added at the beginning, middle, and end of the essays. We extracted a paragraph from essays with a prompt different from the analyzed essay and added it to the essay. This test tries to mimic the behavior of students when they make their responses lengthy by adding irrelevant information.

Add song and cake recipe. Although these perturbations add unrelated content to an essay, they have a very different language structure than written prose in essays. So, this can be used to test a system negatively. Furthermore, it has been observed that students use this strategy in their exams, possibly in an attempt to fool the system[2]. We created two sets of perturbations, one for cake recipe and the other for the song. In both sets, we add unrelated content in the middle of the essays.

[1] https://querobolsa.com.br/revista/redacao-enem-2023-quantos-textos-tiraram-nota-mil-quantos-zeraram

[2] https://g1.globo.com/educacao/noticia/2013/03/queria-testar-correcao-do-enem-diz-jovem-que-pos-receita-na-redacao.html

Add repeated text. For this adversarial strategy, we also created three sets of perturbations. For each set, we repeated the essay content at the beginning, middle, and end of the essays. The motivation for this perturbation is that, according to [Kabra et al. 2022], students sometimes repeat sentences or specific keywords in their responses to make them longer yet not out of context and to fashion cohesive paragraphs [Higgins and Heilman 2014, Yoon et al. 2018].

Delete text. Similar to adding repeated text, we created three sets of perturbations in this strategy. For each set, we removed a paragraph at the beginning, middle, and end of the essays. According to [Kabra et al. 2022], these tests generally break the flow of an argument, delete crucial details from an essay, and decrease wordiness. This perturbation can seriously detract from the coherency and quality of writing and frustrate readers.

Shuffle text. For this perturbation, we randomly shuffle the content of an essay. The motivation for this adversarial strategy is to analyze important aspects of essay scoring, such as coherence and organization, which measure the extent to which the response demonstrates a unified structure and direction of the narrative [Barzilay and Lapata 2008, Tay et al. 2018].

After generating adversarial essays, we evaluated a state-of-the-art automated essay scoring [de Sousa et al. 2024] based on the BERT model [Devlin et al. 2019]. We assessed that model using each ENEM competency through the Quadratic Weighted Kappa (QWK) metric [Cohen 1968] for original and adversarial essays. QWK is a metric commonly used to assess AES models [Yannakoudakis and Cummins 2015]. Table 1 shows the results on the original essays, and Table 2 presents the results on the adversarial essays.

Tables 1 and 2, from C1 to C5, indicate the five competencies of the ENEM, and the total is the final grade for an essay. In Table 2, we highlight the values greater than or equal to the value of the original essays.

Analyzing the values from the two tables, we can see that only the values of adding text at the beginning and adding a cake recipe were not greater than the original essay values, indicating that the AES model was able to identify perturbations in the essays, penalizing their scores. On the other hand, the scores for adding unrelated text in the middle, in the end, and a song were greater than or equal to the values for the original essays. An interesting finding is that, despite adding an unrelated text at the end of an essay, the C5 score was not penalized. Competency 5 of the ENEM is dedicated to elaborating a proposal to solve the problem. The proposal normally appears at the end of an essay, and the AES model could not detect the unrelated content added to an essay. More than that, the final grade of original and adversarial essays had the same QWK value, suggesting that the AES model failed to capture this perturbation.

For the perturbation of repeating a text in the essay, the AES model graded the original and adversarial essays with the same score, mainly in competence four. Competence 4 evaluates the superficial structure of the text, that is, how the sentences and paragraphs are linked through cohesive elements. This way, the AES model should negatively score such responses. Besides, the scores for repeating a text in the middle and at the end of an essay had the same value as the original essays.

Another interesting finding is that deleting some parts of the essay improves its grade in various competencies. As we can see, the scores for deleting a text in the essay

are greater than or equal to the scores of the original essays, including the final score. These results show that the AES model could not identify a break in the flow of an argument when essential parts of the essay were removed.

Finally, and perhaps the most interesting finding, is that shuffling the paragraphs of an essay produces better results than the original essays. This result demonstrates that the AES model could not determine the cohesion and coherence of the essays. That is, the AES model did not identify the transition between the lines of the essays, verifying disconnected ideas that change the meaning substantially.

The source code of the AES model and for generating adversarial essays are publicly available at `https://github.com/liara-ifpi/essay-robustness`.

Table 1. Quadratic Weighted Kappa results on the original essays.

C1	C2	C3	C4	C5	Total
0.44	0.23	0.29	0.24	0.62	0.46

Table 2. Quadratic Weighted Kappa results on the adversarial essays.

