Automated Thematic Coherence Scoring of Student Essays Written in Portuguese

Rafael Pacheco[1], Luiz Rodrigues[4], Lucas Lins[2], Péricles Miranda[2], Valmir Macário[2], Seiji Isotani[5,6], Thiago Cordeiro[4], Ig Ibert Bittencourt[4,6], Diego Dermeval[4] Dragan Gašević[7], Rafael Ferreira Mello[3,7]

[1] Universidade Federal Pernambuco

[2]Universidade Federal Rural de Pernambuco

[3]Centro de Estudos e Sistemas Avançados do Recife (CESAR School)

[4]Núcleo de Excelência em Tecnologias Sociais (NEES), Universidade Federal de Alagoas

[5]Universidade de São Paulo

[6]Harvard Graduate School of Education

[7]Monash University

{pachecorlps,lucas.felins}@gmail.com

{pericles.miranda,valmir.macario,rafael.mello}@ufrpe.br

{luiz.rodrigues,thiago.cordeiro,diego.matos}@nees.ufal.br

{seiji_isotani,ig_bittencourt}@gse.harvard.edu

dragan.gasevic@monash.edu

Abstract. *While Thematic Coherence is a fundamental aspect of essay writing, scoring it is labor-intensive. This issue is often addressed using machine learning algorithms to estimate the score. However, related work is mostly limited to the English language or argumentative essays. Consequently, there is a lack of research on other widely used languages and essay types, such as Brazilian Portuguese and narrative essays. Hence, this paper reports on the findings of a study that aimed to evaluate the value of machine learning algorithms to automatically score the Thematic Coherence of both narratives (n = 400) and argumentative (n = 6567) essays written in Brazilian Portuguese. Expanding on previous studies, this paper evaluated regression models using conventional, feature-based algorithms according to essays' linguistic features. Overall, we found that Extra Trees was the best performing algorithm, yielding predictions with moderate to strong correlations with human-generated scores. Mainly, those findings expand the literature with evidence on the potential of machine learning to estimate the Thematic Coherence of narrative and argumentative essays, suggest an improved performance for the former type.*

1. Introduction

Essay writing has long been an essential task for students at all levels of education. However, unlike multiple-choice assignments, essay scoring is still labor-intensive work.

Automating the scoring process may improve the application of large-scale assessments while providing individualized feedback for students [Burstein et al. 2003].

Within several constructs that are usually analyzed in essay scoring systems, the *Thematic (or textual) Coherence* (TC) evaluates whether the essay makes logical sense – globally (i.e., the overall text) and locally (i.e., within and among close paragraphs) – and relates to the text's adherence to the proposed theme/prompt [Palma and Atkinson 2018]. Consequently, TC has a prominent effect on essay scoring as the adherence to the prompt's topic often has significant implications for the score. Similarly, lacking logical sense will likely impair one's understanding of the essay and lead to poor scoring. TC is an important construct in many exams such as the Brazilian National High School Exam (ENEM) -0 an annual assessment that directly impacts students' chances of getting into higher education and earning scholarships.

Automating Essay Scoring (AES) is a research area that aims to address the issue of TC by using computational models to automate the scoring task [Ferreira-Mello et al. 2019]. Specifically, machine learning has been widely used to realize AES [Costa et al. 2020, Filho et al. 2021, Lima et al. 2018]. This approach enables timely and individualized student feedback and optimizes educators' practices by standardizing and speeding up the scoring process, and thus facilitating the application of large-scale assessments [Burstein et al. 2003]. Moreover, AES might offer detailed feedback on student performance based on the essay's textual characteristics [Khosravi et al. 2022]. Hence, AES-based technology might provide significant benefits to enhance the teaching-learning process.

While researchers have made substantial contributions to AES over the last years, there is a lack of research on scoring essays written in languages other than English [Bai and Stede 2022]. Notably, a few recent studies started to contribute towards automatically scoring essays written in Brazilian Portuguese – often based on ENEM data (e.g., [Marinho et al. 2022, Haendchen Filho et al. 2018, Júnior et al. 2017, Oliveira et al. 2022, Lima et al. 2018]). Nevertheless, those are centered on *argumentative* essays, while other essay types, such as *narrative*. Moreover, to the best of our knowledge, prior research did not explore the automatic evaluation of TC on narrative essays written in Brazilian Portuguese, nor how such scoring performance compares to that of argumentative essays. Additionally, previous research did not inspect which essay features are relevant for TC scoring, which is essential to properly enhance the teaching-learning process with insightful feedback regarding the assigned score.