Adversarial strategy	C1	C2	C3	C4	C5	Total
Add unrelated text at the begging	0.41	0.14	0.15	0.17	0.57	0.40
Add unrelated text in the middle	0.42	0.18	0.24	0.20	**0.62**	0.44
Add unrelated text at the end	0.43	0.22	0.28	**0.24**	**0.62**	**0.46**
Add song	0.38	0.21	0.23	0.20	**0.64**	0.42
Add cake recipe	0.38	0.18	0.22	0.18	0.61	0.41
Repeat text at the begging	**0.44**	0.20	0.23	**0.24**	0.59	0.44
Repeat text in the middle	0.43	0.21	0.27	**0.24**	0.61	**0.46**
Repeat text at the end	0.43	**0.23**	0.28	**0.24**	**0.62**	**0.46**
Delete text at the begging	0.40	0.19	**0.30**	0.22	**0.66**	0.45
Delete text in the middle	0.40	**0.23**	**0.29**	**0.25**	**0.66**	**0.46**
Delete text at the end	0.41	**0.23**	**0.29**	**0.24**	**0.63**	**0.46**
Shuffle text	**0.47**	0.20	**0.29**	**0.24**	**0.64**	**0.47**

4. Conclusion

This paper presented a robustness analysis for automatic essay scoring focusing on the Portuguese language. We used the Essay-BR corpus, which is based on the ENEM competencies, to perform that analysis. Our strategy was to add several perturbations to produce adversarial essays, aiming to check if a state-of-the-art automated essay scoring model can detect any difference between original and modified responses. From the analysis, we have learned that the automated essay scoring model could not identify the perturbations in the essays, producing scores that were even greater than the original responses. We hope that this analysis sheds light on this research area and helps develop more robust strategies for automatically grading essays.

For future work, we intend to develop more perturbations and create a toolkit to facilitate the creation of adversarial essays.

References

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Beigman Klebanov, B. and Madnani, N. (2020). Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213–220.

de Sousa, R. F., Marinho, J. C., Neto, F. A. R., Anchiêta, R. T., and Moura, R. S. (2024). PiLN at PROPOR: A BERT-based strategy for grading narrative essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 10–13, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Higgins, D. and Heilman, M. (2014). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice*, 33(3):36–46.

Kabra, A., Bhatia, M., Singla, Y. K., Jessy Li, J., and Ratn Shah, R. (2022). Evaluation toolkit for robustness testing of automatic essay scoring systems. In *Proceedings of the 5th Joint International Conference on Data Science & Management of Data*, pages 90–99, Bangalore, India. Association for Computing Machinery.

Liu, R., Wang, X., Liu, J., and Zhou, J. (2024). A comprehensive analysis of evaluating robustness and generalization ability of models in aes. In *Proceedings of the 7th International Symposium on Big Data and Applied Statistics*, pages 1–5, Beijing, China. IOP Publishing.

Marinho, J. C., Anchiêta, R. T., and Moura, R. S. (2021). Essay-br: a brazilian corpus of essays. In *XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2021*, pages 53–64, Online. SBC.

Marinho, J. C., Anchiêta, R. T., and Moura, R. S. (2022a). Essay-br: a brazilian corpus to automatic essay scoring task. *Journal of Information and Data Management*, 13(1):65–76.

Marinho, J. C., C., F., Anchiêta, R. T., and Moura, R. S. (2022b). Automated essay scoring: An approach based on enem competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60, Campinas, Brazil. SBC.

Mello, R. F., Oliveira, H., Wenceslau, M., Batista, H., Cordeiro, T., Bittencourt, I. I., and Isotanif, S. (2024). PROPOR'24 competition on automatic essay scoring of Portuguese

narrative essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 1–5, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Oliveira, H., Ferreira Mello, R., Barreiros Rosa, B. A., Rakovic, M., Miranda, P., Cordeiro, T., Isotani, S., Bittencourt, I., and Gasevic, D. (2023). Towards explainable prediction of essay cohesion in portuguese and english. In *Proceedings of the 13th International Learning Analytics and Knowledge Conference*, pages 509–519, Arlington TX USA. Association for Computing Machinery.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing*, 21:104–111.

Tay, Y., Phan, M., Tuan, L. A., and Hui, S. C. (2018). Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the Thirty-second AAAI conference on artificial intelligence*, pages 5948–5955, New Orleans, Louisiana, USA. AAAI Press.

Yannakoudakis, H. and Cummins, R. (2015). Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223, Denver, Colorado. Association for Computational Linguistics.

Yoon, S.-Y., Cahill, A., Loukina, A., Zechner, K., Riordan, B., and Madnani, N. (2018). Atypical inputs in educational applications. In Bangalore, S., Chu-Carroll, J., and Li, Y., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 60–67, New Orleans - Louisiana. Association for Computational Linguistics.