Therefore, this paper proposes assessing several machine learning models to automatically TC score narrative and argumentative essays providing a detailed analysis of the most relevant features for this task. Specifically, we first computed an extensive set of linguistic features from two datasets of essays written in Brazilian Portuguese by prospective college students and elementary school learners. Then, we trained and evaluated several machine learning models to predict human-generated ratings for TC on both datasets. The results revealed that i) Extra Trees' predictions showed moderate to strong correlations to human scoring, ii) results for narrative essays were better than those for argumentative ones.

2. Related Works

Most studies on AES concern text written in English [Bai and Stede 2022]. Despite that, recent research has contributed to AES based on the Brazilian Portuguese language. For instance, Júnior et al. [Júnior et al. 2017] evaluated a feature-based approach to determine scores for analyzing lexical and syntactic errors, achieving 93% precision. Oliveira et al. [Oliveira et al. 2022, Oliveira et al. 2023] researched feature-based models for automatically scoring essays written in Portuguese, focusing on analyzing cohesion. In the first study, the authors trained models on an argumentative dataset to estimate TC, achieving a moderate Pearson correlation of 0.53 and a Mean Absolute Error (MAE) of 26.97 (on a scale of 0 to 200) [Oliveira et al. 2022]. In the second study, Oliveira et al. [Oliveira et al. 2023] explored deep learning and conventional feature-based models to score textual cohesion in Portuguese and English essays and explain their predictions. While deep learning yielded the best results, with a moderate correlation to human-generated scores, conventional models provided greater explainability. Although the scoring criteria in [Oliveira et al. 2023] differed from that of Text Cohesion, it presents promising possibilities for automatic essay scoring in Portuguese. Furthermore, several other studies have begun to address similar issues, as described below.

Two relevant studies have focused on estimating TC in essays written in Brazilian Portuguese. Haendchen Filho et al. [Haendchen Filho et al. 2018] proposed an approach that utilized classification and regression models based on Support Vector Machines (SVM) to automatically score adherence to the theme and argumentative structures of essays. This approach achieved an average error of 0.3440 and a Pearson correlation of 0.7410 for TC. Besides conventional feature-based models, some studies also explored those based on deep learning. Marinho et al. [Marinho et al. 2022] trained feature-based and deep learning models on ENEM data. Concerning TC, they obtained a moderated agreement with the ground truth, based on the quadratic weighted kappa, with the feature-based model. Additionally, they discussed that the possible reason for this finding is the importance of the similarity between the essay and the prompts, providing valuable insights into understanding the model's predictions.

In summary, while there have been some studies on AES in Brazilian Portuguese, the literature is still limited. Although feature-based models have shown promising results in predicting human-generated scores, the insights are mainly based on predictions for argumentative essays. As a result, there is a lack of research on the automatic scoring of narrative essays, and it is unclear how predictive performance compares between argumentative and narrative essays. Therefore, this paper aims to fill this gap by investigating the automatic scoring of both argumentative and narrative essays comparing the predictive performance of different models.

3. Method

This study analyzed the viability of i) using machine learning methods to estimate human-generated ratings of TC scores for narrative and argumentative essays and ii) providing empirical evidence on how the performance of machine learning models compares for narrative and argumentative essays. Accordingly, we sought to answer the following Research Question (RQ):

- **RQ:** *To what degree do machine learning algorithms accurately estimate the coherence of human-generated essay scores?*

The following sections introduce this study's datasets, feature extractors, and model selection procedure.

Table 1. Guidelines human experts followed to assess narrative essays' TC.

Score	Criteria
1	The text only briefly touches on the topic and lacks a clear and logical progression of ideas.
2	The text does not provide enough development of its ideas and lacks a clear progression from one point to the next, relying too heavily on the motivating situation. Moreover, it relies heavily on copied material from the motivating situation and therefore lacks originality.
3	Although the text does present a clear and logical progression of ideas, it relies too heavily on paraphrasing the motivating situation, lacking original insights. The text presents a complete progression of ideas but relies on general, common sense concepts without providing specific details or examples.
4	The text presents a clear and logical progression of ideas that is consistent with the motivating situation. It uses common sense concepts to develop its ideas, while also providing specific details and examples to support its arguments.
5	The text presents a complete and well-developed progression of ideas that go beyond the motivating situation. It showcases a consistent repertoire of concepts that are relevant to the topic at hand and provides insights that are applicable to a broader context.

3.1. Datasets

This study used two datasets of student essays written in Brazilian Portuguese. The **narrative dataset** was developed by the [REMOVE FOR BLIND REVIEW] Program. It comprises 400 narrative essays written by students of mid-school age from public schools in Brazil. The essays should have at least five lines (text with less than five lines was excluded) about a fictional narrative created based on motivational situations/prompts provided by the teachers. This dataset was coded by human experts who scored the essays' TC with a score from 1 to 5 based on the criteria presented in Table 1.

The **argumentative dataset** is based on the Essay-BR corpus [C. Marinho et al. 2022]. It comprises ENEM essays written by prospective university students who received a prompt and were required to write an argumentative essay featuring from 8 to 30 lines. It encompasses a dataset with 6,567 essays from 151 topics that were written between December 2015 and August 2021. Additionally, human experts similarly scored them in terms of *understanding of the proposed theme*. Originally, the scores ranged from 0 to 200. Nevertheless, we normalized them to 1 to 5 to facilitate comparisons with the narrative dataset.

Table 2 summarizes both datasets by presenting descriptive statistics for their Text Coherence (TC) scores and the total number of essays, as well as the mean and standard deviation of both the number of sentences and words per essay.

Table 2. Descriptive statistics of the datasets used in this study. N refers to the number of essays. Sentences and Words are shown as Mean (Standard Deviation).

Dataset	Score	N (%)	Sentences	Words
Essay-BR	0	123 (02%)	8.56 (3.38)	218.86 (64.35)
	1	93 (01%)	8.57 (3.84)	211.31 (113.45)
	2	918 (14%)	9.62 (4.24)	231.58 (79.58)
	3	2440 (37%)	10.88 (3.98)	271.84 (76.79)
	4	2421 (37%)	12.82 (3.32)	323.87 (68.43)
	5	572 (09%)	13.89 (3.24)	342.25 (73.75)
	Overall	*6,567*	*11.60 (3.98)*	*289.68 (83.51)*
Narrative	1	134 (34%)	3.15 (3.13)	139.26 (67.80)
	2	59 (15%)	2.39 (2.13)	114.86 (71.39)
	3	168 (42%)	3.43 (3.76)	142.80 (68.72)
	4	31 (08%)	4.19 (3.72)	156.61 (58.74)
	5	8 (02%)	5.88 (3.04)	194.38 (105.47)
	Overall	*400*	*3.29 (3.38)*	*139.60 (69.87)*

3.2. Feature Extraction

Considering the previous works in the literature on essay scoring and cohesion analysis [Guinaudeau and Strube 2013, Ferreira-Mello et al. 2019, Ferreira Mello et al. 2022, Oliveira et al. 2023, Oliveira et al. 2023], we decided to use linguistic features based on state-of-the-art tools such as Coh-Metrix. It follows a brief description of the features used.

- **Coh-Metrix:** The Coh-Metrix set of linguistic indicators [Graesser et al. 2004, McNamara et al. 2014] is designed to extract features that are associated with text cohesion, linguistic complexity, text readability, and lexical diversity. To conduct our analysis, we utilized the Portuguese version of Coh-Metrix that was presented in [Camelo et al. 2020].
- **Legibility:** These features include various metrics, such as the number of syllables per word, words per sentence, and the number of unique words in the essay. In our analysis, we computed the following metrics: Mean Tokens per Sentences, Mean Syllables per Word, Flesch Reading Ease, Gunning Fog Index, Automated Readability Index, and Word Variation Index.
- **Similarity:** The features in this group aim to measure the similarity between the essay and the prompt. We computed two metrics for this purpose. The first is the cosine similarity between the vector representations of the essay and the prompt, which were generated using Spacy [Honnibal et al. 2020]. The second is the Jaccard similarity index between the sets of words in the essay and the prompt, as well as the keywords from both.
- **Local Coherence:** We extracted six features using the TRUNAJOD library [Palma and Atkinson 2018], which is based on the entity grid model proposed in [Guinaudeau and Strube 2013]. This model is designed to compute the overlap of entities between subsequent sentences.

Together, those generated a set of 103 linguistic features for each essay from each dataset. These sets were the input we passed to train and test the machine learning models, as described below.

3.3. Model Selection and Evaluation

To answer the RQ, we trained different regression algorithms using the features described in Section 3.2. Although the TC scores for both datasets were discrete, we treated essay-scoring systems as regression problems in line with how the literature often approaches them [Basu et al. 2013].

We chose a wide variety of algorithms (e.g., Ensembles, Neural Networks, and SVM, among others) to achieve a representative set of the options available nowadays, which resulted in considering 12 ones. Those were implemented using the following libraries: scikit-learn[1], XGBoost[2] and LGBM[3]. For all algorithms, we used the libraries' default parameters as performing hyper-parameter tuning for all of them was unfeasible given this study's resources.

To evaluate the models, we adopted a 5-fold, Stratified Cross-Validation strategy. For each model, we additionally tested the contribution of three oversampling settings (i.e., none, RandomOverSampler - ROS, and SMOTE [Chawla et al. 2002]) aiming to mitigate issues related to class imbalance (see Table 2). Both oversampling methods were computed using their respective implementations from the imbalanced-learn[4] library. Hence, we performed three cross-validation procedures for each of the 12 algorithms considered, leading to the development of 36 models.

To assess models' performances, we followed literature guidelines [Fernández-Delgado et al. 2019]. For each cross-validation, we extracted the following measures: Pearson correlation, Root Mean Squared Error (RMSE), and MAE. The RMSE, MAE, Acc, MedAE, and F1-Score measures were calculated using the scikit-learn library, and Pearson's correlation was computed using the SciPy library. Furthermore, we followed the guidelines presented in [Ratner 2009] to interpret the linear correlation's coefficient: 0 means no relationship; between 0 and 0.3 means a weak relationship; between 0.3 and 0.7 implies a moderate relationship; between 0.7 and 1.0 means a strong relationship; and 1 means a perfect relationship. Note that the same interpretations apply to negative coefficients. Importantly, note that results were based on the cross-validation test results to maximize external validity [Wohlin et al. 2012].

4. Results

Table 3 presents the cross-validation results. For the narrative dataset, the algorithms that yielded the best performances were: Extremely Randomized Trees (Extra Trees), LGBM, and Random Forest. All of those achieved similar values of RMSE ($\approx$ 0.58), MAE ($\approx$0.55), and correlation ($\approx$0.71). It is important to highlight that the best results were reached without using any oversampling algorithm. For the argumentative dataset, the Extra Trees (also without oversampling) yielded the best results in all measures (RMSE = 0.7065, MAE = 0.6303, correlation = 0.5288). Interestingly, Extra Trees achieved such

[1] `https://scikit-learn.org/`

[2] `https://github.com/dmlc/xgboost/`

[3] `https://github.com/Microsoft/LightGBM/`

[4] `https://imbalanced-learn.org/stable/index.html`

results with no oversampling in both cases. Overall, these results indicate moderate to strong relationships between predictions and human-generated scores, based on Pearson Correlation, for both datasets. Hence, as Extra Trees' yielded the best results for both cases, we consider it as our best model in subsequent analyses.

Table 3. Performance of the regression algorithms in estimating essay's coherence score on the Essay-BR and Narrative datasets.

Algorithm	Oversampler	Narrative			Essay-BR		
		RMSE	MAE	P	RMSE	MAE	P
AdaBoost	None	0.6682	0.6517	0.6521	0.8224	0.7088	0.4797
	ROS	0.7131	0.6853	0.6225	1.1028	0.8378	0.4814
	SMOTE	0.6823	0.6548	0.6447	1.1632	0.8732	0.4534
Bayesian Ridge	None	2.4691	1.1946	0.1586	13.0782	3.4882	0.3723
	ROS	2.2574	1.1595	0.1626	13.2237	3.4884	0.2563
	SMOTE	2.2182	1.1296	0.1990	13.7240	3.5241	0.2477
Decision Trees	None	1.0844	0.6488	0.5235	1.4582	0.8494	0.3047
	ROS	1.0967	0.6798	0.5321	1.4478	0.8579	0.2821
	SMOTE	1.0344	0.6488	0.5462	1.8471	0.9816	0.2668
ExtraTrees	None	**0.5751**	0.5671	0.7093	**0.7065**	**0.6303**	**0.5288**
	ROS	0.6165	0.6116	0.6891	0.7105	0.6334	0.5264
	SMOTE	0.6289	0.6117	0.6776	0.7678	0.6690	0.5153
Gradient Boosting	None	0.5993	0.5627	0.6967	0.7202	0.6336	0.5186
	ROS	0.7061	0.6173	0.6527	1.0373	0.7892	0.4961
	SMOTE	0.6717	0.6025	0.6638	0.9753	0.7638	0.4778
LGBM Regressor	None	0.5787	0.5522	**0.7112**	0.7393	0.6468	0.5053
	ROS	0.6952	0.6094	0.6527	0.8885	0.7204	0.4715
	SMOTE	0.6450	0.5829	0.6776	0.8877	0.7206	0.4659
Linear Regression	None	1.2211	0.8661	0.3362	0.7353	0.6432	0.5044
	ROS	1.5823	0.9715	0.2782	1.3773	0.9300	0.4490
	SMOTE	1.7066	0.9968	0.2471	1.5190	0.9717	0.4333
MLP Regressor	None	1.5209	0.9173	0.3224	1.2611	0.8634	0.3357
	ROS	1.5525	0.9186	0.3306	1.5481	0.9567	0.2937
	SMOTE	1.6928	0.9367	0.3378	1.5102	0.9648	0.3173
Random Forest	None	0.5834	**0.5422**	0.7045	0.7122	0.6348	0.5257
	ROS	0.6391	0.5824	0.6765	0.7616	0.6573	0.4981
	SMOTE	0.6381	0.5891	0.6748	0.8185	0.6922	0.4925
Ridge	None	1.1087	0.8404	0.3917	0.7289	0.6391	0.5095
	ROS	1.6165	0.9883	0.3151	1.3200	0.9126	0.4646
	SMOTE	1.7052	1.0157	0.2963	1.4221	0.9450	0.4591
SVR	None	0.9281	0.7992	0.4441	0.7346	0.6385	0.5072
	ROS	1.0507	0.8365	0.3806	1.0338	0.7842	0.4608
	SMOTE	1.0496	0.8338	0.3781	1.0428	0.7853	0.4455
XGB Regressor	None	0.6427	0.5773	0.6763	0.8183	0.6858	0.4578
	ROS	0.7578	0.6323	0.6257	0.9702	0.7575	0.4132
	SMOTE	0.7150	0.6079	0.6375	1.0126	0.7748	0.3968

5. Discussion

In summary, this study demonstrates the feasibility of using machine learning algorithms to estimate human-generated scores for argumentative essays' TC. Overall, our models yielded moderate to strong correlations with human-generated scores, which is similar to prior research on AES for argumentative essays written in Brazilian Portuguese (e.g., [Júnior et al. 2017]). Particularly for TC, those results align with related work (e.g., [Haendchen Filho et al. 2018, Marinho et al. 2022, Oliveira et al. 2022]). Unlike prior research, our study is not limited to argumentative essays. We also developed and tested models to estimate narrative essays' TC. Thus, this paper expands the literature on AES with empirical evidence on the suitability of estimating the TC of a narrative essay written in Brazilian Portuguese.

Furthermore, related work is mainly concerned with a single essay type. Research on AES for Brazilian Portuguese texts is predominantly concerned with argumentative essays [Júnior et al. 2017, Marinho et al. 2022]. In contrast, this paper revealed that the same features and machine learning models could reach good results for different types of essays. Specifically, the results indicated that machine learning algorithms were better at predicting narrative essays' TC than estimating that of argumentative ones. Thus, this paper also expands the literature with evidence on how the performance of machine learning algorithms compares depending on the essay type.

The findings also indicate that the oversampling algorithms did not increase the results despite the unbalanced nature of the datasets. This is an issue for many text mining problems [Ferreira-Mello et al. 2019], but in many cases, the adoption of content-independent features reduces this issue [Osakwe et al. 2022, Ferreira Mello et al. 2022].

6. Pedagogical Implications

Following our encouraging findings, this section discusses how one might use AES to enhance the teaching-learning process compared to the standard practice.

Overall, the traditional essay scoring process often is as follows:

1. The student writes their essay and submits it for assessment;
2. The educator assesses the essay from scratch, scores it, designs feedback, and sends their considerations to the student;
3. The student receives and analyzes the educator's considerations.

Based on that process, we highlight two key issues. First, the student is limited to waiting while the educator works on step two. Because educators commonly need to assess essays from several students, they cannot provide timely feedback to students. Consequently, students might have to wait for days to receive feedback on their essays. Second, educators often need to start from scratch when assessing students' essays, especially when it comes to evaluating standard issues like whether the text aligns with the prompt and how paragraphs connect to each other, among other factors [Palma and Atkinson 2018]. This holds true for assessing argumentative and other essay types as well, which share similar characteristics. Therefore, educators can greatly benefit from technological tools to optimize the analysis process and provide meaningful feedback based on these similarities.

AES might address the above issues in the following ways. First, once the student submits their essay for assessment, an AES-based system might process the essay and promptly provide feedback to the student. For instance, such feedback might comprise the essay's scoring plus considerations regarding the features it (does not) possess to explain the scoring for the student. Thereby, the student no longer has to wait until the educator has the time to assess their essay. Second, with the help of AES, educators no longer need to start from scratch when assessing student essays. For instance, an AES-based system can suggest a score and highlight the essay's features that need improvement for each student submission based on our research results. The educator can then use this information to make an informed decision on the final score and provide valuable feedback to guide their students. By building upon the recommendations of the AES-based system, the educator can streamline the assessment process and provide more effective feedback to their students.

Based on that context, the essay scoring process can be rethought to include the following steps: (i) The student submits their essay for assessment; (ii) The AES system provides rapid feedback, such as scoring and identifying existing or missing features, to both the student and educator; (iii) The educator analyzes the AES feedback, evaluates the essay based on their own considerations, and provides feedback to guide the student's improvement; (iv) The student receives and analyzes the educator's final considerations. By incorporating intelligent recommendations from an AES system, educators can optimize their teaching practice, and students can benefit from rapid feedback to improve their writing skills. This streamlined approach to essay scoring has the potential to enhance the learning experience for both students and educators.

7. Conclusions

Whereas essay writing is an essential learning activity for students, essay scoring is a labor-intensive task for educators. Accordingly, researchers have investigated the use of Automatic Essay Scoring (AES) to optimize this task, often using machine learning models to estimate essay scores as well as provide insights on essay characteristics that affect its score. However, most studies are limited to essays written in English, which highlights the need for research on other languages. For Brazilian Portuguese, for example, despite initial research efforts, those mostly concerned with argumentative essays. In contrast, narrative essays also play a significant role in teaching-learning.

Therefore, this paper tackles that gap with an empirical study analyzing, in terms of TC, i) machine learning algorithms' potential to automatically score both narrative and argumentative essays, ii) how the performances of the algorithms compare depending on the essay type, and iii) which features affect their scores the most. In summary, our results revealed i) the suitability of estimating narrative essays' scores, ii) that the algorithms performed better for narrative compared to argumentative essays, and iii) that the similarity between the essay and the prompt, vocabulary richness and word/sentence count are the most relevant features for narrative and argumentative essays.

Compared to prior research [Marinho et al. 2022, Haendchen Filho et al. 2018, Oliveira et al. 2022, Lima et al. 2018], this paper expands the literature on AES for the Brazilian Portuguese language in three points. First, it provides empirical evidence on the suitability of machine learning algorithms for estimating narrative essays' TC. Second, it demonstrates how such algorithms' performances compare to those of argumentative essays. Third, it reveals the features affecting the TC of narrative and argumentative essays the most. Accordingly, those findings inform practitioners of the value of using machine learning to rapidly estimate narrative and argumentative essay scores and gather feedback on features (they should have) to improve their TC. Furthermore, our findings inform researchers on how machine learning performance differs depending on the essay type and promising features to be explored in similar studies. Thus, we contribute empirical evidence supporting and informing future research on AES's role in technology-enhanced learning.

Finally, we acknowledge this study has some limitations that must be considered when interpreting its findings. First, the narrative dataset is limited to 400 samples, which might have limited the algorithms' ability to learn further how to estimate their TC. Hence, we encourage future research to build and expand on it with similar data to test our findings' generalization.

Second, although we tested several algorithms, we were unable to explore some more computationally expensive alternatives, such as deep neural networks. Similarly, we could not perform hyper-parameter tuning within the cross-validation model selection due to resource restrictions. Thereby, we call for future research to explore other algorithms, as well as test hyper-parameter tuning, to verify to which extent the predictive performance found in this paper holds and/or might be improved.

Lastly, one must consider the human factors involved in this study. On the one hand, the narrative dataset is based on children's essays, whereas the argumentative essays are comprised of teenagers' ones. Consequently, this distinction might play a role in comparing machine learning algorithms' performances between essay types and the features that determine TC the most. On the other hand, essay scoring itself is a subjective task. Accordingly, two human experts might likely score the same essay differently. Similarly, training high-performing AES models is challenging because datasets will likely reflect such subjectivity. Hence, we recommend future research to explore how those human factors affect AES.

References

Bai, X. and Stede, M. (2022). A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *International Journal of Artificial Intelligence in Education*, pages 1–39.

Basu, S., Jacobs, C., and Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Burstein, J., Marcu, D., and Knight, K. (2003). Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.

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

Camelo, R., Justino, S., and de Mello, R. F. L. (2020). Coh-metrix pt-br: Uma api web de análise textual para a educação. In *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação*, pages 179–186. SBC.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Costa, L., de Oliveira, E. H. T., and Júnior, A. C. (2020). Corretor automático de redações em língua portuguesa: um mapeamento sistemático de literatura. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1403–1412. SBC.

Fernández-Delgado, M., Sirsat, M. S., Cernadas, E., Alawadi, S., Barro, S., and Febrero-Bande, M. (2019). An extensive experimental survey of regression methods. *Neural Networks*, 111:11–34.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

Ferreira Mello, R., Fiorentino, G., Oliveira, H., Miranda, P., Rakovic, M., and Gasevic, D. (2022). Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in portuguese. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 404–414.

Filho, A. H., Concatto, F., do Prado, H. A., and Ferneda, E. (2021). Comparing feature engineering and deep learning methods for automated essay scoring of brazilian national high school examination.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Guinaudeau, C. and Strube, M. (2013). Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria. Association for Computational Linguistics.

Haendchen Filho, A., do Prado, H. A., Ferneda, E., and Nau, J. (2018). An approach to evaluate adherence to the theme and the argumentative structure of essays. *Procedia Computer Science*, 126:788–797.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Júnior, C. R., Spalenza, M. A., and de Oliveira, E. (2017). Proposta de um sistema de avaliação automática de redações do enem utilizando técnicas de aprendizagem de máquina e processamento de linguagem natural. *Anais do Computer on the Beach*, pages 474–483.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., and Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074.

Lima, F., Haendchen Filho, A., Prado, H., and Ferneda, E. (2018). Automatic evaluation of textual cohesion in essays. In *19th International Conference on Computational Linguistics and Intelligent Text Processing*.

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

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

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 *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 509–519.

Oliveira, H., Miranda, P., Isotani, S., Santos, J., Cordeiro, T., Bittencourt, I. I., and Mello, R. F. (2022). Estimando coesão textual em redações no contexto do enem utilizando modelos de aprendizado de máquina. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 883–894. SBC.

Osakwe, I., Chen, G., Whitelock-Wainwright, A., Gašević, D., Cavalcanti, A. P., and Mello, R. F. (2022). Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence*, 3:100059.

Palma, D. and Atkinson, J. (2018). Coherence-based automatic essay assessment. *IEEE Intelligent Systems*, 33(5):26–36.

Ratner, B. (2009). The correlation coefficient: Its values range between+ 1/- 1, or do they? *Journal of targeting, measurement and analysis for marketing*, 17(2):139–142.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.